[AI Nexus x SNU Silicon Valley Immersion Program] The AI Inflection Point - Where AI Technology, Economic Transformation, and Human Questions Collide in the Age of Agentic Intelligence

Sunghee Yun

Co-Founder & CTO @ Erudio Bio, Inc.
Co-Founder & CEO @ Erudio Bio Korea, Inc.
Co-Founder & Leader & Chair of Silicon Valley AI Nexus
CGO / Global Managing Partner @ LULUMEDIC
Advisor to Korean American Semiconductor Professional Alliance
KFAS-Salzburg Global Leadership Initiative Fellow
Visiting Professor @ Sogang University
Advisory Professor @ DGIST

About Speaker

- *Co-Founder & CTO @ Erudio Bio, Inc., San Jose & Novato, CA, USA* 2023 ∼
- *Co-Founder & CEO @ Erudio Bio Korea, Inc., Korea* 2025 ∼
- *Co-Founder, Leader, and Chair of Silicon Valley AI Nexus, USA* 2024 ∼
- *CGO / Global Managing Partner @ LULUMEDIC, Seoul, Korea* 2025 ∼
- *AI-Korean Medicine Integration Initiative Task Force Member @ The Association of Korean Medicine, Seoul, Korea* 2025 ∼
- *Advisor to Korean American Semiconductor Professional Alliance (KASPA)* 2026 ∼
- *KFAS-Salzburg Global Leadership Fellow @ Salzburg Global Seminar, Austria* 2024 ∼
- *Adjunct Professor, EE Department @ Sogang University, Seoul, Korea* 2020 ∼
- *Advisory Professor, EECS Department @ DGIST, Korea* 2020 ∼
- Global Advisory Board Member @ Innovative Future Brain-Inspired Intelligence System Semiconductor of Sogang University, Korea 2020 ∼
- Technology Consultant @ Gerson Lehrman Gruop (GLG), NY, USA 2022 ∼

- Co-Founder & CTO / Head of Global R&D / Chief Applied Scientist / Senior Fellow @ Gauss Labs, Inc., Palo Alto, CA, USA 2020 $\sim$ 2023
- VP / Fellow @ SK hynix 2020 $\sim$ 2021
- Senior Applied Scientist @ Amazon.com, Inc., Vancouver, BC, Canada 2017 $\sim$ 2020
- Principal Engineer @ Software R&D Center, Samsung Electronics 2016 $\sim$ 2017
- Principal Engineer @ Strategic Marketing & Sales, Samsung Electronics 2015 $\sim$ 2016
- Principal Engineer @ DT Team, DRAM Dev, Samsung Electronics 2012 $\sim$ 2015
- Senior Engineer @ CAE Team, Memory Business, Samsung Electronics 2005 $\sim$ 2012
- PhD - Electrical Engineering @ Stanford University, CA, USA 2001 $\sim$ 2004
- Development Engineer @ Voyan, Santa Clara, CA, USA 2000 $\sim$ 2001
- MS - Electrical Engineering @ Stanford University, CA, USA 1998 $\sim$ 1999
- BS - Electrical & Computer Engineering @ Seoul National University 1994 $\sim$ 1998

Highlight of Career Journey

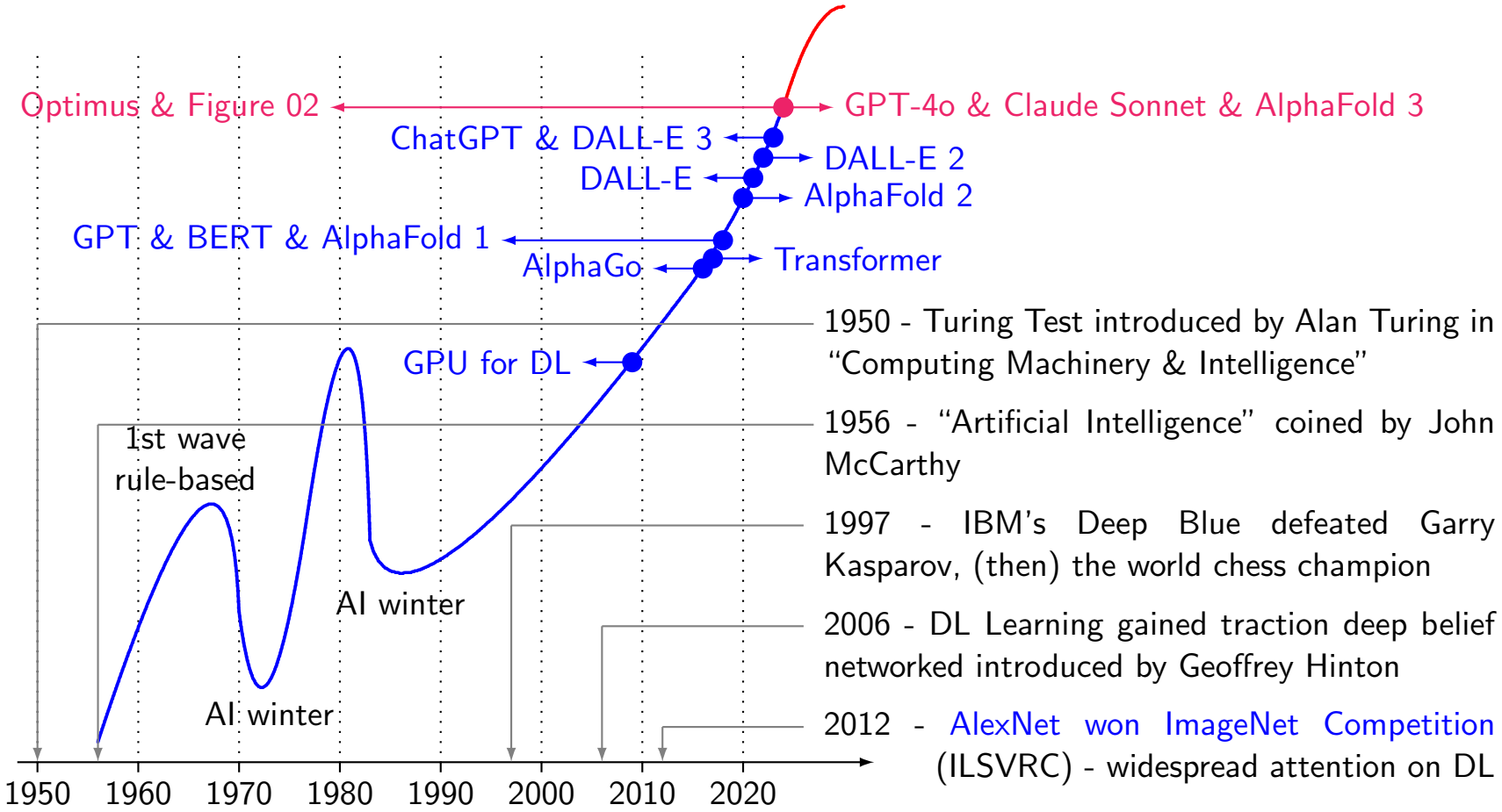
- BS in Electrical Engineering (EE) @ Seoul National University
- MS & PhD in Electronics Engineering (EE) @ Stanford University
  - *Convex Optimization - Theory, Algorithms & Software*
  - Advisor - *Prof. Stephen P. Boyd*
- Principal Engineer @ Samsung Semiconductor, Inc.
  - *AI & Convex Optimization*
  - collaboration with *DRAM/NAND Design/Manufacturing/Test Teams*
- Senior Applied Scientist @ Amazon.com, Inc.
  - *e-Commerce AIs* - anomaly detection, deep RL, and recommender system
  - *Jeff Bezos's project - drove $200M* in sales via Amazon Mobile Shopping App
- *Co-Founder & CTO / Global R&D Head & Chief Applied Scientist* @ Gauss Labs, Inc.
- *Co-Founder & CTO* @ Erudio Bio, Inc.
- *Co-Founder & CEO* @ Erudio Bio Korea, Inc.

Unpacking AI for KFAS Scholarship Program Students

Artificial Intelligence

AI History

History



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 – 2025

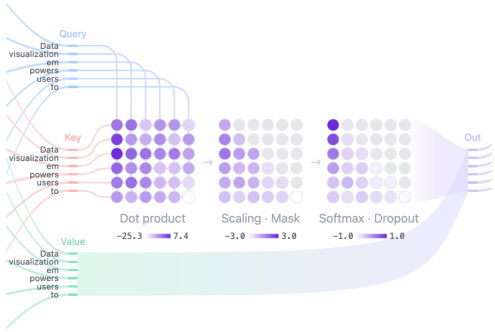
Deep learning revolution

- 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] CV: computer vision, NN: neural network, CNN: convolutional NN, RL: reinforcement learning

Transformer changes everything

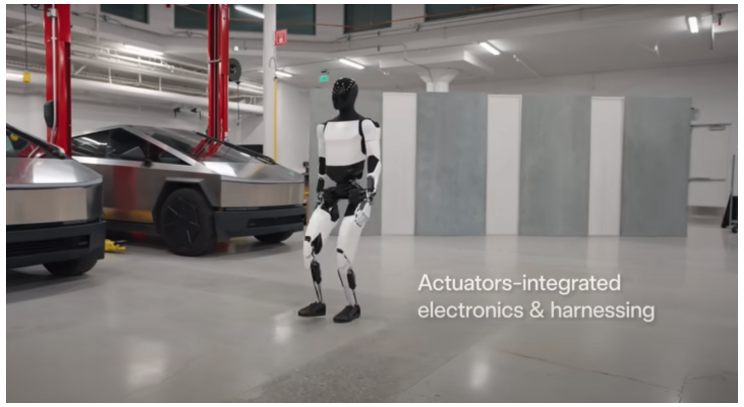
- 2017 – 2018 - Transformers & NLP breakthroughs[2]
  - *Transformer (e.g., BERT & GPT) revolutionized NLP*
  - major advancements in, *e.g.*, machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
  - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting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

