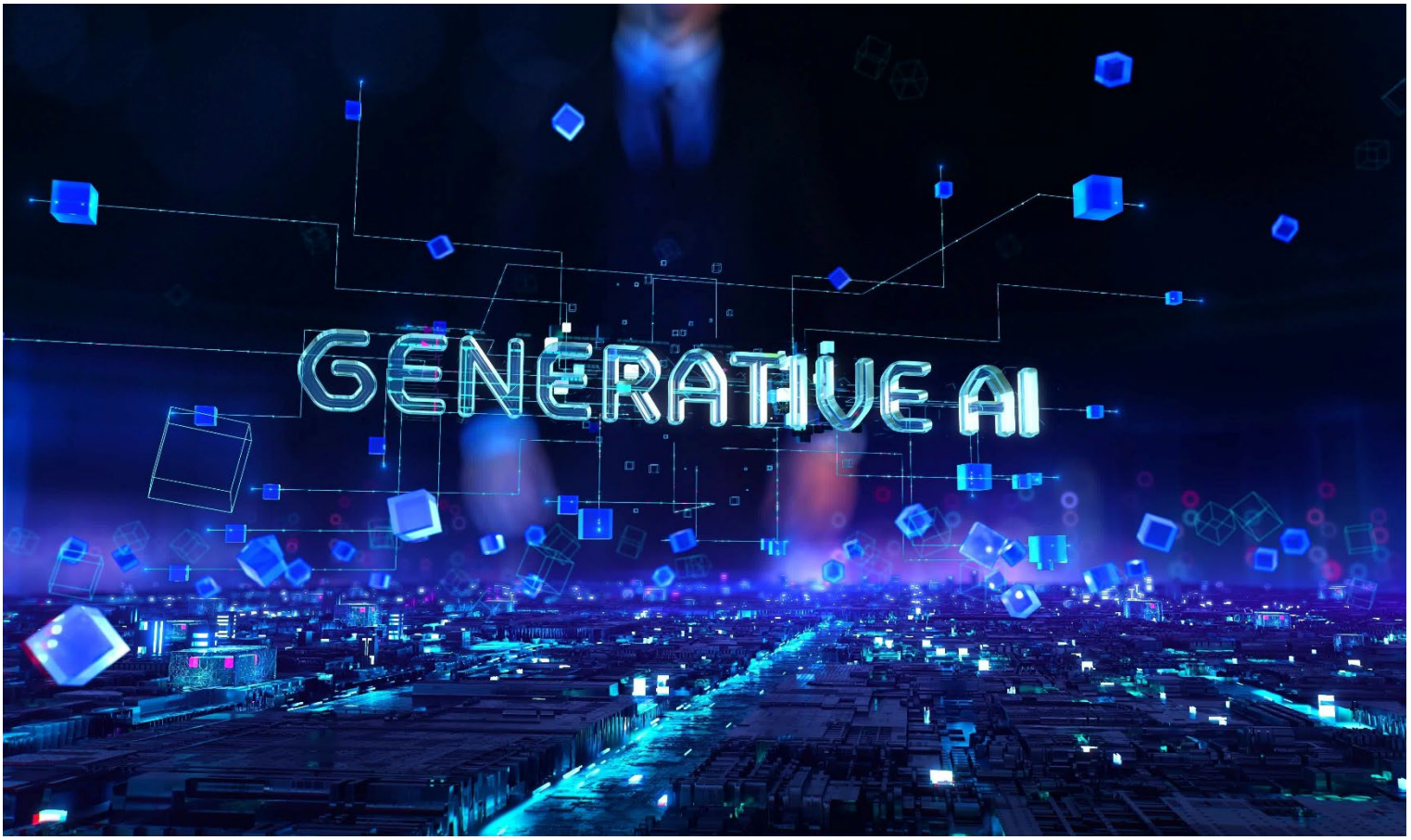
Citi GPS: Global Perspectives & Solutions

UNLEASHING AI

The AI Arms Race

Generative AI represents the latest inflection point in the evolution of AI. Although AI is not new, what is distinctive about Generative AI is the tremendous potential it holds to transform work across industries and boost overall productivity. We look at the opportunities for AI not only in the technology sector but across the supersectors. We also look at the race to dominate in the AI space by analyzing research papers and patents.



September 2023

Citi Research Analysts

Andrew Baum, MD
Global Head of Healthcare Research

Fatima Boolani
U.S. Software

Laura (Chia Yi) Chen
Asia Semiconductors & Components

Christopher Danely
U.S. Semiconductors

Patrick Donnelly
Life Science Tools & Diagnostics

Steven Enders, CFA
U.S. Back Office Software

Tiffany Feng
HK/China Consumer

Surendra Goyal, CFA
India IT Services, Media & Education Analyst

Aaron Guy
EMEA REITs

Simon Hales
EMEA Beverages

Nick Joseph
Head of U.S. REIT and Lodging Team

Ronald Josey
U.S. Internet

Andrew Kaplowitz
U.S. Multi-Industry, Conglomerate, and E&C

Peter Lee
Asia Semiconductor & IT Hardware

Paul Lejuez, CFA, CPA
U.S. Retail & Food Retail

Andre Lin, CFA
Asia Hardware

Carrie Liu
Taiwan Technology/Hardware

Craig Mailman
U.S. Retail & Industrial REITs

Atif Malik
Semiconductor Capital Equipment & Specialty Semiconductor

Asiya Merchant, CFA
U.S. Hardware

Itay Michaeli
U.S. Auto & Auto Parts

Takayuki Naito
Japan Electronic Components

Arthur Pineda
Head of Pan-Asian Telecom Research

Jenny Ping
EMEA Utilities & Renewables

Tyler Radke
U.S. Software

Michael Rollins, CFA
U.S. Communications Services & Infrastructure

Masahiro Shibano
Japan Precision & Semi Prod Equipment

Ashwin Shirvaikar, CFA
U.S. Payments, Processors & IT Services

Thomas A Singlehurst, CFA
Head of European Media Research

Alastair R Syme
Global Head of Energy Research

Martin Wilkie
Head of European Capital Goods Research

Alicia Yap, CFA
Head of Pan-Asia Internet Research

Oscar Yee
Head of Pan-Asia Materials Research

Steven Zaccone, CFA
U.S. Retailing/Hardlines

Judy Zhang
Co-Head of Pan-Asia Banks Research

Expert Commentary

Pantelis Koutroumpis
Director, Oxford Martin Programme on Technological and Economic Change
Oxford Martin School

Ian Goldin
Professor of Globalization and Development
Oxford Martin School

Helen H Krause, CFA
Head of Data Science Insights
Citi Global Data Insights

Yehuda Dayan
Data Scientist
Citi Global Data Insights

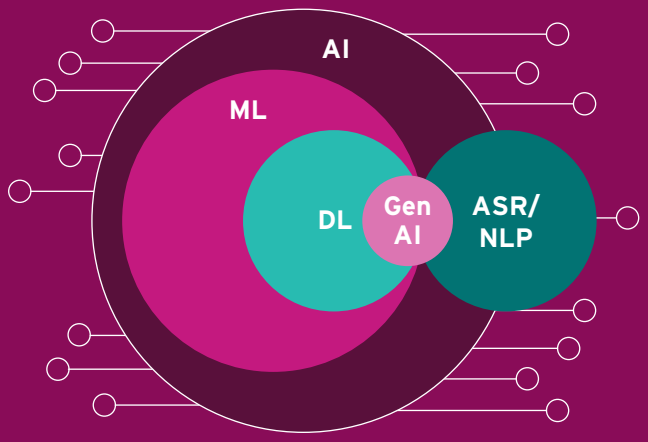
Brian Yeung
Data Scientist
Citi Global Data Insights

Amit B Harchandani
European Technology Sector Expert

Rob Garlick
Head of Innovation & Technology
Citi Global Insights

THE RISE OF AI

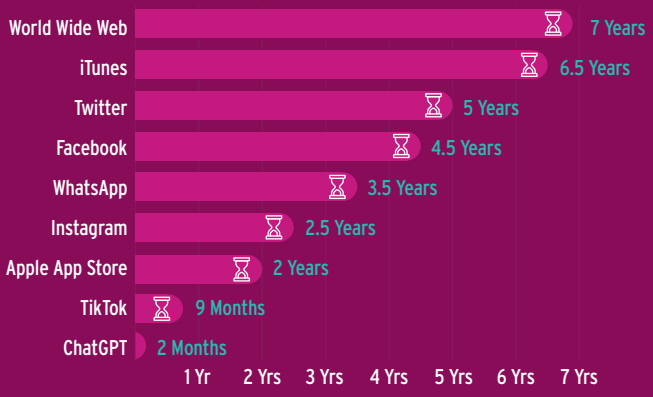
GENERATIVE AI: THE LATEST INFLECTION POINT IN AI



Artificial Intelligence · Machine Learning · Deep Learning
Generative AI · Automatic Speech Recognition, Natural Language Processing

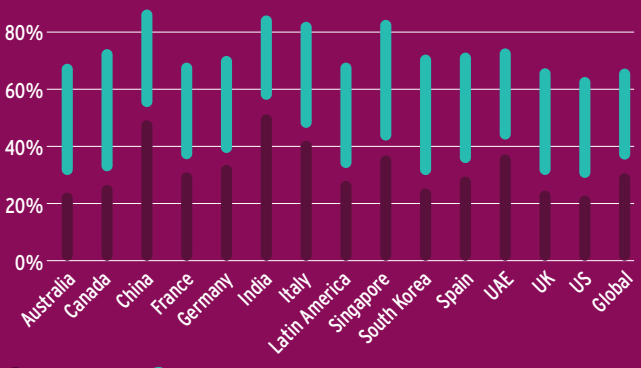
Source: Citi GPS, IDC

CHATGPT'S RAPID GROWTH AT LAUNCH
(Time to reach 100 million users worldwide)



Source: Citi GPS, SimilarWeb, Open AI

AI IS BEING USED ACROSS THE WORLD

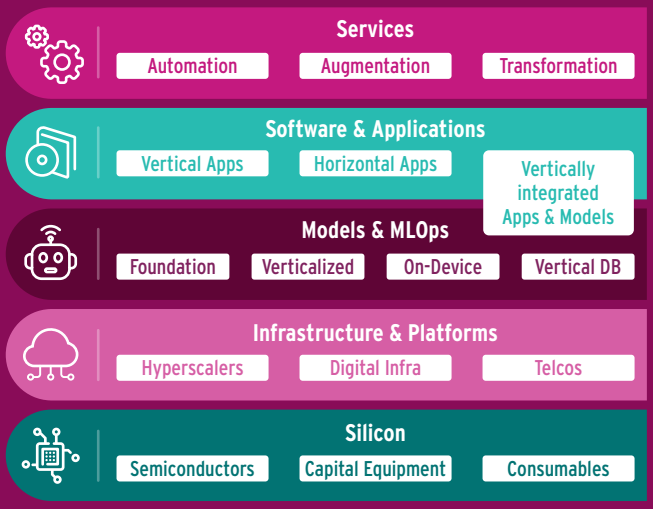


Deployed AI · Exploring AI

Source: IBM

OPPORTUNITIES FOR ENABLERS

Generative AI Technology Value Stack



GENERATIVE AI: ORDER OF IMPACT ACROSS SUPERSECTORS

Note: The Tech & Comms supersector is shown separately as it is the enabling supersector.

KEY ADOPTION CHALLENGES

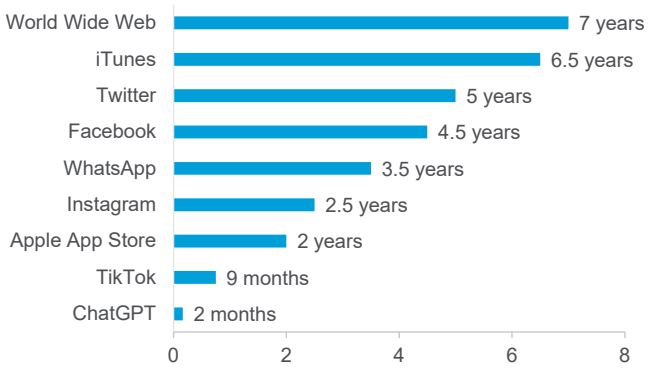
Contents

Executive Summary

The big picture

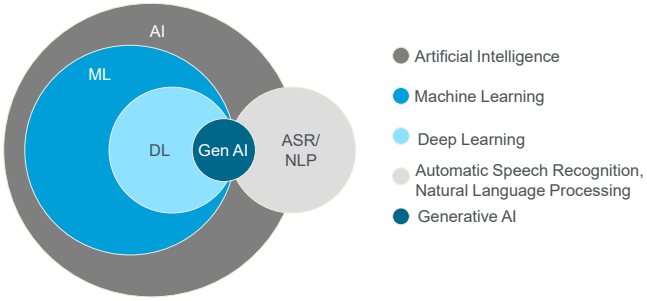
Generative AI has burst onto the scene following the launch of ChatGPT in 2022. As a concept, artificial intelligence (AI) is not new, and Generative AI represents the latest inflection point in the evolution of AI. However, what is distinctive about Generative AI is the tremendous potential it holds to transform work across industries and boost overall productivity. Taking a more holistic view, Generative AI might not only bring the power of AI itself to the masses but in fact accelerate the wider democratization of innovation. We believe it is a game changer.

Figure 1. When Launched, ChatGPT Was the Fastest-Growing Consumer Application in History (Time to Reach 100 Million Users)



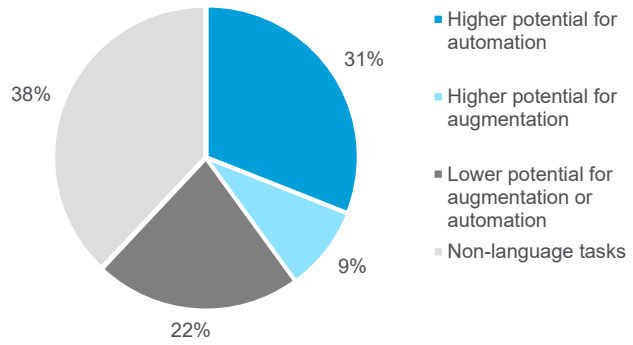
Source: Citi Research, SimilarWeb, OpenAI

Figure 2. Generative AI Potentially Represents the Latest Inflection Point in the Evolution of AI



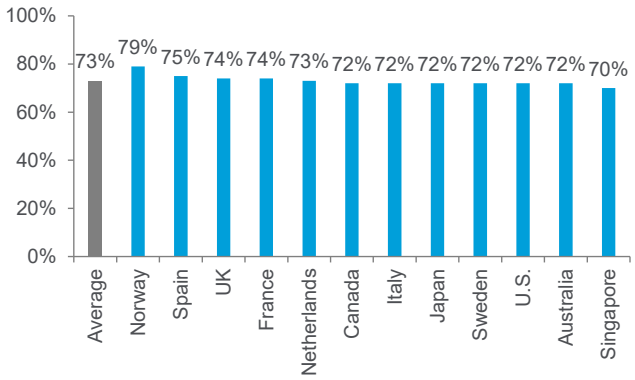
Source: Citi Research, IDC

Figure 3. Generative AI Holds Tremendous Potential to Transform Work (Share of Tasks That Could be Automated/Augmented by Its Adoption)



Source: Citi Research, Accenture

Figure 4. Consumers Have High Trust Levels for Generative AI-based Interactions (Share of Consumer Who Trust Generative AI Content)



