

**[Silicon Valley AI Nexus x KASC Colloquium Keynote]
Tech Sovereignty and the Korea-US Alliance**

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 Group (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 ~ 2023
- VP / Fellow @ SK hynix 2020 ~ 2021
- Senior Applied Scientist @ Amazon.com, Inc., Vancouver, BC, Canada 2017 ~ 2020
- Principal Engineer @ Software R&D Center, Samsung Electronics 2016 ~ 2017
- Principal Engineer @ Strategic Marketing & Sales, Samsung Electronics 2015 ~ 2016
- Principal Engineer @ DT Team, DRAM Dev, Samsung Electronics 2012 ~ 2015
- Senior Engineer @ CAE Team, Memory Business, Samsung Electronics 2005 ~ 2012
- PhD - Electrical Engineering @ Stanford University, CA, USA 2001 ~ 2004
- Development Engineer @ Voyan, Santa Clara, CA, USA 2000 ~ 2001
- MS - Electrical Engineering @ Stanford University, CA, USA 1998 ~ 1999
- BS - Electrical & Computer Engineering @ Seoul National University 1994 ~ 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 for AI Nexus Scholarship Program Students

● Tech Sovereignty and the Korean-US Alliance	- 5
– The Broken Assumptions	- 6
– Last Friday in San Francisco	- 10
– Distributed Tech Sovereignty	- 14
– To the Bridge Builders	- 18
● Appendices	
– Silicon Valley AI Nexus	- 21
– Artificial Intelligence	- 32
– AI history & recent significant achievements / market indicators	
– AI Agents	- 49
– big data → ML/DL → LLM & genAI → agentic AI	
– LLM as extremely effective knowledge-transfer representation learner	
● Selected references	- 65
● References	- 67

Tech Sovereignty and the Korean-US Alliance

The Broken Assumptions

Tech sovereignty - textbook definition

- standard definition – a nation's ability to access, shape, and trust the critical technologies its security, economy, and society depend on
- why it became THE question? – AI, semiconductors, and biotech are now *strategic assets*, not mere industries
 - compute, memory, energy, and data play the role that oil, steel, and shipping lanes played in the last century
- textbook trilemma – every nation must choose
 - build domestic capability – slow & ruinously expensive
 - align with allies – powerful, but interdependent
 - accept dependence – cheap today, hostage tomorrow
- *Tonight's claim – this framing hides an assumption!* that the *unit* of sovereignty is one nation
 - “let me show you why that assumption is broken – with evidence”

Nobody owns the stack - the evidence

- walk one chip through its life – no single nation appears at every layer
 - design software (EDA) – Synopsys & Cadence (+ Siemens EDA): an effective US oligopoly; every advanced chip on earth is designed with American CAD tools
 - lithography – EUV = ASML, the Netherlands: one company, one country, no alternative
 - equipment – Applied Materials, Lam, KLA (US) & Tokyo Electron (Japan)
 - fabrication – TSMC, Samsung, SK hynix: manufacturing excellence lives in Asia, where manufacturing culture runs deepest
 - memory – Korea holds the keys (HBM) (together with Micron)
 - frontier models – the US and China
- even the US – the most powerful tech nation – had to *induce* Samsung, SK hynix, and TSMC to build fabs on its soil (CHIPS Act)
 - the strongest player on the board still cannot make the whole board
- *conclusion – the full stack belongs to no one – and never will*

A reality check - the frontier is a two-player game

- everyone wants sovereign LLMs – Korea, EU, Japan, India, the Gulf: national foundation-model programs everywhere
- honest scoreboard – frontier-class models today come from US and China – full stop
 - not for lack of talent or will: \$10B+ training runs, compute access, data scale, and talent gravity compound at the top
- national models still matter – language, culture, public-sector security, deployment speed
 - but a national model $\neq$ frontier sovereignty
- *realistic question is not “how do we build our own GPT?”*
 - *but “how do we become indispensable to whoever builds it – and earn a seat at the governance table?”*
 - this is exactly what happened last Friday!

Last Friday in San Francisco

The K-AI Summit - Security Alliance becoming Tech Alliance

- *24-Jul-2026, The Midway, San Francisco* – South Korea President Lee hosted 150 leaders, in real time *turning 70-year security alliance into technology alliance*
 - Jensen Huang (NVIDIA) · Sam Altman (OpenAI) · Dario Amodei (Anthropic) · Hock Tan (Broadcom) × the Chairmen of Samsung, SK, Hyundai, and NAVER
- *“San Francisco K-AI Declaration”* – Korea as an *irreplaceable* node of the global AI supply chain; “AI for Everyone” for 50M citizens; sharing AI’s benefits with the world
- *~\$950B in initiatives, 6 MOUs signed*
 - SK × NVIDIA: \$500B+ LOI – AI factories & next-gen memory; 2GW Vera Rubin / HBM4 data center by 2027



Silicon Valley AI Nexus was in the room!

- *not* something we watched on the news – this community was there
 - *AI Nexus stewarded significant share of the invitation-only seats: founders, researchers, and investors from this community, present at historic moment for Korea-US technology relations*
- *lesson - bridges are not built by governments alone*
 - they are built by people who belong to both sides – which is exactly what you are training to become



One voice can reach the stage

- *three days after the summit, this ran in Korea's leading economic daily*
 - my op-ed on what that night meant, seen from inside Silicon Valley (*The Korea Economic Daily*, Jul 28, print B7)
 - memory bottleneck · physical AI · bio AI · the diaspora as the bridge
- *the point is not the article – it is the path*
 - community → stage → national conversation: bridges carry traffic in both directions

그날 샌프란시스코에선 'K-AI'의 무대가 열렸다

기고 / 윤성희 애플다오바이오코리아 대표

**韓, 실리콘밸리서 AI 비전 선언
메모리-전력 병목 해결 열쇠로**

지난 24일 저녁 샌프란시스코의 한 행사장. 엔비디아의 젠슨 황, 오픈AI의 샘 올트먼, 엘트론의 다리오 아모데이 최고경영자(CEO)가 한 자리에 모였다. 이재명 대통령과 우리 기업 총수들도 함께했다. 이날 인공지능(AI) 해결 경쟁의 심장부에서 한국이 비전을 선언하고, 세계가 경청했다. 필자는 스탠포드에서 전자공학 박사 학위를 마친 뒤 삼성전자와 SK하이닉스에서 반도체를, 아마존에서 전자싱커와 AI를 경험했다. 한국과 실리콘밸리를 오가며 스트리트뷰도 촬영했다. 하지만 이날의 장면은 여전히 낯설면서도 흥분했다. 세계에서 가장 바쁜 사람들이 한국이 마련한 자리에 모인 건 외교적 예우가 아니다. 한국이 AI의 핵심 국가라는 증거다.

오랫동안 실리콘밸리에서 한국은 메모리 가격을 얘기할 때 등장하는 '물류론 공급자' 정도였다. 그런데 국

기가 달라졌다. 요즘 자네 자리에서는 '메모리가 병목이고 전력이 모자란다'는 얘기가 회자다. 추론의 시대에 병목이 연산에서 메모리-전력으로 옮겨 온 것이다. 그 병목의 열쇠를 쥔 나라가 한국이며, 이는 우리 기업들이 수십 년 축적으로 배려 낸 성과다. 한국은 이제 공급망 조연이 아니라 AI 혁명의 속도를 결정하는 핵심 변수다. 남은 과제는 이 두 위대 국가의 지속적 경쟁력으로 바꾸는 일이다. 삼성전자와 SK하이닉스의 과감한 투자가 이치고 그 열에서 스타트업이 자라나도록, 정부가 인허가·인재의 토양을 다져줘야 한다.

더 중요한 것은 다음 승부처다. 챗GPT가 연 첫 말이 '말하는 AI'였다면, 실리콘밸리의 관심은 이미 퍼지컬 AI, 즉 '말하는 AI'로 옮겨 가고 있다. 이 승부는 최고의 모델을 넘어, 그것을 훈련시키고 검증할 현실의 제조 환경을 갖춰느냐에서 걸린다. 자동차·조선·배터리·로봇을 아우르는 한국의 밀집된 제조 생태계는 세계가 쉽게 복제할 수 없는 거대한 학습장이다.