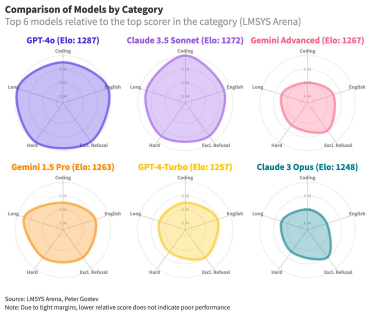
Lots of breakthroughs in AI technology and applications in 2024

- proliferation of advanced AI models
  - GPT-4o, Claude Sonnet, Claude 3 series, Llama 3, Sora, Gemini
  - *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*



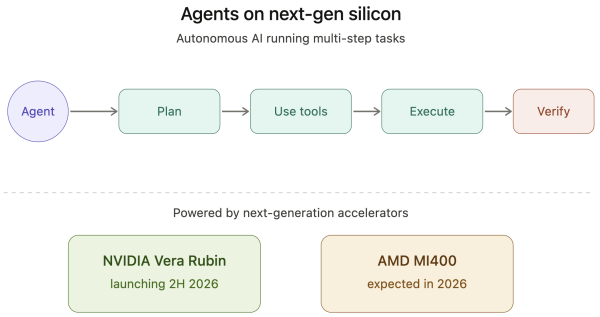
Major AI Breakthroughs in 2025

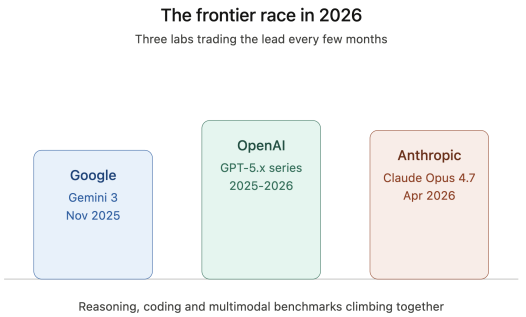
- next-generation foundation models
  - GPT-5 (Aug 2025) and Claude 4 demonstrate strong reasoning abilities
  - open-source models (Llama, DeepSeek, Qwen) closing the gap
- hardware innovations
  - NVIDIA Blackwell Ultra (B300) shipped in late 2025, with Rubin announced for 2026
  - AMD's MI350 series accelerators challenging NVIDIA's market dominance
- AI-human collaboration systems
  - agentic AI going mainstream – systems autonomously executing multi-step tasks
  - multimodal interfaces enabling more natural human-AI collaboration
  - AI systems increasingly explaining their (reasoning) and recommendations



Where AI Is Heading in 2026 (as of May 2026)

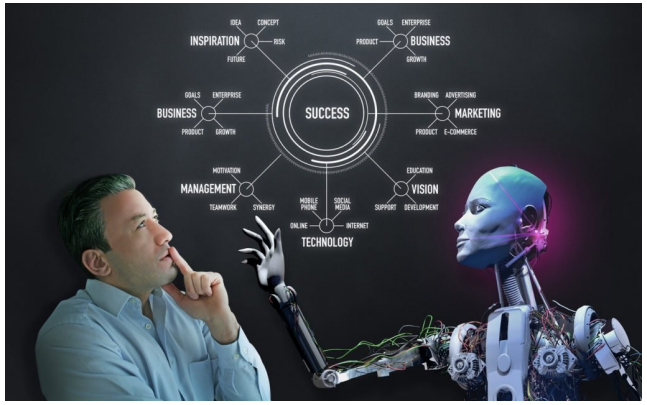
- frontier models reaching new performance ceilings
  - Gemini 3, GPT-5.2 / 5.x series, Claude Opus 4.7 $\rightarrow$ multimodal benchmarks higher
  - intensifying competition among Google, OpenAI, and Anthropic
- hardware scaling and diversification
  - NVIDIA's Vera Rubin platform launching in 2H claiming $\sim$5x faster than Blackwell
  - AMD MI400 series expected in 2026, continuing to challenge NVIDIA's dominance
- agentic AI going mainstream
  - AI agents autonomously executing long-horizon, multi-step tasks
  - expected expansion into enterprise workflows across software, finance, and research
  - growing focus on AI safety, reliability, and self-verification as capabilities scale

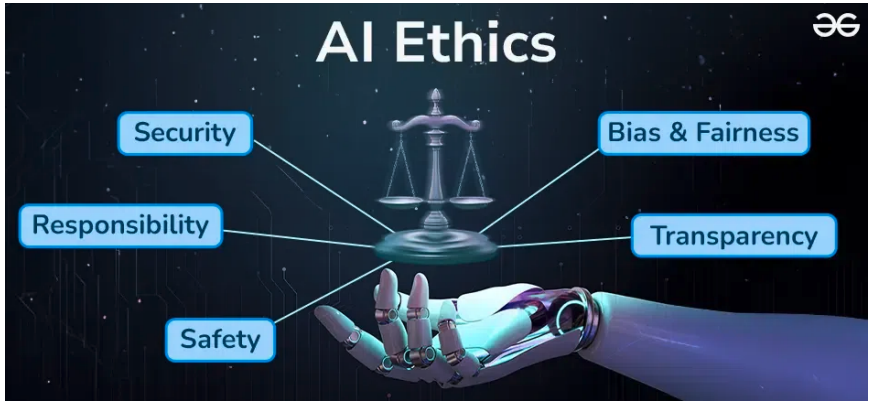




Transformative impact of AI - reshaping industries, work & society

- 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*

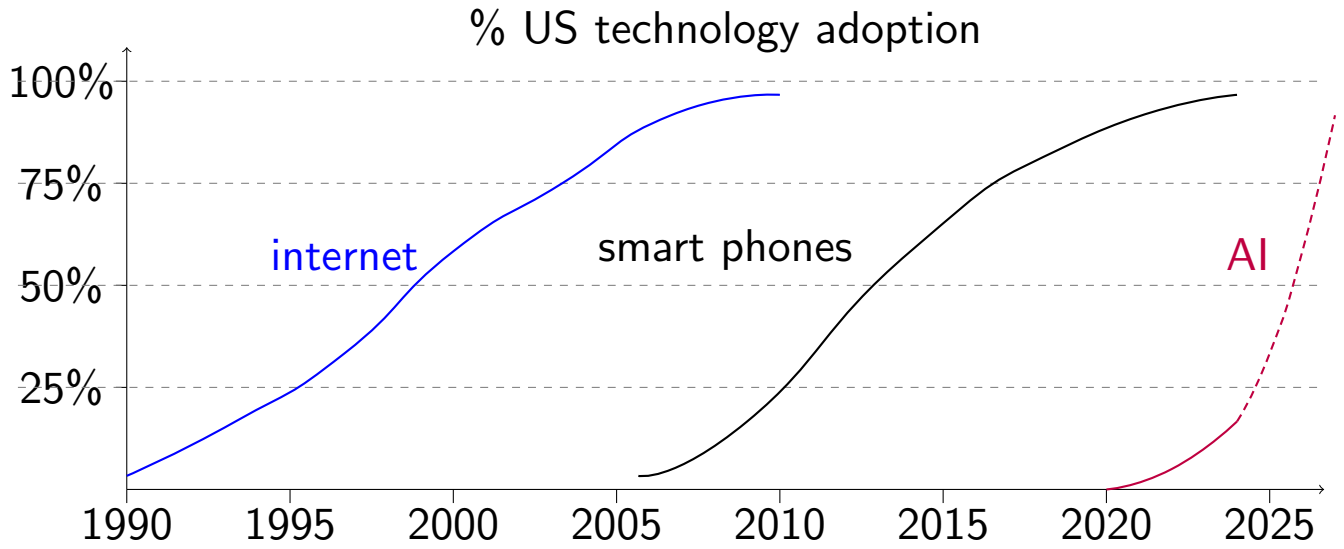




Measuring AI's Ascent

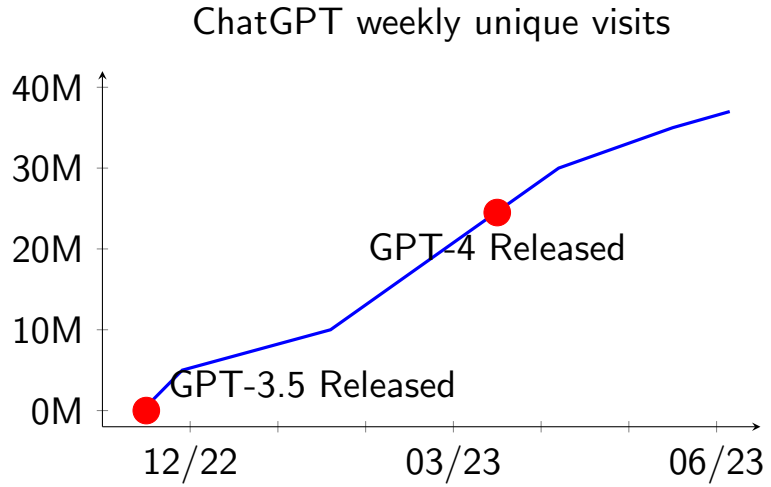
Where are we in AI today?

