TECHNOLOGIES
Bring the future to humanity

State of AI 2025,

31 January 2025

Introduction

This report is intended to summarize the advancements in AI and its media perception in the last 12 months (January- January 2025): the progress in research, specific AI news in the media and emerging trends (catalysts) which will impact the near future. Our understanding of the AI advancements and the trends allow us to make trend predictions we believe will happen in the next year (last 9 slides). We will review accordingly in one year's time.

Author

Andrea Isoni

Chief AI Officer

Physicist and PhD from Imperial College, Former Founders Factory (FF), Chief AI Officer at AI Technologies. Experienced in leading edge AI algorithms and how to translate them into business applications (including pioneering Human Machine Interface systems). Advisor to Waed Saudi Aramco and writer in AI ('Machine Learning for the Web' published by Packt in English, Chinese and Korean and more than 10,000 copies). Committee member of ISO and IEEE for the AI standards and International AI speaker (London, New York, Dubai, Saudi Arabia). AI Newsletter 'Thoughts about AI by a Human' (2.8k+ subscribers).

TECHNOLOGIES
Bring the future to humanity

2024

Notes on the Report

*

This year data is considered until 30 January 2025: so January 2024 to January 2025 included.

*

AI Leaderboard used: https://huggingface.co/spaces/ArtificialAnalysis, https://lmarena.ai/?leaderboard, https://opendfm.github.io/MULTI-Benchmark/#leaderboard, https://artificialanalysis.ai/leaderboards/models, https://www.wolfram.com/llm-benchmarking-project

*

Definitions (only very technical ones. It is expected the reader know what GPU means etc.):

-pre-training: a process where a model is first trained on a large, general dataset.This allows the model to learn general features from the pre-training data

-post-training: set of processes and techniques applied to a model after it has been initially trained on a dataset. It focuses on refining and optimizing the model to improve performance, efficiency and meets specific practical requirements.

-multimodal: AI systems that simultaneously process and integrate various data types.

-SOTA: state of the art, it refers to a model achieving best scores on benchmarks.

*

The BONUS prediction is half a joke, half true. We believe many more will grasp some coding terms and that may naturally spill out into common conversations.

Technologies

2024

How to read the Report

‘secular’ trends + ‘catalyst’ trends → ‘contingent events’ + trend prediction

* ‘Contingent Events’ means events or reasons that are happening or just happened this year (left box). They are ‘signals’ that the prediction (detailed on the right box) may likely happen soon.

* ‘Catalyst Trends’ means trends and events developing in the last 5 years or less. Each trend has immediate effects, it may not last and it can accelerate or slow down depending on political or other reasons. These trends impact some predictions (indicated with numbers at the bottom of each trend prediction page).

* ‘Secular Trends’ means trends that are developing in the last decade or more. Each trend is slow to take effect but durable over time.These trends are always impacting predictions (indicated with numbers at the button of each trend prediction page).

Pages’ guide

- Secular trends from page 8 to page 12.
- Catalyst trends from page 13 to page 22
- Trend predictions, from page 23 to page 31. We consider and give predictions for the following sector: hardware progress, industry adoption, consumer adoption, cybersecurity, agents and robotics. Each page describes what it is actually happening now, reference the catalysts trends acting on the sector and then our educated guess what it is suppose to happen in the next year.

2024

Catalyst Trends

Download 2024 report here

01

Governance AI: EU AI regulations are coming by 2024.

02

Both synthetic data and human generated are increasing and contributing to AI training.

03

Cloud adoption growth.

04

AI/crypto hardware demand and energy consumption

05

Supply chain localisation and workforce

06

AI investment continues.

07

Adoption of genAI in music and images, videos

Checks

01

Happening. Regulations are coming but slower than expected.

02

Happening. Especially synthetic data this year had found better usage in models.

03

Happening. Due by AI adoption or foreseen adoption new data centers were planned..

04

Partially happening. Lot of demand of AI hardware (GPUs etc.) but energy consumption was a topic of discussion but not a real contingent issue.

05

Happening. US and other countries tend to build factory within their territory or near their borders.

06

Happening. AI funds were concentrated in large rounds but the trend is continuing.

07

Happening. Adoption in visual arts and music is continuing but now more perceived like ‘just tools’ so not really an impacting trend (in terms of behavioural shift).

TECHNOLOGIES
Bring the future to humanity

2024

Predictions

Download 2024 report here

01

Adoption of ‘simple/‘small’ models in industry in regulated industries or functions. Sophisticated models only if low costs of regulations. First AI drone in battlefield.

02

Startup raised $100m+, valuation $1b+, bought for <50% or shut down.

03

AI fake detection fails often. 15%+ of new images with a watermark.
A genAI cybersecurity startup will be acquired for more than $100M.

04

ChatGPT and competitors used at least once by 30% of US workforce. Top 3 non first world (excl. China) have 20% combined traffic on chatGPT and competitors. US workforce not decreasing. 1% GDP due to AI.

05

Leading LLM will be from BigTech. RLHF will be matched by other methods. A model not transformed based reached chatGPT 3.5 levels. A lab will reach chatGPT 3.5 but with <100GB of GPU memory, <20B parameters

Outcomes

01

Partially right. Highest AI adoption was in functions like fraud detection, risk analysis, diagnostic, admin (which requires ‘smaller’ models) but it happened in regulated industries (finance, health). Yes various sources confirmed Ai drive drone in Ukraine (likely Anduril)

02

Right. Adept AI reached $1b+ valuation and was bought back by Amazon for <$500m. The company was out of our possible candidates but still it did happen.

03

Partially right. Yes AI fake detection have been proven wrong except for video which will remain detectable if fake for a while. Wrong on watermarking new images. Yes Cisco bought Robust Intelligence for $40Mm.

04

Right. 31% US adults used chatgpt at least once Aug-Sep 2024 (statista). Per similarweb, India, Brazil and Indonesia did 14.5% of traffic in Dec 24 and ~24% Jul-Se 24. No decrease in workforce. Goldman Sachs gives AI 1.5% annually over 10 years.

05

Right. OpenAI O1 still on top. RLHF has been matched by DPO. Codestral Mamba (state space architecture) outperformed chatGPT 3.5. Many models outperformed chatGPT 3.5 and < 100GB GPU memory and <20B.

TECHNOLOGIES
Bring the future to humanity

2024

Best AI advancements

01

‘Mixtral of Experts’ --> Method to scale large language models efficiently by activating only a subset of the model's parameters for each input. Seminal work used in many top model, like DeepSeek R1.

02

‘The Llama 3 Herd of Models’ --> Introduced open weights models multi-staged pre-training and post-training with Direct Preference Optimization (DPO) .Llama 3.2 included multimodal.

03

‘DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning’--> Open weights model reaching SOTA performance close to OpenAI O1 but will less resources and better speed.

04

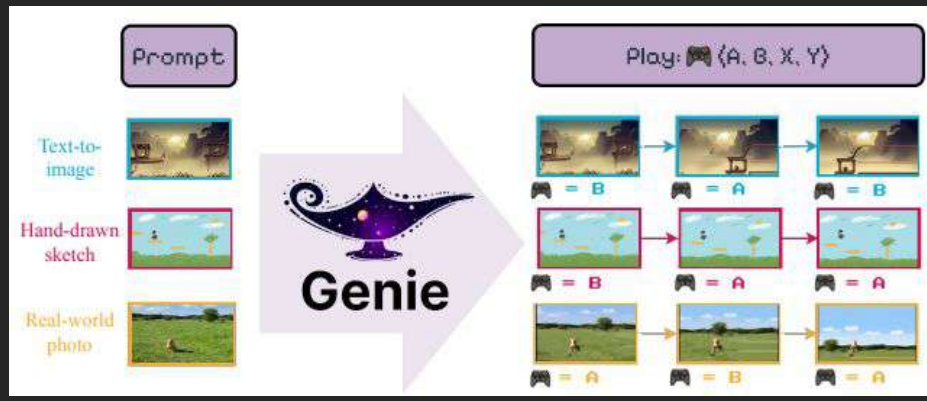
‘Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction’-->A new approach that outperforms SOTA diffusion transformers in visual in-painting/editing with scaling properties similar to LLMs.

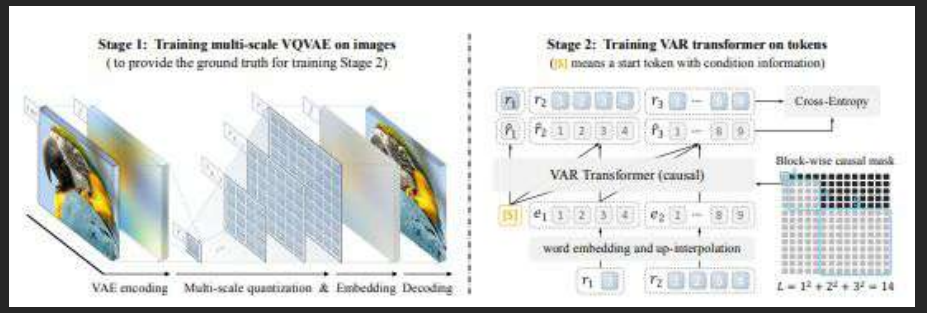
05

‘Genie: Generative Interactive Environments’--> Method to generate action-controllable virtual worlds from unlabelled videos by text and image prompts.

06

NVIDIA’s NVLM: Open Frontier-Class Multimodal LLMs’--> Hybrid approach to multimodality that allow to score best in different tasks from OCR to image and text tasks.





deepseek

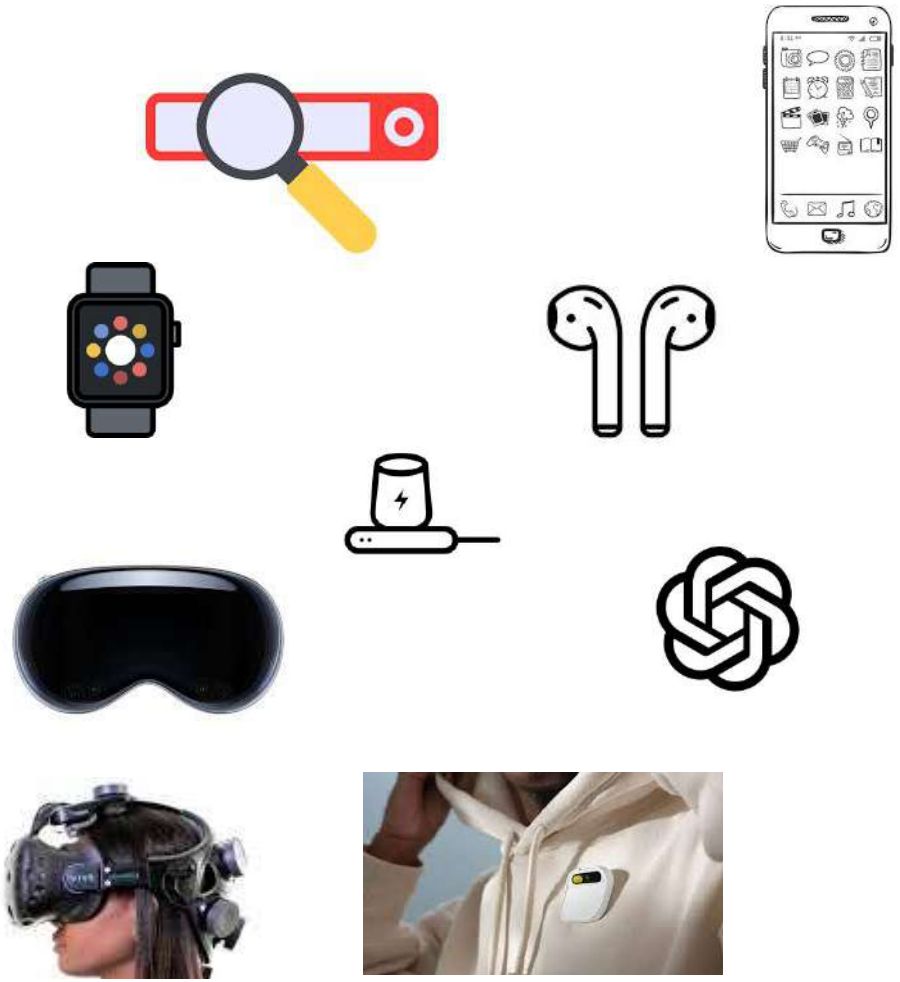
Technologies

Secular Trends 1

Human Machine Hybridization

Humans are progressively integrating with technology, hardware and software.