바이오 AI의 잠재력도 크다. 임상에 들어간 신약 후보는 실종률 90%에 달하는데, 한국은 남다른 조건을 지녔다. 전국민 건강보험으로 축적한 데이터가 있고, 건강검진이 일상으로 자리 잡은 덕에 국가 단위로 쌓인 건강 관련 사람들의 데이터까지 있다. 무엇이 건강을 상태인지 AI가 배울 수 있다는 뜻이다. 세계 최고 수준의 바이오 의학·공생·생산 능력에 이 데이터가 더해지면 '신약 발굴 엔진'의 마지막 조각이 채워진다.

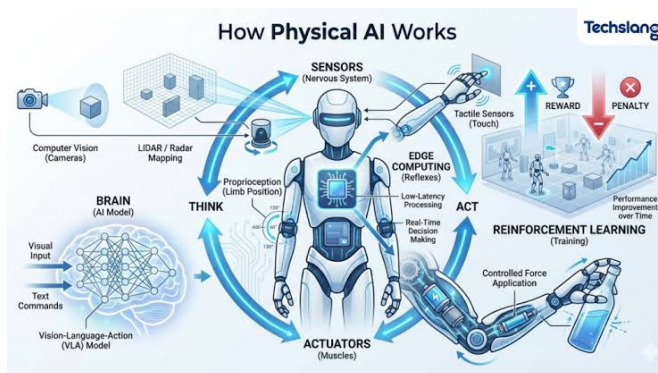
샌프란시스코 AI 선언은 시작일 뿐이다. 자본의 깊이, 심사의 전문성, 실패를 용인하는 문화도 속재료 남는다. 그렇기 그 연결은 이미 시작됐다. 실리콘밸리의 한인 전문가들이 한국의 기업·연구기관·정부와 교류하며 통로를 만들어 왔고, 이번 선언은 여기에 정부의 무게를 더했다. 앞으로 정부의 꾸준한 관심과 실리콘밸리의 뜨거운 의지가 오랫동안 지속되길 기대한다.



Distributed Tech Sovereignty

Why Korea? - sovereignty through indispensability

- *Korea never tried to own the whole stack* – made its layers unownable by anyone else
 - *memory* – in the inference era, the bottleneck moved from compute to memory bandwidth & energy; decades of accumulation $\Rightarrow$ Korea holds a chokepoint
 - *physical AI* – “AI that works” needs factories, ships, batteries, and robots to learn in; Korea’s dense manufacturing ecosystem = a learning field the world cannot replicate
 - *bio AI* – 9 of 10 drug candidates fail in trials; national single-payer health data + routine-checkup “healthy-normal” data = an asset almost no nation has
- *this is the model* – be the partner nobody can route around
 - sovereignty as *indispensability!* (not independence)



A Proposal - Distributed Tech Sovereignty

- *idea-space so far*
 - EU “technological sovereignty” – already supranational, but regional self-interest
 - US “digital solidarity” (State Dept, 2024) – partnership-based self-determination, but it *abandons* the word sovereignty
 - “CERN for AI” (Marcus 2017 → EU today) – pooled institutions, not a redefinition
- *Tonight I propose - Keep the Word – Change its Unit*
 - *Distributed Tech Sovereignty* – sovereignty over critical technology is a property no single nation can hold; it *emerges* from a trusted network of nations, each indispensable in its own layer, governing the stack together – and for general intelligence, the ultimate sovereign is humanity itself
- *like distributed system* – no node owns the whole; the network does
 - availability through redundancy · trust through protocol · resilience through diversity
- *corollary!* – the question shifts from “how do we own it all?” to “which network do we belong to, what layer do we anchor, and who writes the protocol?”

The Korea-US alliance as sovereignty infrastructure

- *reread last Friday through this lens* – the summit was not Korea buying access; it was two nations *wiring the network*
 - US anchors - frontier models, risk capital, research universities
 - Korea anchors - memory, advanced manufacturing, deployment & data
- 70 years of security alliance $\Rightarrow$ the trust protocol already exists
 - the technology alliance runs on rails the security alliance built
- open problems the network must solve
 - export controls vs. open science · cross-border data governance
 - who sets AI safety standards · how benefits flow beyond the two nations (“AI for Everyone,” globalized)
- *sovereignty through alliance* – trust is the scarcest resource in the AI supply chain; alliances are how trust is manufactured at scale

To the Bridge Builders

What you inherit - and what I hope you'll argue about!

- the governance choices being made right now will outlive their makers
 - who gets compute · whose data trains the models · which safety standards win · how AI's wealth is distributed
- you are AI natives – these are not policy abstractions for you; they are lived experience
 - and you will live with them longer than anyone on this stage
- *questions for the flash talks & panel*
 - is distributed sovereignty compatible with national democratic accountability – who votes on the network?
 - if humanity is the ultimate sovereign of general intelligence – who speaks for humanity?
 - what would make the Korea-US tech alliance *worth inheriting*?

Appendix

Silicon Valley AI Nexus

Silicon Valley AI Nexus (or just AI Nexus)

- pioneering community of professionals dedicated to building privacy-preserving AI solutions, products, and systems
- comprehensive expertise across AI domains
 - biotechnology, healthcare, and medical research
 - industrial applications and data centers
 - cloud infrastructure, storage solutions, mobile technologies
 - customer service platforms, multi-agent systems
 - RAG implementations, vector databases, agentic AI frameworks
- vision
 - *shaping future where AI innovation and privacy protection go hand in hand*
- active community with [homepage](#) & KakaoTalk collaboration platform for members

Our journey - forum history

- Nov-Dec 2024 - “The AI Strikes Back” & “Free Your Data”
 - Prof. Jung Hee Cheon (homomorphic encryption revolution)
- Jan 2025 - “The AI Knight Rises”
 - [Sunghee Yun](#) @ Erudio Bio on deep learning to flourishing societies
- Feb 2025 - “Silicon Citadel”
 - Chanik Park @ MangoBoost on AI data infrastructure
- Mar 2025 - “Blockchain Awakens”
 - Daejun Park @ a16z crypto on decentralized AI
- Apr 2025 - “Advancing Humanity”
 - Stanford Medicine team on bio/medical AI
 - co-hosting with K-BioX
- May 2025 - “The Autonomous Alliance”
 - Microsoft, GitHub, Uclone, SK Hynix on AI agents

Our journey - forum history

- Jun 2025 - “Silicon Companions”
 - Altos Ventures on robotics & smart devices
- Aug 2025 - “The Human-Centric AI Revolution”
 - address legal and ethical issues related to AI
- Nov 2025 - “The AI Silicon Race”
 - Korea-US Innovation Leadership at K·ASIC



Strategic partnerships & ecosystem

- *Perpetual Partnership with KOTRA Silicon Valley as Strategic Alliance*
- 2026 co-hosting partners
 - K-ASIC (Korea AI & IC Innovation Center)
 - K-BioX (biotech innovation)
 - KOTRA Silicon Valley (trade & investment)
 - Consulate General of the Republic of Korea at San Francisco (diplomatic support)
 - KABANC (Korean American Bar Association of Northern California - legal expertise)
- building bridges between Silicon Valley innovation and Korean institutional networks
- creating comprehensive support ecosystem: technical, legal, business, diplomatic



Community & engagement

- membership requirements
 - attend one! AI Nexus event!
- member benefits
 - networking with AI professionals across all domains
 - knowledge sharing and collaboration opportunities
 - direct access to world-class speakers and experts
- forum format - 5pm-8pm, typically Wednesdays at premier Silicon Valley venues
- venues - Stanford, KOTRA, SK Hynix, Altos Ventures, K-ASIC, and more
- active community engagement and professional development



Three Core Tenets

- *Explore & Lead*
 - world-class forums, the AI & Humanity Council, and thought leadership shaping AI's most consequential questions
- *Elevate Community*
 - mentorship, startup showcases, professional connections, and leadership opportunities that amplify every voice
- *Give Back to Society*
 - policy engagement, academic partnerships, and public discourse ensuring AI's benefits reach beyond Silicon Valley

AI Nexus encourages and embraces Interest Groups

- Bio / Medical / Data Group (7)
 - Chanik Insuk Jeff Jongmin Minha Mitch Sunghee led by Chanik
- Karaoke Group (4)
 - Hayden Jeff Sunghee Yeseul led by Sunghee
- AI Hands-on Group (5)
 - Chanik James Kyoung Whan Minha Sunghee led by Minha
- East Bay Group (7)
 - Daniel Eugene Hyunjung Linda MiSook MJ Sunghee led by Hyunjung