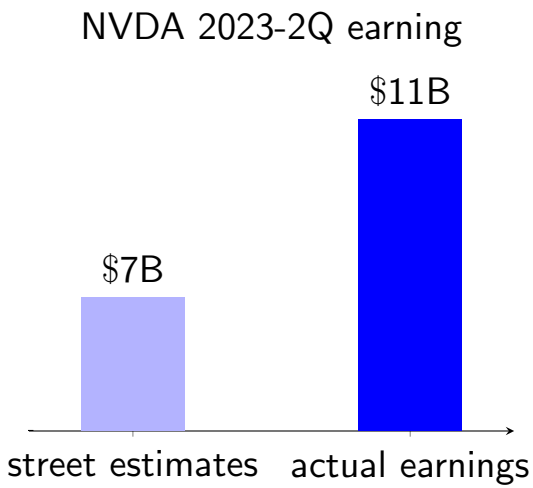
- sunrise phase - currently experiencing dawn of AI era with significant advancements and increasing adoption across various industries
- early adoption - in early stages of AI lifecycle with widespread adoption and innovation across sectors marking significant shift in technology's role in society



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]

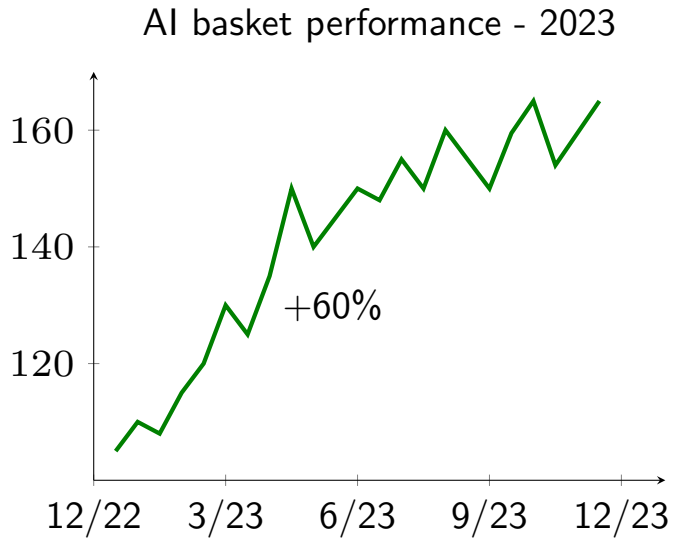


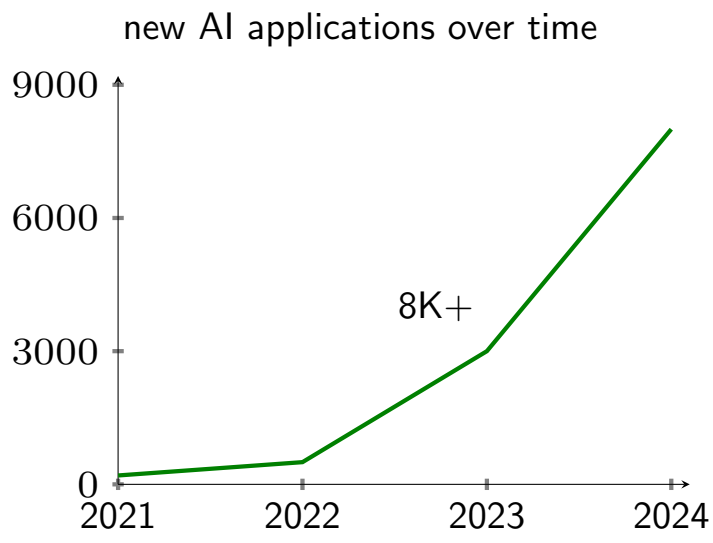


[3] source - Bloomberg

Explosion of AI ecosystems - AI stock market

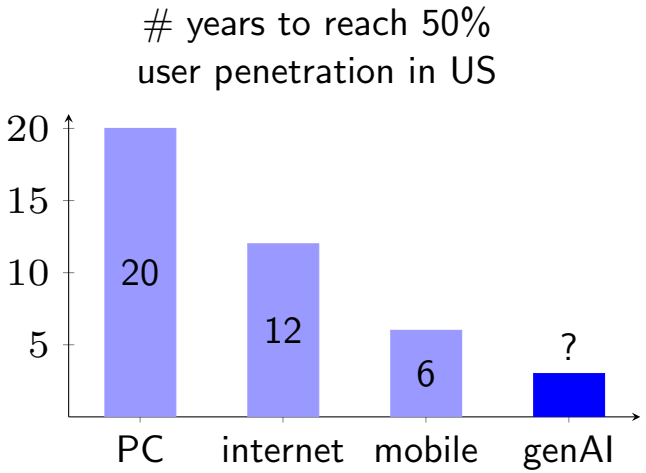
- *AI investment surge in 2023 - portfolio performance soars by 60%*
  - AI-focused stocks significantly outpaced traditional market indices
- *over 8,000 new AI applications* developed in last 3 years
  - applications span from healthcare and finance to manufacturing and entertainment

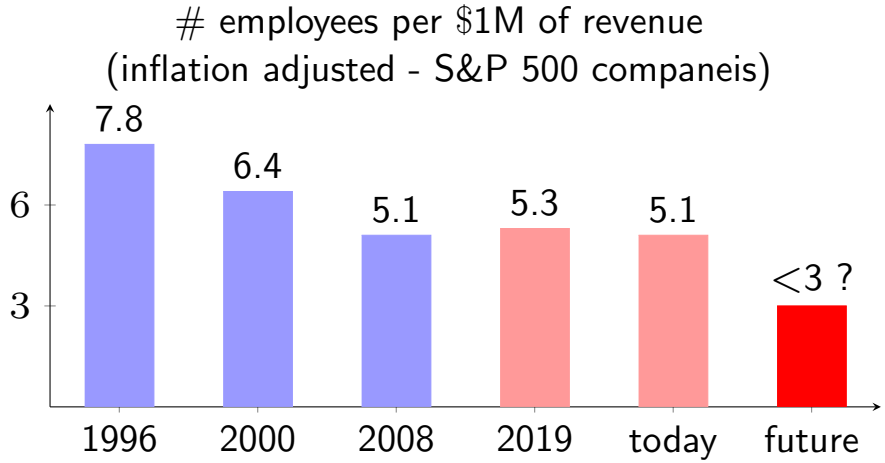




AI's transformative impact - adoption speed & economic potential

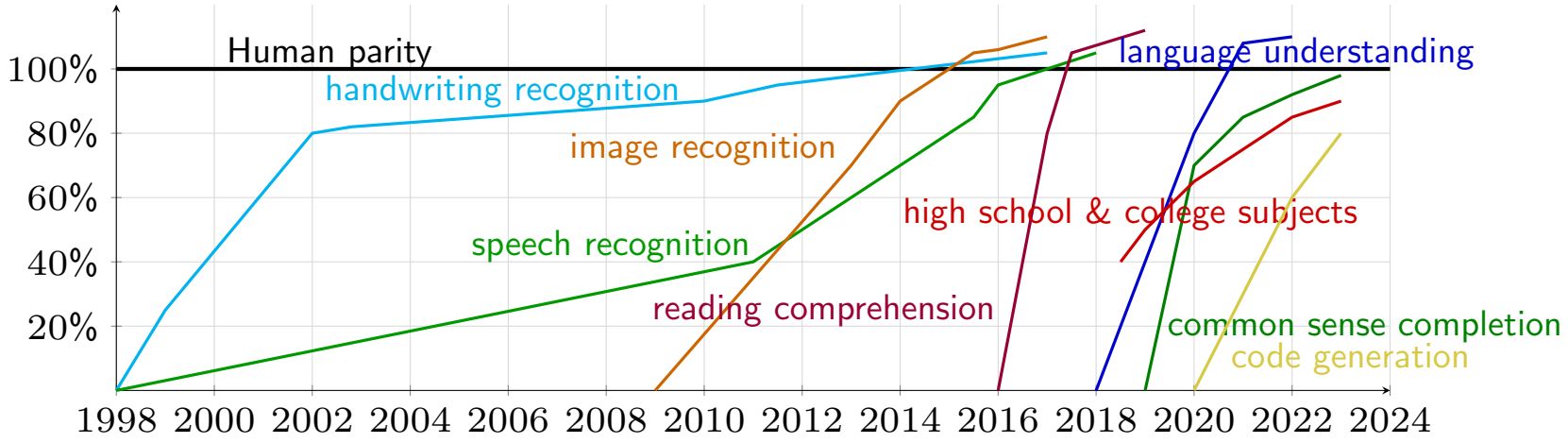
- adoption - has been twice as fast with platform shifts suggesting
  - increasing demand and readiness for new technology improved user experience & accessibility
- AI's potential to drive economy for years to come
  - 35% improvement in productivity driven by introduction of PCs and internet
  - greater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





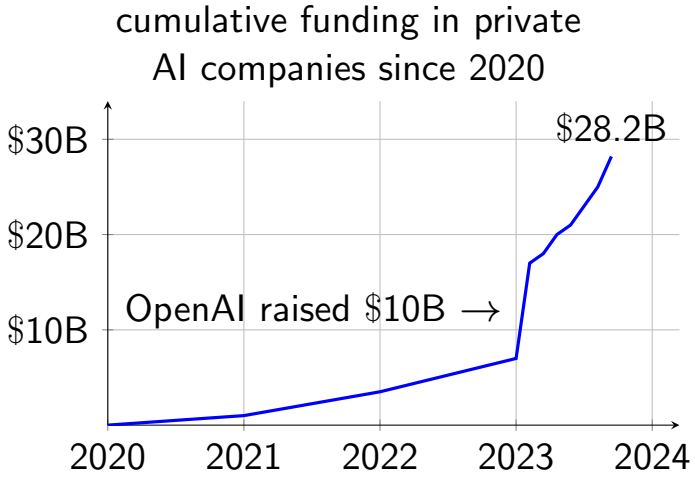
AI getting more & more faster

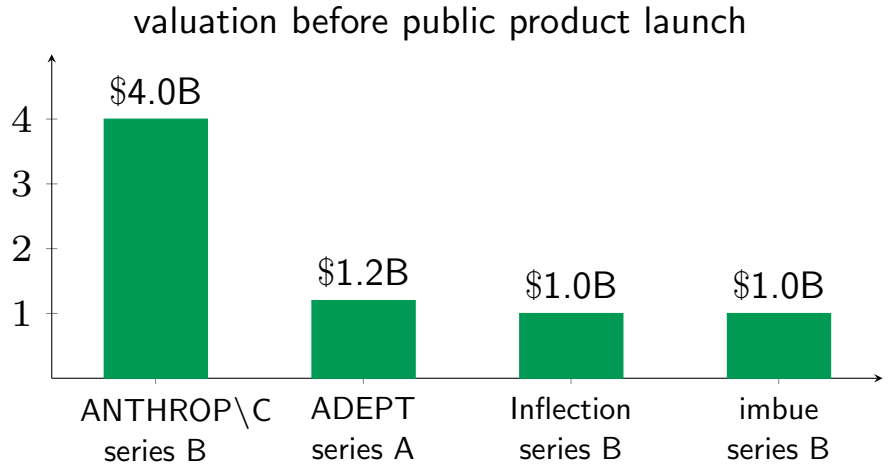
- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached 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