Source: Citi Research, Capgemini

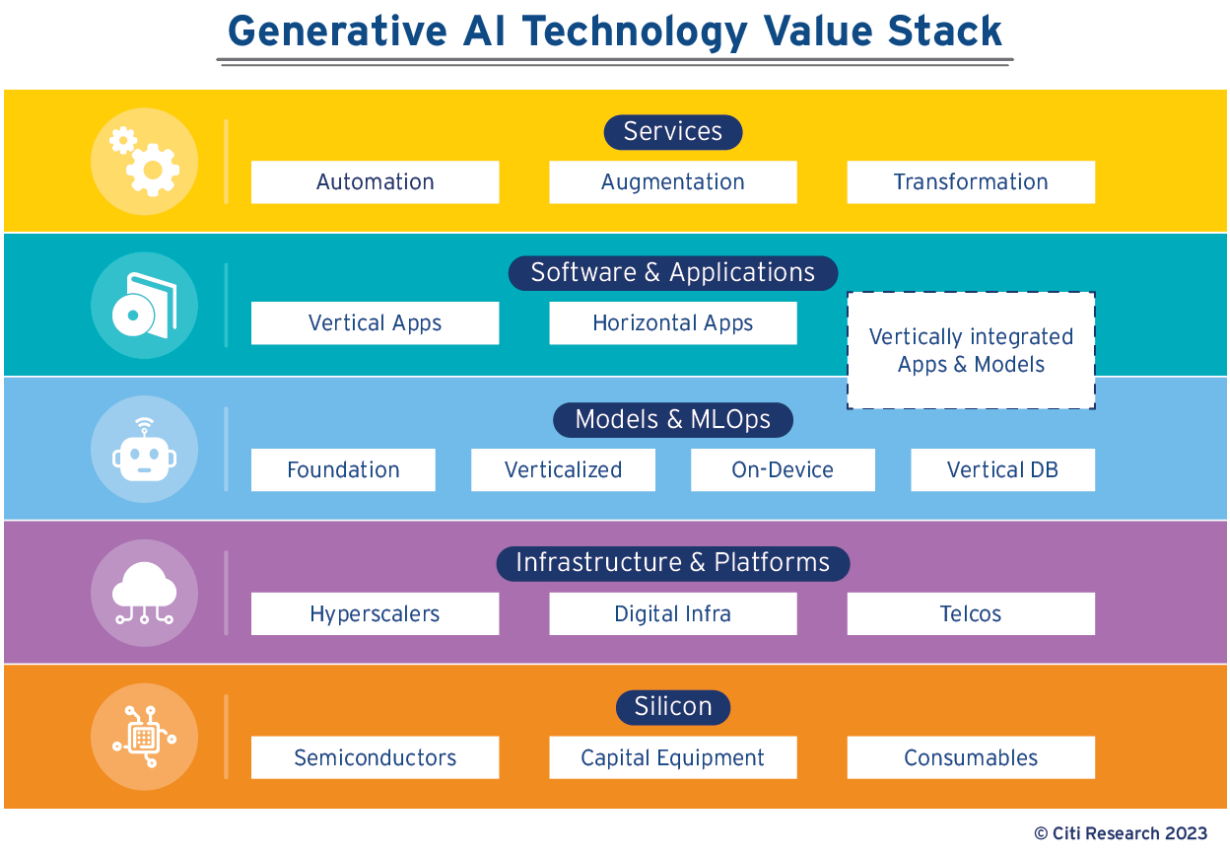
History, opportunities, challenges, and regulation

Although it feels like Generative AI has come out of nowhere, the report the looks at the rise of AI starting in the 1950s and through its significant growth over the past decade. We then explore the potential opportunities and challenges brought by Generative AI. Key challenges include those centered around bias, inequality, authenticity, infringements as well as the more debated one underpinned by existential concerns. The emergence of Generative AI has unsurprisingly seen AI as a broader topic become a firm focus for policy makers around the world. However, the regulatory path taken so far has varied. Given the stakes involved, we believe policy and governance evolution will play a defining role.

Opportunities for enablers

The first wave of potential opportunities for Generative AI is centered on the technology value stack, as outlined in Figure 5. Historically, the Silicon layer has been the de facto foundation of almost all technological shifts in the technology value stack, and Generative AI is expected to drive significant growth for compute (i.e., processing power), networking, and memory chips. However, as we look at the whole technology value stack, we see opportunities in each layer. In the Infrastructure & Platforms layer, we see the hyperscalers/cloud providers racing today to build the underlying infrastructure that enables Generative AI applications and services, but over time we expect to see higher or more differentiation. When it comes to Models and Machine Learning Operations (MLOps), the open-source community is likely to be a key driver of innovation. Moving further up the stack, we believe nearly all software companies will be impacted in some form by Generative AI, and company-specific execution will be critical. Lastly, we believe Generative AI represents a step forward from ongoing AI/automation initiatives at the Services layer.

Figure 5. Generative AI Technology Value Stack

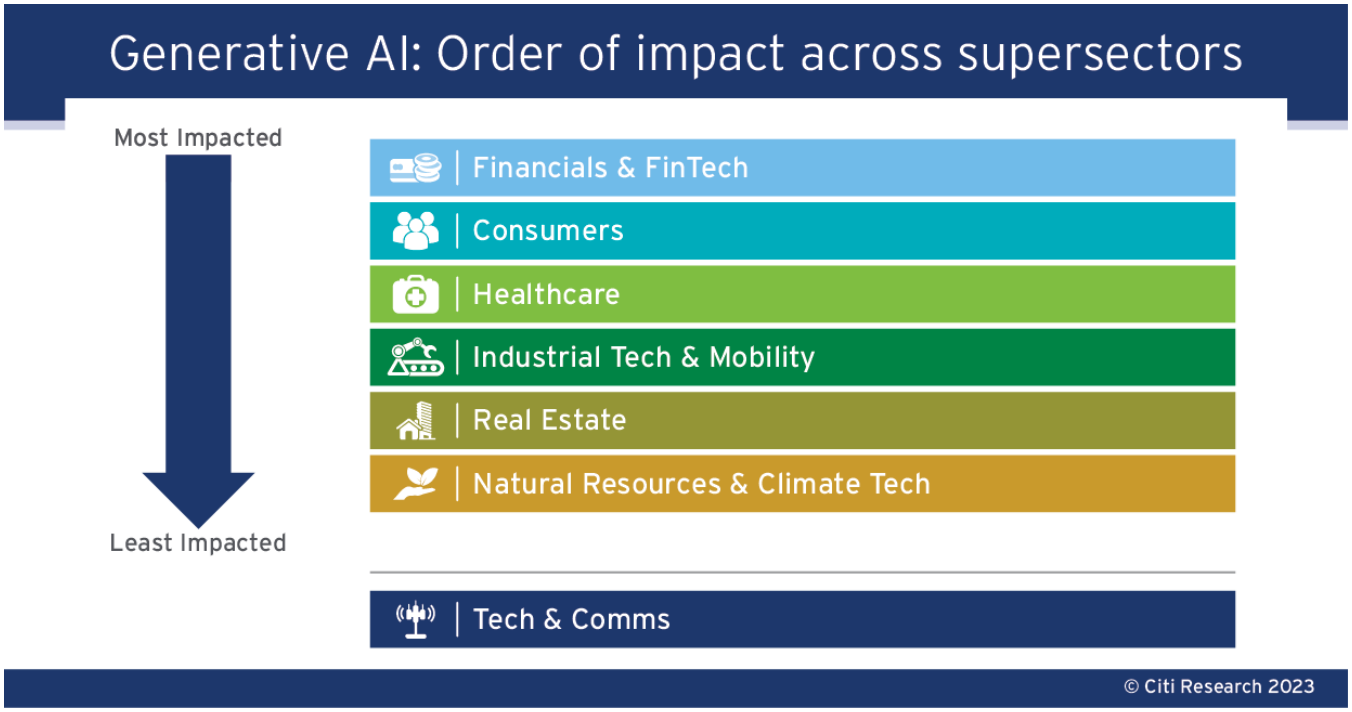


Source: Citi Research

Opportunities extend across sectors

Opportunities are not just limited to the technology value stack —they are also spilling into sectors outside of technology. To capture the full picture, we extended our analysis to look at the impact of Generative across six supersectors (Figure 6). We do this through a two-stage framework to assess risk/reward and apply that broadly across companies and sectors. Our analysis finds the Financials & FinTech supersector to be the most likely to be impacted overall, followed by the Consumer sector. At the other end of the spectrum, Natural Resources & ClimateTech at this stage look the least likely ones to be impacted.

Figure 6. Generative AI: Order of Potential Impact Across Supersectors



Note: We show the Tech & Comms supersector separately as it is the enabling supersector
Source: Citi Research

Patents and papers indicating we are in an "AI Arms Race"

What does the future look like for Generative AI? One way to investigate the global trends and growth is from the perspective of investment in technological innovation. We do this by analyzing the number of AI-related patent applications over time and across countries. Research papers are also telling, with the total cumulative AI research output increasing 1,300% between 2003 and 2021. Given the importance of AI as a foundational technology, the race is on between countries for scientific and technological dominance.

AI and Our New Renaissance

Ian Goldin
Director of the Oxford Martin Programme on Technological and Economic Change
Oxford Martin School, University of Oxford

This section draws on Ian Goldin and Chris Kutarna's *Age of Discovery: Navigating the Storms of Our Second Renaissance*, Bloomsbury, London, 2016.

The growing potential for AI to become a widely applicable form of general-purpose technology could create an era defining period of disruption with wide ranging social, political and economic implications. Generative AI could be as revolutionary a technology as the printing press, allowing people everywhere to enhance their writing and creativity and to share knowledge and ideas more widely and cheaply. The development of the printing press and exponential spread of ideas led to the Renaissance which fundamentally reshaped scientific, artistic, and religious views, facilitating an Age of Discovery which reshaped the world economy and had far reaching consequences on all continents, which continue to reverberate today. The development of the internet and World Wide Web means billions of people globally share ideas. This global pollination allows individual genius to be discovered wherever it is, and for collaboration which builds on diverse insights to flourish. Generative AI offers the potential to catapult creativity, science, and collective intelligence to higher levels.

Large language models and simultaneous translation means barriers between people speaking different languages will be lowered. This will increase access and personalization allowing for significant improvements in education and health outcomes. While 95% of scientific studies are published in English, under 5% of the world's population speak English as their native language. Drawing on the collective intelligence of people everywhere not only means that there are many more brains engaged in problem solving and innovation, but also that because they are more diverse the potential for disruptive breakthroughs is much greater.

Our new Renaissance offers an extraordinary opportunity to address some of the greatest challenges facing humanity. The growing potential of AI means new cures for cancer, Alzheimer and other terrible afflictions are more likely to be found, as are the means to generate low-cost clean energy and develop crops which can withstand climate change.

Five hundred years later, we still celebrate the outstanding achievements of the Renaissance. But it ended in tears for many, with religious wars and the brutal rise of slavery and imperial power. In Europe, it was associated with the rise of fundamentalism, with the challenge to the status quo leading to the Bonfire of the Vanities, burning, and banning of books and inquisitions. Then, as now, the growing concentration of wealth and potential of new technologies to create fabulous wealth for some and unemployment for others was a source of growing tensions and anger. Then, as now, those losing jobs (scribes in the 15th century and media folks today) were heard more often than those in newly created jobs (bookbinders and printers in the 15th century and programmers and knowledge workers today). Then as now, place and dynamic cities became more important as the knowledge economy accelerated, leading to a growing resentment against metropolitan elites. Then as now, connectivity and globalization led to the spreading of diseases and pandemics, including those which killed millions of native Americans. And then as now the power to use technologies to create false narratives and spread fake news became a tool for fragmenting societies and distrust of experts.

But the Renaissance teaches us that none of this can be taken for granted, as new technologies require a social license to operate. The challenge of our New Renaissance is to ensure that AI works for all.

The Rise of Artificial Intelligence

Pantelis Koutroumpis
Director, Oxford Martin Programme on Technological and Economic Change
Oxford Martin School

"We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with?"
-- Alan M. Turing, Computing Machinery and Intelligence, 1950.

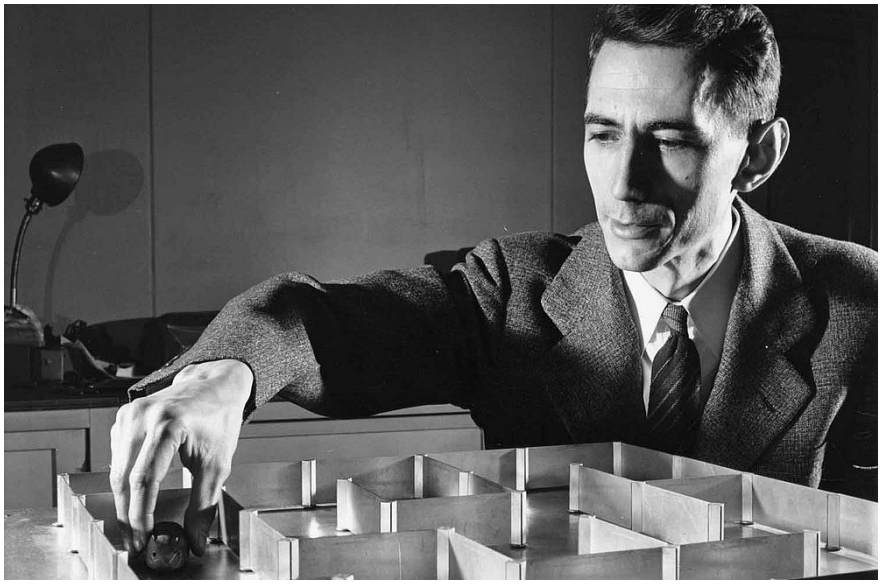
Can machines think? More than 70 years ago, Alan Turing posed this question in his seminal paper "Computing Machinery and Intelligence". To answer this question, he famously proposed a game for thinking machines, the Turing test or imitation game, where an interrogator asks the same questions to a human and a computer and tries to find out which one is the human.

Artificial Intelligence has been growing ever since but more significantly during the past decade. The recent advances in Large Language Models (LLMs) have taken the world by surprise and awe. Impressively, the social challenges associated with this progress have not changed significantly ever since Turing proposed this question.

In 1950, the computer science pioneer Claude Shannon from Bell Labs introduced the first artificial intelligence application, a wheeled mouse named Theseus that methodically explored the surroundings of a 25-square maze and found its way out.[1] Shannon wanted his "mouse" to navigate a labyrinth and escape it by learning its structure, resembling the thread used in Greek mythology to mark the hero's path. The mouse itself was a magnet on wheels but underneath the maze a complex web of switches and relays allowed it to move and learn through trial and error. Inspired by Shannon, Micromouse competitions have been running since the 1970s building on the advances of technology.[2]

Theseus "…inspired the whole field of AI" as "this random trial and error is the foundation of artificial intelligence."
-- Mazin Gilber, Google, Director of Engineering, Telecommunications, Orchestration, Analytics, and Automation.

Figure 7. Claude Shannon Showing the First AI Application



Source: Oxford Martin School

A few years later, in 1958, Frank Rosenblatt from the Cornell Aeronautical Laboratory built Perceptron Mark I, the first image classification computer. The goal of this computer was to identify objects even when the images were taken across different orientations, sizes, colors, and backgrounds.[3]

[1] Daniel Klein, "Might Mouse," *MIT Technology Review*, December 19, 2018.
[2] Micromouse Online, Homepage, accessed September 5, 2023.
[3] Cornell Aeronautical Laboratory, Inc., "The Perceptron: A Perceiving and Recognizing Automation," January 1957.

Following this invention, Rosenblatt spurred controversy in the AI field when he openly supported that the perceptron, which was the algorithm used to classify objects in his computer, would become “the embryo of an electronic computer” that “will be able to walk, talk, see, write, reproduce itself, and be conscious of its existence.”[4] In spite of the criticism, Rosenblatt also introduced the term "back-propagating error correction" in 1961, a key theoretical foundation of modern neural networks, although he did not know how to implement this in his computer.

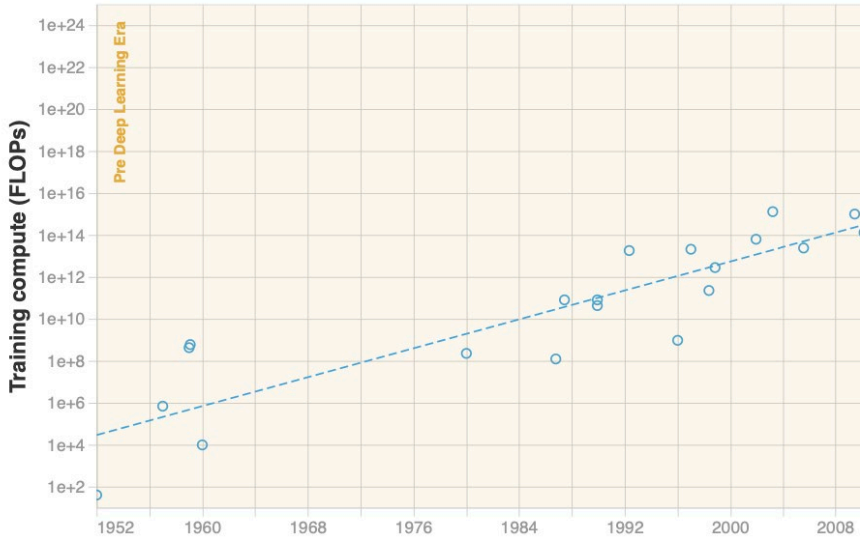
The promise of computing technologies and modelling capabilities led other prominent researchers to follow Rosenblatt in his predictions. In 1958, political scientist Herbert Simon and computer scientist Allen Newell predicted that “within ten years a digital computer will be the world’s chess champion”, and “within ten years a digital computer will discover and prove an important new mathematical theorem.” Soon after, in 1965, Simon moved even further, supporting that “machines will be capable, within twenty years, of doing any work a man can do.”[5] In 1967 Marvin Minsky, a leading AI scholar, predicted that “Within a generation... the problem of creating ‘artificial intelligence’ will substantially be solved.”[6] He then followed with an interview in 1970 stating that “from three to eight years we will have a machine with the general intelligence of an average human being.”