Collaboration with Institutions & Universities

- MOUs
 - SNU
 - *AI Nexus × SNU College of Engineering Silicon Valley Immersion Program*
 - *Winter Internship 2026*
 - asked by
 - KIC SV
 - SVreboot
 - XL8
 - KASPA
 - Korea Univ.
 - Sungkyunkwan Univ.
 - KAIST
 - Korean American Semiconductor Professional Alliance (KASPA)
 - *etc.*

Collaboration with Institutions & Universities

- Co-hosting with
 - K-BioX
 - KOTRA SV
 - KIC SV
 - K-ASIC
 - Korea-America Student Conference (KASC)
 - Consulate General of the Republic of Korea at San Francisco
 - Korean American Bar Association of Northern California (KABANC)
 - *etc.*

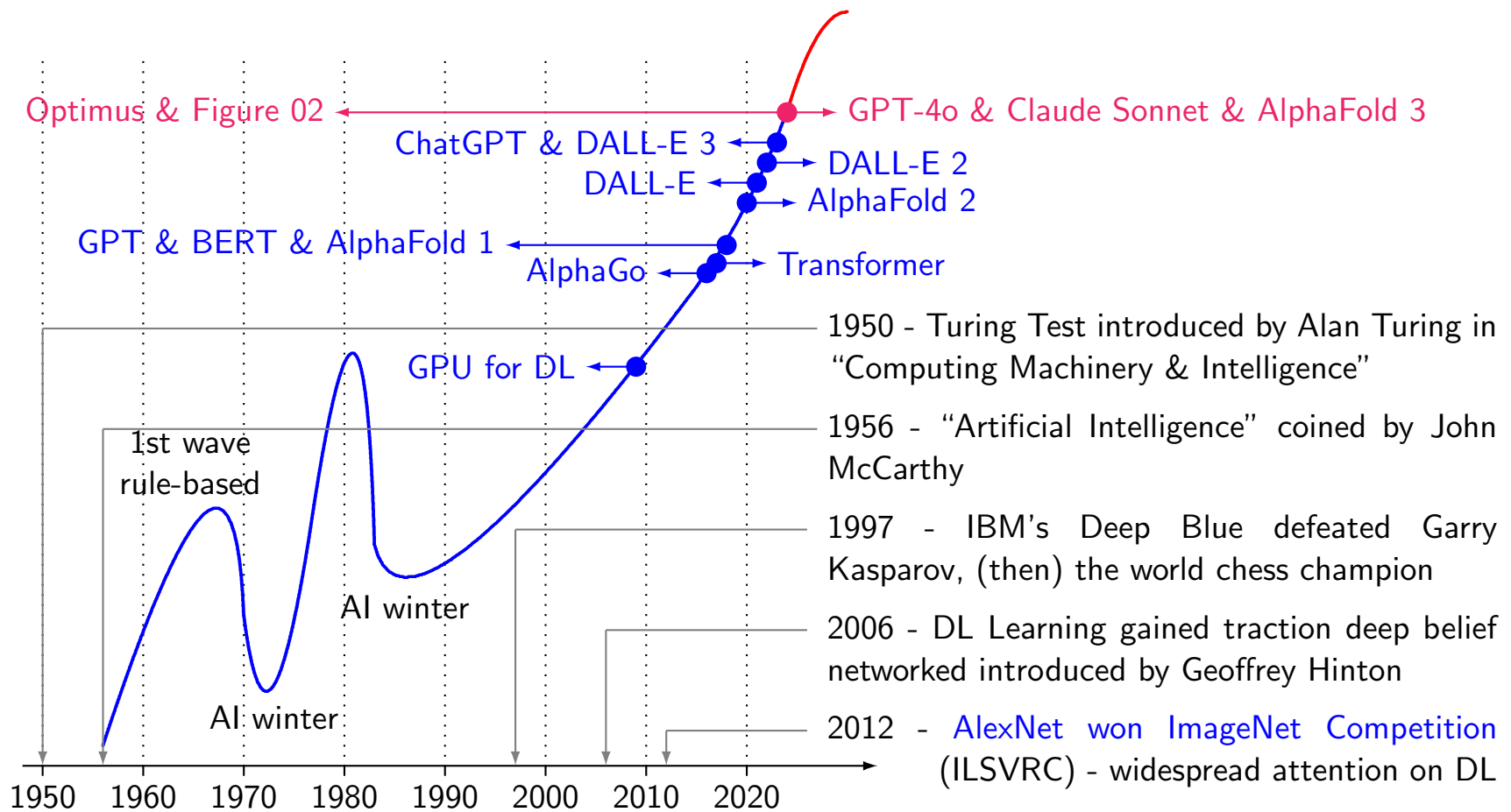
AI Nexus Committee

- Board of Directors
 - *Sunghee Yun* - Co-Founder & Char
 - *Minha Hwang* - Co-Founder & AI Master
 - *Sue Kim* - Co-Founder & Marketing Leader
 - *Youngwook (Hayden) Song* - Co-Founder & Legal Affairs Advisor
 - *Bo Hyoung (Jeff) Lee* - Co-Founder & Legal Affairs Advisor
 - *MiSook Chung* - Co-Founder & Community Engagement Chair
- Vision & Principles Committee
- Tech & Social Media Committee
- Program & Content Committee
- the full list of committees can be found [here](#)!

Artificial Intelligence

AI History

History



Significant AI Achievements - 2014 – 2025

Deep learning revolution

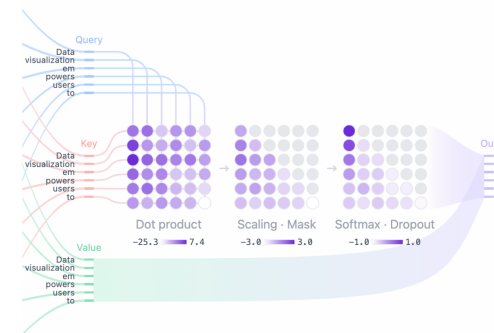
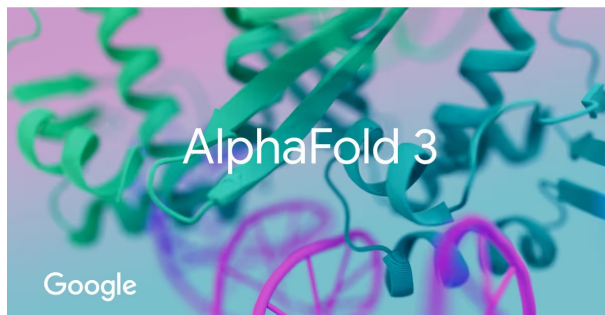
- 2012 – 2015 - DL revolution¹
 - CNNs demonstrated exceptional performance in image recognition, *e.g.*, *AlexNet's victory in ImageNet competition*
 - widespread adoption of DL learning in CV transforming industries
- 2016 - AlphaGo defeats human Go champion
 - DeepMind's AlphaGo defeated world champion in Go, extremely complex game *believed to be beyond AI's reach*
 - significant milestone in RL - AI's potential in solving complex & strategic problems



¹CV: computer vision, NN: neural network, CNN: convolutional NN, RL: reinforcement learning

Transformer changes everything

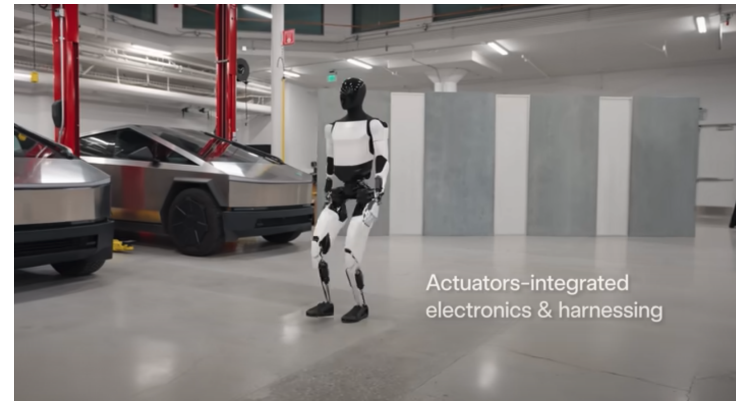
- 2017 – 2018 - Transformers & NLP breakthroughs²
 - *Transformer (e.g., BERT & GPT) revolutionized NLP*
 - major advancements in, e.g., machine translation & chatbots
- 2020 - AI in healthcare – AlphaFold & beyond
 - DeepMind's *AlphaFold solves 50-year-old protein folding problem* predicting 3D protein structures with remarkable accuracy
 - accelerates drug discovery and personalized medicine - offering new insights into diseases and potential treatments