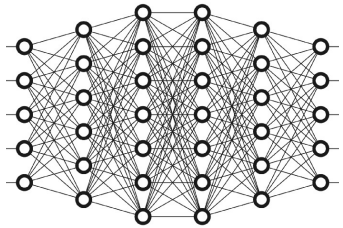


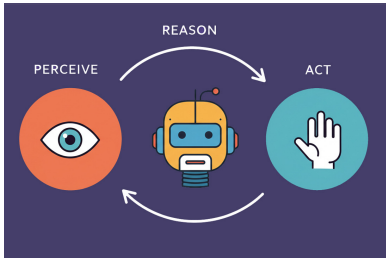


AI Agents

AI progress in 21st century in keywords

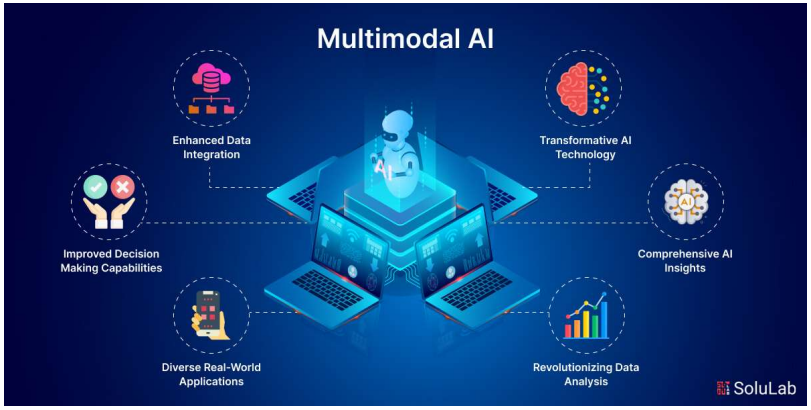
- 2010 $\sim$ Big Data
- 2012 $\sim$ Deep Learning
- 2017 $\sim$ Transformer - Attention is All you need!
- 2022 $\sim$ LLM & genAI
- 2024 $\sim$ AI Agent (Agentic AI)

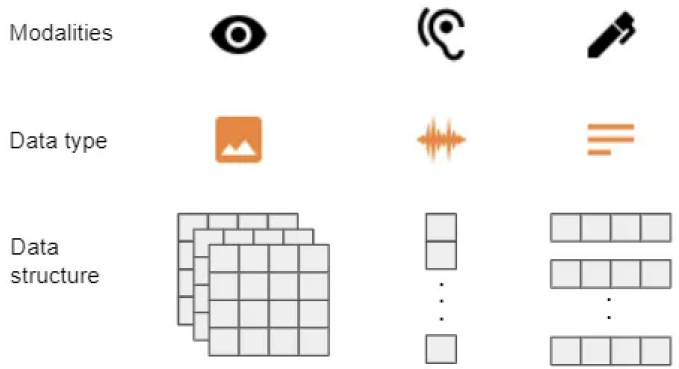




Multimodal learning

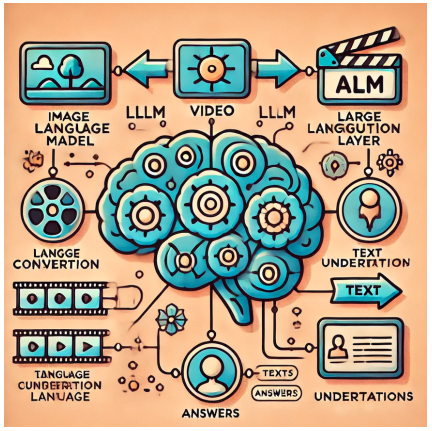
- understand information from multiple modalities, $e.g.$, text, images, audio, video
- representation learning methods
  - combine multiple representations or learn multimodal representations simultaneously
- applications
  - images from text prompt, videos with narration, musics with lyrics
- collaboration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





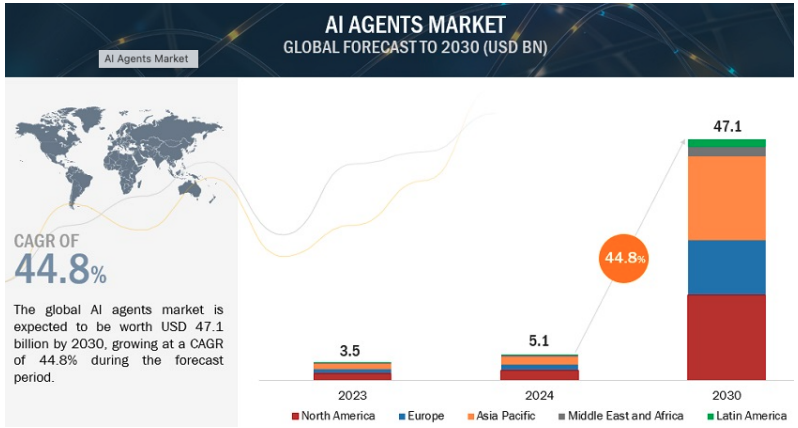
Implications of success of LLMs

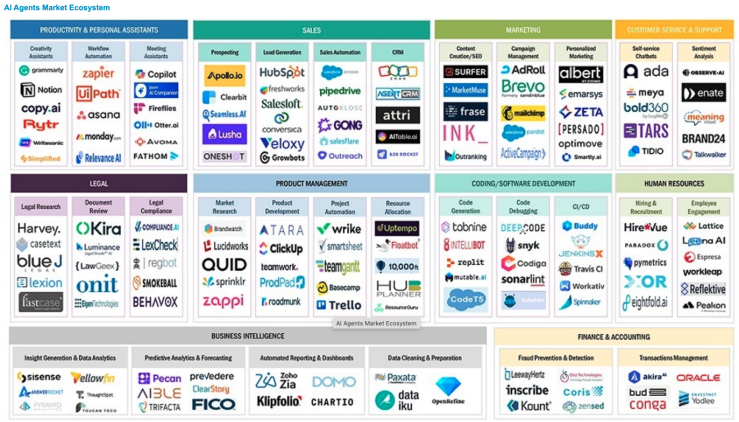
- many researchers change gears towards LLM
  - from computer vision (CV), speach, music, video, even reinforcement learning
- *LLM is not only about NLP . . .* humans have . . .
  - evolved to optimize natural language structures for eons
  - handed down knowledge using *this natural languages* for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



Multimodal AI (mmAI)

- 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



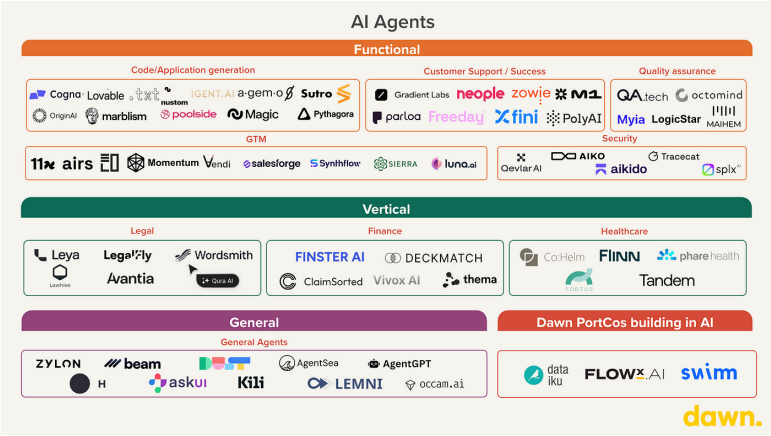


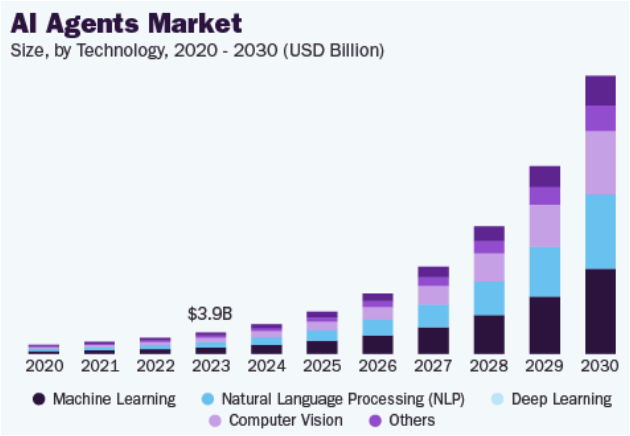
mmAI Technology

- 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

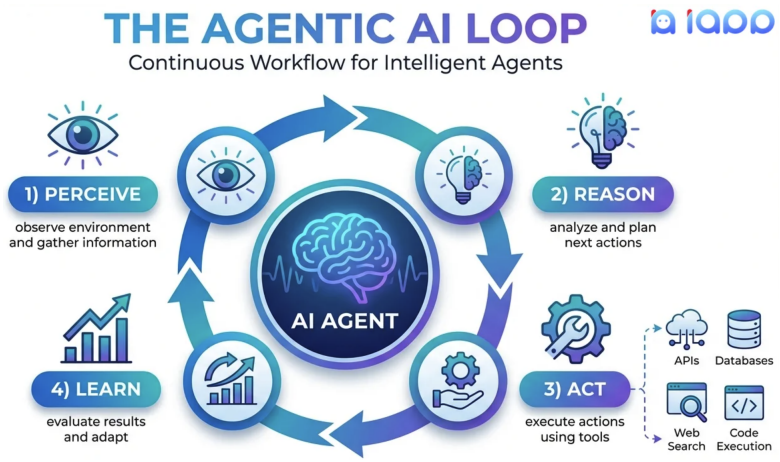
- foundation
  - integrate multimodal AI capabilities for enhanced interaction & decision-making
- components
  - perceive environment through multiple modalities (visual, audio, text), process using LLM technology, generate contextual responses & take actions
- capabilities
  - understand complex environments,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





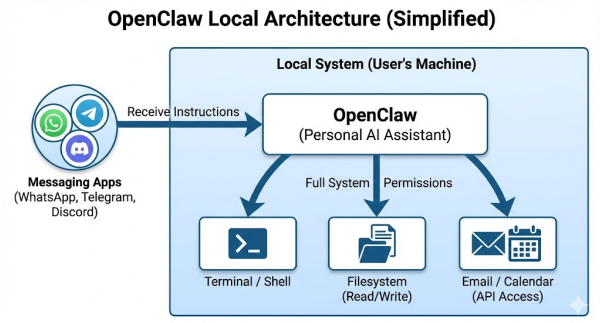
What makes AI "Agentic"?