In the late 1960s and 1970s, the AI field experienced a number of setbacks which led researchers to introduce the term AI winter. Most of the efforts in the 1960s revolved around military and intelligence agencies that needed increased speed and decision-making accuracy during the Cold War.[7] One of these efforts involved the automatic translation of Russian documents into English. Despite the initial optimism, in 1966 the Automatic Language Processing Advisory Committee (ALPAC) concluded that machine translation was slower, more expensive and more inaccurate than humans.[8] A number of theoretical and practical issues led the field to the abandonment of "connectionism" which was linked to neural networks and perceptrons in favor of symbolic reasoning in the late 1960s. The setbacks continued in the 1970s with the UK parliament Lighthill report concluding that AI had failed to achieve its "grandiose objectives". In the U.S., a shift to mission-oriented direct research, rather than basic undirected research led to further cuts and frustration among researchers.

[4] Mikel Olazaran, “A Sociological Study of the Official History of the Perceptrons Controversy,” *Social Studies of Science*, Vol. 26, No. 3, August 1996
[5] Herbert A. Simon, *The New Science of Management Decision* (New York, Harper & Row, 1960)
[6] Marvin Minsky, *Computation: Finite and Infinite Machines* (New Jersey, Prentice-Hall Inc., 1967)
[7] Dafydd Townley, “Intelligence Agencies Have Used AI Since the Cold War — But Now Face New Security Challenges,” *The European Financial Review*, May 22, 2023.
[8] John Hutchins, “The History of Machine Translation,” PDF, accessed August 12, 2023.

Figure 8. The Pre-Deep Learning Era in Artificial Intelligence

The AI winter in the 1960s and 1970s along with inflated expectations led the AI field to a pause. After the mid-1980s, new models emerged and continued the progress.



Source: Epoch AI

During the 1980s and 1990s the field gradually evolved producing new models that covered a diverse set of applications including the Neocognitron for handwriting and pattern recognition (1980), NetTalk for text to audio transformation (1987), ALVINN for autonomous vehicle navigation through a camera and a laser (1988) and Long Short Time Memory (LSTM) as a foundational breakthrough in neural network research (1991). Although the field experienced a stable progress during this period, large production systems using AI technology rarely made explicit references to it. As Nick Bostrom, professor of philosophy in Oxford, stated in 2006 "a lot of cutting-edge AI has filtered into general applications, often without being called AI because once something becomes useful enough and common enough it's not labeled AI anymore." Because of this progress, the computing resources that supported AI systems kept a steady improvement pace — doubling every 21 months from 1952 to 2010 and aligning the field's progress with the well-known Moore's Law, which measures transistor density in semiconductor chips at a similar rate (Figure 8).

The performance of AI systems has increased rapidly since 2010

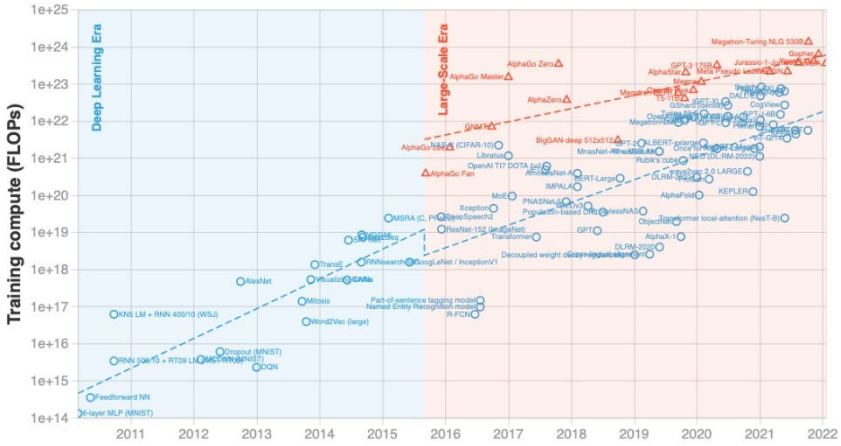
Analyzing information about the computing resources used to train 123 milestone Machine Learning systems from Epoch AI and the analysis performed by researchers, we show that a sharp discontinuity in computing resources emerged around 2010.[9] The rate of computing used post-2010 doubled every six months in FLOPs (Floating Point Operations Per Second) achieving a three-fold increase compared to 18-month doubling times in the previous period. The reason behind this change is largely attributable to the transition to Deep Learning models since 2010. These models construct complex non-linear relationships between their inputs and outputs through a layered composition of their features. Combining features from lower layers led to dramatic performance improvements compared to models with shallower architectures.

[9] Epoch AI, "Studying Trends in Machine Learning", accessed August 12, 2023; Jaime Sevilla et al., "*Compute Trends Across Three Eras of Machine Learning,"* PDF, accessed August 12, 2023.

The Deep Learning revolution was underpinned by the successful implementation of the connectionist models (which were abandoned in the 1960s) and some key concepts that were introduced at the same time, including back-propagation. The combination of core inputs — including human ingenuity, computers, algorithms, data quantity, and quality — led to a sharp transformation of the field. Since the late 2000s, Deep Learning models outperformed "shallow" ones in several machine learning competitions. Already in 2011 DanNet was the first model to achieve superhuman performance in visual pattern recognition, outperforming traditional methods by a factor of three. In 2012 AlexNet won the large-scale ImageNet competition by a significant margin.

Figure 9. The Deep Learning Era

After 2010, the rate of computing resources used to train AI models rose by a factor of three, doubling every six months, compared to an 18-month doubling rate in the previous period.



Source: Epoch AI

The large-scale model era

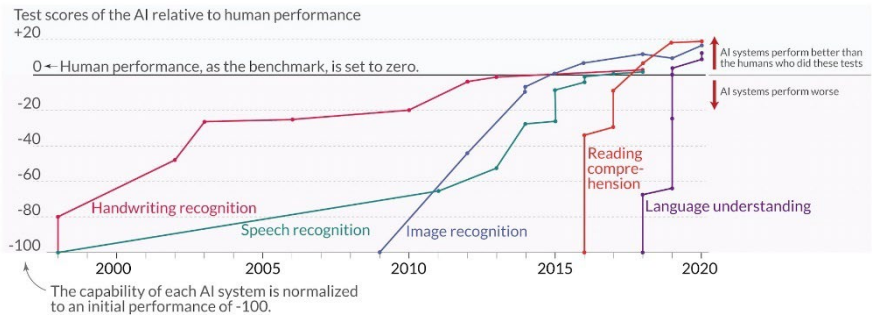
The profound rise in performance by the post-2010 models, led firms to invest heavily in AI applications. This drove some of the earliest efforts to use AI and reach super-human results in closed-world situations like the game of Go. Training compute resources used for some of the most advanced AlphaGo models (Zero and Master) were only matched with huge LLMs five years later (Figure 9). Within this period, some important discoveries took place. Before 2017, the field of Natural Language Processing (NLP) was lagging computer vision in terms of its "human-level" performance metrics (image and handwriting recognition, Figure 10). In June 2017, researchers from Google published a groundbreaking paper, titled "Attention is all you need," that introduced a significant advance in the use of the attention mechanism as the main improvement for the Transformer model. With this, long sequences of text that were used for translation or prediction of the next word, would not rely on the last state of the encoder, as it was usually done with RNNs (Recurrent Neural Networks) but instead they would extract information from the whole sequence.

Several important NLP models used these advances including BERT and GPT-2, allowing reading comprehension to improve dramatically along with language understanding soon after (Figure 10). Transformer models were also used in other applications including models that predict protein folding, text-to-image (used for Latent Diffusion Models in conjunction with content delivery networks (CDNs) and Diffusion models) and are likely to become a general-purpose mechanism underpinning most AI applications.

Armed with these improvements in the foundation models, AI firms increased the computing resources to improve their performance (Figure 10) and a new set of models emerged in the Large-Scale Era using more than 100 times the resources than the Deep Learning Era models were trained on. The breadth of applications was vast, including models that outperformed Go masters, to protein-folding predictions and large language models. The algorithms underpinning these models paved the way to the term Generative AI allowing a huge range of possibilities to emerge.

Figure 10. Superhuman Benchmarks Reached by AI

Since 2020, ML models achieved supe-human performance in image, speech, handwriting recognition, reading, and language understanding



Source: DynaBench, OWID

The Landscape of Generative AI

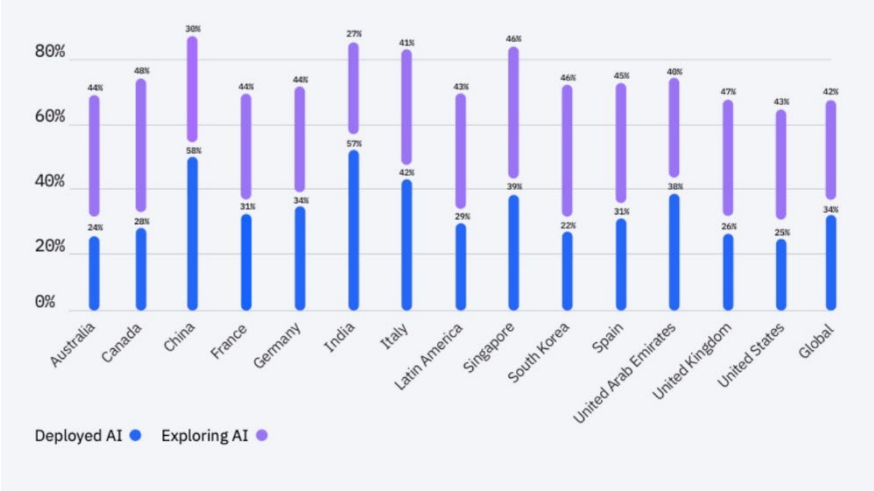
Vast majority of firms expect to use AI in coming years

The progress in the recent developments in artificial intelligence has attracted strong interest from many firms aiming to integrate content generation and decision making in their processes. In a 2022 AI report, IBM measured the proportion of firms that have already deployed and plan to use AI in the coming year. China and India were leading in these metrics with 58% and 57% of the firms having already deployed AI respectively, followed by Italy and Singapore.[10] Overall, the report found that vast majority (>70%) of firms expected to use AI in the coming years (Figure 11).

[10] IBM, "IBM Global AI Adoption Index 2022," May 2022.

Figure 11. AI Used Across the World

AI is becoming a global technology. The leading countries that have deployed AI in 2022 are China and India (58% and 57% of the firms) followed by Italy and Singapore. Even those countries lagging behind are exploring AI use in the coming years.



Source: IBM

The early winners in the AI race

Looking into the market performance of the leading firms in AI, one can observe that the growth in the S&P 500 index in the first half 2023 largely came from firms that produce core components of AI technologies. In the first half of 2023, a majority of returns in the S&P 500 index came from the seven biggest stocks, which were driven by optimism around AI.

Beyond the usual suspects, like the Big Tech firms, the equipment vendors supplying the necessary computing infrastructure outperformed in that time period, as every firm that intends to deploy AI will need to use these resources. Analysts of AI markets expect the growth of the AI market to continue as the breadth of applications is likely to fuel demand for more and faster graphic processing units (GPUs). These components are not only used for inference purposes (answering questions on existing models) but also during the development phase of a model (training), and this demand scales linearly with headcount. As a result, there is no sign that the GPU shortage we have in 2023 will abate in the near future.[11]

The AI Productivity Boost

"*Generative AI could provide complementary tools to knowledge workers. These would be creating new tasks (for educators, nurses, creative workers, tradespeople, and even blue-collar workers) and providing inputs into better decision marking for knowledge work.*"
-- Daron Acemoglu, MIT, Economics Professor

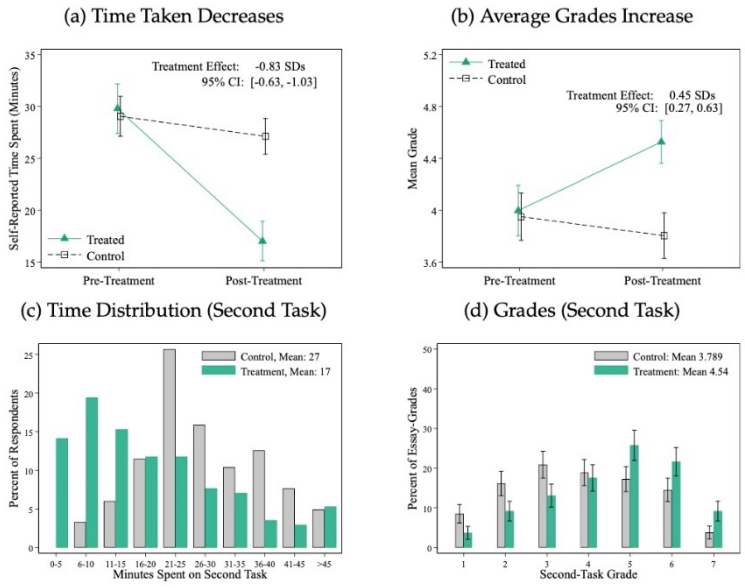
Can AI tackle a pressing economic paradox? The famous quip by Paul Krugman, Nobel laureate in Economics, that "productivity isn't everything, but in the long run it is almost everything" runs at the heart of every innovation. Despite its importance and the impressive technological change that has taken place in the recent past, productivity growth has been slowing down for decades across advanced economies.[12] It is not surprising that industry leaders and academics often refer to the potential of AI technologies as a way to end this downward trend. While there is still no definitive answer, several researchers suggest that this new wave of large language models is very promising, based on preliminary results from their studies.

[11] Guido Appenzeller, Matt Bornstein, and Martin Casado, "Navigating the High Cost of AI compute," Andreesen Horowitz, April 27, 2023
[12] Ian Goldin et al., "Why Is Productivity Slowing Down," Oxford Martin School Working Paper No. 2012-6, May 9, 2021.

In a recent paper, researchers from MIT investigated the effects of ChatGPT in the context of mid-level professional writing tasks.[13] Assigning college-educated professionals to incentivized writing tasks and randomly exposing half of them to ChatGPT, they found that those who used the Generative AI technology decreased their completion times by 0.8 standard deviations and increased their output quality by 0.4 standard deviations. The inequality between workers also decreased, as ChatGPT benefited low-ability workers more thus compressing the productivity distribution across the sample. The authors noted that ChatGPT mostly substitutes for worker effort rather than complementing worker skills, and restructures tasks towards idea-generation and editing and away from rough-drafting. Beyond the hard numbers, the researchers also found that exposure to ChatGPT increased job satisfaction and self-efficacy but it also heightened the concerns about automation technologies (Figure 12).

Figure 12. ChatGPT Impact on Writing Tasks

ChatGPT increases the productivity of college-educated professionals in writing tasks, decreases the inequality, and benefits low-ability workers the most.



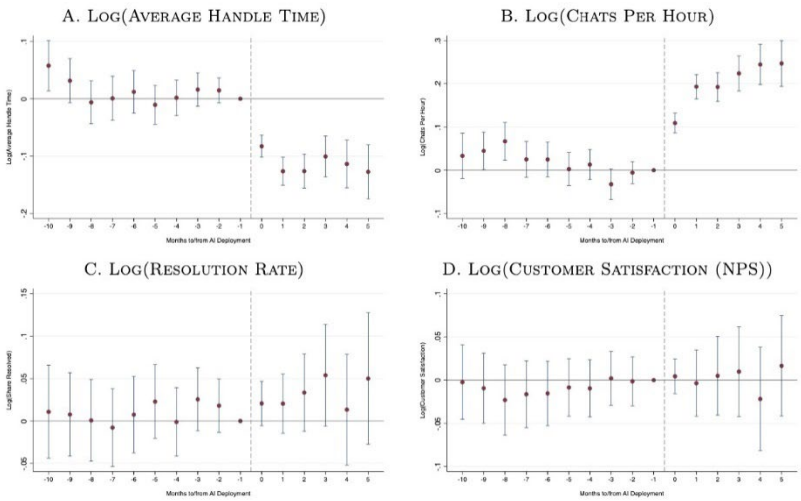
Source: Noy and Zhang (2023)

[13] Shakked Noy and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence, *Science*, July 13, 2023.