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

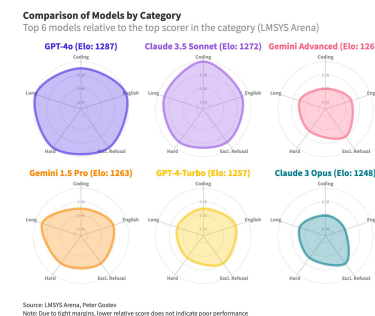
Lots of breakthroughs in AI technology and applications in 2024

- 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*



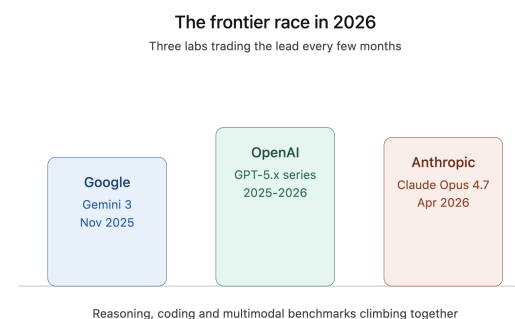
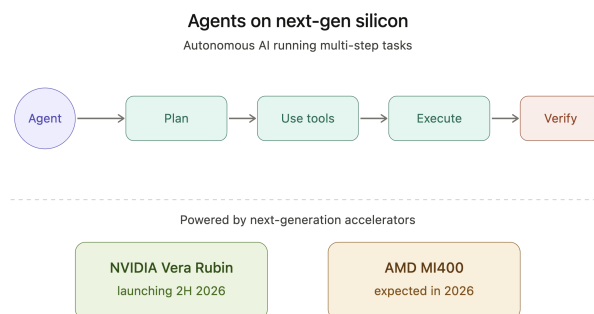
Major AI Breakthroughs in 2025

- 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



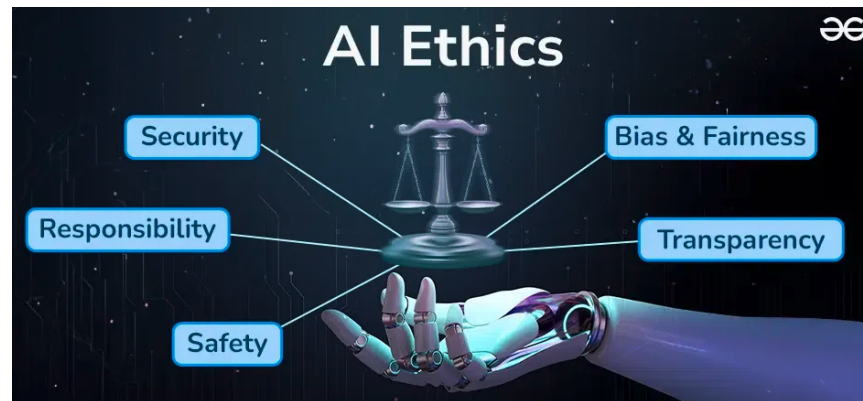
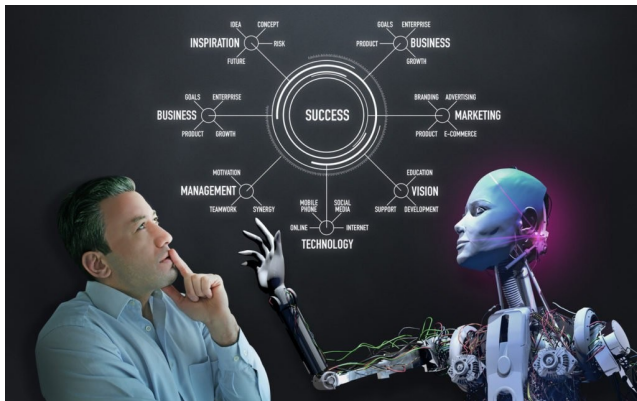
Where AI Is Heading in 2026 (as of May 2026)

- frontier models reaching new performance ceilings
 - Gemini 3, GPT-5.2 / 5.x series, Claude Opus 4.7 → multimodal benchmarks higher
 - intensifying competition among Google, OpenAI, and Anthropic
- hardware scaling and diversification
 - NVIDIA's Vera Rubin platform launching in 2H claiming $\sim 5\times$ 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

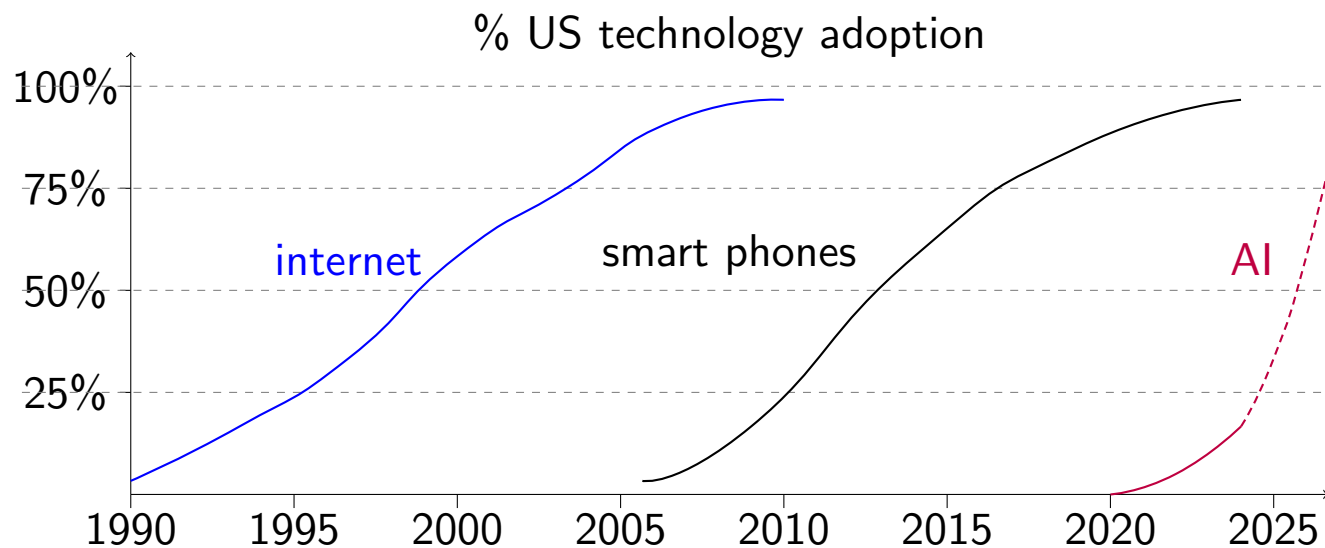
- accelerating human-AI collaboration
 - not only reshaping industries but *altering how humans interact with technology*
 - AI's role as collaborator and augmentor redefines productivity, creativity, the way we address global challenges, *e.g.*, *sustainability & healthcare*
- AI-driven automation *transforms 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

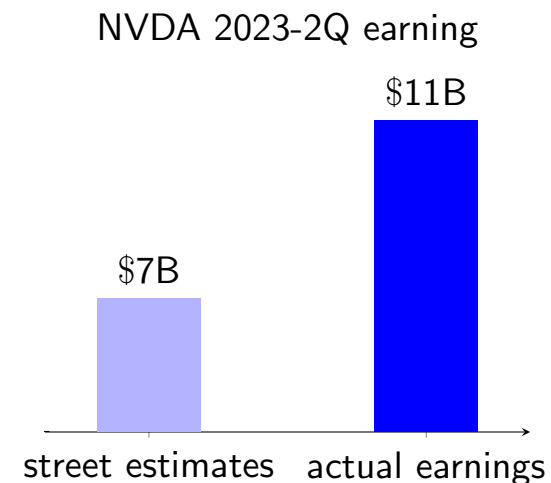
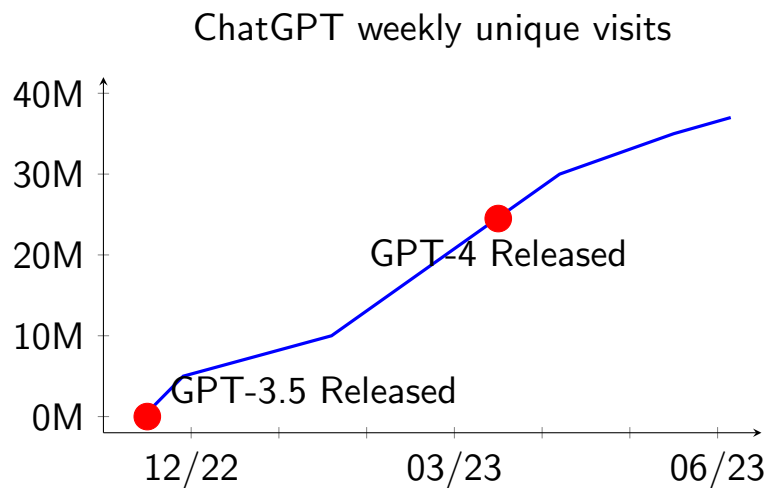
Where are we in AI today?