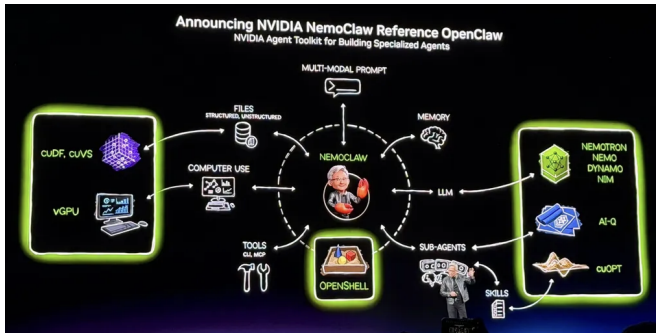
- old AI responds to prompt; agentic AI *pursues goal*
- core loop
  - perceive → plan → (reason) → act → observe → repeat
- four traits - autonomy, tool use, memory/state, long-horizon planning
- enablers
  - tool/function calling, retrieval, code execution, multi-agent orchestration
- shift - *"answer my question" → "accomplish my objective"*



Cutting-edge AI agent tools - open source

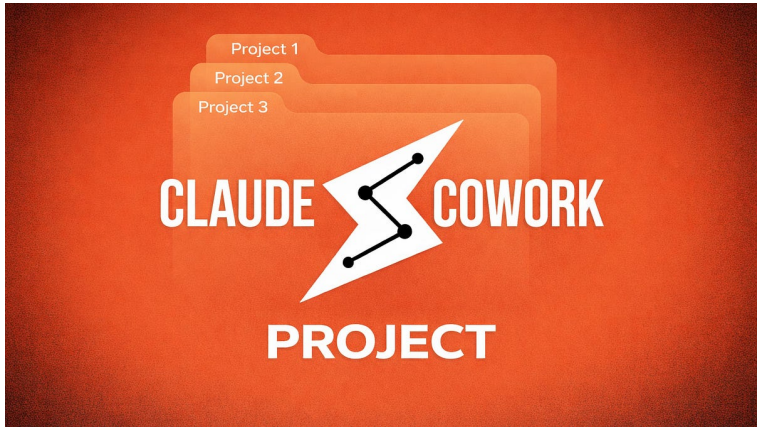
- OpenClaw (Peter Steinberger) - open-source, runs locally, connects LLMs to real software
  - reads/writes files, runs shell commands, browses web, sends email, controls APIs
  - 350k+ GitHub stars (by May 2026) — most-starred GitHub software project
  - skill-based architecture - SKILL.md folders, shareable on ClawHub
  - works through chat apps - Slack, Telegram, WhatsApp, Discord, iMessage, *etc.*
  - model-agnostic - Claude, GPT, Gemini, or local via Ollama
- NVIDIA NemoClaw - security/privacy layer on top of OpenClaw
  - one-command install of Nemotron models + OpenShell secure runtime
  - network & filesystem isolation, local inference so no data leaves the device KKR

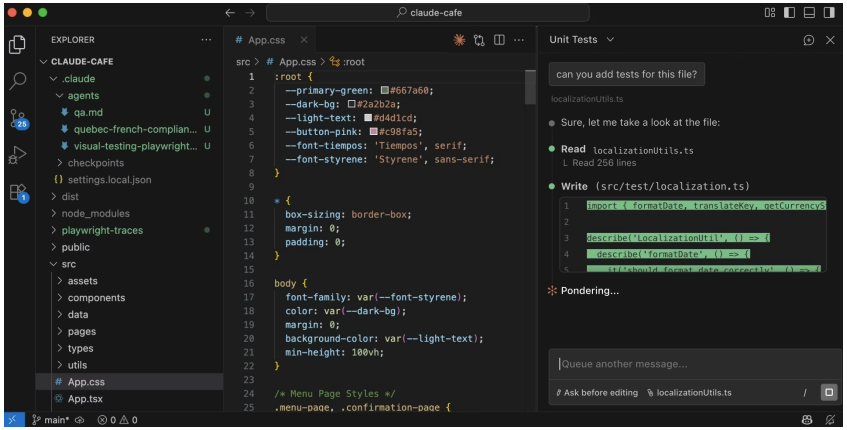




Cutting-edge AI agent tools - Anthropic

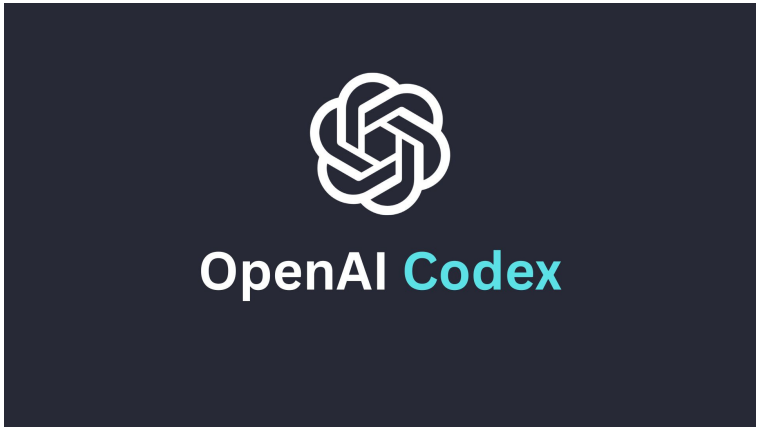
- Claude code
  - CLI/IDE coding agent; subagents, hooks, plugins, auto mode, routines
- Claude cowork
  - desktop tab; file-system access, scheduled recurring tasks, plugin marketplace
- managed agents
  - multi-agent orchestration; cloud-deployable agent templates TrendForce
- vertical bundles already shipping
  - legal, small business, marketing ops, finance

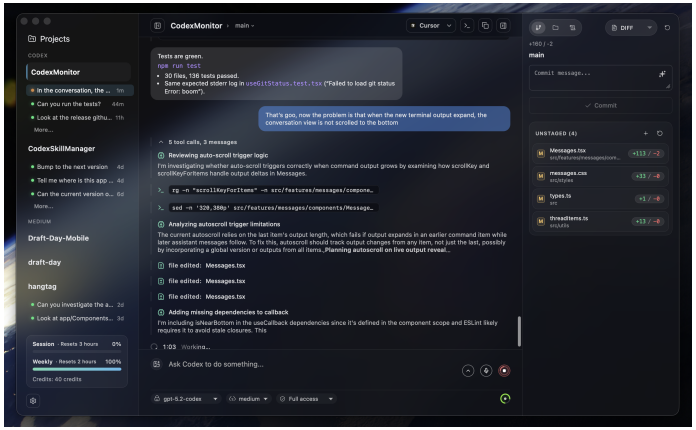




Cutting-edge AI agent tools - OpenAI

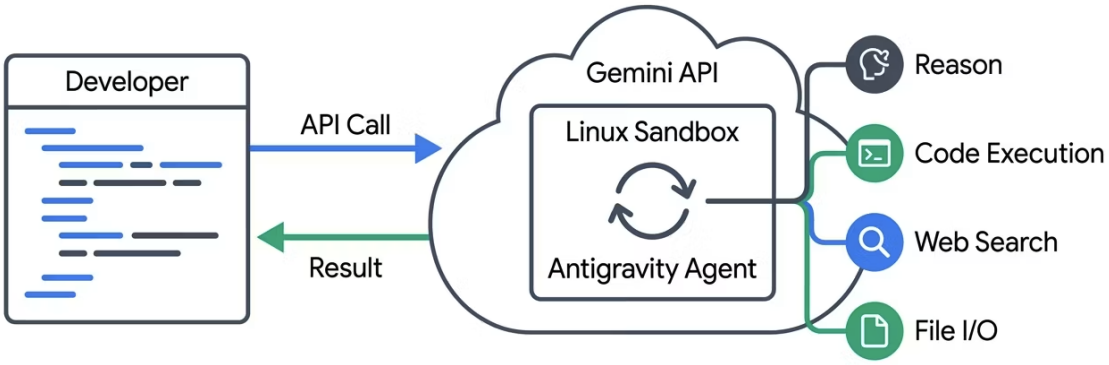
- OpenAI Codex
  - agentic coding tool - CLI, IDE, ChatGPT, desktop & now mobile
  - 2026 shift - from code editor → full "agent workspace"
  - multi-agent parallelism - runs several tasks in separate sandboxes while you review
  - powered by GPT-5.5 - tightly coupled, not model-agnostic (unlike Claude Code / OpenClaw) InfoQ
- Codex Security
  - dedicated agent that finds & fixes vulnerabilities