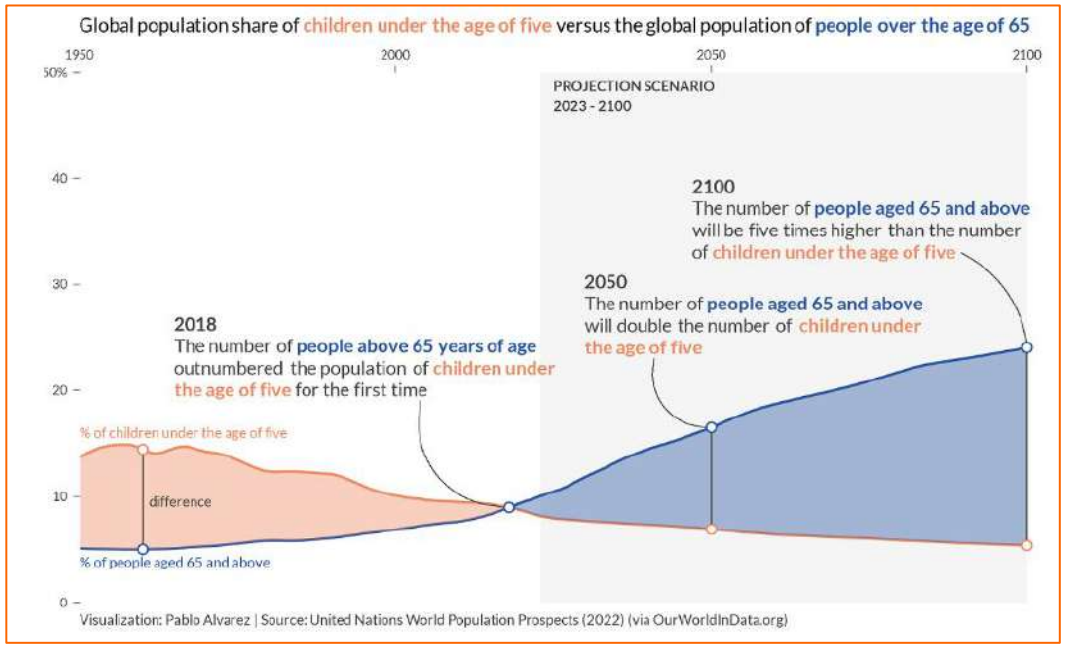
- Mobile phones, search engines (2000s)
- Smart Watch (2010s), iPods
- Smart Home (2015s)
- **Now (2020s), (chatgpt) AI assistants, VR headset, good proportion of GenZ watch Youtube at 1.5x/2x**
- Future.. wearable pins, Brain Computer Interface

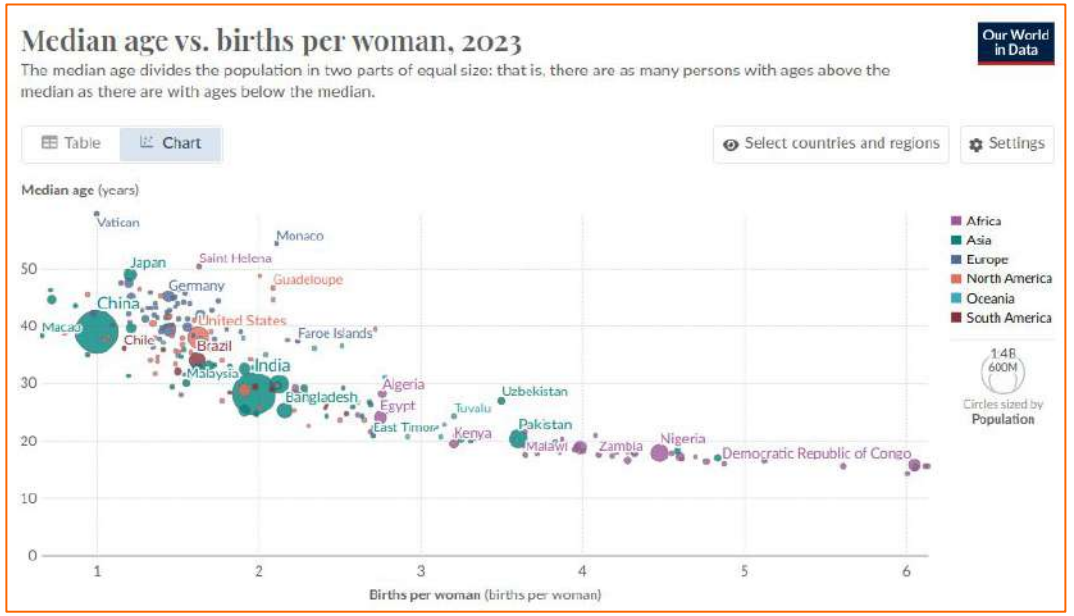


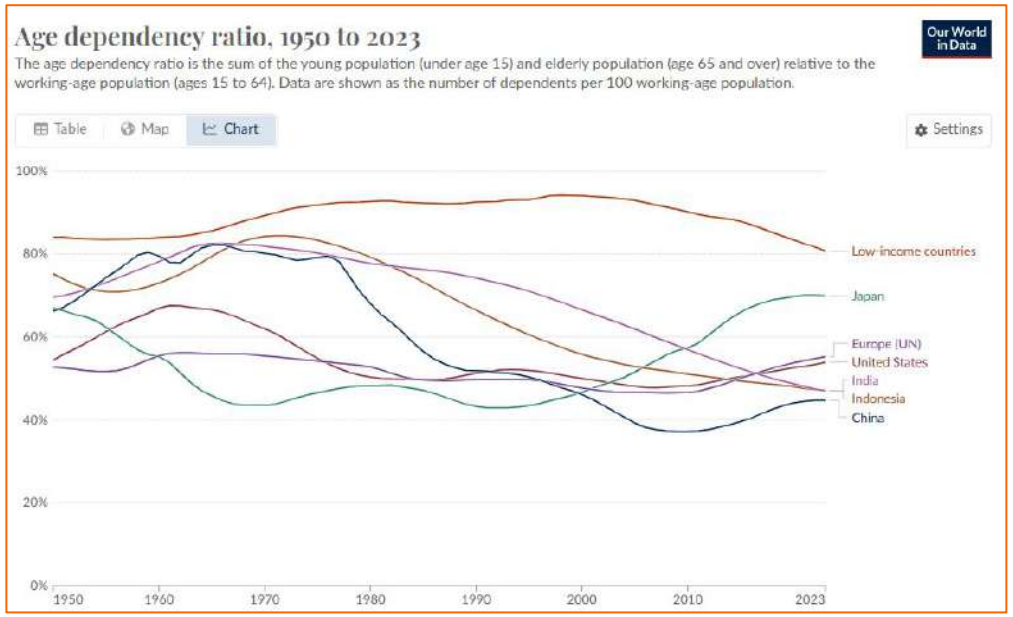
Secular Trends 2

Aging

Most countries have an increasingly older population with births rates less than 2. Age dependency ratio (bottom right) is rising in Western countries but stable or decreasing in China and low Income countries.