Looking at a different type of office worker, a paper by Erik Brynjolfsson released in 2023 looked into the productivity results of adding Generative AI to customer support agents.[14] The AI system used in this study was based on a GPT family language model that was fine-tuned to focus on customer service interactions. The authors found that the employees with access to this tool managed to increase their productivity by 14%, measured by the number of issues they resolved per hour. The results on the distributional impact of AI seem to align with the ones on college-educated professionals, as the study also found that the greatest productivity impact was on novice and low-skilled workers, with minimal effects on experienced and highly skilled workers. The authors found the AI model disseminated potentially tacit knowledge of more able workers and helped newer workers move down the experience curve. In addition, they showed that AI assistance improved customer sentiment, reduced requests for managerial intervention, and improved employee retention.

Figure 13. Generative AI on Customer Support Agents

AI tools are round to increase the productivity of customer support agents by 14%, but the effects are mainly driven by novice and low-skilled workers.



Note: These figures plot the coefficients and 95 percent confident interval from event study regressions of AI model deployment using the Sun and Abraham (2021) interaction weighted estimator.
Source: Citi GPS

Implementation Issues

Although increased efficiencies are great, most businesses run on legacy infrastructure with multi-year transitions needed to incorporate next-gen tools. Many businesses want to integrate these new tools with their data, in on-premise or domain models, but this takes time. There is also a rush to access high-end compute, data center capacity, and time with leading AI providers. Of course, all of this will require more resources and that often involves zero-sum investment decisions (i.e., taking investment dollars from other areas) which are never easy. For example, the recent announced pricing of Microsoft's Copilot AI product came in higher than expected, and while this may highlight the value it offers, firms will have to find the resources to deploy products widely.

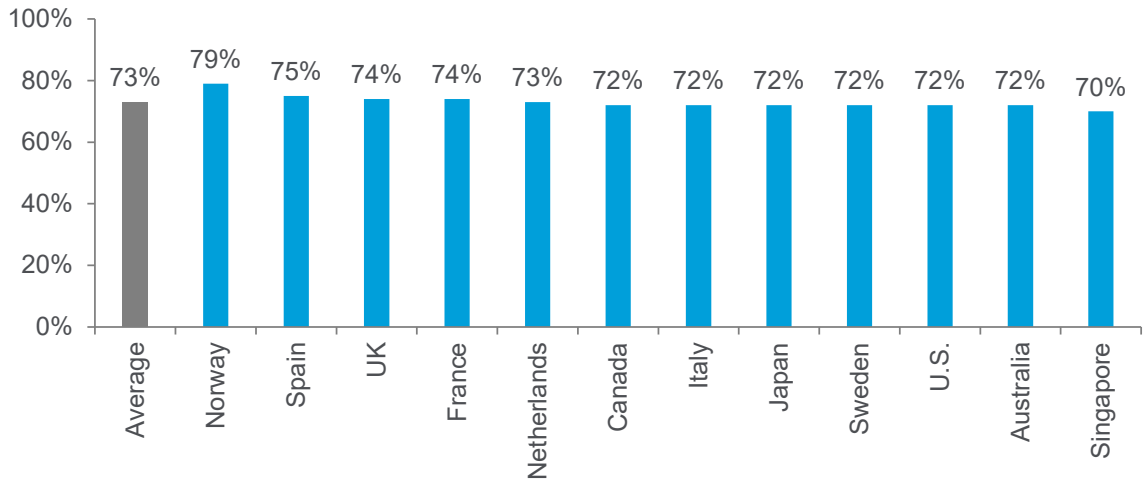
[14] Eric Brynjolfsson, Danelle Li, and Lindsey R. Raymond, "Generative AI at Work," National Bureau of Economic Research, Working Paper no. 31161, April 2023.

Change is usually incremental as companies experiment, learn, and iterate. It takes time for businesses to develop processes and people to capitalize on opportunities. In the case of AI, some may be fearful of implementing a technology they do not trust or understand, let alone a technology that could substitute their roles. In addition to company guardrails for responsible AI, many industries are highly regulated, and AI tools will need to have adequate risk management, transparency, or explainability, and in time also auditability. For example, while in theory, AI offers significant opportunity in medicine, a prognosis needs to be explained. Humans-in-the-loop, as happens with aircraft autopilot or fact checkers, may be therefore needed, which would in turn will slow implementation.

Generative AI Acceptance and Trust

The fact that ChatGPT has become one of the fastest-growing consumer applications in history illustrates that consumers have enthusiastically embraced Generative AI. We also believe they have high trust levels for Generative AI-based interactions, supported by multiple studies. One such study from Capgemini — based on a sample of ~8,600 respondents across multiple countries — suggests that 73% of respondents trust content written by Generative AI.

Figure 14. Share of Consumers Who Trust Content Written by Generative AI



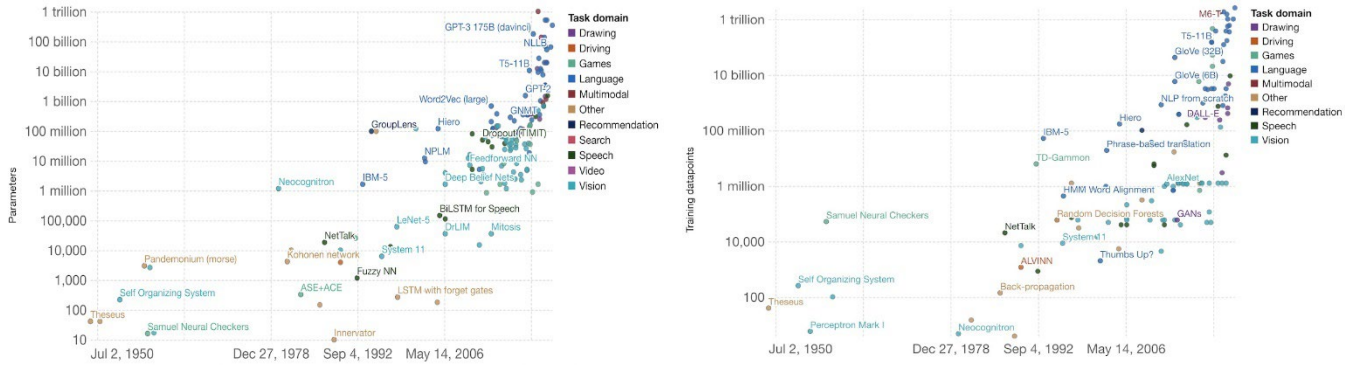
Source: Capgemini, Citi GPS

The Next AI Era: Promises and Challenges

"Suddenly, anyone could fine-tune the model to do anything, kicking off a race to the bottom on low-budget fine-tuning projects."
-- Anonymous researcher from Google

Computational needs are increasing exponentially. The phenomenal achievements of AI models in the recent years are largely driven by significant changes in the core inputs used to train these models. As a result, in the Large-Scale Era starting in late 2015, the compute doubling times appear to be almost twice as high (10-months) compared to the pre-2010 era.[15] During the same period, the number of parameters used in large-scale models exploded by 10-fold every year, in line with the increase in the size of input datasets (Figure 15).[16] In contrast, the average GPU improvements over the same period had a doubling rate in computer performance (measured as FLOPs/$) every 2.5 years.[17]

Figure 15. Rise in the Number of Parameters and Training Data for Recent AI Models



Source: Epoch, Our World in Data

Compute costs for models are rising

Following these recent trends, researchers used data from existing models to predict the total cost of the future models and found that they will soon become impossible to sustain — exceeding the U.S. GDP by the end of 2030.[18] In a separate study on the same subject, OpenAI estimates that by the end of the decade the costs to train large scale models will reach $500 million — significantly different from the estimates from independent researchers (Figure 10).[19]

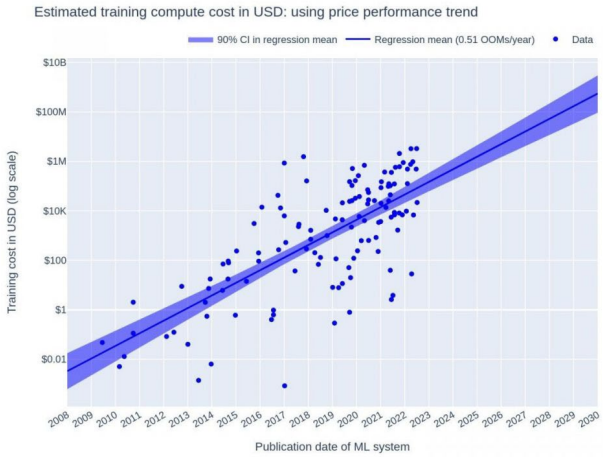
The reasons for these large disparities emerge due to the approaches used in the extrapolation process of each study. If we assume stable progress in the GPU trend per dollar or the algorithmic improvements, we end up naively extrapolating previous trends. The report by OpenAI on the other hand, explicitly reports a "best-guess forecast" assuming that "the growth in costs will slow down in the future," which means that their results "should be interpreted with caution".

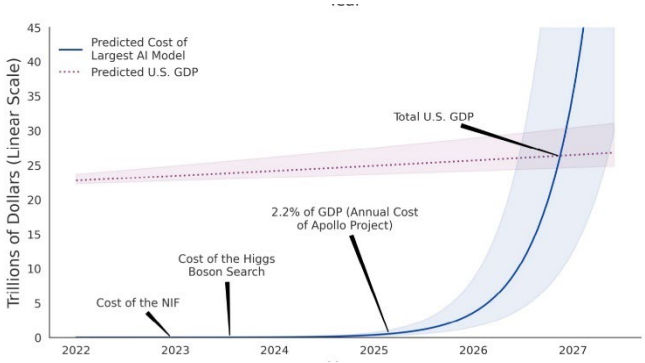
[15] Jaime Sevilla et al., "*Compute Trends Across Three Eras of Machine Learning,*" PDF, accessed August 12, 2023.
[16] Julien Simon, "Large Language Models: A New Moore's Law?", Hugging Face, October 26, 2021.
[17] Marius Hobbhahn and Tamay Besiroglu, "Trends in GPU Price-Performance," EPOCH, June 27, 2022.
[18] Andrew Lohn and Micah Musser, "AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?" Center for Security and Emerging Technology, January 2022.
[19] Ben Cottier, "Trends in the Dollar Training Cost of Machine Learning System," Epochai.org, January 31, 2023.

In the naive OpenAI scenario, costs of training will surpass $233 billion in 2032 which seems to align with the reports from other researchers. In both cases this rise in computing resources is a signal for caution about the future capabilities of large models. If performance needs to increase, researchers should also consider other ways to achieve this, either by improving the algorithms they use or by focusing on the quality of the data that are fed in their models.

Figure 16. The Rise in Compute Costs

Depending on the extrapolation process, researchers estimate that the cost to train large-scale models will soon become unsustainable for most firms. Some support that this time will come before the end of the 2020s and others in early 2030s





Source: Lohn and Musser (2022), Epochai.org

Open-Source Code and Competition

The Linux moment for AI and the reasons behind it

Open-source code has played a significant role in the development of machine-learning models from the early days of the Deep-Learning Era. Both in terms of datasets and models (including Vicuna and Stable Diffusion) the open-source community has provided support to the leading firms in the AI domain. In 2023 we already saw a long list of open-source applications that seem to compete with OpenAI's ChatGPT (like Vicuna with 13 billion parameters, Figure 17). In this section we outline some of the key elements that facilitated this process.

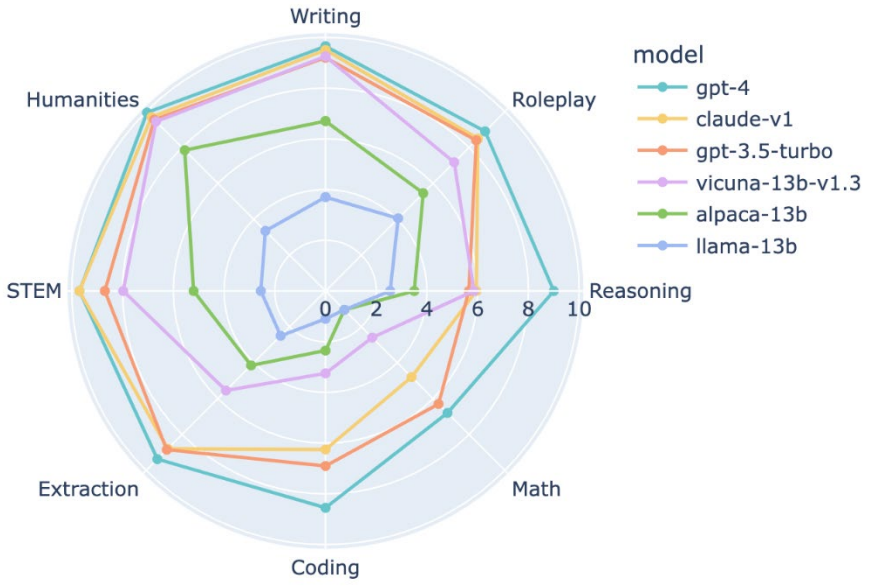
First, software developers can easily share their code on Hugging Face, a company that made its name from its Transformers library built for NLP applications and its platform that allows users to share machine learning models and datasets. Its users span from researchers at universities and software developers to employees at Big Tech firms, giving Hugging Face a fertile ground for the exchange of ideas, experimentation, and development.

Second, notable algorithmic improvements — like LoRA (Low-Rank Adaptation of Large Language Models) by Microsoft researchers — deal with the problem of fine-tuning large language models instead of retraining them on the new corpus. LoRA can use the existing weights of a prohibitively expensive general-purpose model (like Meta's Llama) to specific applications by vastly reducing the number of trainable parameters. This is achieved by directing the Transformer attention blocks of large-language models to the specific inputs allowing this process to reach impressive results, almost on par with full model fine-tuning at a fraction of the cost and time.

Third, the benchmarking process in the AI community has allowed researchers and firms to get a better view of their progress against well-known targets, although this now coming to a saturation point where an 80% or 90% accuracy is not meaningful. This is why several researchers propose to introduce newer and more comprehensive benchmarks to evaluate their models.

Figure 17. Competition in the Chatbot Area

Vicuna, an open source chatbot based on Llama, outperforms other proprietary chatbots and scores close to the market leader in a range of tasks.



Source: LMSYS

Despite the commercial and open-source competition, open-source models are likely to continue to develop in a symbiotic rather than antagonistic way with the leading firms, providing more efficient algorithms and leveraging the global talent pool of researchers and software engineers that even large technology firms lack.

Do we need more data to train AI models?

More data to train AI models or better data quality?

With the increase in computing resources in the Large-Scale Era, vast amounts of data have been added for training purposes (Figure 15). Data is almost every digitized bit of information, but in this setting, we refer to online available human generated (for text), annotated (for images), and curated in formation, including Wikipedia articles, IMDb images, various videos, online news articles, coding material, and general interest fora. Building on these resources, firms and researchers can train and improve their models to produce more accurate results as their algorithms and computational inputs grow. However, this process is not without limits.

First, the issue of data quality appears to become more prominent as we move towards more advanced models. In a recent paper, researchers showed that by using a much smaller corpus, based on "textbook quality data," AI models can achieve much higher performance improvements compared to the training results from similar or much larger online coding fora.[20] In this sense, the quality parameters of the inputs appear to have a dramatic impact largely overcoming the issues with scale and costs that large models faced.

Second, the progress of large models for language and text-to-image generation have fueled predictions about the state of online content in the near future. Some predict that within the next decade, at least 50% of online content will be generated by or augmented by AI.[21] The optimism of these predictions relies solely on the cost reductions associated with text and image generation and often fails to understand the importance of human generated content that is needed for the improvements of these models. In 2023, researchers from Oxford and Cambridge uncovered a deep issue with large AI models when these are fed with their own outputs as a training dataset. Based on their findings they tried to answer the question: What will happen to GPT-n once LLMs contribute much of the language found online? They observed that this self-feeding process will lead to a model collapse due to irreversible defects that gradually trim the tails of their predictions compared to other models trained on original human-generated content. They called this effect "the curse of recursion" when models that inbreed end up effectively forgetting and missing the diversity of outcomes that we would expect from human-level intelligence.[22]

In addition to the challenges noted above, we also recognize:

1. **Bias:** Generative AI systems are based on large amounts of training data, which means that the results can be susceptible to bias or inaccuracies in the training data, particularly if it is built on the gender, racial, and myriad other biases of the internet and society more broadly. Additionally, results can lack human/logical reasoning.