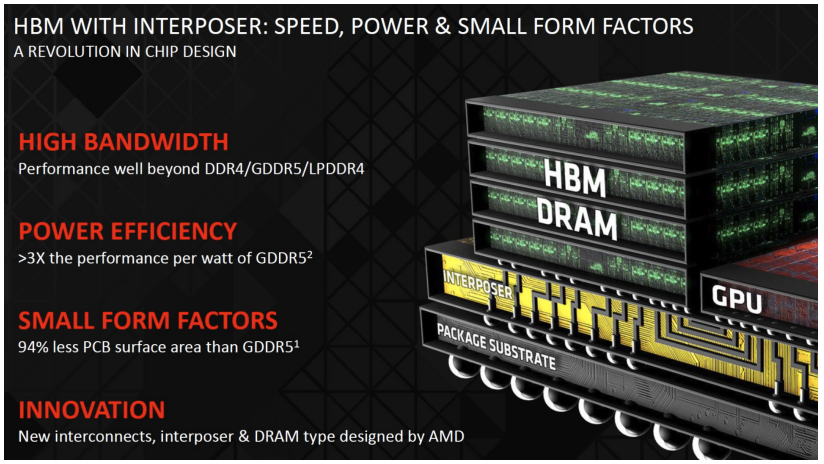
Cutting-edge AI agent tools - Google

- Antigravity 2.0
  - agent-first development platform; desktop app + CLI + SDK NVIDIA
- Gemini API Managed Agents
  - one API call spins up agent that reasons, uses tools, executes code
- Jules
  - AI agent for GitHub - debugging, pull-request prep NVIDIA Newsroom
- Gemini Spark
  - 24/7 personal agent on Gemini 3.5 Flash, wrapped in Antigravity
  - connects to Canva, OpenTable, Instacart, Workspace via MCP NVIDIA Blog



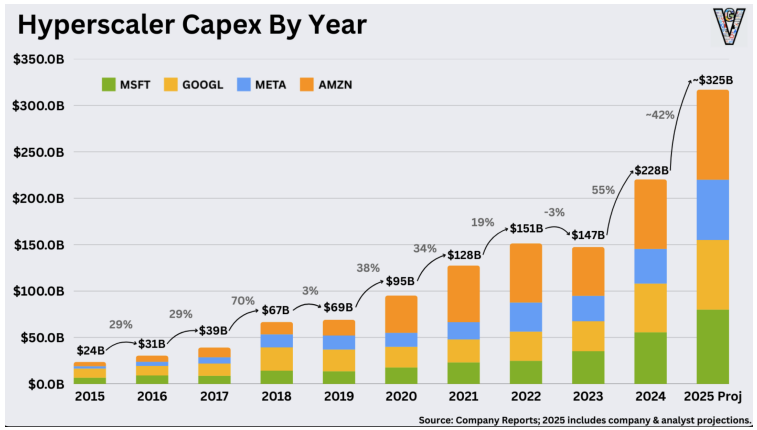
Agentic stack - LLM is engine, but not whole system

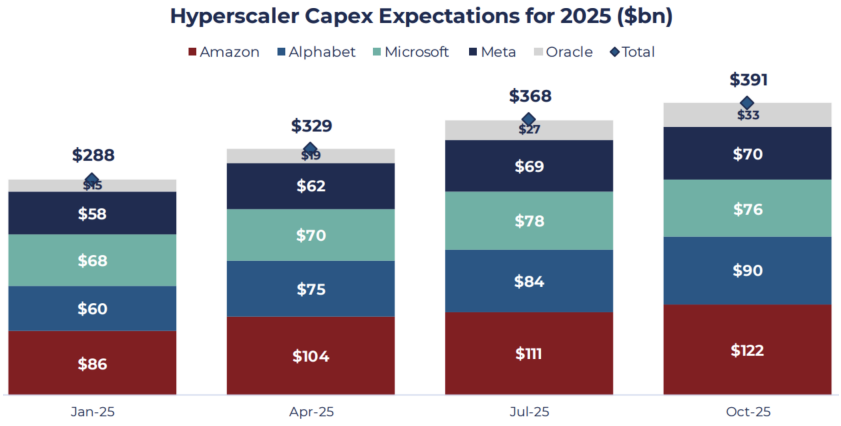
- *LLM - reasoning engine, not the system*
- stack
  - planner/orchestrator, memory (short/long-term), tools/APIs, environment interface
- patterns
  - ReAct, reflection/self-critique, planner–executor, multi-agent
- interoperability protocols emerging, $e.g.$, MCP, agent-to-agent
- *value migrating from model → system design*



Trillion-dollar gravity well - capital, talent, geopolitics

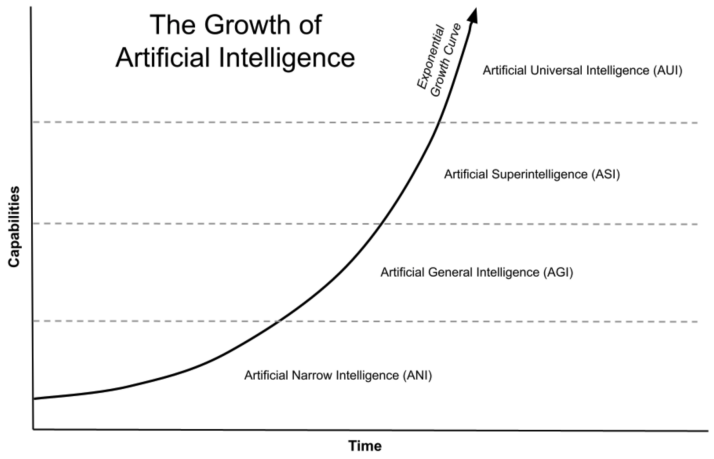
- big 5 hyperscalers $\sim$ \$725B AI capex in 2026 $\sim$ Switzerland's GDP
- trajectory - \$256B (2024) $\rightarrow$ \$443B (2025) $\rightarrow$ \$725B (2026)
- *2026 is the first trillion-dollar year of compute capex in history*
- \$6.7T global data-center capex by 2030 ($\sim$70% AI) (McKinsey forecasting)
- *geopolitics*
  - export controls, chip sovereignty, national AI budgets

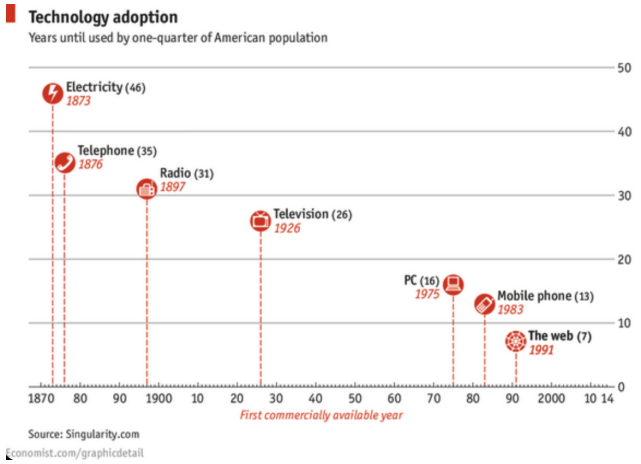




What took decades now takes months

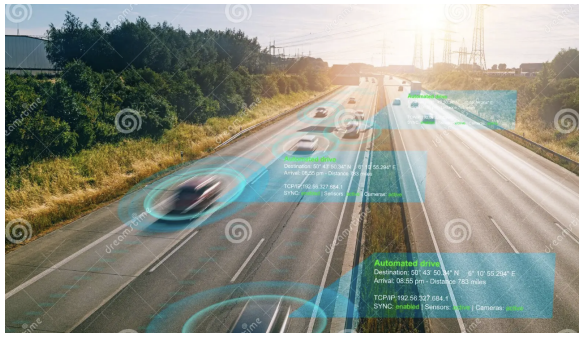
- AlexNet (DL) → AlphaGo → Transformer → GPT (LLM) → Agentic in a decade
- adoption collapsing too
  - genAI penetration in fraction of PC/internet time
- *frontier you train on today will move by graduation*
- durable skill ≠ any one tool - it's relearning the frontier
- *what previously took decades now compresses into months!*





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

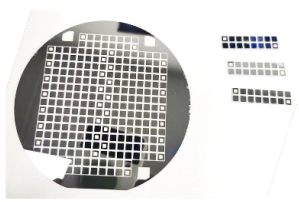


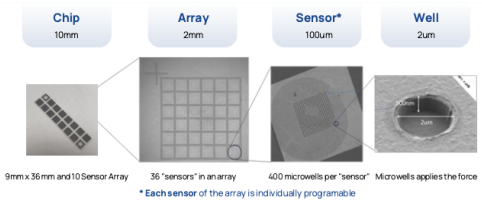
Appendix

Erudio Bio

Powering AI-driven medicine with ground-truth binding data

- problems we solve
  - 90% of drugs fail in clinical trials due to poor early-stage prediction
  - multiplexed diagnostics suffer from false positives and cross-reactivity
- *Erudio Bio's Innovation*
  - *VSA* platform uses patented *"dynamic force spectroscopy"* to generate 1000x more high-quality binding data from single sample ($\sim 10\mu$L)
  - measuring not just presence, but *strength* and *kinetics* of molecular interactions
- *dual business model*
  - diagnostics - *multi-cancer biomarker detection* with clinical institutions & hospitals
  - *drug discovery - bioTCAD$^{\mathrm{TM}}$ platform* providing ground-truth labels to train & validate pharma AI models, reducing preclinical cycles





Validated technology, proven team, clear path to market

- validated impact
  - *$1M Gates Foundation Grant* (2025) to democratize drug development for global health
  - partnerships with top research institutions (KRIBB, KAIST)
- unique team - *Stanford-trained founders* combining
  - semiconductor TCAD expertise & force spectroscopy innovation (20+ years)
  - AI & optimization leadership (Samsung, Amazon, SK hynix, Gauss Labs)
- market entry
  - *Korea → (Asia hub &) US* strategy with 2026 regulatory milestones
  - expanding *pharma partnerships and B2G*