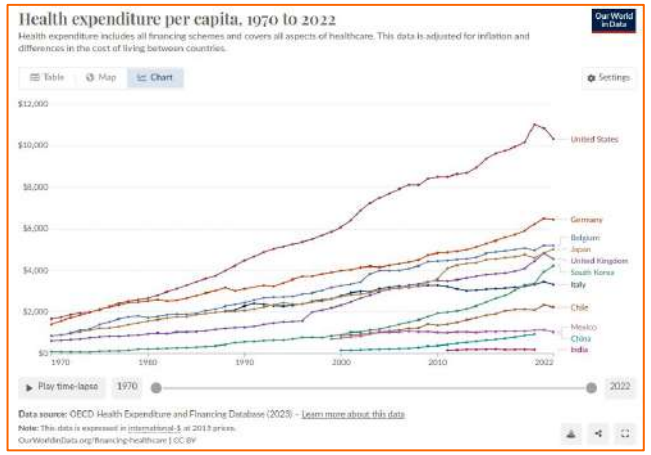


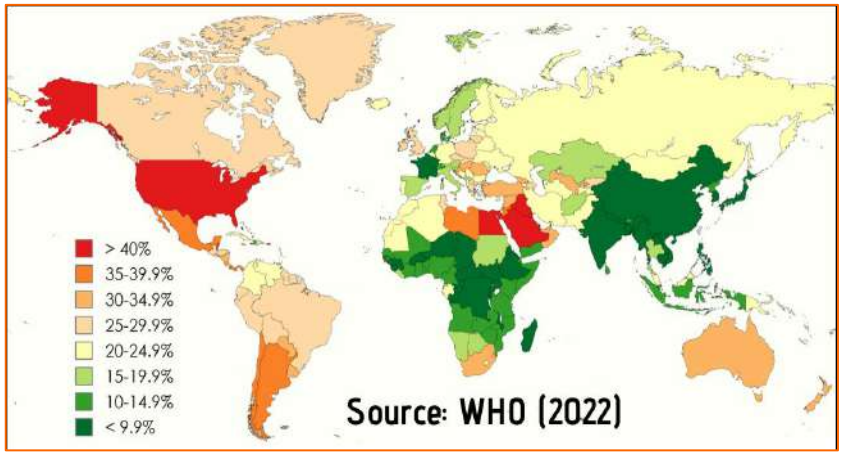


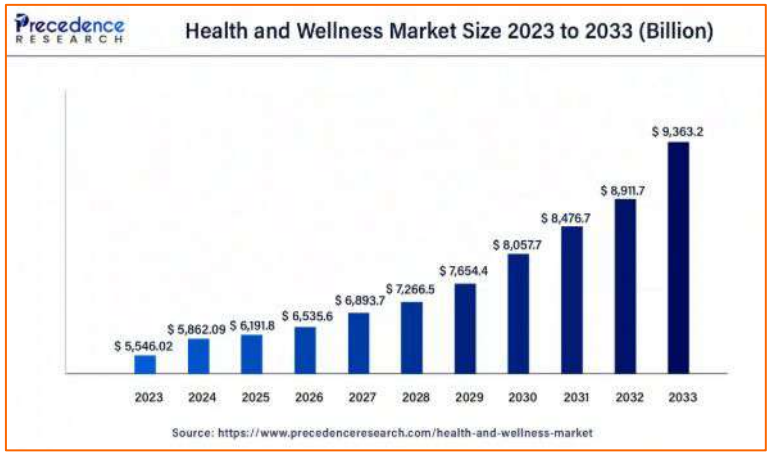


Secular Trends 3

Obesity and Health





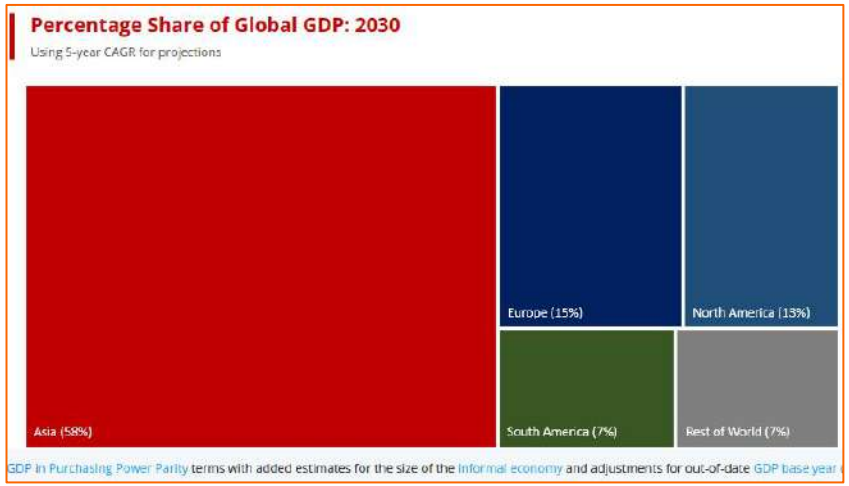


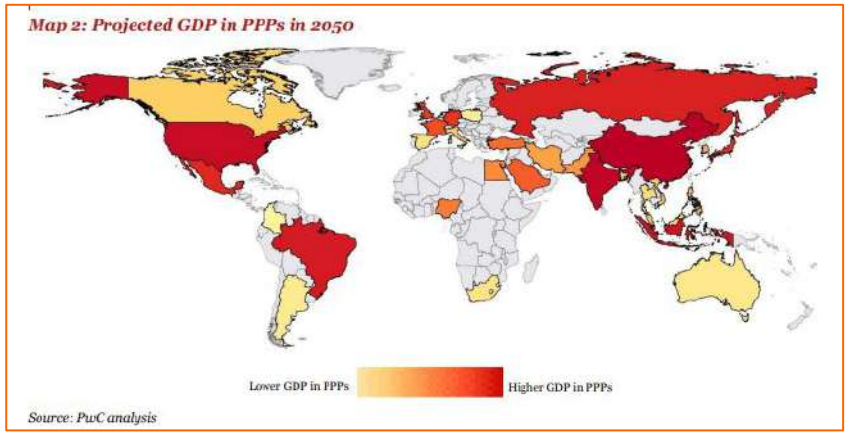
- From 1990 to 2022, global obesity rates more than quadrupled in girls (1.7% to 6.9%) and boys (2.1% to 9.3%), with increases seen in almost all countries. In adults, obesity rates more than doubled among women (8.8% to 18.5%) and nearly tripled in men (4.8% to 14.0%) between 1990 and 2022. (NCD Risk Factor Collaboration, NCD-RisC in collaboration with the World Health Organization, WHO)
- 79% of adults with overweight and obesity will live in Low- and Middle-Income Countries (LMICs) by 2035. It is projected that the number of adults living with obesity will rise to 1.53 billion in 2035 (worldobesity.org).
- Qatar, Egypt and USA have rate of obesity> 40%, China >8% and Japan has the lowest rate of obesity among the OECD member countries at 3.2% WHO, 2022).
- Health expenditure per capita over the last decades always increased (OECD).
- Global health and wellness market size keeps growing and it accounted for USD 5,862.09 billion in 2024 and is expected to be worth around USD 9,363.2 billion by 2033, at a CAGR of 5.34% from 2024 to 2033. (Precedence Research).

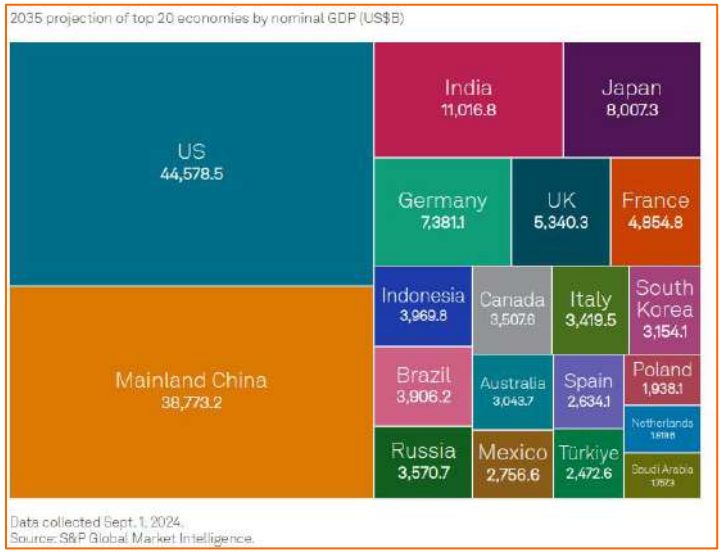
TECHNOLOGIES
Bring the future to humanity

Secular Trends 4

Asian growing economies





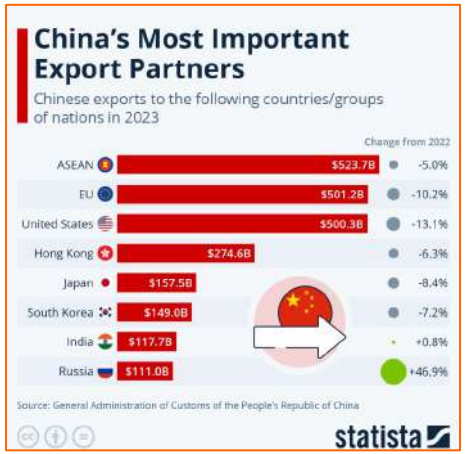


- Asian share of global GDP is projected to reach around 58% by 2030, up from 55% currently. EU share is expected at 15% and USA 13% (worldeconomics.com).
- S&P Global projects 4.2% expansion in 2025 and 4.1% in 2026, slightly lower than earlier forecasts of 4.4% for both years (S&P).
- By 2035, emerging markets will contribute about 65% of global economic growth. By 2050, emerging markets' old-age dependency ratio is expected to reach 35%, still well below the 50% expected for high-income countries (S&P).
- China's current share of world GDP at PPPs stands at almost 18%,and it is projected to increase to 20% by 2050 while India is projected to .rise steadily to over 15% by 2050 (PWC).
- Six of the seven largest economies in the world are projected to be emerging economies in 2050 including China, Japan, India and Indonesia (PWC).

TECHNOLOGIES
Bring the future to humanity

Secular Trends 5

World polarization



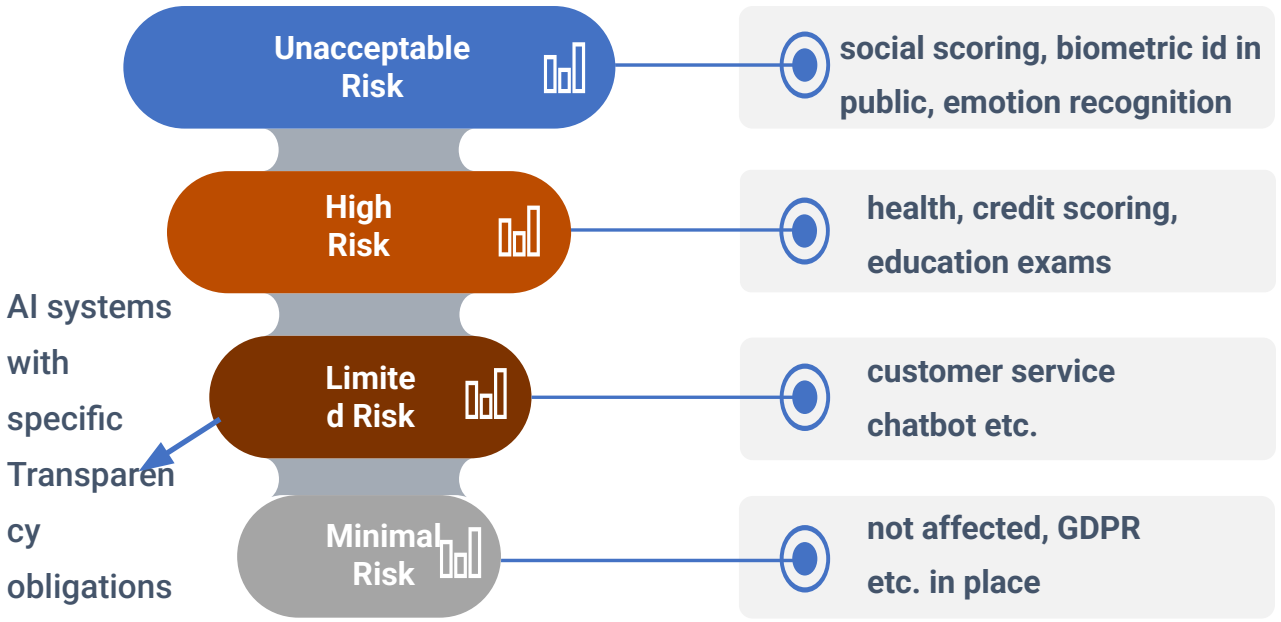
With China competing with US in all fronts, the world is polarised:

- China has been bolstering its investment in the BRICS countries as a counterweight to the G7 (various sources).
- The European Union is navigating a complex position between the US and China, debating whether to pursue equidistance or alignment. Equally, generally Asian and MENA countries and specifically South East Asian Nations (ASEAN) countries have consistently emphasized their desire to avoid taking sides in the US-China competition: both countries are ASEAN's largest trading partners (various sources).
- Taiwan faced significant incidents in 2024: 650% increase in cyber attacks on the telecommunications sector, 57% increase on defense supply chain targets and 70% increase on transportation targets. Record number of military aircraft breached Taiwan's ADIZ (various sources).
- Technology has become a key battleground, with both countries implementing restrictions and investing heavily in critical sectors like AI, quantum computing, and semiconductors. Famous example is the US Chip Act (various sources).
- The new Trump administration is expected to continue the tension with China (tariffs, trade wars and other restrictions).

TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 1

governance AI: EU (and others) AI regulations in 2024



Limited Risk

- **Limited risk regulations (general purposes):**
- would apply in 24 months from Journal Publication
- technical documentation available for clients of the AI system
- 'summary' of content and data used in training
- policy in place to respect copyright law

High Risk

- **High risk regulations:**
- would apply in 24-36 months from Journal Publication
- all requirements of limited risks
- AI model evaluation records
- AI model assessment and mitigation of risks
- management and response policy of incidents
- cybersecurity measures and adversarial testing

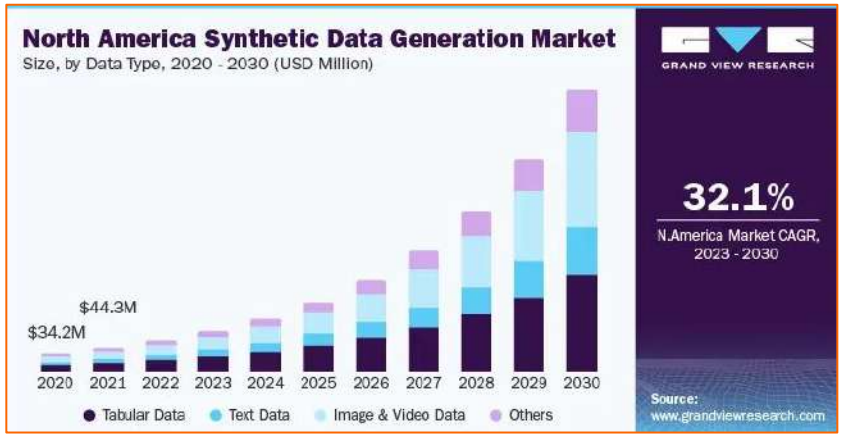
-Most regulations follow a similar structure of the EU AI Act: the focus is on high-risk uses of AI (e.g., healthcare diagnostics, education etc.) to ensure accountability and prevent harm. Using AI in this category means developing and maintaining model evaluation records, incident response and cyber security measures.
-The definition of 'risk' may be different in different jurisdictions: in California is large-scale AI systems that require over $100 million in training costs or pose significant risks (e.g., mass casualties or material damages exceeding $500 million, EU depends on sectors.
-China vs EU: China has 'vertical' regulations focused on specific technologies (AI) while EU has a 'horizontal' framework applicable to all industries using AI. Also Chinese approach requires 'traceability' of output back to source while EU approach requires disclosing data information.

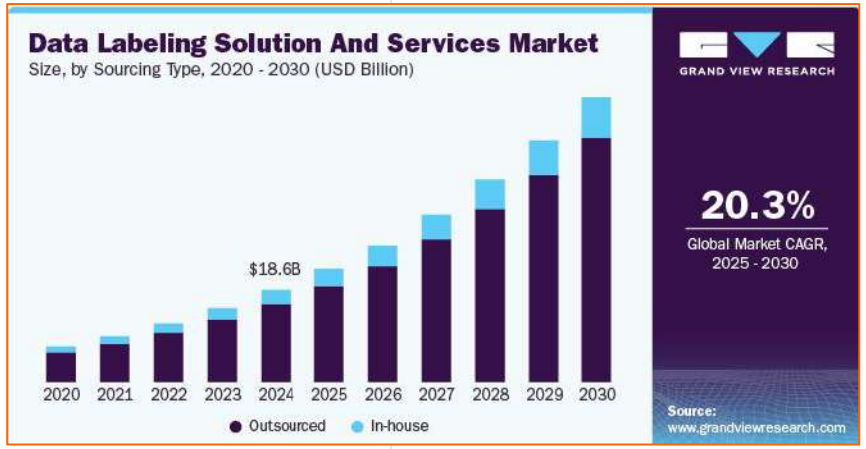
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 2

Synthetic VS human data generation

Humans still produce significant data but synthetic data is increasing as well.





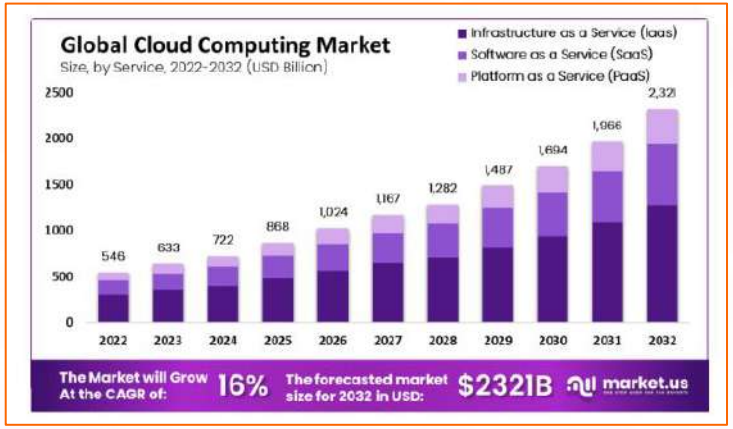
- By 2025, the global volume of data ‘created (includes data that is newly generated, captured, copied, or consumed) is projected to rise to 181 zettabytes. Videos account for over half of internet data traffic. Other estimates suggest it could reach 394 zettabytes by 2028 (digitalsilk.com).
- Global data labeling market was valued at USD 18.63 billion in 2024 and is projected to grow at a CAGR of 20.3% to 2030. The market is growing due to the increasing demand for AI and ML across industries.Most of the labelling is outsourced. Scale AI deal with OpenAI confirmed the trend. (grandviewresearch.com).
- Various estimates project synthetic data market size will grow CAGR 32-35% between 2025-2030. Tabular and Visual data are expected to grow more than other types.

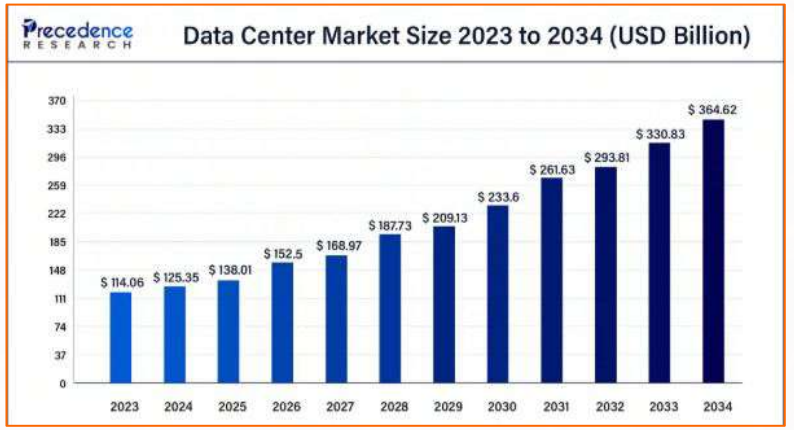
TECHNOLOGIES
Bring the future to humanity

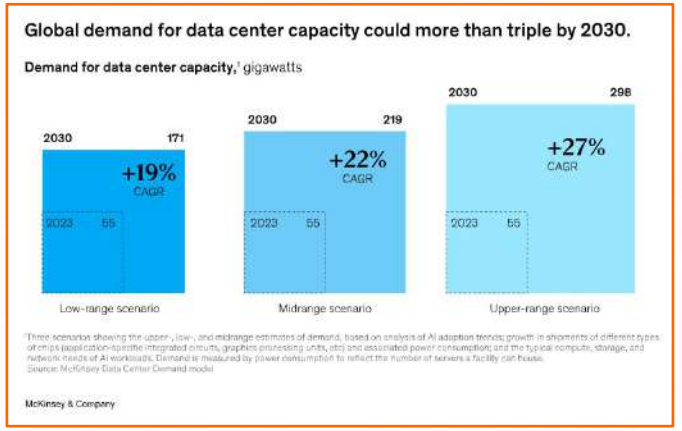
Trends Catalyst 3

Cloud and data centers growth

Cloud and data centers are growing and becoming more 'sophisticated' (AI, automation, data platforms etc.) with specialised services on top of the infrastructure. Sovereign (public) cloud demand keep contributing.