- 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

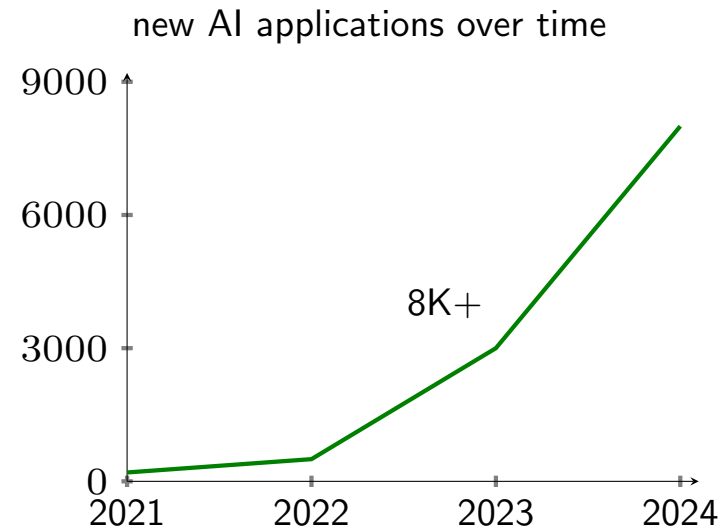
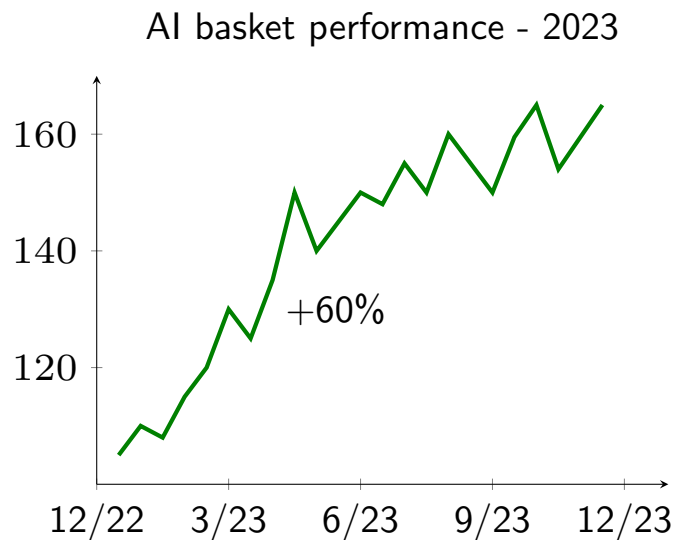
- took only *5 months for ChatGPT users to reach 35M*
- NVIDIA 2023 Q2 earning exceeds market expectation by big margin - \$7B vs \$13.5B
 - surprisingly, *101% year-to-year growth*
 - even more surprisingly *gross margin was 71.2%* - up from 43.5% in previous year³



³source - Bloomberg

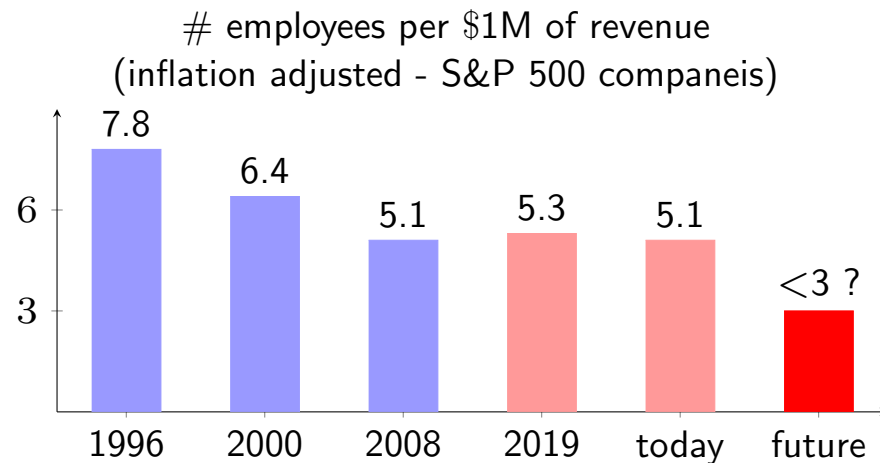
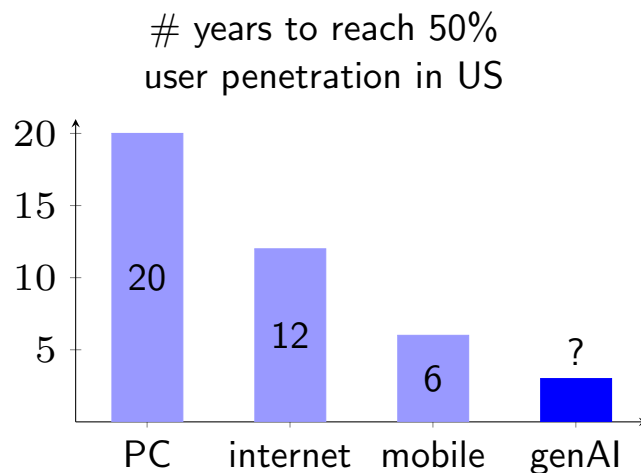
Explosion of AI ecosystems - AI stock market

- *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



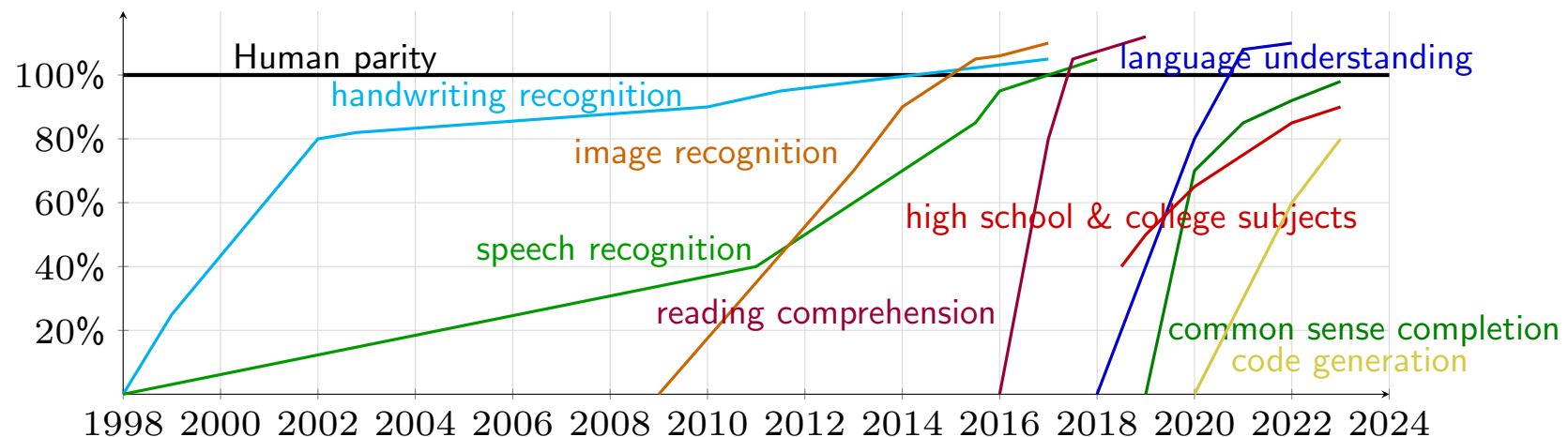
AI's transformative impact - adoption speed & economic potential