Gates Foundation

Erudio Bio

반도체 첨단정밀과학으로 의료진단의 혁신을 이끄는 생명공학 기업

Biological Assays Struggle with Scale & Accuracy

Data is expensive

- so we make decisions with *incomplete* picture
- status quo
  - limited, small-scale testing confirms diagnosis
  - outcome only as good as doctor's ability to determine which tests, limiting the picture
  - cross reactivity prevents larger scale testing
- Erudio creates
  - *comprehensive, large-scale* testing will drive diagnosis without assumptions
  - increased scale enables enhanced scientific discovery leading to
    - *better patient care*
    - *reduced time to diagnosis*
    - *cost reduction*



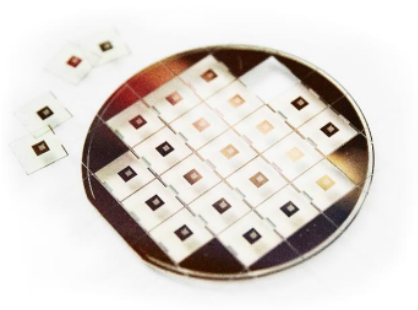

Erudio Bio starts revolution with Gates Foundation's support

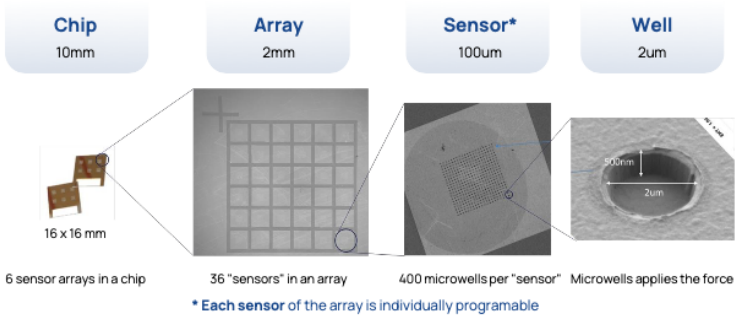
- more data
  - comprehensive data from *single biological sample*
  - *multiplexed analysis* of nucleic acid, protein, cells, and more!
  - *multi-omic platform*
- actionable data
  - combined quality score from all data sources for comprehensive & conclusive assessment
- earlier data
  - complete data early to drive accurate decision making



Versatile Smart Assay (VSA) Platform

VSA technology





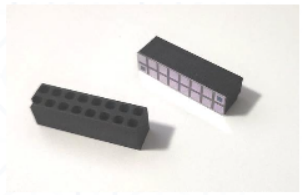
- generates *1000x more data* than the prevailing technology
  - scalable multi-omic microarray sensor
- *21 patents* in US, Canada, EU, and China
- indicates how good the data is in real time
  - patented "dynamic force spectroscopy" and "powerful Bayesian inference" method provides our data *quality score* to know their accuracy for actionable data
- AI software extracts a detailed, interpretable picture for quick diagnosis
  - leads to *AI knowledge discovery* resulting in *data-driven diagnosis*

Enabling comprehensive data acquisition

- antibodies - versatile tools in biology
  - can engineer to target virtually *anything* we want
  - problem
    - indiscriminate interactions severely limits use of antibodies – *cross-reactivity*
    - error-prone results due to *non-specific binding*
- solution - comprehensive data with *dynamic force spectroscopy*
  - comprehensive binding strength to distinguish specific from non-specific binding
  - *quality score* discerns noise from useful data to enable multiplexing



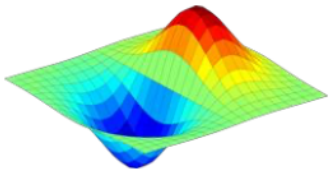

VSA's business models



Consumable chip & flowcell

Instrument

Consumable reagent kit



Software AI/ML & SaaS

- VSA platform
  - instrument - recurring revenue with high margin
  - modular licensable software - AI based data interpretation and feature extraction
- SaaS
  - subscription based pre-validation of reagent database
  - AI feature extraction and knowledge discovery

When Erudio's VSA meets AI - Gates Foundation Grant

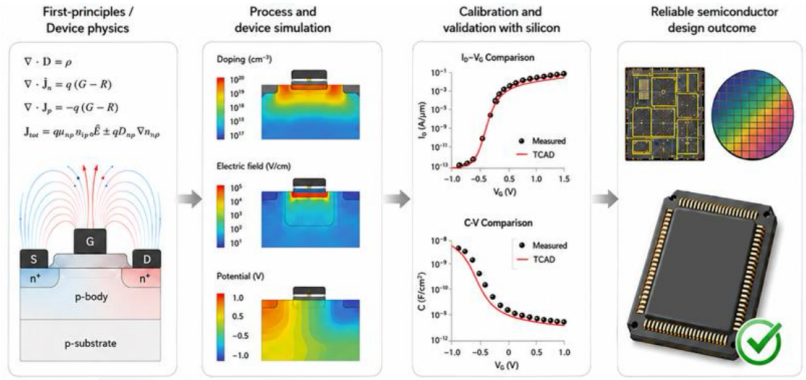
Erudio Bio wins $1M Gates Foundation Grant - scaling bioTCAD

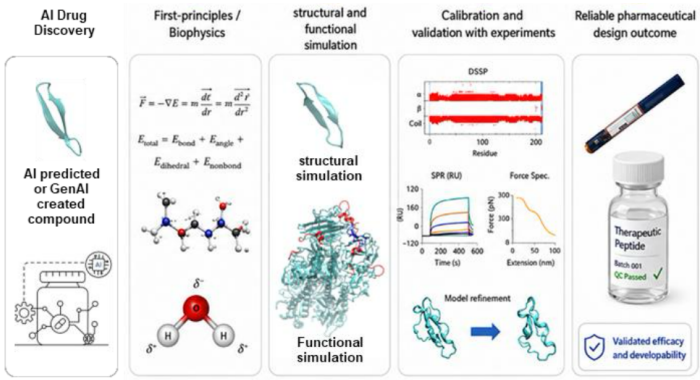
Gates Foundation

- *$1M Grant Award (2025)*
  - Gates Foundation recognizes Erudio Bio's potential to transform drug development for global health
- mission alignment - democratizing medicine by making preclinical drug design faster, yet reliable & accessible
  - lowering development costs for diseases affecting LMICs
  - addressing the 90% clinical trial failure rate that drives up drug costs
- funded project - develop *bioTCAD$^{\mathrm{TM}}$ platform for lead optimization of drug discovery*
  - expand force spectroscopy measurements across high-burden disease targets
  - train AI models with kinetics-resolved binding data (on/off rates, unbinding forces)
  - *enable pharma/biotech to prioritize candidates earlier with higher confidence*

bioTCAD - hybrid AI & physics-based drug development

- (old school) AI/ML excels at pattern recognition across large chemical libraries
  - identifying candidate hits at scale – valuable role established and growing
- AI alone *cannot reliably explain binding physics or predict behavior in unexplored and novel chemical space* – *capability and credibility gap*
- bioTCAD combines *AI and measurement-backed, physics-grounded simulation*
- applies *the* principle that made semiconductor TCAD trustworthy
  - validate model parameters to YOUR experimental measurements to be reliably right





Erudio Bio's Business Models

Erudio Bio Applications

- drug development
- clinical diagnostic
  - medicine is already 20% of the world's economy and growing at 5% per year
- biodefense
  - GWOT to great power competition
  - need to defend against near-peer adversaries
  - flexible, efficient solution needed from CBRND to readiness

Teams

Team & advisory board

- team
  - Kee-Hyun Paik, Ph.D. (CEO) - chip, microfluidics, instrumentation
  - Sunghee Yun, Ph.D. (CTO) - AI, optimization, business development, software
  - Susanne Baumhueter, Ph.D. - biology, immunology, project management
- advisory board
  - Michael Cola - CEO of AEVI Genomic Medicine ($62B sales to Takeda)
  - Tim Germann - CCO of Carterra Bio
  - Karyn Eliot - retired CIA Sr. Executive
  - Phil Ferro - virologist, formerly DoD, DoS, HHS and White house
  - Bill Chen - Former hedge fund and VC professional with national security, FFRDC
  - Ronald W. Davis - Director of Stanford Genome Tech Center ($15B+ exits)
  - Michael Snyder - Prof. Genetics, Dir. Stanford Human Genome Center
  - William J. Greenleaf - Prof. Genetics and Applied Physics, Stanford University


SAMSUNG
ELECTRONICS


amazon


SK hynix


Stanford
Computer Science

Some Important Questions around AI

Some important questions around AI

- why human-level AI?
- what lies in very core of DL architecture? what makes it work amazingly well?
- biases that can hurt judgement, decision making, social good?
- AI ethics & legal issues
- consciousness
- utopia vs dystopia
- knowledge, belief, reasoning
- risk of anthropomorphization

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, . . .