2. **Inequality:** The economics associated with the operation of large language models could trigger further increases in inequality between those who have access to these capabilities and those who do not. One could take it a step further and argue that the emergence of a select few companies as gatekeepers of all AI development could effectively monopolize the industry and trigger a centralization of collective intelligence.

3. **Authenticity:** While language models have become increasingly more coherent, we observe that they have also become more fluent at generating factually incorrect statements and fabricating falsehoods. Recent research suggests that authenticity levels tend to improve with increases in the size of models (i.e., in the number of parameters), but we believe more needs to be done here, particularly given the levels of consumer trust in Generative AI interactions as well as the need for high levels of accuracy in certain industries such as financial services.

[20] Suriya Gunasekar et al., “Textbooks Are All You Need,” Microsoft Research, June 20, 2023.
[21] MarketsandMarkets, “Generative AI Market by Offering, Application, Vertical, and Region: Global Forecast to 2028,” April 2023.
[22] Ilia Shumailov et al., “The Curse of Recursion: Training on Generated Data Makes Models Forget,” PDF, May 31, 2023.

4. **Infringements:** Training of Generative AI models/platforms requires access to data lakes and question snippets — billions of parameters that are constructed by software processing huge archives of images, text, and other forms of input (depending upon the type of the model). The AI platforms recover patterns and relationships, which they then use to create rules, and then make judgments and predictions, when responding to a prompt. This process comes with legal risks, including intellectual property (IP) infringement.

5. **Existential concerns:** The transformative potential of Generative AI has understandably triggered broader concerns around the existential threat posed by AI, with multiple experts taking contrasting views — particularly against a varied regulatory backdrop. We do not believe it is our place to be taking a view on this debate other than to simply state (what might seem quite obvious to some) that we do not believe this debate is likely to go away any time soon.

Regulatory and Copyright Concerns

Over the past decade several proposals for regulating AI have been put forward, either by regulators or the private sector. In 2017, Bill Gates supported the introduction of a robot tax as a response to the — so far unfulfilled — expectation that robots would excessively substitute low-skilled manufacturing workers.[23] Since then, the rising concerns for citizen surveillance in China through image recognition AI technologies and other means, have alerted governments in Europe and the U.S.[24] In 2023, academics and market leaders proposed a pause on AI technologies due to the existential risk that they believe it represents for humans.[25] These proposals about AI applications appear to be both unclear and hard to enforce.

Further issues have recently surfaced due to the copyright issues for the data used to train large language models and text-to-image applications. Without any change in copyright legislation in the U.S., in June 2023 OpenAI and Microsoft were sued in a class action lawsuit for $3 billion over alleged privacy violations of their chatbot.[26] This is not the first time these firms face similar issues. In 2022, GitHub programmers filed a similar lawsuit for scraping their code without their consent.[27]

"*On artificial intelligence, trust is a must, not a nice to have.*"
-- Margrethe Vestager, Executive Vice President for a Europe Fit for the Digital Age

The regulatory front is also changing across the world. Perhaps the most advanced regulatory framework is currently proposed by the EU Commission and was amended to become even more demanding from its original version in June 2023.[28] It is likely that some of its core ideas will be followed globally as is common with the Brussels effect.

[23] Quartz, "The Robot That Takes Your Job Should Pay Taxes, Says Bill Gates," February17, 2017.
[24] Zeyi Yang, "The Chinese Surveillance State Proves that the Idea of Privacy is More "Malleable" Than You'd Expect," *MIT Technology Review*, October 10, 2022.
[25] Future of Life Institute, "Paul Gian AI Experiments: An Open Letter," March 22, 2023.
[26] Chloe Xiang, "OpenAI and Microsoft Sued for $3 Billion Over Alleged ChatGPT 'Privacy Violations'," *Vice*, June 29, 2023.
[27] Chloe Xiang, "GitHub Users File a Class-Action Lawsuit Against Microsoft for Training an AI Tool With Their Code" *Vice*, November 4, 2022.
[28] Foo Yun Chee and Supantha Mukherjee, "EU Lawmakers Vote for Tougher AI Rules as Draft Moves to Final Stage," *Reuters*, June 14, 2023.

EU lawmakers initially proposed a risk-mitigation strategy in their AI Act with a focus on biometric surveillance, but they soon moved towards harsher measures that include a mandate on companies developing Generative AI applications for full disclosure of copyrighted material used to train their systems. Their proposal also includes requirements for companies working on high-risk applications to perform a fundamental rights impact assessment and also evaluate the environmental impact of their applications (for inference and training), which ties to the recent increase in computing resources for AI.

In a preliminary assessment of leading proprietary and non-commercial LLMs, researchers from the Stanford Institute on Human-Centered AI found that most of the leading applications lag significantly behind the requirements of the draft EU AI Act.[29] In particular, they pointed out that many market-leading foundation models have limited disclosure for copyrighted material, uneven reporting of energy use, inadequate disclosure of risk-mitigation strategies and an absence of evaluation standards or auditing. systems. This suggests that market leaders need to work further to meet the compliance requirements put forward by the EU AI Act.

Global Regulatory Landscape Mixed

The emergence of Generative AI has unsurprisingly seen AI as a broader topic become a firm focus for policy makers around the world. However, the regulatory path taken so far has varied — aside from the EU, which has been one of the first government bodies to formally propose a law specifically aimed at controlling the use of AI, much of the regulation introduced in other regions has been principle-based and non-mandatory. Given the stakes involved, we do believe that policy and governance is only going to get complicated ahead, but the pace and direction of its evolution will play a defining role.

- **U.S.:** Despite the U.S. being at the forefront of the AI boom, both in terms of technological development and deployment, concrete federal legislation governing AI has yet to be enacted. Perhaps the most prominent issuance of regulation at the federal level is the Blueprint for an AI Bill of Rights, proposed by the Biden administration in October 2022. While federal legislation is still developing, regulation at the local and state level is more advanced. New York City, for example, has been one of the first municipalities to create a law explicitly referencing AI.

- **EU:** The EU has used existing legislation to rule against the improper use of AI. For instance, the General Data Protection Regulation (GDPR) requires corporations to inform individuals when AI has been used to process personal data. More recently, Italy used GDPR rules to become the first European country to temporarily ban OpenAI's ChatGPT following privacy concerns. The EU has also gone one step further than many other countries and regions around the world when it comes to implementing new AI regulation. The European Commission published its first proposal for an AI regulation in April 2021, and the Council of the European Union adopted its common position ("general approach") on the AI Act in December 2022. In June 2023, the European Parliament (EP) voted to adopt its own negotiating position on the AI Act, triggering discussions between the three branches of the EU — the European Commission, the Council, and the Parliament — to reconcile the three different versions of the AI Act, the so-called "trilogue" procedure.

[29] Rishi Bommasani et al., "Do Foundation Model Providers Comply with the Draft EU AI Act?" Stanford University HAI, accessed August 17, 2023.

- **UK:** Although the UK government has acknowledged the risks posed by AI, it has been keen to emphasize the benefits of, and its support for, the technology. In March 2023, the government released a white paper on how it plans to regulate AI. One of the more important developments from this publication is that the government intends to rely on existing regulators, rather than a new entity, to devise appropriate measures to mitigate AI risks. While no new legislation has been proposed, the 90-page report does outline five key principles regulators should consider when placing guardrails on AI: (1) safety, security, and robustness; (2) transparency and explainability; (3) fairness; (4) accountability and governance; and (5) contestability and redress. In June 2023, the UK government announced it will host the first major global summit on AI safety. This is due to take place in late Autumn 2023.

- **China:** Authorities in China have been interested in regulating AI since the country devised the New Generation Artificial Intelligence Development Plan back in 2017. The way China approaches AI regulation will likely be consistent with its approach to regulating other areas of prominent technology, such as internet or social media, where it operates strict censorship to control the flow of information. Specific to Generative AI, the Cyberspace Administration of China (CAC), seen as the country's leading AI regulator, notably released draft measures to address concerns it has with the technology. Titled as the Administrative Measures for Generative Artificial Intelligence Services, the set of measures obliges tech companies in China to register Generative AI products with the agency alongside a risk and security assessment before it is available for public use. Following the release of the draft, the government opened the measures up for public consultation (this ended in May) to understand the different aspects of regulation that might be needed. Furthermore, in early June China announced that the 2023 legislation plan of the State Council will include the submission of a "draft AI law," among the 50 or so other measures up for review by the National People's Congress (NPC).

- **India:** As the world's most populous country, the largest democracy, and one of the largest and fastest growing economies in the world, India's influence on AI usage continues to grow. In 2019, the government released its National Strategy for AI. The report highlights how five sectors — healthcare; agriculture; education; smart cities and infrastructure; and smart mobility and transportation — stand to benefit the most from AI technologies. As recently as April 2023, the Ministry of Electronics and IT published a statement saying that it was not planning to regulate AI, pointing to the technology's positive impact on the economy.

Sector Opportunities from Generative AI

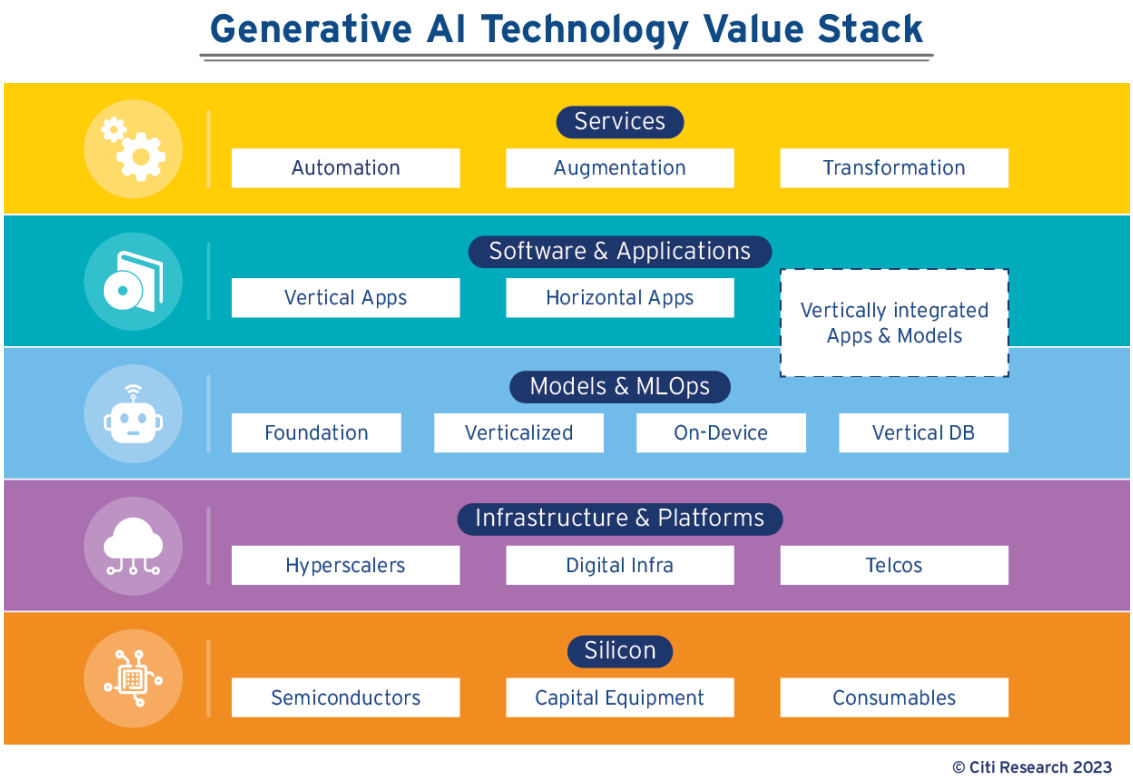
After looking at the history, the opportunities, and the risks associated with Generative AI, we now drill down further into the individual subsegments that are benefitting from the emergence of Generative AI. We start by looking at the technology value stack, and how Generative AI works its way from the Silicon layer all the way up to the services layer. Next, we look at how Generative AI is likely to affect the six supersectors and find Financials & FinTech to be the most affected.

The Generative AI Technology Stack

Citi Research Global Technology Team

We see five layers within the AI technology value stack, as outlined in Figure 18. Historically, the Silicon layer has been the de facto foundation of almost all technological shifts, and Generative AI is expected to drive significant growth for compute (i.e., processing power), networking, and memory chips. In the Infrastructure & Platforms layer, we see the hyperscalers/cloud providers racing today to build the underlying infrastructure that enables Generative AI applications and services, but over time we expect to see differentiation. When it comes to Models and Machine Learning Operations (MLOps), the open-source community is likely to be a key driver of innovation. Moving further up the stack, we believe nearly all software companies will be impacted in some form by Generative AI, and company-specific execution will be critical. Lastly, we believe Generative AI represents a step forward from ongoing AI/automation initiatives at the Services layer.

Figure 18. Generative AI Technology Value Stack



Source: Citi Research

Silicon

Atif Malik

Christopher Danely

Laura (Chia Yi) Chen

Masahiro Shibano

Peter Lee

Takayuki Naito

The Silicon layer has historically been the de facto foundation of almost all technological shifts, including the rise of machine learning and deep learning in the past. The emergence of Generative AI is therefore no exception. We expect Generative AI will drive significant growth across the supply chain, led by greater demand for compute, networking, and memory chips. In this context, we also note initiatives by the hyperscalers to make their own custom Silicon for compute. The picture at this layer would be incomplete without talking about the semiconductor capital equipment and consumables sector, which we view as the "picks and shovels of the wider AI gold rush," and hence good proxies for the emergence of Generative AI.

Compute Looks to be the Key Beneficiary

Within the semiconductor sector, Citi Research believes compute will be the key beneficiary from the emergence of Generative AI due particularly to demand for graphic processing units (GPUs), accompanied by the adoption and continued co-existence of custom application-specific integrated circuits (ASICs).

Heterogenous computing and different types of compute chips: Although Generative AI is relatively new, the process of sorting and understanding massive amounts of data (training) and making predictions (inferencing) has been around for decades. Training workloads have typically been performed using GPUs where processes are parallelized, while inferencing workloads have typically used central processing units (CPUs) where workloads are processed serially. As Moore's law slowed over the past decade, chipmakers have turned towards a heterogenous computing approach utilizing multiple types of compute chips including not only GPUs and CPUs, but also others such as application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs). The objective of this heterogenous computing approach is to optimize various workloads, balancing the trade-offs between performance and energy efficiency.

Figure 19. Compute Chips: CPU vs. GPU vs. ASIC vs. FPGA

Type of AI Chip	CPU	GPU	ASIC	FPGA
Definition	A general-purpose processor that can handle a variety of tasks.	A specialized processor that is designed to handle parallel processing tasks, making it well-suited for AI workloads.	A custom-built chip that is designed specifically for a particular task or set of tasks, making it very efficient at those tasks.	A chip that can be programmed to perform custom logic functions, making it highly flexible and adaptable.
Speed	Medium to high	High	High	Medium to high
Power consumption	Low to medium	High	Low	Medium to high
Cost	Low to medium	High	High	High
Use case	**CPU**	**GPU**	**ASIC**	**FPGA**
Large-scale training		√	√	√
Small-scale training	√	√		
Real-time inference	√	√	√	√
Low-power inference	√	√	√	√

Source: Citi Research

We see a particularly strong inflection for GPUs: GPUs remain at the forefront of performance benchmarks when it comes to both training and inference. In addition, they offer the ability to grow one's AI infrastructure at scale (particularly for larger models), this while retaining their interoperability benefits. Therefore, we see a particularly strong inflection for them.

- **Adoption and continued co-existence of ASICs:** An ASIC is a customized integrated circuit designed for a specific use case. Typically, designing custom chips requires a lot of engineering and financial resources, as companies will have to continuously innovate to keep pace with computing demands. However, the economics may potentially be much more appealing for hyperscalers who have the scale and, in the case of AI, may look into specialized hardware as an alternative to general purpose hardware, which then is customized via software. As such, both GPUs and ASICs will likely be used in the push to build the necessary infrastructure for AI, with ASICs potentially being used to primarily infer smaller and more specialized models and GPUs for both training and inference of larger and often more complex models.