- 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



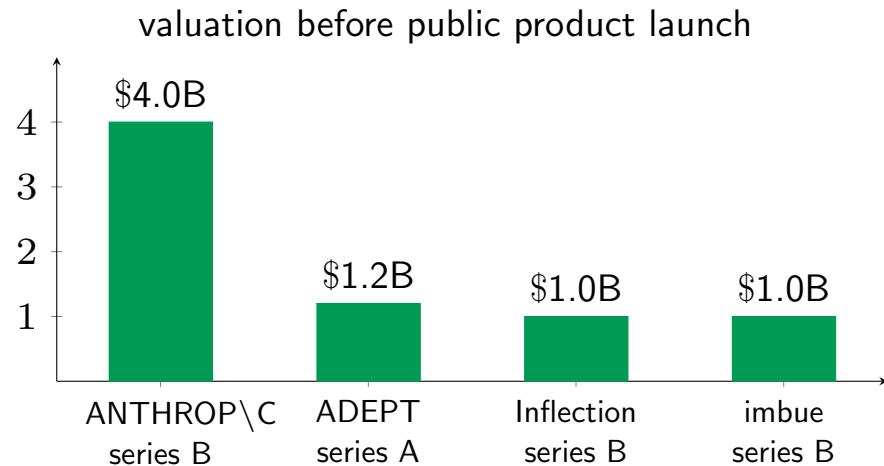
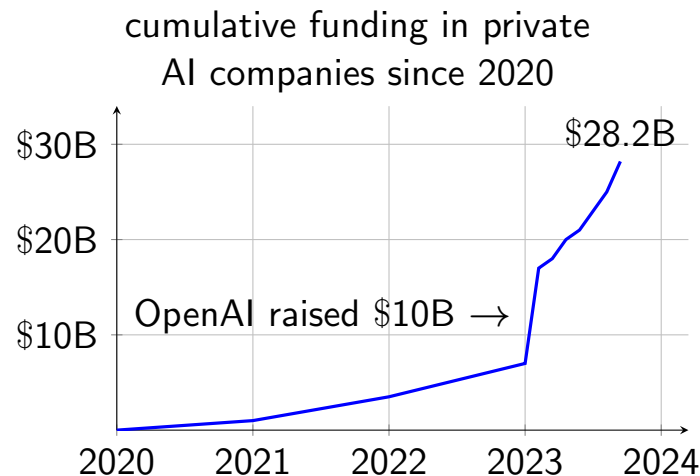
AI getting more & more faster

- 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

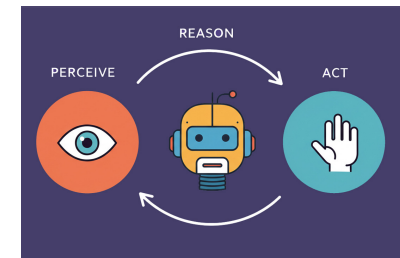
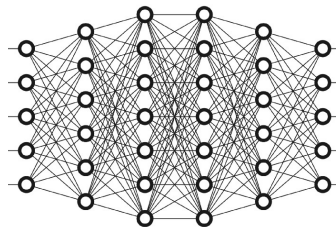
- *explosive growth* - cumulative funding skyrocketed reaching staggering \$28.2B
- OpenAI - significant fundraising (=\$10B) fueled rapid growth
- *valuation surge* - substantial valuations even before public products for stellar companies
- *fierce competition for capital* among AI startups driving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



AI Agents

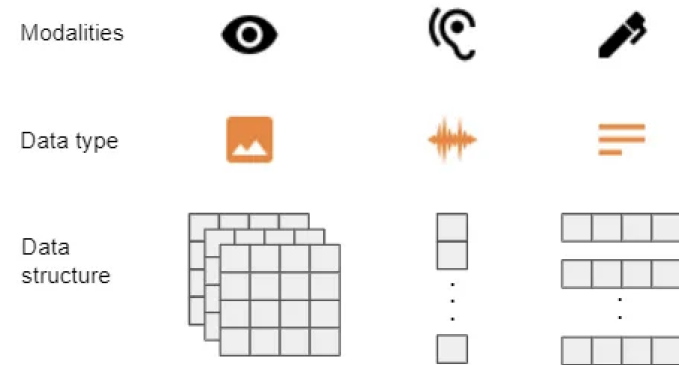
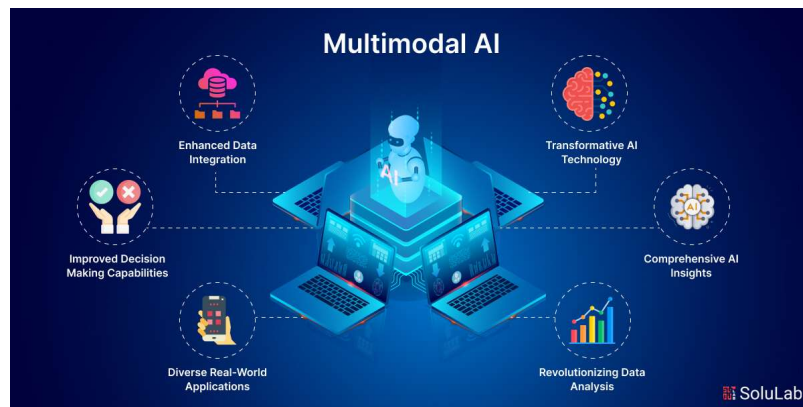
AI progress in 21st century in keywords

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



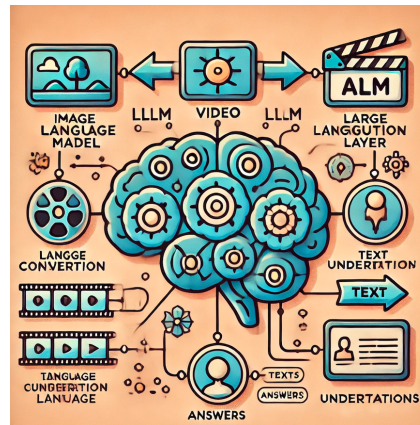
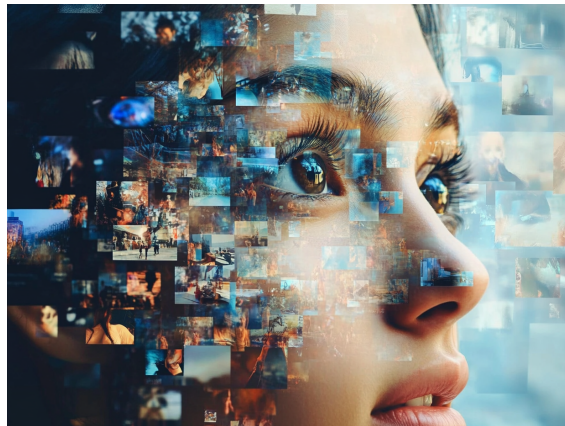
Multimodal learning

- 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)



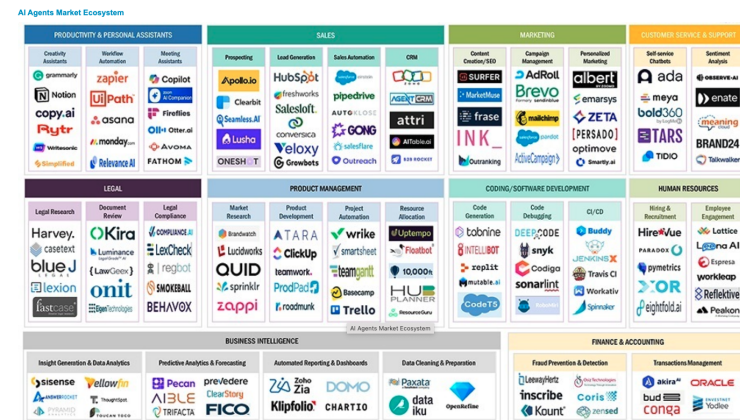
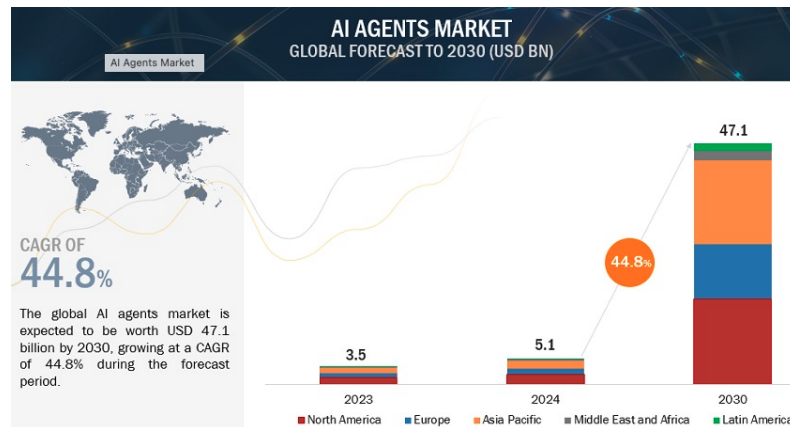
Implications of success of LLMs