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

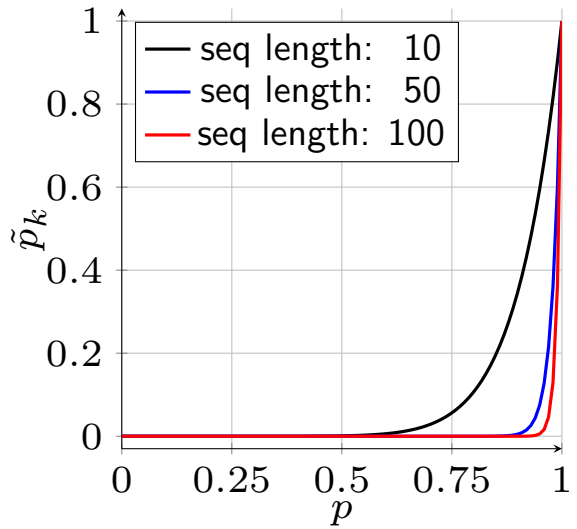
$$\rho_i = \frac{\tilde{p}_i - \tilde{p}_{i-1} p_i}{\sqrt{\tilde{p}_{i-1}(1-\tilde{p}_{i-1}) p_i (1-p_i)}}$$

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

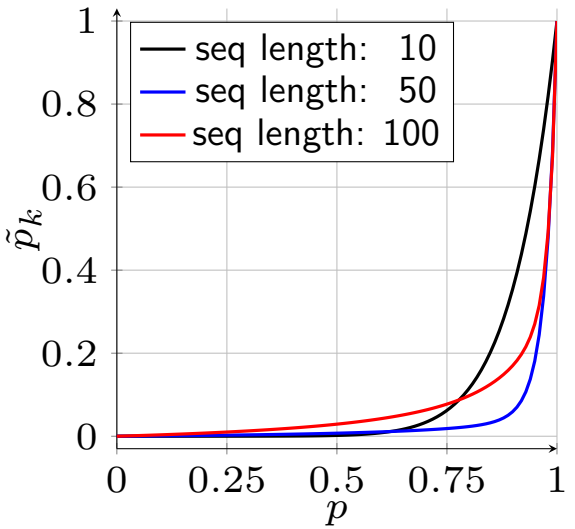
Dramatic improvement of LLM near saturation

- 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



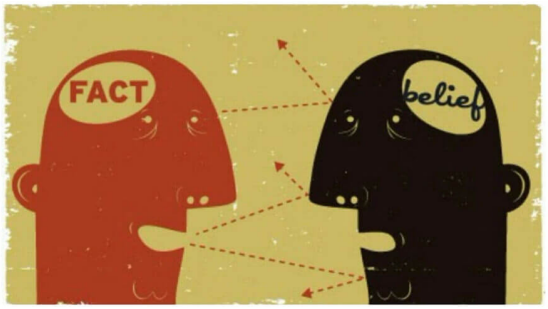
when tokens are dependent

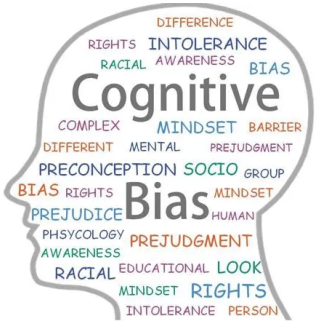


Biases

Cognitive biases attributed to humans

- 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





Biases of LLMs

- LLMs subjec to
  - 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

AI Ethics

Ethical issues related to AI

- 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

AI related Legal Issues

Legal issues with ethical consideration

- 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 & humanitarian 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?
  - can AI have *desire* to survive (or save earth)?

Utopia vs Dystopia

Utopia vs dystopia

- 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, immoral, and unethical) people possessing social, economic & political power
  - *social good and welfare impaired* by either exploiting AI or ignorance of (inner workings of) AI
- should concern
  - choice or balance among utilitarianism, humanitarianism & values
  - amend or improve laws/regulations
  - ethical issues caused by AI

Knowledge, Belief, and Reasoning

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

What categories of questions do they belong to?

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

LLMs . . .

- 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

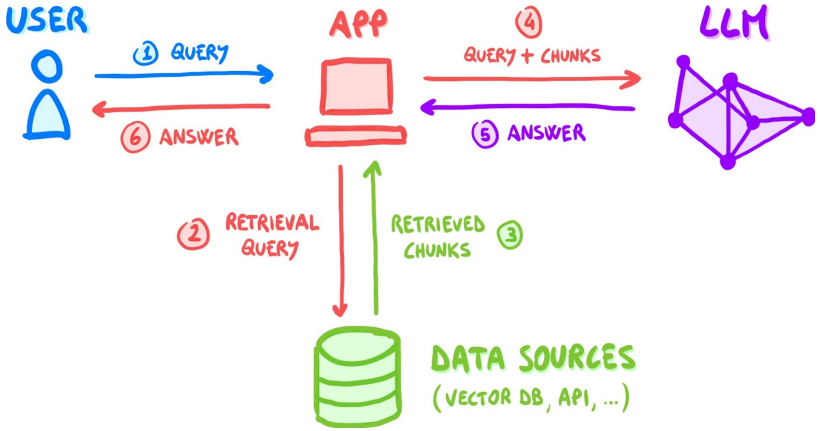
  - making vulnerable to anthropomorphism
- examine it by throwing questions such as
  - *"does LLM have knowledge and belief?"*
  - *"can it reason?"*

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 vs 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, *e.g.*, for question-answering, news article summarization, screenplays generation, language translation



How ChatBot works?

- 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 perspectives
  - cognitive scientific & philosophical perspectives

Laymen's perspectives on knowledge, belief & reasoning

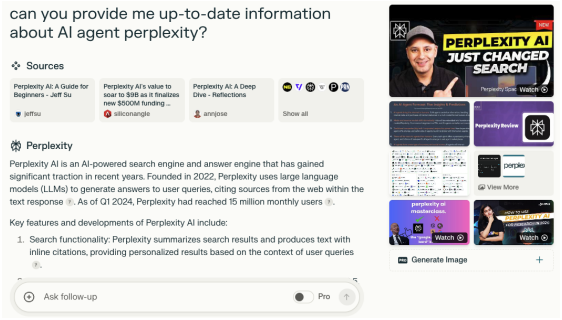
- does (good) LLM have knowledge?
  - Grandmother: looks like it cuz when instructed *"explaing big bang"*, it says
    " The Big Bang theory is prevailing cosmological model that explains the origin and evolution of the universe. . . . 13.8 billion years ago . . . "
- does it have belief?
  - Grandmother: 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!

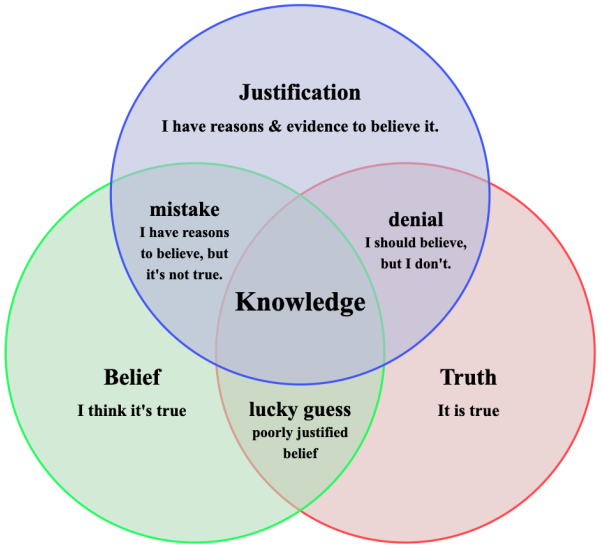
Belief

- nothing can count as *belief about the world* we share unless
  - 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*!



Knowledge in philosophical and cognitive scientific sense

- does LLM have knowledge?
  - Sunghee: *I don't think so!*
- why?
  - we say we have "knowledge" when
    "we do so against ground of various human capacities that we all take for granted when we engage in everyday conversation with each other."
  - 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 words following "Tom Cruise's mother is."
  - so *we cannot say it really knows the fact!*



Belief in philosophical and cognitive scientific sense

- for the discussion
  - do *not* concern any specific belief
  - but concern *prerequisites for ascribing any beliefs to AI system*
- so does it have belief?
  - nothing can count as belief about the world we share unless

    it is against ground of the ability to update beliefs appropriately in light of evidence from that world, essential aspect of the capacity to distinguish truth from falsehood
  - LLM does not have this ground, essential consideration when deciding whether it *really* had beliefs.
- Sunghee: so *no, LLM cannot have belief!*

Reasoning in philosophical and cognitive scientific sense

- 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. . . . "
- however, remember, once again, what we just asked it to do is *not* "deductive inference"
  given the statistical distribution of words in public corpus, what words are likely to follow the sequence, "humans are immortal and Socreates is human therefore."
- Sunghee: so *no, LLM cannot reason, either!*
- but, LLM
  - pretends to reason, and from which capabilities, we can benefit!
  - also, 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 for most people to accept (after understanding) previous arguments!*
- maybe, o.k. for laymen to (mistakenly) anthropomorphize LLM(-embedded systems)
- however, *imperative for (important, smart, and responsible) AI researchers, scientists, engineers & practitioners* to have rigorous understanding in these aspects especially when
  - advise and be consulted by law makers, policy makers, journalists, and various stakeholders responsible for *critical business decisions (in private sectors) and public policies (in public sectors)*
  - collaborate with or/and help professionals in liberal arts, such as *philosophy, ethics, law, religion, literature, history, music, cultural studies, psychology, sociology, anthropology, political science, economics, archaeology, linguistics, media studies, natural sciences, fine arts, . . .*

  — to address negative soceital and economic impacts

Moral

- AI shows incredible utility and commercial potentials, hence should
  - make informed decisions about trustworthiness and safety
  - avoid ascribing capacities they lack
  - *take best utilization of remarkable capabilities of AI*
- today's AI 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, law & policy makers, journalists, . . .*

Selected References & Sources

Selected references & sources

- Robert H. Kane "Quest for Meaning: Values, Ethics, and the Modern Experience" 2013
- Michael J. Sandel "Justice: What's the Right Thing to Do?" 2009
- Daniel Kahneman "Thinking, Fast and Slow" 2011
- Yuval Noah Harari "Sapiens: A Brief History of Humankind" 2014
- M. Shanahan "Talking About Large Language Models" 2022
- A.Y. Halevry, P. Norvig, and F. Pereira "Unreasonable Effectiveness of Data" 2009
- A. Vaswani, et al. "Attention is all you need" @ NeurIPS 2017
- S. Yin, et. al. "A Survey on Multimodal LLMs" 2023
- Chris Miller "Chip War: The Fight for the World's Most Critical Technology" 2022

- CEOs, CTOs, CFOs, COOs, CMOs & CCOs @ startup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California, USA

References

References

[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