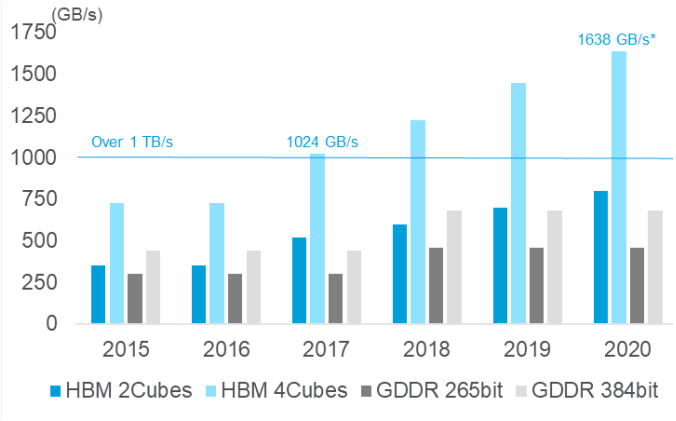
Generative AI Requires Memory

In addition to compute, demand for High Bandwidth Memory (HBM) and Double Data Rate 5 (DDR5) memory is likely to increase from the growth in AI computing.

- **HBM:** HBM is a high-speed system in package memory technology that uses stacks of vertically interconnected DRAM chips and wide (1024-bit) interface to enable higher storage capacity and data bandwidth than memory products using conventional wirebonding-based packages. While the processing speed of the fastest graphic DRAM is 600 gigabytes per second (GB/s), the comparable metric for HBM is up to 1,638 GB/s. This makes is very suitable for AI processing.

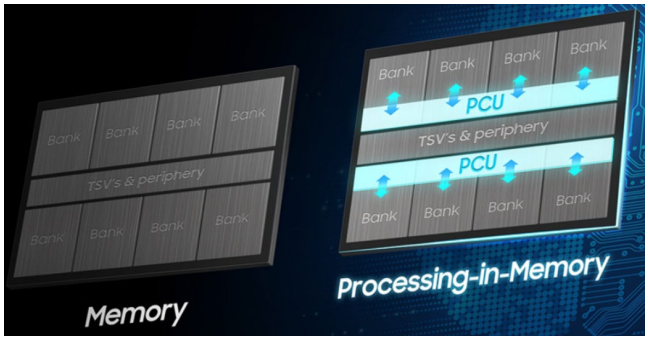
DDR5: With the advancement of AI models, the memory requirement in some GPUs have quadrupled over the past four to five years. This increase should drive DDR5 to become a mainstream product by 2024.

Figure 20. HBM vs. GDDR Bandwidth Comparison



Source: Company Reports, Citi Research

Figure 21. HBM-PIM (Processing-in-Memory)



Source: Company Reports

Networking Chips and IC Design Services Can Also Benefit

In addition to compute and memory chips, enterprises and hyperscalers need to evolve their network infrastructure to support the exponentially increasing amount of data processing associated with the emergence of Generative AI. With data centers becoming less and less collections of computers and more and more fleets of computers that are operated by a large operating system, there is a clear need to enable accelerated computing.

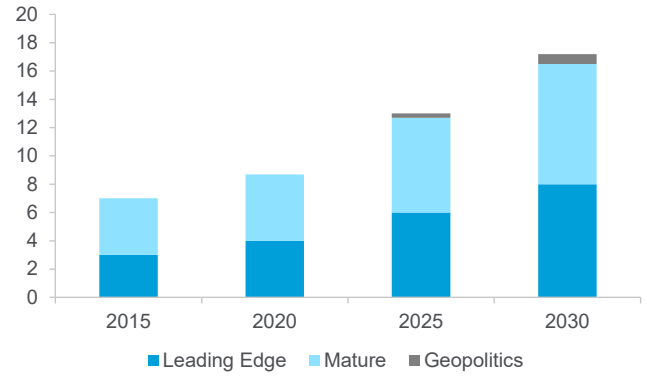
The development of compute chips, such as GPUs and ASICs aimed at AI workloads, as well as other specialized hardware, is in turn driving the development of new hardware and software architecture for IC design.

Picks and Shovels of the Wider AI Gold Rush

The picture at the Silicon layer would be incomplete without talking about the semiconductor capital equipment and consumables players. We view these as the "picks and shovels of the wider AI gold rush" and hence also as beneficiaries from the emergence of Generative AI, particularly in EMEA and Japan.

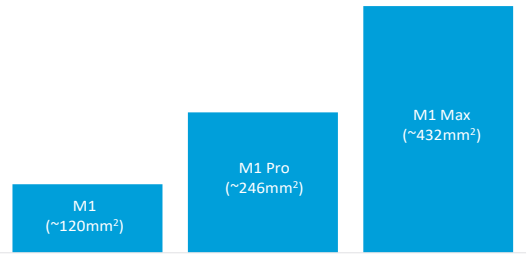
Despite a cyclical pullback in 2023, Citi Research sees industry wafer starts growing at a compound annual growth rate (CAGR) of about 8% over 2020-25 versus around 4% over 2015-20 driven by two broad demand vectors: digitalization and electrification, as well as technological shifts. Further out they see a 6% CAGR over 2025-30, factoring in geopolitical forces. This forecast captures the wider tailwind from AI within the two broad demand vectors and as such, the emergence of Generative AI as an additional tailwind that should further strengthen the secular credentials of sector.

Figure 22. Semiconductor Industry: Wafer Starts



Source: Company Reports, Company Reports and Citi Research Estimates

Figure 23. Technological Shifts: Apple M1 Chip Size Evolution



Source: Citi GPS, Citi Research, AnandTech (October 20021)

Infrastructure & Platforms

Alicia Yap, CFA

Andre Lin, CFA

Arthur Pineda

Asiya Merchant, CFA

Carrie Liu

Michael Rollins, CFA

Ronald Josey

The Infrastructure & Platforms layer in the technology value stack primarily incorporates major hyperscalers, e.g., large cloud service providers, that provide services such as computing and storage at enterprise scale, as well as other digital infrastructure providers encompassing data center-oriented firms in the U.S. and telecom operators in Asia.

Hyperscalers Are Racing Today Expect Differentiation over Time

Given the myriad of enterprise use-cases, major hyperscalers/cloud providers are competing to build the underlying infrastructure to support the adoption of Generative AI applications and services. With competition ramping and LLMs proliferating, we expect differentiation over time as hyperscalers compete within a matrix of trade-offs for their solutions, balancing parameters, domain-specific training, integration opportunities, performance, accuracy, speed, and pricing of Generative AI platforms.

- **Major hyperscalers/cloud providers are competing to build the underlying infrastructure** to support the adoption of Generative AI applications and services.

- **Hyperscalers could see multiple monetization opportunities** as newer offerings are already showing signs of tangible revenue acceleration, driven by the ramp in Generative AI investment in and capital expenditure. Monetization opportunities include model as a service (MaaS), application programming interfaces (APIs) and software development kits (SDKs), plugins and partnerships, and hosting and consumption fees.
- **Differentiation over time will result in a federated approach from customers.** The major Cloud providers will compete to host Generative AI workloads and we expect customers will favor a federated approach, to swap specific microservices, depending on use-case-specific needs for lower pricing versus features/speed.
- **Open-sourced platforms are racing in parallel to build out AI infrastructure.** We expect major platforms will continue to develop increasingly sophisticated Generative AI infrastructure and toolkits, thereby accelerating innovation through abstraction layers that can simplify enterprise adoption.

Models and Machine Learning Operations (MLOps)

Alicia Yap, CFA

Ronald Josey

Tyler Radke

The Models and MLOps layer encompass all types of models facilitating Generative AI from large generic foundation models to verticalized ones, associated entities such as hubs (which can simply be thought of as marketplaces for models), as well as other facilitating elements related to MLOps. The importance of this layer cannot be emphasized enough. Simply put, if the entire Generative AI value stack was a solar system, then the model would be its star.

Consistent with software evolutions in the past, we see the open-source community driving innovation in this layer. A key trend to monitor is the rise of verticalized models. As well as models enabling on-device AI.

Open-Source Community Driving Innovation

There has been an exponential acceleration in Generative AI innovation driven by open-source marketplaces, open-source LLMs, and open-source vector databases. Innovations around model scaling, instruction tuning, and on-device AI are starting to take hold and may help enterprises more readily buy into trialing Generative AI capabilities, as these innovations allow enterprises to search/ask questions of their data without feeding more data into hyperscalers/researchers, that they could use to train more models.

Figure 24. Large Language Models: Open-Source vs. Closed-Source

Hugging Face Top 6 Ranked Models	Date Added	Compute Infrastructure - Hardware	Fine-Tuned from Model	Model Details and Best Use Cases
Falcon-40B-Instruct	May-23	AWS SageMaker: 64 A100 40GB GPUs	Falcon-7B	Not tuned for a particular purpose. Acts as a "base model" and was trained on chats
CalderaAI/30B-Lazarus	May-23	-	Alpaca & Vicuna	Text generation and natural language understanding (NLU)
Falcon-40B	May-23	AWS SageMaker: 384 A100 40GB GPUs	-	Not tuned for a particular purpose. Acts as a "base model" and was trained on chats
LLaMA-30b-supercot	Apr-23	-	LLaMA with SuperCOT-LoRA	LangChain prompting
LLaMA-65B	Feb-23	-	-	Not tuned for a particular purpose. Acts as a "base model" and was trained on chats
GPT4-X-Alpasta-30b	May-23	-	Alpaca-LoRA	Text generation and NLU
Notable Proprietary LLMs	**Data Added**	**Parameters**	**Company**	**Best Use Cases**
PaLM 2	May-23	*~340B	Google	Multilingual translation, code generation, reasoning tasks
GPT-4	Jan-23	*1T	OpenAI	Conversational AI, text summarization, NLU
ChatGPT	Nov-22	-	OpenAI	Chatbot development, customer service, NLU
OPT-175B	May-22	175B	Meta	Content creation, data analysis, question answering
GPT-3	Sep-21	175B	OpenAI	Text generation, sentiment analysis, text classification

Note(*): The parameters for PaLM 2 and GPT-4 are estimated.
Source: Citi Research

Rise of Verticalized Models and On-Device AI Are Trends to Monitor

- **Verticalized models aimed at specific use cases**: We are seeing more research/companies iterate on top of base models for specific use cases. These iterations use proprietary training data and creating new, smaller, and more efficient models.

- **Differentiation and outperformance versus foundation models.** As foundation models proliferate and models scale (i.e., parameter count), parameters could become less important than adapting each model for verticalized use-cases. Verticalized models have outperformed foundational models with enhanced capabilities and better accuracy. By customizing models on unique company and industry training data, we believe companies can drive differentiation in LLMs over time and reduce commoditization risk.

- **Models enabling on-device AI:** Similar to the trend of smaller, more use-case specific models, hybrid AI/on-device AI hopes to offer scalability, performance, and personalization at reduced costs, with the added benefit of on-premise/on-device security. If model size and prompt (i.e., question) are below a certain threshold, then inference can run on a device (currently estimated to be >1 billion parameters, though models with >10 billion are slated to work eventually). Larger models can use a hybrid approach to work across devices and the cloud. Bringing Generative AI capabilities closer to the source can also enable per-user alignment and tuning. As models become more user-case specific, the added benefit of a model that is on-device means that the model can run and train locally without exposing data to hyperscalers.

Software and Applications

Amit B Harchandani

Fatima Boolani

Steven Enders, CFA

Tyler Radke

Beyond a few application use cases that have already seen strong adoption, we suspect narratives in the application layer will take more time to play out relative to the underlying infrastructure layer. We expect nearly all software companies will be impacted in some form, creating a larger emphasis on company-specific execution in navigating the rapidly changing landscape. At the highest level of analysis, we see the opportunity for the application space to take on a larger share of organizations budgets, enabling new monetization opportunities based on the increased value add.

Some of the most common use cases for Generative AI are customer service, content creation, data augmentation, code generation, and research and development. The integration of Generative AI capabilities into software, applications, and workflows represents another key driver of disruption.

- **Data Warehouses/Analytic Platforms:** Data/analytics software is among the most exposed software categories to Generative AI as it "sits' closer to the underlying infrastructure and could benefit from the increased compute/data intensity of LLMs. We believe the sub-category could benefit broadly with increased prioritization of data/analytics projects as organizations modernize environments to be able to leverage LLMs. In addition, we see a potential acceleration of "data democratization" tailwinds with the proliferation of search-based/low complexity tools in the hands of more knowledge workers.

- **Front Office/CRM Software/Digital Commerce:** Generative AI has several compelling potential use-cases within the front office space that can help drive efficiencies and better customer relations/engagement based on the large volume of data. Industry research firm Gartner estimates suggest that by 2025, 30% of outbound marketing messages from large organizations will be synthetically generated, which is up from less than 2% in 2022. Use cases within e-commerce, like human image generation for modeling and showing alterations of different poses/ages/diverse representations, are also becoming commonplace. Gartner data also indicates that customer experience is currently the most common primary focus of Generative AI investments, cited by 38% of respondents polled.

- **Back Office Software:** Across our sub-categories — HR Software, Collaboration Software, and Financial Software — we see varying levels of clarity in how Generative AI will be or already is incorporated into feature sets. HR software vendors are incorporating Generative AI into their platforms with use cases like recruiting, talent management, and core HR processes. On the other hand, collaboration vendors are still relatively early in integrating Generative AI into their platforms, with existing workflows primarily targeting process automation or adding a level of intelligence to the software. Enterprise Financials is likely to be one of the slower areas of uptake due to sensitivity around core financial data.

- **Cybersecurity and Infrastructure Software:** In the realm of cybersecurity, Generative AI presents meaningful opportunities but also poses threats. Threat actors will likely look to lean further into Generative AI to drive attach speed and augment attack capabilities. Security vendors are leveraging Generative AI to elevate analyst capabilities and streamline threat intelligence. As a result, we view Generative Ai as a force multiplier and a positive for cyber vendors. We also see it as a solution for the dire talent shortage that exists today in the Security industry.

Services

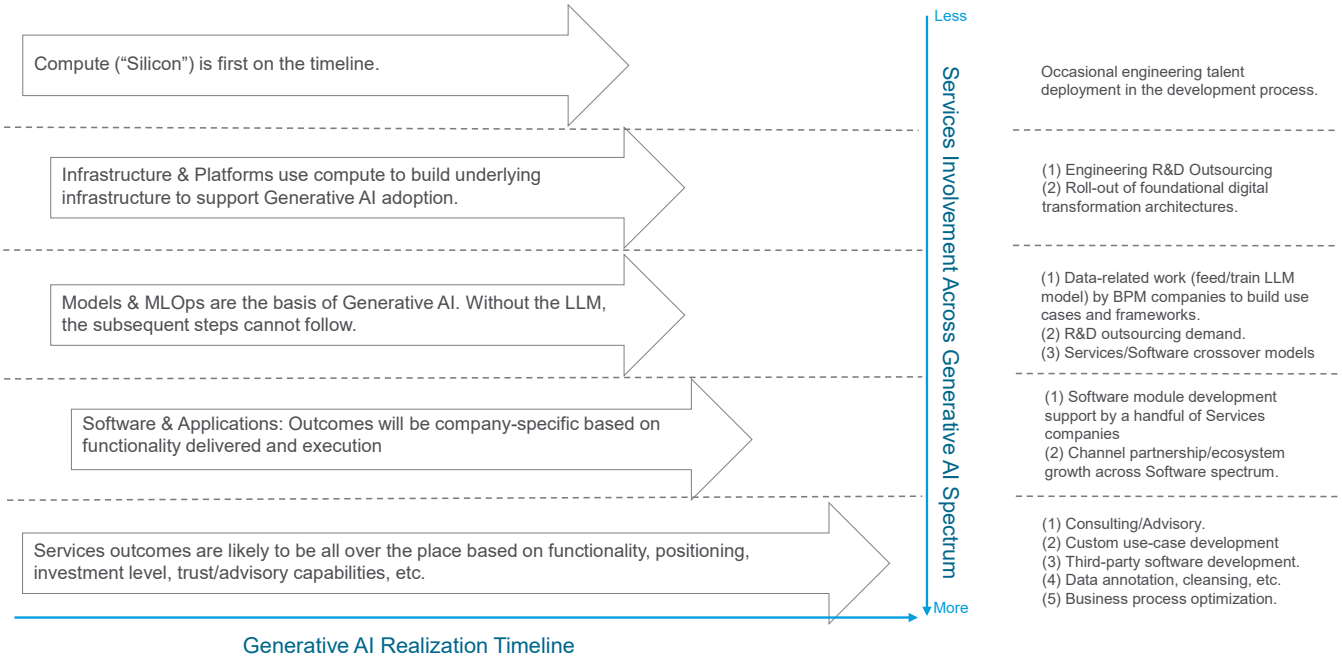
Amit B Harchandani

Ashwin Shirvaikar, CFA

Surendra Goyal, CFA

Generative AI represents a step forward from ongoing AI/automation initiatives across the IT & Business Process Management (BPM) Services space. However, productivity savings could be more meaningful. As with other advancements in the past, there will be share shifts. Over time, companies will likely build their own solutions/business units, mirroring the trajectory we saw with the shift to the cloud. Jobs will likely be eliminated but these are likely to be replaced by others. Therefore, on balance, we believe Generative AI is neutral to net positive for the overall space. However, the majority of impact in the case of the Services layer is delayed because it is Services that makes technology adoption a reality.