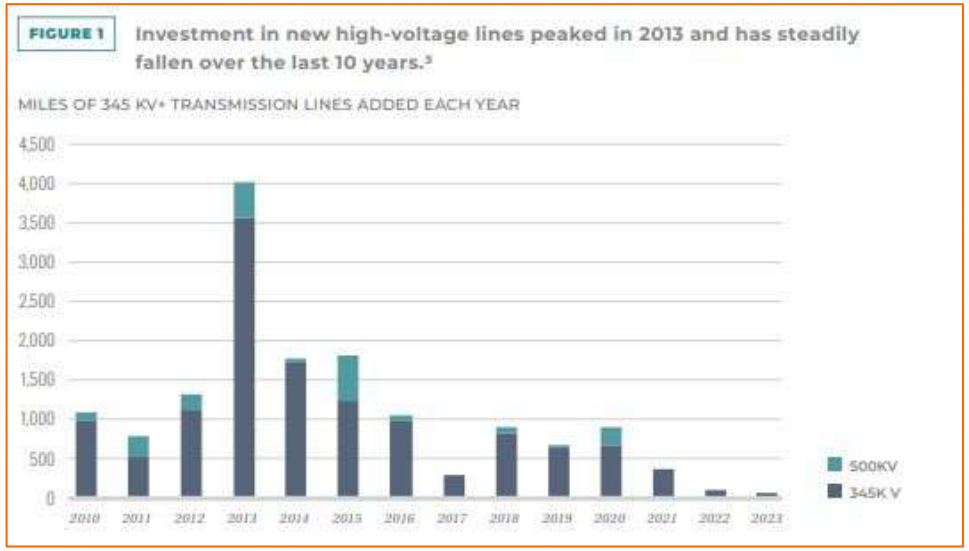


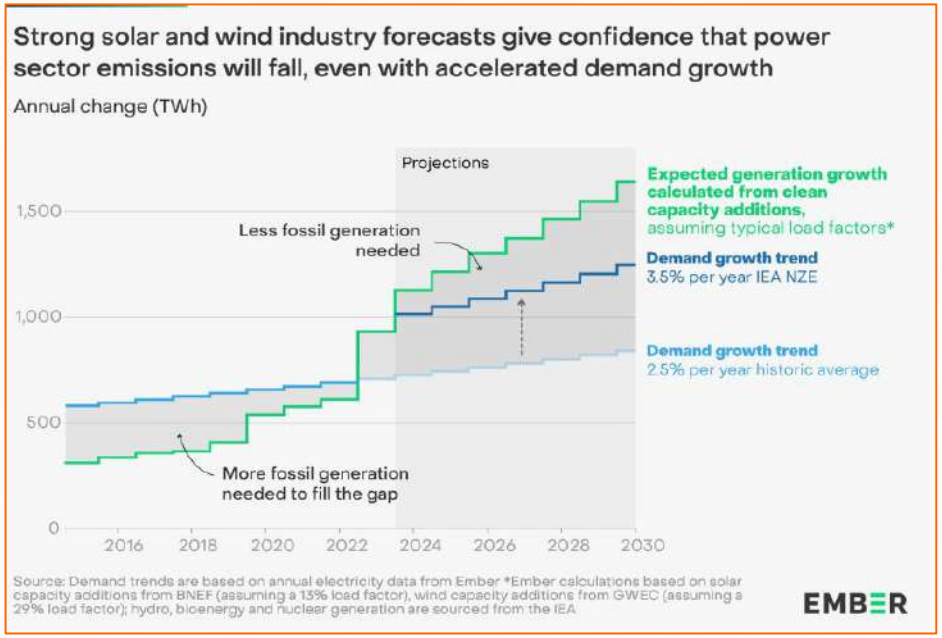
- Global data centers hardware/software sales reached $282 billion in 2024, with public cloud infrastructure now accounting for $156 billion of the total (Synergy Research Group). The global data center market is expected to reach $364.62 billion by 2034, growing at a CAGR of 11.4% from 2024 to 2034, faster in Asia Pacific slower in USA (6%). (Precedence Research).
- 62% of data centers is currently in USA (~5K), 5% in China, UK and Germany have 9% each. North America is 40% of global cloud market.
- Demand for AI-ready data center capacity is projected to rise at an average rate of 33% per year between 2023 and 2030 (McKinsey).
- BigTech (Google, AWS, Meta, Microsoft etc.) invested more than $180 billion in data center expansions and related in 2024 (constructiondive.com). Also, BigTech spent more than twice 'training' models compared to running them for their users (New Street).
- Worldwide spending on public cloud services is forecast to reach $805 billion in 2024 and double by 2028, with a CAGR of 19.4% (idc). AI platforms are expected to be the fastest-growing technology segment in the cloud market, with a five-year CAGR of 51.1% (impossible cloud). Global Cloud Computing Market s expected to be worth around 2.9n by 2033, growing at a CAGR of 16.8% (market.us).

TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 4

Energy supply and consumption





AECG

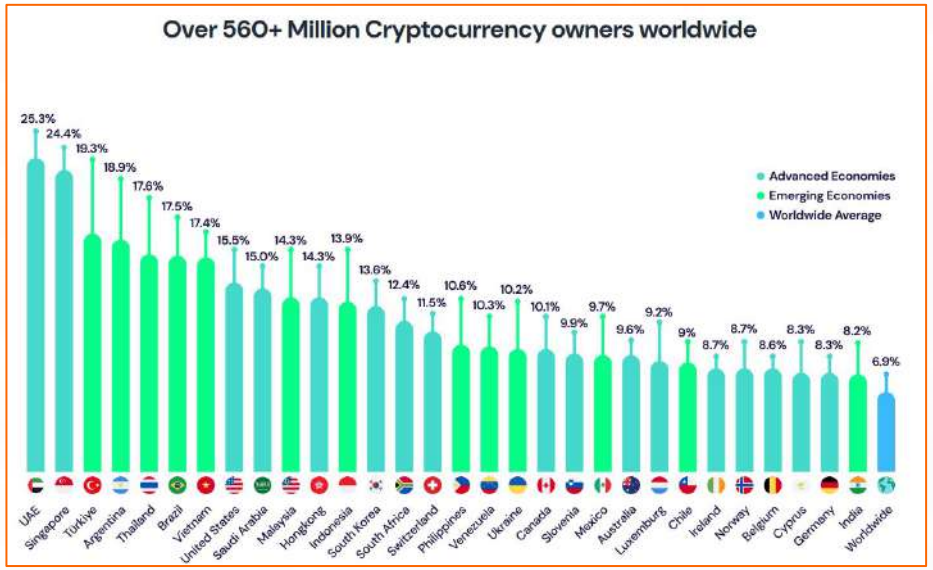
As energy overall supply increase, the grid to manage the distribution becomes the bottleneck:

- World supply of Oil and LNG is still projected to over the demand until 2030. USA still larger producer (~10%) but expected peaking in 2026 (EIA reports). Coal is set to decrease in the West, potentially increase in Asia (China consumes 60% of world coal).
- Still long term 40-60% energy supply comes from fossil (by 2050) rest renewable/low emissions (EIA reports).
- EU & UK are pushing on renewables (UK closed leven its last nuclear power plant) further increasing electricity demands.
- Electricity production growth increased by 28% due to data centers and air cooling (Ember). Even if demand rises up to 3.5% yearly, production of renewable should rise more (Ember). Battery minerals supply should be enough for the next few years (various sources).
- The main issue is transmission grid stability (BDEW-German Association of Energy and Water Industries, GridStrategies, PMJ, AECG), especially long range high voltage lines are underinvested and inadequate if solar/wind supply increases.

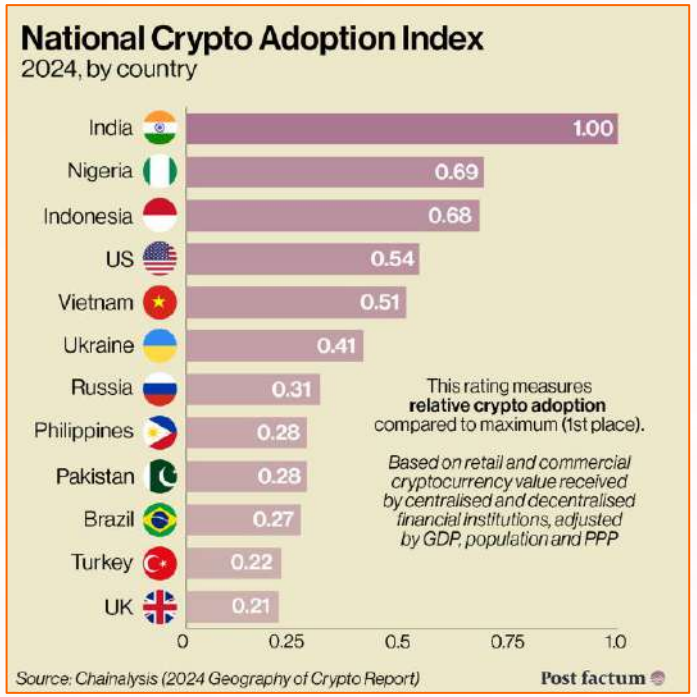
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 5

Crypto Ecosystem Development (and AI)



triple-a.io



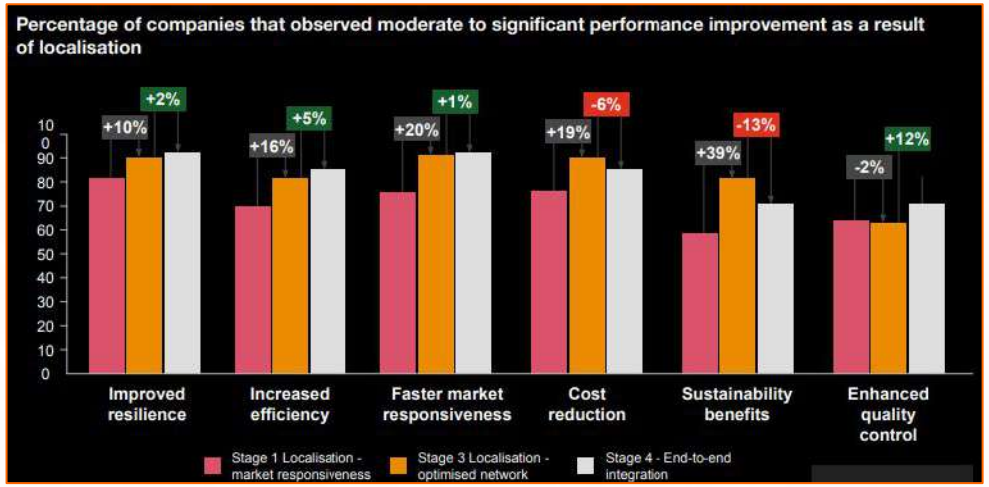
As AI, crypto and its applications are increasing:

- Bitcoin crossed the 100K USD and keep bring new media level of attention to the whole crypto ecosystem. Adoption PPP and GDP adjusted is higher in developing countries (India, Nigeria, Indonesia etc.).
- As of 2024, global cryptocurrency ownership rates are estimated at an average of 6.8%, with over 560 million cryptocurrency users worldwide. 61% are male and 34% aged between 25-34 (triple-a.io).
- In 2024 ETFs on Ethereum (July) and Bitcoin (January). New ETFs approval are rumoured in 2025 (Litecoin, XRP, Solana) etc. pushing crypto demand even higher.
- New cryptos are emerging on the AI space to solve various ‘decentralised’ issues, share of computing power (ex. Render), NTFs to authenticate data, train AI models (ex. Bittensor, ASI), build and share ‘agents’ and revenue from its usage (ex. Virtuals, Griffain).

Trends Catalyst 6

Supply chain localisation, sovereign cloud and workforce

Countries trying to be less vulnerable are localising supply chains and implemented ‘industrial policies’.



After the Pandemic and given the political worldwide instability , supply chain is getting localised or shifted to ‘nearshoring’ countries:

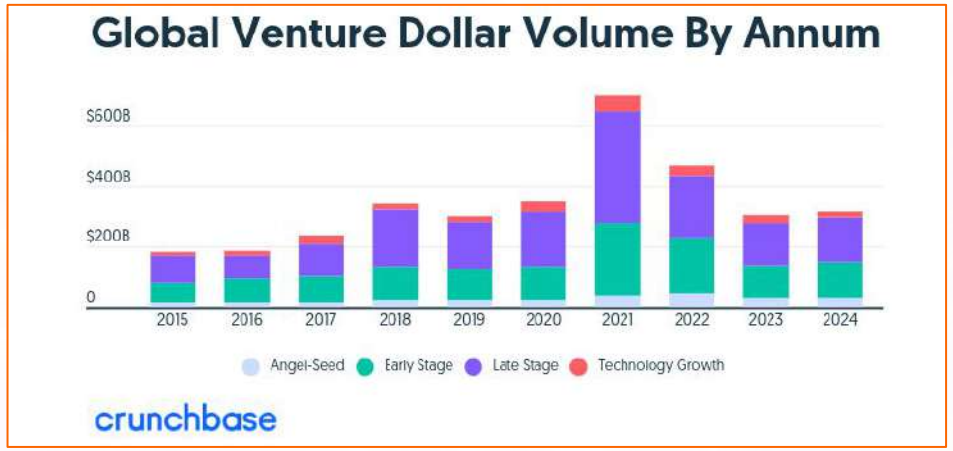
- US and other countries bans on export of chips and related technology: TSMC and Lenovo have announced plans to expand local production capabilities in the US and Saudi Arabia, respectively.
- According to a PWC survey, 82% of companies reporting improved resilience, 77% reported cost reduction through localisation and a moderate localisation enhances supply chain performance across all drivers.
- Sovereign cloud: AI and data regulation favours data storage and processing within a country/economic area. It is expected a double digit growth in the sovereign cloud market over the next several years (broadcom).
- According to bls.gov, US workforce is employed 13% in professional services, 13% healthcare and social, 12% public sector, 8% manufacturing. Employment growth is expected in healthcare, public sector and manufacturing.