- many researchers change gears towards LLM
 - from computer vision (CV), speech, 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

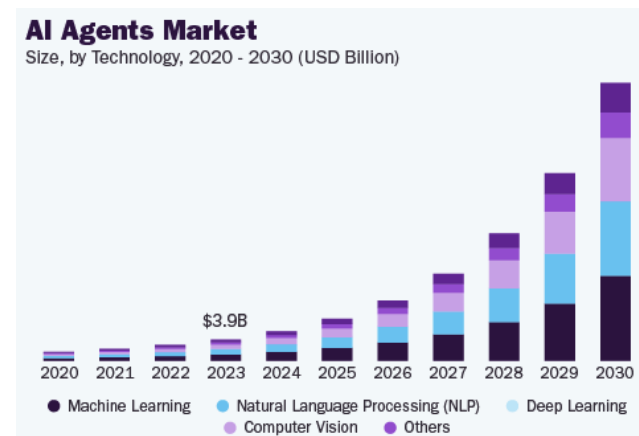
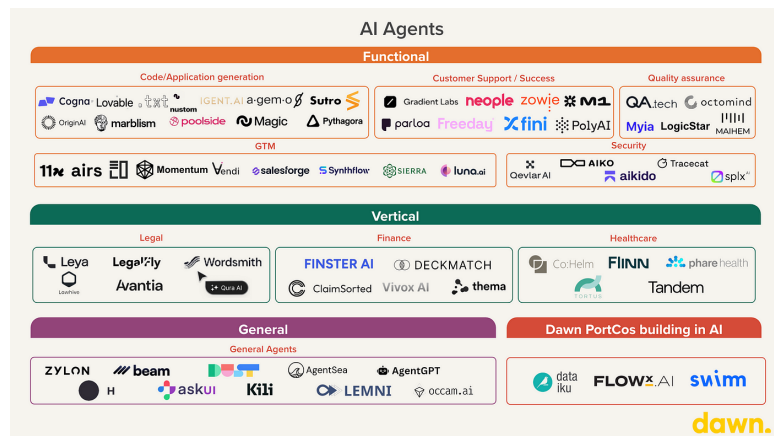


mmAI Technology

- 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

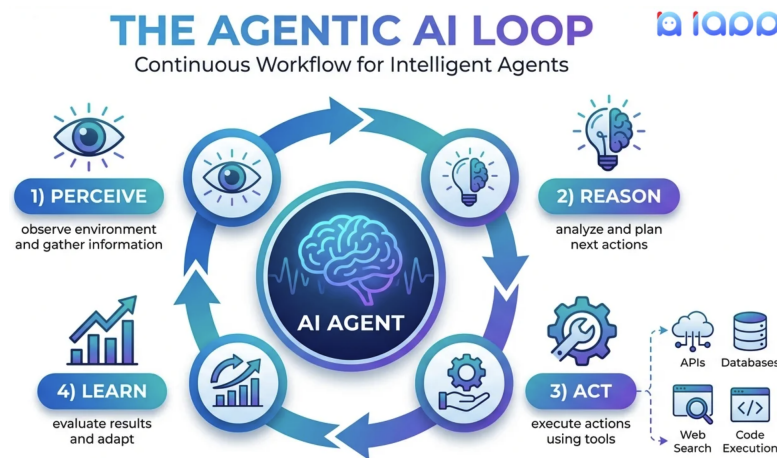
AI agents powered by multimodal LLMs

- 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



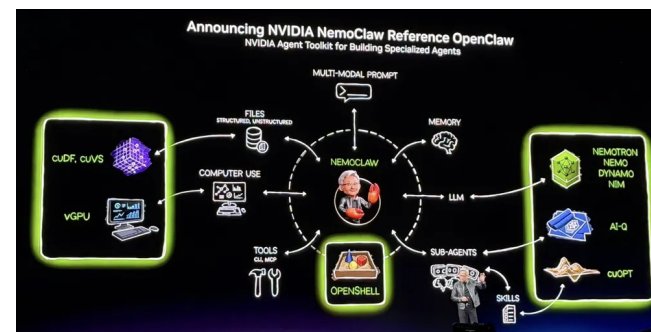
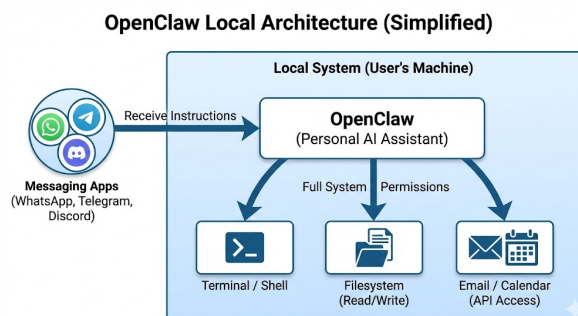
What makes AI “Agentic”?

- 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”*



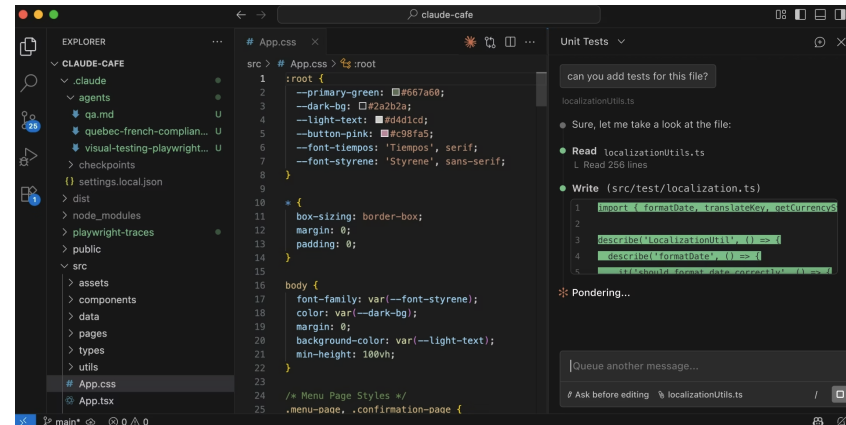
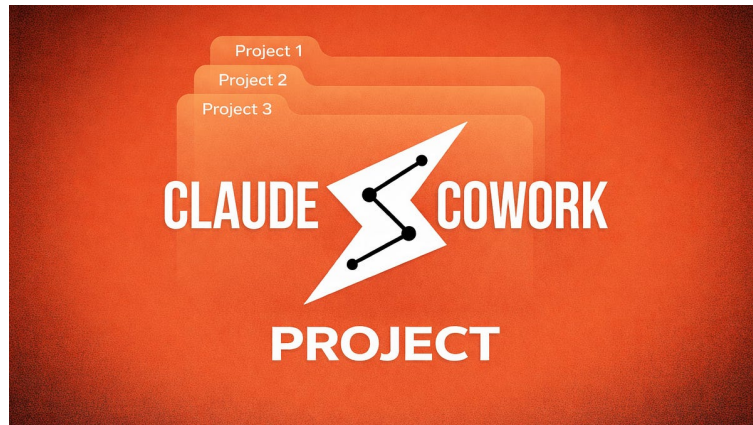
Cutting-edge AI agent tools - open source