Figure 25. Generative AI Impact on Services



Source: Citi Research

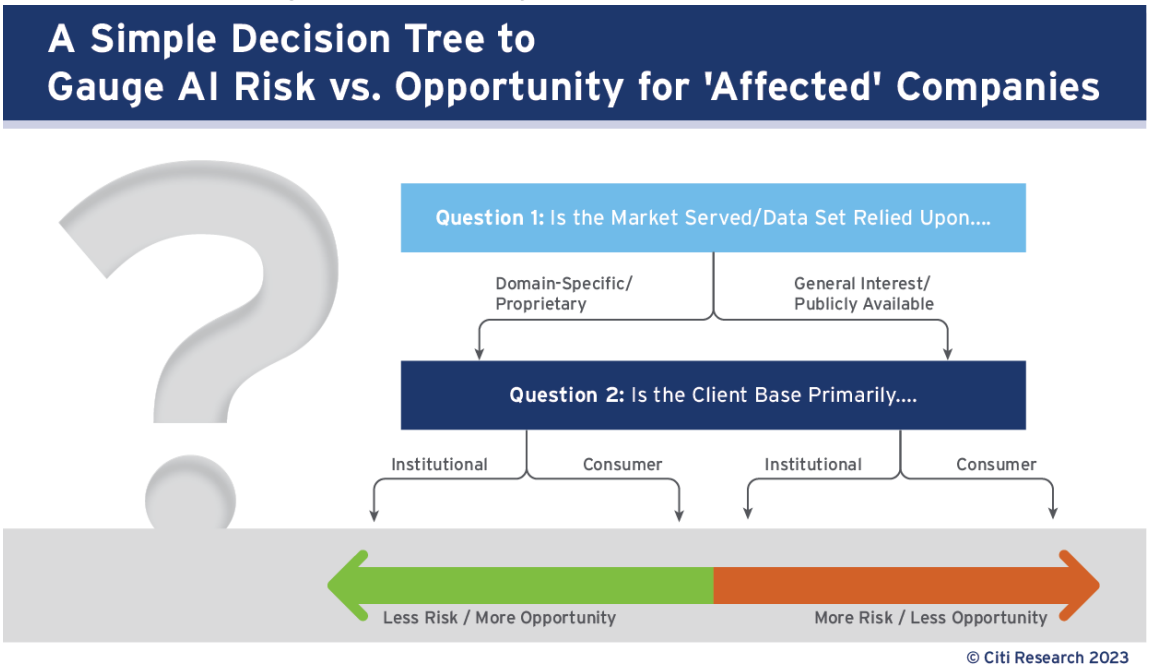
Generative AI: Assessing the Risk/Reward Outside Technology

Amit B Harchandani

Thomas A. Singlehurst, CFA

Although the majority of talk surrounding Generative AI focuses on the opportunities it creates, there are also likely to be risks created. To get an overall picture of the how sectors outside of the Technology stack could be affected by Generative AI, we devised a two-stage framework to assess risk/reward, which we believe can be broadly applied across businesses and sectors.

Figure 26. A Simple Decision Tree to Gauge AI Risk vs. Opportunity for "Affected" Companies



Source: Citi Research

Two-Stage Framework to Assess Risk/Reward

The first stage of our framework involves (1) mapping the risk from Generative AI, based on both first-order and second-order implications, (2) determining the implications for the cost base, and (3) gauging end-industry risk. The goal of this exercise is to identify the degree which companies within a sector are affected versus classifying companies into buckets of “winners” and “losers.”

In the second stage, we use a simple two-step decision tree to gauge risk/reward for those companies within a sector deemed to be affected from the emergence of AI after the first stage. Factors in the decision tree are based on the principles that (1) a relevant database trumps a good algorithm, and that (2) consumers move faster than institutions.

Stage 1: Mapping AI Opportunity and Risk

The mapping exercise in the first stage of our framework results in an AI risk score based upon four inputs. This score then helps us to differentiate between companies within a sector which are more and less affected by AI.

- **First-order vs. second-order considerations:** The distinction between first-order and second-order considerations is closely related to short-term versus medium/long-term considerations. An easy way to look make this distinction is by looking back at history. When software first emerged in the 1970s/80s, it undoubtedly was thought to be game-changing firms employing white collar professionals, such as bankers, lawyers, and accountants. But five decades later, these firms all benefited from the productivity saves generated by computing.
- **Opportunities vs. risks:** We should also consider that for some of the companies, even those perceived short-term to be at risk, the disruption of AI-based technology may end up being positive for their prospects. Likewise, what at first seems like a potential windfall, may end up being unhelpful.
- **We use these distinctions to come up with an AI score** based on four inputs that we describe in more detail below. The score is based on a simple average (i.e., with equal weighting) of the numerical scores allocated to each of the inputs.
  - **First-order implications:** What do we anticipate will be the first impact of AI? Often, this is the near-term or direct implication rather than how we think AI will play out longer-term.
  - **Second-order implications:** Here, we try to capture the second order/longer-term effects of AI. In some cases, the implications may be more significant but in others it may be less so.
  - **Cost implications:** Looking at the cost implications specifically is important. While for some companies, the emergence of AI-based technology will require additional investment, for others it will potentially drive greater cost efficiency via automation. This needs to be captured within any scoring system.
  - **End-industry risk:** Although we likely already captured some of the end-market risk in the second-order implications, we think it important as a long input to capture not only any significant adverse exposure to sectors that may be disrupted, but also to take into account things like whether a particular company's client base is more skewed to organizations/institutions or individual consumers.

Stage 2: Gauge Risk/Reward for Those Deemed to be "Affected"

Once we determine an AI score for companies within a sector, we then use a simple two-step decision tree to gauge risk/reward for those deemed to be affected from emergence of AI after the first stage. By doing this, we can hopefully determine if the risk is positive or negative, and the speed at which the risk is likely to appear. We base this exercise on two underlying principles: (1) a relevant database trumps a good algorithm, and (2) consumers move faster than institutions.

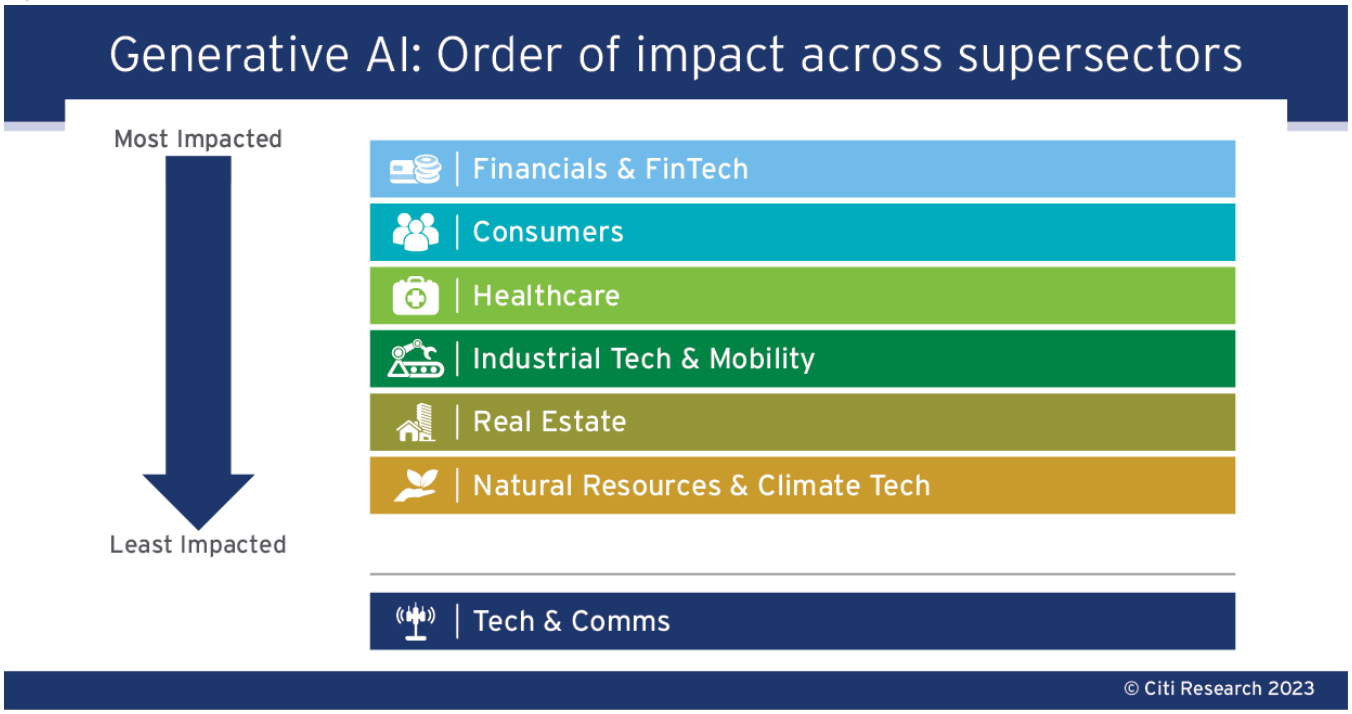
- **Principle 1: A relevant database trumps a good algorithm:** However good a company's search/analytics technology is, if it is training on a database that is sub-par, the quality of the technology's output will be compromised. This reminds us that it is not only important to consider the size of a database but also to look at its relevance and accuracy. Applying this principle to our analysis, we need to look at whether the industry is "domain-specific" when determining whether a company is likely to be adversely affected.
- **Principle 2: Consumers move faster than institutions/enterprises:** In general, consumers are more likely to be quick to adopt new generative AI-based tools than institutions/enterprises. Based on this, institutional user bases will likely take longer to disrupt.

Reflecting on both points, we think it is fair to assume that, at the very least, institutional user bases will take longer to disrupt and, possibly more likely, for those companies' serving institutions, there is more of a chance of being the intermediary when it comes to AI deployment. Considering both principles, we create a simple decision tree for investors as shown in Figure 26.

Generative AI: Impact Across Supersectors

With the two-stage framework defined, Citi Research analysts around the world were asked to assess the adoption curve for Generative AI by applying the framework to their company coverage lists across the seven global supersectors. The results of the exercise can be seen in Figure 27 and shows the order of impact from Generative AI across the supersectors. The most impacted supersector is Financials & FinTech (with variations across sectors) followed by Consumer. Natural Resources & ClimateTech looks to be the least impacted, sitting at the bottom of the spectrum.

Figure 27. Generative AI: Order of Impact Across Supersectors



We show the Tech & Comms supersector separately as it is the enabling supersector
Source: Citi Research

Financials & FinTech

Ashwin Shirvaikar, CFA

Judy Zhang

The tenor and frequency of AI conversations within Financials & FinTech depends on the specific sector. The interesting overall takeaway is that questions being asked around AI are more like "What can banks/brokers/asset managers/insurance companies do with Generative AI?" rather than "How will Generative AI disrupt the industry?" This is partly because end-markets in this sector are largely regulated therefore regulators will have a say in the speed and design of AI take-up. Also, financial services companies are already spending a good deal of money upgrading their technology away from legacy tech, which could limit widespread adoption in the near term.

Key opportunities from AI/Generative AI adoption?

Current use-cases from Generative AI and AI more broadly in Financials & FinTech tend to focus on improving customer service models or optimizing costs. However, Citi's Financial Services and FinTech team believes Generative AI has potential to (1) help democratize investing and markets, (2) improve algorithmic trading, (3) enable more robust data utilization; and (4) allow for better analyzing and pricing of insurance risk. Other opportunities include:

- **Improved customer experience:** Providing personalized services, using chatbots and virtual assistants to provide 24/7 support and handle customer inquiries, and helping human agent and services industry participants improve the effectiveness of their conversations.
- **Fraud detection and prevention:** Identifying fraudulent activities in real time.
- **Business risk management and decision-making:** Helping financial institutions make informed decisions in terms of macro and market risk assessment.
- **Regulatory compliance:** Automating compliance checks and monitoring.
- **Digital operations transformation:** Speeding up the sectors digital transformation journey.
- **More robust data utilization and more effective algorithms:** AI could help drive more robust data and utilization of data that would create more effective/efficient algorithms that investors utilize.
- **Improved investor education:** Lowering the barrier to entry for less sophisticated investors.
- **Democratization of wealth management:** Filling the education gap between advisors and retail investors who are new to alternative investments.
- **Property & Casualty insurers:** play a key role in helping clients (policyholders) Lowering the cost of capital and protecting client's economic output by invest in AI for revenue-generating or cost-efficiency programs.
- **Better pricing of risks:** Deploying AI to better analyze and price insurance risks.

Key risks to existing business models from AI/Generative AI adoption?

There are functional risks around the use of AI in the Financials & FinTech sector, although there is also potential for competitive risk from not investing in AI. Key risks specific to the Financial & FinTech sector include:

- Generative AI as a **driver of fraud**, especially identity fraud.
- **Reputational and regulatory factors** that arise if AI strategies are not properly executed.
- **Market manipulation** from AI and "meme stock"-like events.
- **Replacement of traditional financial advisor models** with AI-driven "advice."
- **Underperformance by asset managers** from not properly synthesizing AI into the investment process.
- Inadvertently/unintentionally **running afoul of privacy rules**.
- **Malfunctioning of AI.**

Consumer

Simon Hales

Tiffany Feng

Paul Lejuez, CFA, CPA

Steven Zaccone, CFA

AI tools are increasingly being discussed and deployed across business operations within Consumer Staples companies, who are seeing these tools as levers to improve speed and efficiencies of operations. This includes areas like speed of innovation, how to use data more efficiently in manufacturing, supply chain and marketing. AI is also increasingly being appreciated for its dual potential to help companies reach sustainability goals (it can help with discovering new materials, but also with reducing wastage and energy consumption in production). In areas like Traditional luxury, the industry remains a late adopter of technology, and the AI debate continues to be marginal, but in Jewelry, companies see potential for AI in areas like design and modeling but not yet in cutting and polishing. However, it may be used to help with grading, which would improve the efficiency of certification. Interestingly, Hotel companies are actively engaging in AI discussions, while Consumer Retail companies are exploring ways to use generative AI to handle customer inquiries, facilitate text-to-shop, and finding items in stores, as well as increasing efficiency for inventory and supply chain management.

Key opportunities from AI/Generative AI adoption?

- **Mass customization and personalization:** Generative AI can be used to generate content in multiple languages and thousands of versions without having to significantly increase headcount and or budget.
- **Product authenticity at POS:** Generative AI has the potential to fight against counterfeit products, with machine learning helping to identify subtle irregularities such as in shape, color, texture, label used, etc. This is particularly relevant for high-end spirits and cosmetics offerings.
- **Facial recognition:** In-store facial recognition can drive improved personalization of the customer experience in certain countries.

- **Innovation:** Leveraging digital technologies enables R&D departments to innovate "better, faster, and more effectively." This is partly achieved as AI enables companies to understand the world in a more predictive way. AI tools accelerate scientific discovery by taking on multiple tasks that previously required labor-intensive lab work.

- **Cost cutting:** Generative AI tools should help to speed up day-to-day admin operations such as communications, marketing campaigns, coding, and analysis without impacting the size of workforce.