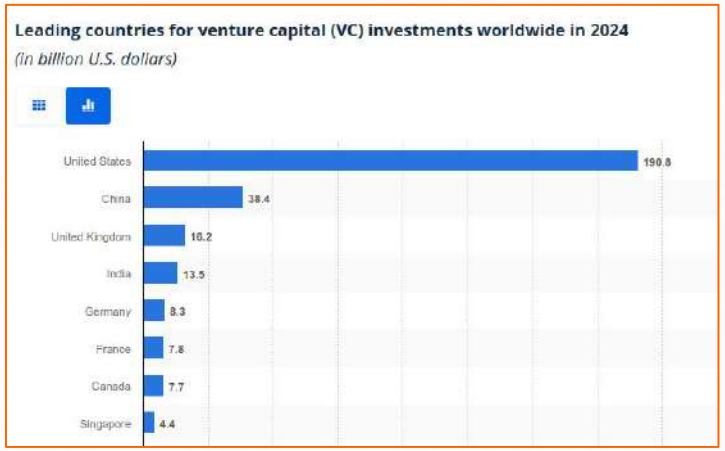
TECHNOLOGIES
Bring the future to humanity

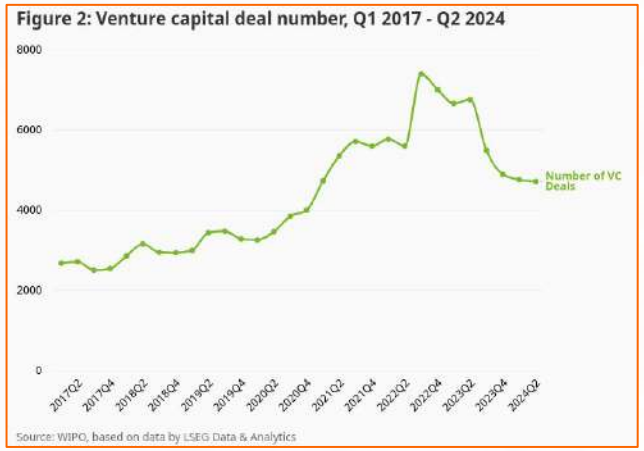
Trends Catalyst 7

VC investment

Due to rising interest rates and other factors, tech investments have been stable or concentrated in mega rounds.







statista

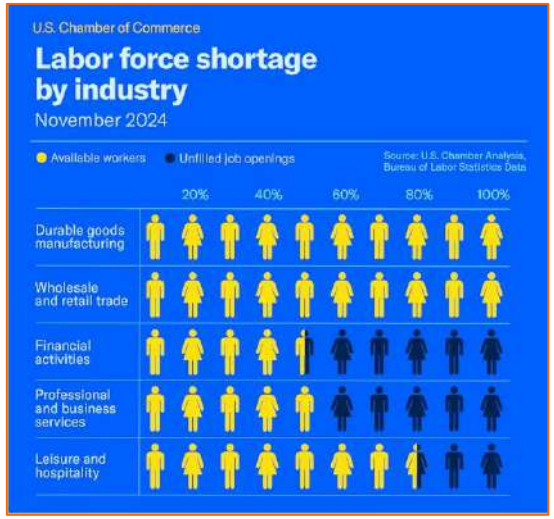
Venture capital investments in tech startups continue to slow down:

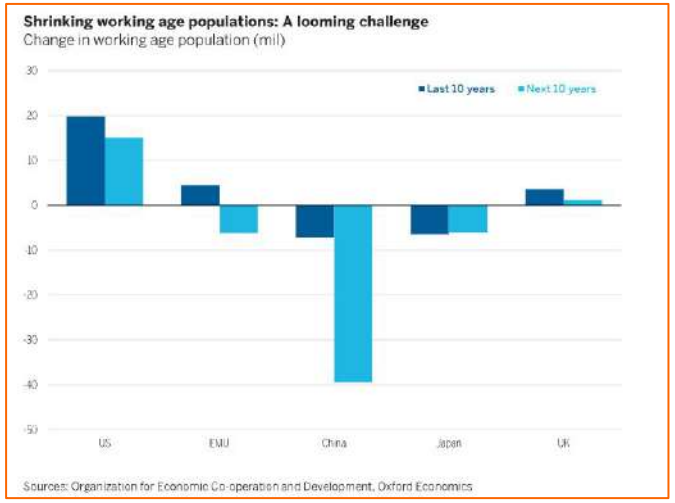
- Global VC investments in 2024, depending on sources, either slightly declined or increased compared to 2023 in terms of total deals values at around 315-33B (wipo, crunchbase, dealroom). Vast majority of global VC money is still spent in US deals.
- Per Crunchbase, AI funding was about 30% (100B) of the total VC funds and the rest of funding went to sectors impacted by AI. 58B (19% of 315B or 58% of 100B) went to mega rounds (1B+) similar but up from 2023 where 15% went to mega rounds. This confirms 'concentration' of capital into few shortlisted winners: xAI (12B), databricks (10B), openAI (6.6B), Waymo (5.6B), Anthropic (4B), Anduril (1.5b), G42 (1.5B), CoreWeave (1.1B) and Wayve (1.1B). Unsurprisingly, number of deals per Wipo are decreasing.
- Seed stage deals may have increased compared to 2023 (Bain reported an increase, Crunchbase,Dealroom reported a flat growth).

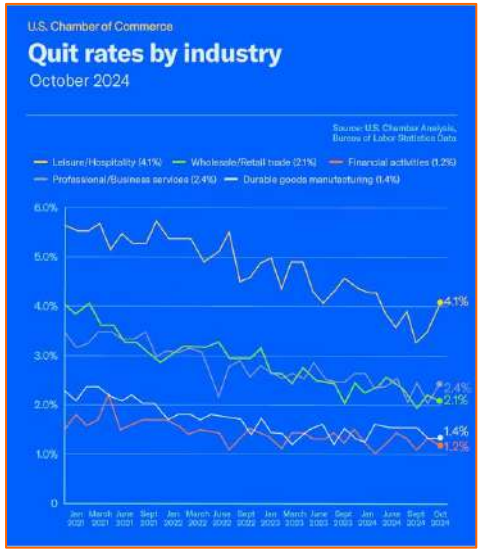
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 8

Labour shortages







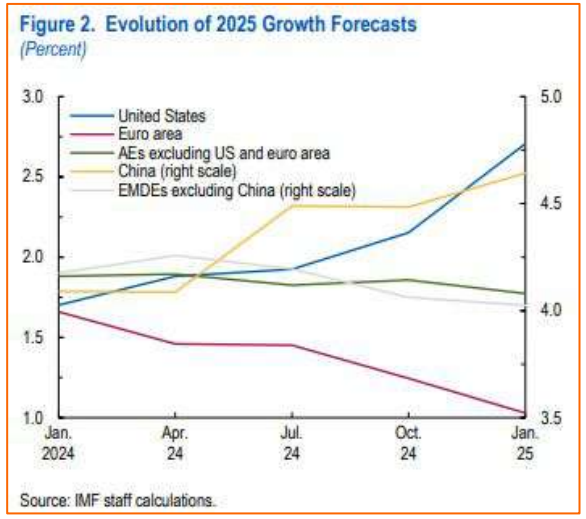
Labour shortages are expected especially in certain sectors (healthcare, manufacturing, hospitality etc.)

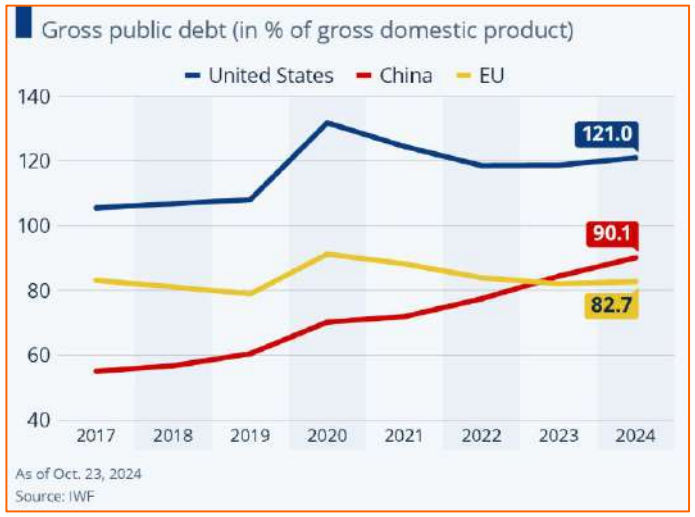
- Large baby boomer generation now retiring and being followed by the smaller Gen X and the elongated profile of Gen Y and Gen Z (i.e., Gen Y and Gen Z are large but births were spread out over a long period), it may induce relative labor scarcity to be one of the enduring themes of the next decade. (Wellington Management).
- In 2024, the United States has about 1.2 job openings for every unemployed person (McKinsey). Across all industries, hiring rates have continuously outpaced quit rates. (uschamber.com).
- In USA, the healthcare sector is expected to face acute worker shortages, particularly in nursing and specialized medical roles. Mercer projects a deficit of over 100,000 healthcare workers in the U.S. by 2028 (https://www.shrm.org).
- In the USA, By 2030, the manufacturing industry could have 2.1 million unfilled jobs, potentially costing the U.S. economy $1 trillion (https://www.shrm.org).
- By 2030, more than 85 million jobs could go unfilled globally due to skills shortages and aging population (kornferry.com).