- 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



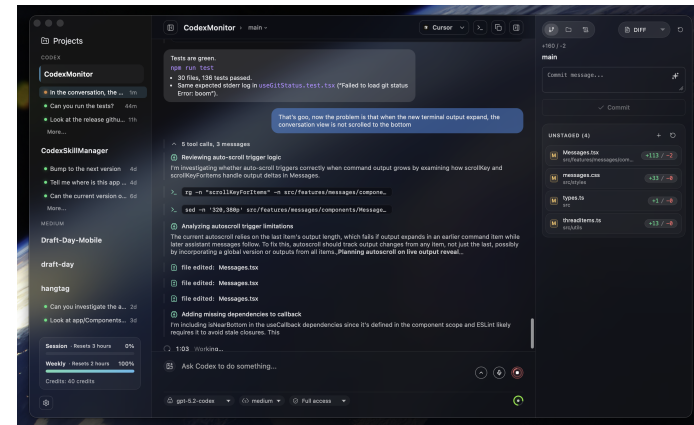
Cutting-edge AI agent tools - Anthropic

- 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



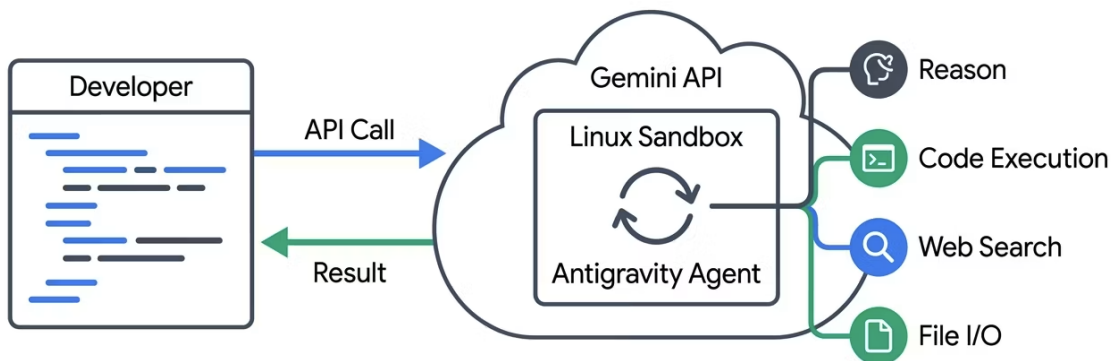
Cutting-edge AI agent tools - OpenAI

- 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



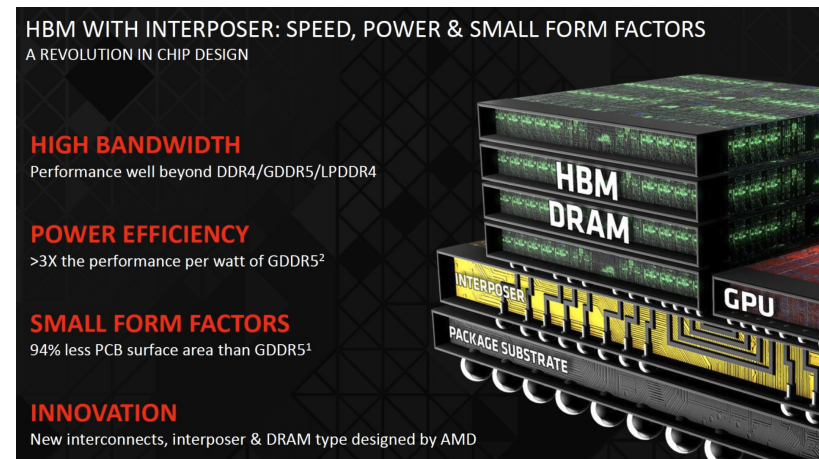
Cutting-edge AI agent tools - Google

- 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



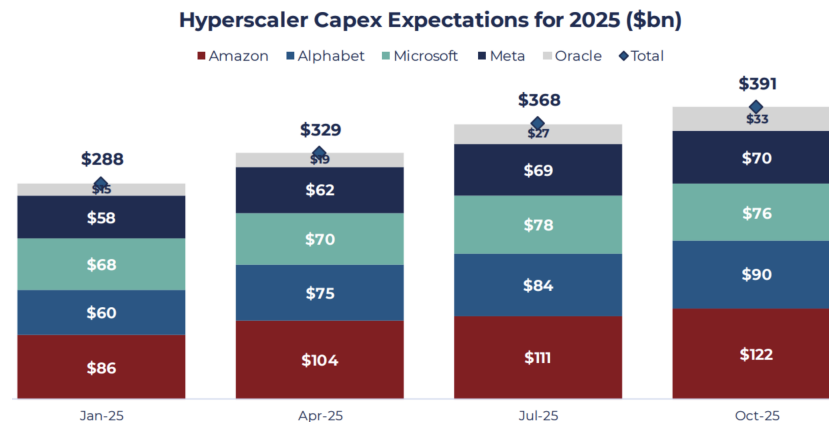
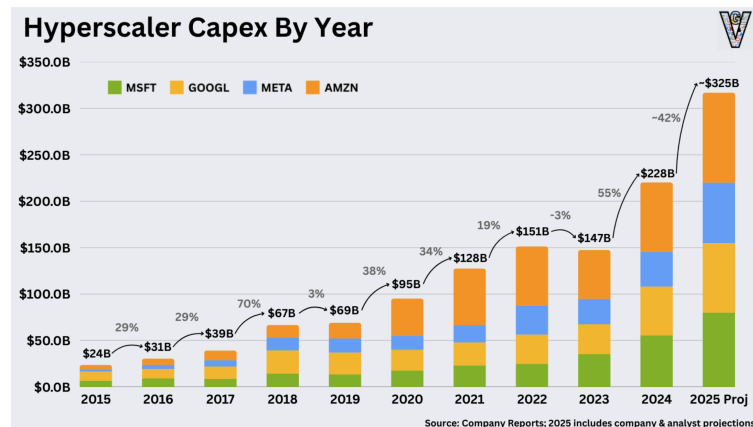
Agentic stack - LLM is engine, but not whole system

- *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*



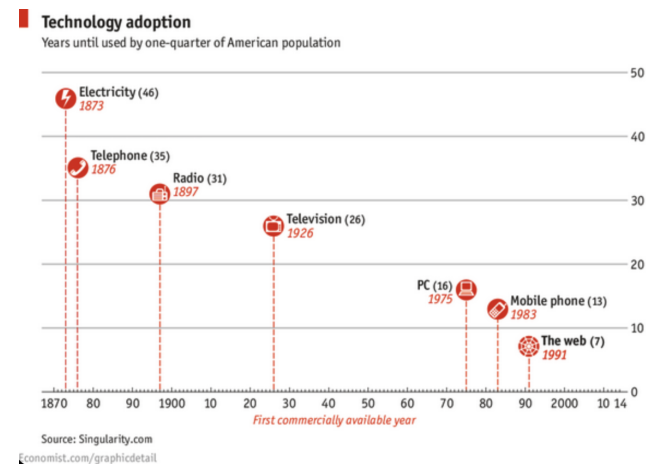
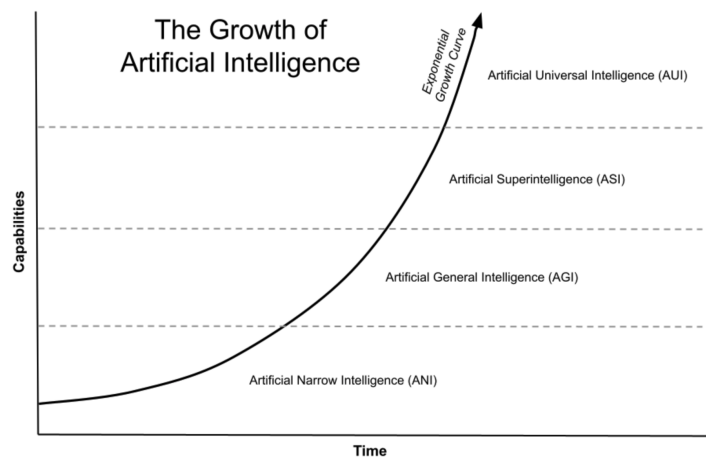
Trillion-dollar gravity well - capital, talent, geopolitics

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



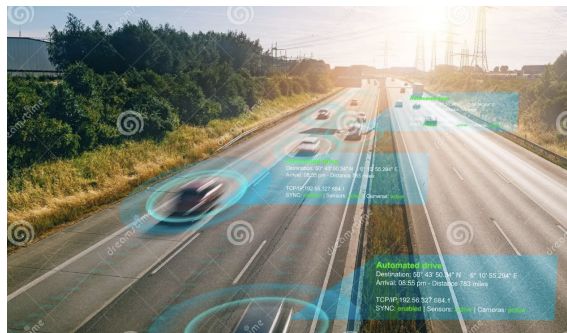
What took decades now takes months

- 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 $\neq$ 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



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

- [HGH⁺22] Sue Ellen Haupt, David John Gagne, William W. Hsieh, Vladimir Krasnopol'sky, 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.
- [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.
- [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