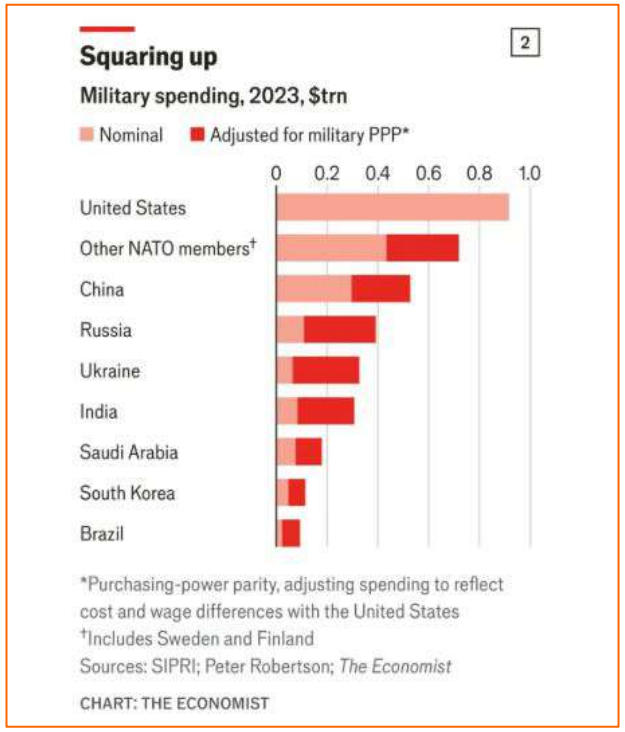
TECHNOLOGIES
Bring the future to humanity

Trends Catalyst 9

Public spending & GDP/debt ratios





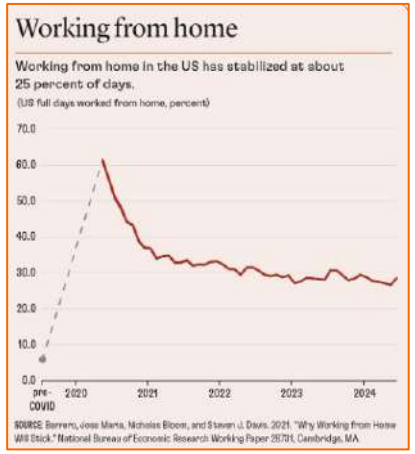


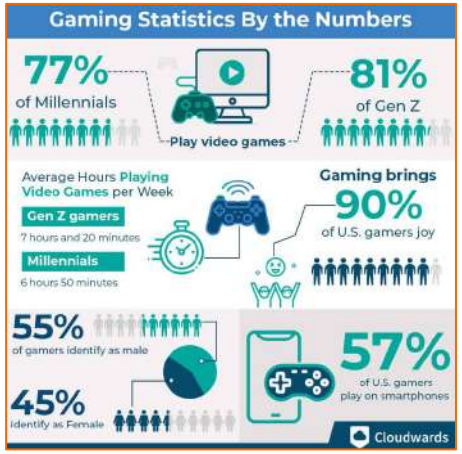
Defence budget is increasing while GDP/debt ratio is stable, marginally improving:

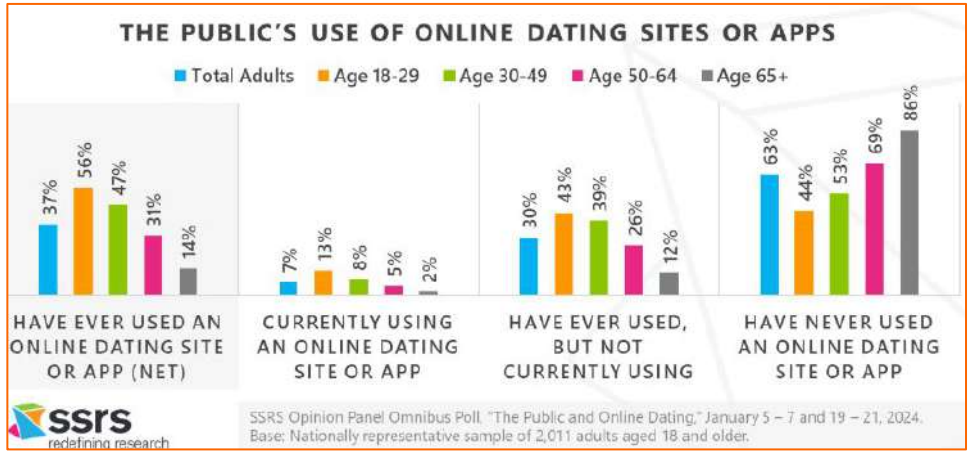
- Global military expenditure reached an all-time high of $2.44 trillion in 2023, marking a 6.8% increase from 2022 (sipri.org).
- The U.S. remains the world's largest defence spender, with a budget of $916 billion in 2023, accounting for over 40% of global military spending. (sipri.org) The U.S. defense budget is expected to reach $852.2 billion in fiscal year 2025, a 3.3% increase over fiscal year 2024 (senate.gov).
- China's defense spending has increased by over 50% in the last decade, from $132 billion in 2014 to $234 billion in 2024 (aei.org). China announced a 7.2% increase in its defense budget for 2024, continuing a trend of growth outpacing economic expansion (forecastinternational.com). Although still the largest defence budget, USA adjusted PPP (purchasing-power-parity) is ~22% larger than other NATO members and ~40% larger than China (economist.com, see image above).
- Public debt keeps increasing as % of GDP in China and US, stable in EU. (IMF, see image above). Inflation is expected above 2% in Western countries (various sources). GDP is expected to grow in Asia, China but not in EU (IMF).

Trends Catalyst 10

Virtualisation of living





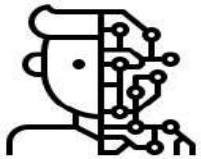


New generations increasingly prefer using virtual environment compared to physically going out:

- Millennials (Born 1981-1996) and Generation Z (Born 1997-2012), ~50% view online experiences as meaningful substitutes for in-person interactions. 40% of Gen Z socializes more in video games than offline (zdnet).
- By 2025, the gaming industry is expected to generate ~$211B in revenue, with mobile gaming contributing $116B (EY). The esports audience reaching 640.8M in 2025 (fortunly).77% of Millennials and 81% of Gen Z consumers play video games (cloudwards.net).
- Top-tier e-sports players (League of Legends, Dota 2) can earn up to $2 million/year including sponsorship and streaming and the average player make $500,000/year. In comparison, a top football player (Premier League, Liga etc.) earns $2-4 million or more (tribuna.com).
- 37% of US adults used online dating at some point. A majority of adults (61%) believe relationships that begin on dating sites are as successful as those that begin in person. 37% of current users are aged 18-29, 38% are aged 30-49; 19% are aged 50-64 and only 6% are aged 65+ (ssrs).
- 22% of the US workforce will work remotely by 2025 and 14% of Americans work from home all of the time, according to Pew Research Center. 41% of employees work remotely on a hybrid basis (usatoday.com). US employees works remotely 25% of the time (IMF).
- Smartphone and internet overuse may decrease cognitive abilities (https://bit.ly/3PXY9ev) and similar for AI usage (https://bit.ly/3CFyi7Z).

AI capital, energy and economy

Trend Prediction 1



AI investment will keep growing and optimisation improvements should make energy consumption under control.

- Deepseek (V3, R1)et al. showed competitive results without prohibitive budget (<$25M, final run) hence less energy but post-training (RLHF, others) is crucial.
- Blackwell GPU will to be up to 8x more efficient for inference (from standalone to large configurations).
- In AI training, data center power optimisation can lead up to 50% energy reduction. https://bit.ly/40FK0HV
- Llama 65B reported ~4 Joules per output token, hence 0.00037 kWh per 333 tokens (in 20s). Assuming 1h for a human to write the same, the average energy consumption is ~6 kWh.https://bit.ly/42uChyU, https://bit.ly/40P9OCB
- Lumen Orbit raised $10M to build data centers in space to 'spare' electricity and operation costs on Earth.
- GenAI services keep reducing the costs. GPT-4 inference in Mar 2023 was $36/M tokens, in Aug 2024, was $4/M tokens.

→

By 2026

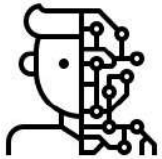
- VC interest in the early stage will grow 15%+ compared to previous year (rounds <$15M).
- A startup with data labelling capabilities reaches $30B+ valuation (or equivalent if bought/IPO).
- There are continuous improvements of hardware, data centers and software in terms of efficiency. There is no reason to assume new efficiencies will keep coming and the combined (compounded) effect should make the AI increasing energy demand manageable as AI usage increase.
    - new data centers construction due to AI will cool off but grid stability will emerge as the bottleneck in the news.

→Triggered from Catalysts: 1,3,4,6,7,9

Technologies

AI Hardware

Trend Prediction 2



Hardware performances are still dominated by Nvidia but a new mid-market with clients looking for cost/performance balance can emerge.

GPUs

- Nvidia Blackwell, released in Jan 25, in training models, should be up to 4x more powerful than H100-200 in large configurations (10+ GPUs) but up to 2.5x in a small configuration (1-8 GPUs). Regarding inference, it should be up to 5x in small configurations and up to 10-30x in large configurations. The energy efficiency should be up to 8x.
- Using MLPerf 4.1 benchmarks (Aug 2024): latest AMD MI300X is on par with Nvidia. H100/H200, new generation Nvidia Blackwell performed up to 4x better than H100 (large configurations).
- Intel lag behind in GPU performances.
- Many BigTech (Meta, Microsoft etc.) plan to build chips, as Google and AWS have their own already for years (internal use).

Inference only

- Cerebras reported the best performances to date compared to competitors Groq, Graphcore, Sambanova et al.

Other startups raised funds >$100M.for GPU/Inference chips.

→

By 2026 :

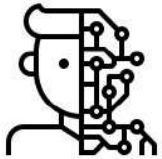
- Nvidia is still significantly ahead in term of performances and efficiencies especially in large configurations but you pay a premium:
  - Nvidia will remain the go to chips for large clusters and big orders (clients)
- Due to small labs (e.g. Deepseek) showing excellent performance with ‘relatively’ old chips and small configurations a new ‘mid-market’ can emerge, considering the cost/performance ratio of AMD GPUs:
  - AMD will have 10%+ of the market for single or small GPU configurations
- A new chip player (i.e. not Intel, Nvidia and AMD) with a different technology will announce orders worth >$500M.

→Triggered from Catalysts: 3,4,5,6,7

Technologies

AI adoption in cybersecurity and legal issues

Trend Prediction 3



AI significantly contributed to the growth of malicious activities and it will trigger counter measures that also leverage AI.

- 1 hour internet shut down costs $450m (US, merchantmachine).
- From 2023 to 2024 attacks are increasing: +15% ransomware, +53% DDoS (qrator), +190% in phishing (netskope). Bots are 73% of the internet traffic, 46% of social media (Arkose Labs).
- +225% automotive attacks in 3 years, almost all remote attacks. Also, in 2024 there were ~26k autonomous vehicles (market.us).
- Cybersecurity spending is 7-11% CAGR (statista, market.us).
- AI safety product industry is emerging: Virtue AI, Protect AI, etc. aim to both protect an AI system and avoid returning 'bad' output.
- Detecting deepfakes is hard (see 2024 predictions). Most socials requires content to be labelled if it is genAI produced, OpenAI put C2PA metadata in its genAI content. Invisible watermarking via NFT may be promising (ex. Imatag, Xeal).
- Agentic AI contributes to attack and defence.Ex: AI botnets scan 1k of sites, find weaknesses and then attack. In defence, AI systems autonomously detect, analyze, and respond to threats.

→

By 2026 :

- There will be a serious remote vehicle attack using AI which will raise awareness on cars vulnerability.
- A safety AI startup will be acquired by a large tech player for more than $100M.
- AI regulations only related to automotive overall will increase in western countries and China.
- At least one new product for watermarking combining.blockchain, AI and cryptography will be presented but it will fail to make significant traction.

→ **Triggered from Catalysts:** 1,2,6,7,9,10

Technologies

Industry AI adoption

Trend Prediction 4

Industry adoption is progressing at speed but in specific use cases (as predicted last year). Rest of industries experience slow adoption and are still searching good use cases.

- Organizations using genAI in one or more functions rose to 72% in early 2024 (McKinsey). AI spending will continue (but 'internal' i.e. no consultants) but still searching for ROI (Wharton and GBK report 2024). Large US banks take ~1 year to roll out an AI solution (various) while IDC reports an average of 8 months.
- Finance and Healthcare reached 71% and 66% adoption (statista) but in 'specific functions' (i.e. 'small' AI solutions as we predicted last year) like fraud detection, risk analysis (finance) and diagnostics, admin automation (healthcare). Highest adoption is in marketing, operation and procurement.
- AI adoption's main barrier is the difficulty of estimating the value of AI projects (Gartner). EY reports AI projects have positive returns.
- USA and China score 77 and 66 at the IMF AI Preparedness Index
- Deepseek and other research showed the importance of RLHF and other post-training techniques which rely on data labelling.
- 300M jobs may at risk of being automated (Goldman Sachs), AI education bills have been proposed at US congress in 2024.

→

By 2026 :

- Due to AI regulation and the raising cost of using AI, the importance of data labelling, Knowledge Process Outsourcing (KPO) is set to grow:
  - US KPO will grow faster than predicted (Precedence research: $110.69B in 2024 expected a 13.7% growth).
- We keep seeing 'small/narrow' AI solutions the most adopted in specific functions while large ('sophisticated') models adoption in processes will be slow especially in regulated industries (as we predicted last year. Note: this does **not** include individual professionals using genAI at work).
- US admin workforce (about 10-12% of the total) will not lose more than 1% due to AI.
- Average AI roll out will still be around 8-12 months in large companies.
- US will establish a nation-wide AI education program to reskill workforce.

→ **Triggered from Catalysts:** 1,2,3,6,8,9,10

Technologies

Consumer AI adoption of LLMs/GenAI tools

Trend Prediction 5

Adoption of genAI tools will increase overall. We will 'talk' a lot about 'agentic' but adoption will be limited and driven by the crypto space or individual/lab projects.

- 31% US adults (18-64 y/o) used Chatgpt at least once Aug-Sep 2024 https://bit.ly/3CvhOPQ Note: about 60% of an advanced country workforce is highly impacted by AI (IMF, 2024 data).
- December 24 data (similarweb). Monthly visits: ChatGPT 3.69 , Claude, Copilot 73M each, Gemini 261M. Traffic: 14% USA, 10% India, 4% each UK Brazil and Indonesia. In the top 5 countries ~85-90%+ traffic goes to ChatGPT. Indonesia and US have the most 'diverse traffic' with 85-87% traffic for ChatGPT, 11% Gemini, 2-3% Claude.
- The smarter you are, the smarter you still are in using AI (compared to less smart employees). https://bit.ly/4aGTgQx
- Software developers using LLMs benefitted the most up to 126% more projects per week (Nielsen).
- Individuals or labs are experimenting with 'agentic' solutions, supervising their own LLM(s): AI scientist (by Sakana AI), AI game developer (https://bit.ly/40H6EzF), crypto trading analyst AIXBT.
- Custom GPTs (agents) are >3M+ but it is unclear the demand/function (similar for Claude Compute and others).

→

By 2026 :

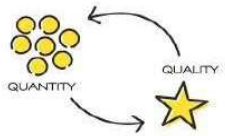
- ChatGPT and competitors combined are used at least once by 40% of the US workforce.
- With cheaper inference and more AI demand, top 3 non first world (and excl. China) countries with large, outsources workforces (like India etc.) will have a combined 25%+ of traffic of chatGPT and competitors.
- AI Agents remain still too complex to set up for non-coder (and a bit of entrepreneurial skills) while basic AI adoption is low.
  - 2% of news x accounts (>100K followers) will be agents
  - at least 50 agents' accounts (>100K followers) will be built on blockchain.
- Major AI foundation model players will re-focus on winning the performance battle more than their agentic market:
  - adoption of blockchain based agents (GRAFFAIN, Virtuals etc.) will be higher than GPTs, Claude Compute.

→ **Triggered from Catalysts:** 2,3,6,8,10

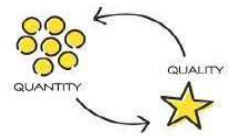
Technologies

Models' Advancements: Text

Trend Prediction 6



Advancement in AI performances will be a mixture from pre-training and post-training methodologies. Broadly 'larger should be better' (up to a T number of parameters or less) but the focus of the year will be 'optimisation' even at scale.



- DeepSeek R1 results (close performance to top model OpenAI O1/O3, with less money, a fraction of the people, <200, significant less GPUs) show 'small' players still can perform and the race is not over for new incumbents. And RL is key.
- From the R1 paper 'advancing beyond the boundaries of intelligence may still require more powerful base models and larger-scale reinforcement learning' (i.e. raw power to compute).
- Last year showed that 'new' architectures (Mamba, ModernBERT) can perform a good levels if re-adapted.
- Stil,l models are undertrained (lack of data) but new methods to develop quality synthetic data had a degree of success (Phi models by Microsoft, Cosmopedia by HuggingFace). Sharing data/resources via blockchain is attempted (ex. Bittensor).
- Post-training techniques start to emerge as 'differentiators' to model performance (DPO, RLHF, CoT etc.) using 'specific data'.
- HuggingFace detailed a process for generating synthetic 'Cosmopedia'. OpenAI GPT-5 may use synthetic data up to 70%.

→

By 2026 :

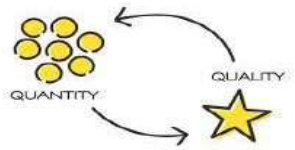
- The best 'overall' pre-trained model will be by a BigTech or Lab (>$100M in funding) with:
  - 40% or more data will be synthetic data
  - not be only transformer based (from MoE to MoArchitectures/Agents)
  - $40M or more in training (final run)
  - average scores max 15% better than the top <$10M models and <50B parameters (final run)
  - It may be proprietary first. If so, within 6 months an open weight will almost reproduce it.
- A top performing 'post-trained' model will rank top 2 on specific benchmark hard to reproduce with:
  - a 'mixture/hybrid' of post-training techniques
  - high scores in 1 hard metric (e.g. Math)
  - 'proprietary data' and by either a 'not' foundation AI model lab/BigTech or open data by a community in blockchain
- a data labelling firm will enter in the race with a 'data proprietary' model ranking top 10 in the benchmarks.

→ **Triggered from Catalysts:** 2,3,4,5,6,7,10

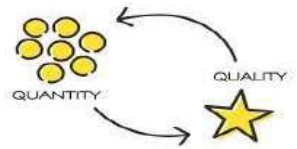
Technologies

Models' Advancements: Audio/Images/Video

Trend Prediction 7



Advancement in AI performances, especially from open source (weight) will lower the barrier to entry for new players.



- txt/img to video: Open AI Sora and chinese Kling 1.5 top the leaderboard. Short movie Heist (Google Veo 2) show great results but still complexity in usage ('1000s of generations'). Movie directors used AI in 'post production' (ex. The Brutalist).
- Txt to image: Midjourney V6, Stable Diffusion 3LT, Recraft v3, Flux perform best on leaderboards. Gemini 3, DALL-E 3, Grok all perform similarly. Deepseek Janus-Pro outperforms DALL-E 3. Small models (Moondream 2B) have decent results and can run almost everywhere (i.e. mobile).
- Multimodal: chinese models InternVL2.5 and Qwen2-VL, are already better than GPT-4o and Gemini 1.5Pro. Qwen2.5-VL just released is even better and open weights.
- Gamefactory and Google Genie2 show that you can create a video game environment with few prompts.
- Google NotebookLM (converts text into podcast and audio overviews) got traction. AI VTuber Neuro-Sama got a record on Twitch.

→

By 2026 :

- On the top 5 models' leaderboard for video or images there will be a new Lab/company coming from an emergent economy country (excl. China).
- A mobile video game built with genAI tools will go in the news.
- A text to image model not diffusion based will achieve SOTA Top 10 on the leaderboard.

→ **Triggered from Catalysts:** 2,3,4,6,7,10

Technologies

AI adoption in science

Trend Prediction 8

Biotech researchers are already experimenting with AI in the last years. We expect to see fruition on the AI developments combined with the economy of GLP-1 drugs.

- In USA, there is significant interest in reducing drug costs (ex. https://www.costplusdrugs.com) insurance and healthcare system in general (https://bit.ly/4aIOimc).
- GLP-1 drugs (e.g Wegovy, Ozempic) show results from diabetes to weight loss with cascade effects: soft drinks, alcohol and salty snacks will fall by as much as 4% through 2035 (Morgan Stanley), new clothes due to the slimmer size (times.com).
- Machine learning has been applied to GLP-1 with drug improvements (ex. https://bit.ly/40W3mdb, https://bit.ly/3WE1Qd2)
- New crypto 'communities' are emerging to built models and simulate proteins (bittensor/macrocosmos.ai subnet 25. (https://github.com/macrocosm-os/folding)
- Google released AlphaFold 3: it allows prediction of interactions between proteins and small molecules. And it has a free server for researchers to experiment.
- Research shows an LLM performing protein engineering. (.https://bit.ly/3PXjRz8).
- Microsoft showed a genAI assistant to design inorganic materials, MatterGen.

→

By 2026 :

- A GLP-1 variant which AI software/model contributed to optimise/discover/design will start FDA approval process.

→Triggered from Catalysts: 1,3,4,5,7

Technologies

BONUS: We will talk Pynglish → From 2024

Trend Prediction *Bonus* (semi-serious one)

Non coders will start picking up some code language.

- Developers are adopting copilot and alike AI software development tools quickly. https://survey.stackoverflow.co/2023/#ai
- Interns and others software related professionals (project managers etc.) learn becomes proficient in coding faster using AI tools. https://bit.ly/498yVSn
- Demand for developers overall will not shrink (apart from front end).

→

By 2027 :

- 10% of non-coders professionals will at least try genAI tools to make their one software scripts.
- non-coders will pick up developers jargon in common life.
- An authoritative English dictionary (like Oxford English Dictionary) will add a software word as a new recognised word.

→ **Triggered from Catalysts:** 6,7,8

Technologies

CONTACT US

andrea@aitechnologies.co

Andrea Isoni, Chief AI Officer, PhD

AI Technologies