- **Manufacturing and supply chain:** AI can be used to simulate the manufacturing process, which can simplify the process and improve its sustainability (i.e., more energy efficiency, less waste). Moreover, supply chains can be more flexible and reactive to shortages, with AI able to uncover alternative ingredients or simplify products by reducing the number of components without impacting their quality or effectiveness. Digital twin applications can also aid in the design process of the manufacturing process.

- **Marketing and sales planning:** AI can improve marketing effectiveness to make campaigns perform better; for instance, providing insights to improve logos, placements of pictures, and sound in advertisements. (See this interview with KH Data Head, mediapost.com.) AI can improve data insights for sales planning purposes. AI can be used to assess data at more granular customer levels, for instance neighborhoods or postcodes, which can then be used to target stores and customize for different neighborhoods.

- **Chatbots:** Companies can improve online customer service through various social media platforms (hotels and the tourism industry have been early adopters).

- **Customer satisfaction:** Operational efficiency and customer experience in the tourism industry could be improved by AI in areas like: (1) tailor-made services catering to customers' preferences (i.e., personalized itinerary based on different tourism destinations and activity suggestions, rooms set up prior to arrival); (2) precise marketing to target groups through data analysis and improve search engine optimization (i.e., high-quality content generation); (3) real-time customer support in pre-booking and during the trip and timely identification and resolution of issues; (4) more accurate demand forecasts and pricing strategies; (5) digitalized experience (i.e., smart robot service, online check-in and check-out); (6) automating tasks to relieve staff workloads and reduce cost.

- **Hedging:** Using AI to predict commodity price movements can potentially help make better hedging decisions.

Key risks to existing business models from AI/Generative AI adoption?

- Building trust in tools that may not always be fully understood and accepting recommendations that may contradict prior beliefs. The risks associated with generative AI include consumer data vulnerability and potential bias in models.

- Adopting generative AI would mean incremental costs incurred through entering into partnerships or developing in-house upskilling.

- Data privacy and liabilities/copyright issues from AI-generated content.

- Brands making decisions or implementing solutions based on AI algorithms should be careful to avoid the introduction of unfair bias.
- In luxury product design, some aspirational elements are always important and cannot be totally replaced by AI. It may also be challenging to execute the transformation of the current digital platform to adopt AI while making sure that it is net-profit positive for the business.

Healthcare

Andrew Baum, MD

Patrick Donnelly

Major pharma companies are becoming more vocal in talking about their investment and the potential of AI across multiple parts of their value chain. The potential for AI use is seen across commercial, drug development, trial, and discovery. Other use cases include trial design protocols, automated administrative efforts, and patient/physician education. In Health Tech, the use of AI has consistently been a topic of discussion and while there could be benefits for its use in certain administrative use-cases, its use in a clinical setting draws skepticism.

Key opportunities from AI/Generative AI adoption?

- Within large pharma, we think the best use cases for AI broadly are in drug discovery and design, patient selection and recruitment, and optimization of sampling/sales calls. Within nontherapeutic healthcare, the sub-sectors payment integrity, utilization management, risk coding, Stars score improvement, and revenue cycle management are key areas of early adoption.
- Generative AI in particularly can be extremely helpful in (1) patient communication, (2) health coaching (e.g., reminding patients to perform healthy habits), and (3) note taking/data input into electronic health records (effectively reducing time physicians spend on admin tasks).

Key risks to existing business models from AI/Generative AI adoption?

- Within pharma, there is little disruptive risk to large cap pharma given the requirements to establish efficacy and safety through comprehensive clinical trials and pre-clinical testing. AI should allow improvement of discovery hit rate and reduce clinical attrition. We also anticipate AI to accelerate clinical trials through improved patient selection and trial recruitment. Generative AI may accelerate the downsizing of both physical and remote sales reps, improving margins.
- Healthcare data is highly regulated. If generative AI cannot be trained on vast amounts of healthcare data (due to regulatory burden), this would limit effectiveness and adoption.

Industrial Tech & Mobility

Martin Wilkie

Andrew Kaplowitz

Itay Michaeli

The broader debate around using advanced technology in the Industrial Tech & Mobility space has been around for a long time, specifically using digitalization as an ongoing growth and margin driver. Commentary on AI is at a very early stage. The Aerospace & Defense sector is looking at AI to assist with data fusion and manipulation, as well as for advanced cryptography. In the Autos & Mobility sector, AI will likely lead to a greater focus on autonomous vehicles but can also be instrumental in driving product development, manufacturing, and customer-facing services.

Key opportunities from AI/Generative AI adoption?

- **Generative AI**
  - **Generative AI in low code/no code industrial systems:** Industrial processes are increasingly digitized, with industrial IoT platforms allowing end-users to create industry-specific and process-specific apps. Generative AI could significantly broaden the ability to create code, massively opening up the market for analyzing data on the industrial IoT platform.
  - **Interactive applications:** Generative AI is expected to contribute to the development of robots that better understand and interact with humans.
  - **Customer facing:** Some customer choice applications (e.g., buying a car) could be an opportunity.
- **AI more broadly**
  - **Process optimization (production, supply chain):** Production processes and supply chains are increasingly complex, with the amount of data available to make an optimization decision increasingly too large (and ever-changing) for legacy optimization models. A combination of machine learning and quantum computing could vastly improve these optimization decisions and increase supply chain efficiency. On design, digital twin applications (for example as automakers build out new EV platforms) can also be areas that benefit from this optimization.
  - **Infrastructure needed for data center growth:** AI will require a lot of compute power, meaning data center infrastructure, in energy use, power management, cooling – and adjacent markets like heat reuse – will become increasingly important as the datacenter infrastructure needed to power AI is rolled out.
  - **Data manipulation:** Defense systems could use AI to drive data fusion, distillation and faster decision making.
  - **Accelerated autonomous vehicle development:** AI can be leveraged to further accelerate the deployment and scaling of autonomous vehicles under various domains and business models.

Key risks to existing business models from AI/Generative AI adoption?

- End-markets like production, installation, construction, extraction, and transportation are seen as having below-average risk of impact, and therefore the broader risk to business models within Industrial Tech & Mobility is likely less impactful than in some other sectors.
- The debate is in the early stages, and we think will form part of the broader debate on cybersecurity and data sovereignty for the Industrial IoT. In certain mission-critical and highly regulated applications, for example aerospace, certification and safety is a key factor — the lack of repeatability and non-auditable decisions may make certification very difficult.

Real Estate

Nick Joseph

Aaron Guy

Craig Mailman

The real estate sector is generally less exposed to AI versus other sectors. However, for data centers, the debate is whether they will be positive or negatively impacted by the rise of Generative AI deployments. Demand for space and power are likely to rise, especially early in the adoption cycle. But there is a risk that accelerating computing will dilute the returns of existing data centers or drive some of them into obsolescence.

Key opportunities from AI/Generative AI adoption?

- **Data centers:** Accelerating high-performance computing adoption associated with AI workloads is a potential positive for data center demand for at least the first few years of this adoption cycle. Although there is a risk that existing general computing workloads are cannibalized initially by the accelerated IT infrastructure, spending is more likely to be immediately focused on an expanding array of learning and inference Generative AI models.
- **Smart buildings:** Smart buildings, which automate features like HVAC, lighting, alarms, leak sensors, and security could be improved by AI to lead.
- **Chatbots:** Further advancements in the use of AI for chatbots could help consumer-facing applications at buildings such as hotels and residential.
- **Efficiency:** AI could facilitate enhancements to the supply chain by predicting product demand and ultimately optimizing logistics, while driving productivity improvements in areas such as lease writing, valuation, due diligence, and legal.

Key risks to existing business models from AI/Generative AI adoption?

- **Data centers may need further investments:** For data center firms, high power density requirements needed with accelerated computing may not be within the design parameters of existing data centers, possibly increasing the need for investment in existing builds facilities or requiring new data center builds.
- **Office space:** There are conflicting views on the impact to office space — with bulls considering the growing office demand from AI companies as a positive in the near term, and bears concerned that AI may displace office using jobs in the long term.
- **Travel & Expenses:** Optimizing travel and expenses which could impact travel and hotels.
- **Regulation:** Regulation is always a potential risk and concern with new technology. Regulation is more likely to impact the users and hosts of AI workloads, rather than the digital real estate managed by the data center firms. A limited amount of regulation can help to unlock the black-box problem attached to AI-created insights and could also help to broaden the interest in AI-based solutions.

Natural Resources & ClimateTech

Alastair R Syme

Jenny Ping

Oscar Yee

Q1. How has the wider AI debate evolved over the past six months following the rise of generative AI?

The Natural Resources & ClimateTech supersector has its roots in heavy industry rather than customer-facing roles. AI is therefore not yet on the radar screen of its companies as they do not benefit initially from workforce-based productivity gains.

Key opportunities from AI/Generative AI adoption?

- **Resource-efficiency:** The use of digital has been key in helping the oil and gas and mining industries have found resources — e.g., the rise of the seismic industry in the 1980s and 90s. Generative learning should help enhance that discovery.

- **Asset-efficiency:** A deepening of electrification (EVs, replacing fossil fuels with electric), more intermittent renewables, more interconnectivity, more smart appliances, more distributed energy (e.g., rooftop solar), and a greater role for storage assets are creating a rich dataset of the changing way we produce and consume energy. This dataset can help manage demand — e.g., the use of smart appliances – as well as manage the role of existing plant (e.g., a gas power plant).

Key risks to existing business models from AI/Generative AI adoption?

- **Demand risks:** As efficiency in the system increases and the global system wastes less, demand is likely to continue to fall. Undoubtedly efficiency brings with it a question of demand. Pressure from increased renewables penetration and distributed generation is increasing the amount of investment networks must make, but better optimization through AI can conceivably see some of this investment pushed out or mitigated entirely. In theory, therefore, networks are a key beneficiary of AI in this super-sector.

- **Regulation:** Only to the extent that what we have referenced above requires widespread sharing of data. Networks in particular are often localized assets, e.g., at a U.S. state or EU country level, and rules and regulations at this level may limit the ability to innovate efficiently.

The AI Arms Race

Pantelis Koutroumpis
Director, Oxford Martin Programme on Technological and Economic Change
Oxford Martin School

What does the future look like for Generative AI? One way to investigate the global trends and growth is from the perspective of investment in technological innovation. We do this by analyzing the number of AI-related patent applications over time and across countries. Research papers are also telling, with the total cumulative AI research output increasing 1,300% between 2003 and 2021. Given the importance of AI as a foundational technology, the race is on between countries for scientific and technological dominance.

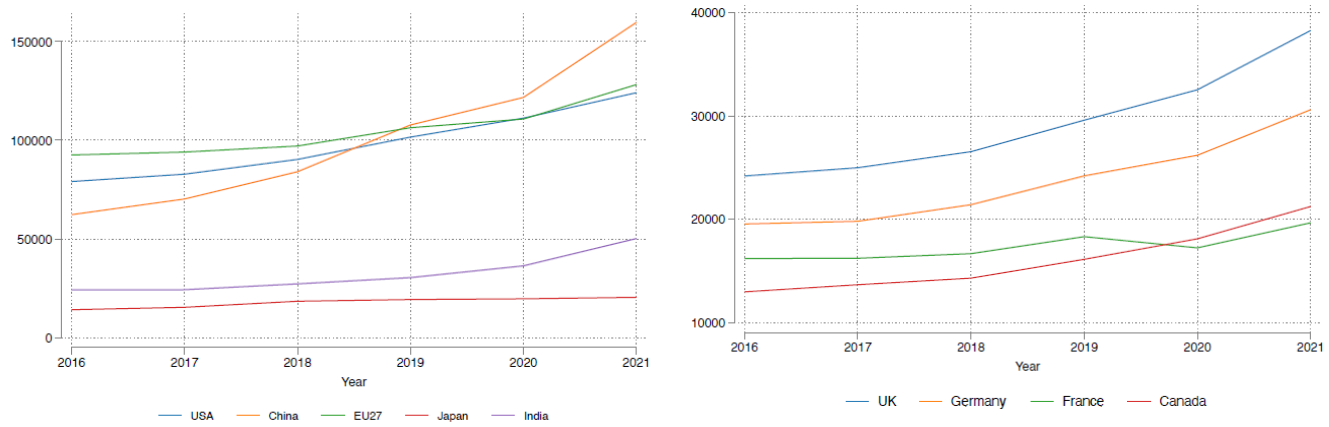
Research Outputs for AI

The research outputs for AI are following an unprecedented pace, as nations vie for leadership. According to our analysis of journal publications and conference papers tracked by the OECD AI database and using OpenAlex and Scopus data, the total cumulative AI research output in 2003 was less than 1 million papers and in 2021 it rose to almost 13 million papers, an increase of 1,300%. The flows have evidently grown as well, reaching 1 million papers in 2021 compared to less than 200,000 back in 2000.

The importance of AI research as a foundational technology has led to a race across countries for scientific and technological dominance. In the early 2000s, the U.S. had a clear advantage over other countries, but this situation started to change around 2010. First, China emerged as an AI research powerhouse, overtaking the U.S. in total publications per year in 2006 and later the EU (as a single entity including the UK back then) emerged as a frontrunner in 2008. In fact, from 2014 to 2018 the EU was the global leader in total research outputs (including papers, conferences, and books.) in the field of AI before being replaced by China again. Since then, China tops the research outputs followed by the EU and U.S. which have very similar counts (Figure 28).

Beyond the leading positions, other large countries have developed AI research over this period. Japan, as a technological leader in electronics, has maintained a stable path, increasing its outputs over time. However, India has exhibited a very impressive record since 2014, when it overcame Japan and since 2018 it held the fourth place in global research outputs. As of 2021, India produced 42% of the U.S. output with an increasing trend and possibly a higher output in the coming years. Looking beyond the top countries, other large economies have steadily emerged with significant research over the past decade. The United Kingdom is the leading country in Europe with almost 40,000 publications in 2021 followed by Germany with ~30,000 and France with ~20,000 during the same year. Canada is also pushing forward steadily over the past decade overpassing France in 2021 by a small margin.

Figure 28. Global Race for Research in AI

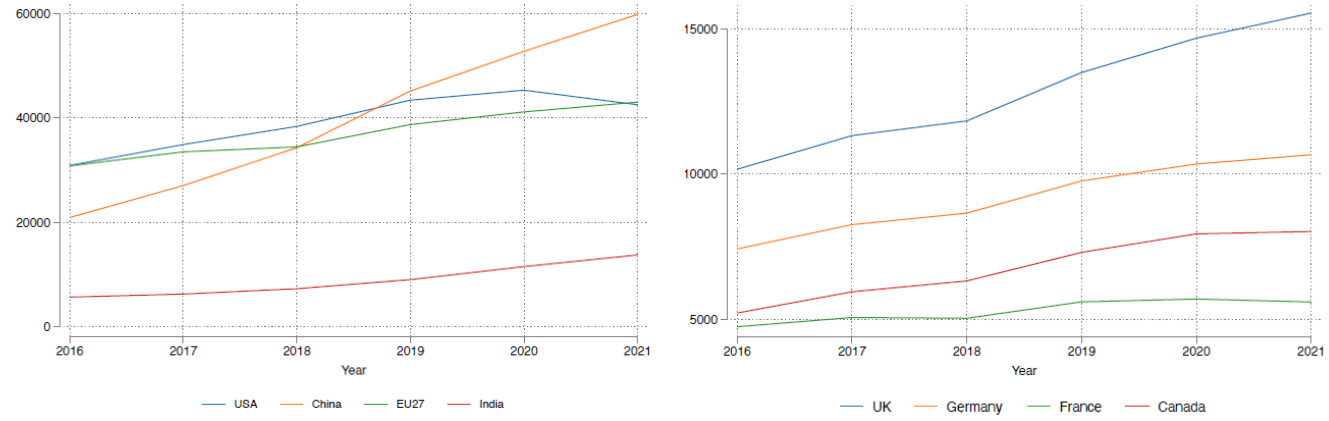


Source: OECD Open AI

Do research counts matter? Ironically, AI is a technology that can produce human-like text on its own, augmenting and facilitating research outputs, at least in the sense of lower quality contributions. In this sense, the total counts described in the previous section might only indicate a general trend rather than actual leadership over specific domains. For this, we look into the high impact research outputs in more detail rather than counting all papers and books that originate from each country.

Using inputs from Scopus, scientific publications are ranked based on the Field-Weighted Citation Impact (FWCI), which is the ratio of the total citations actually received by a scientific publication and the total citations that would be expected based on the average of the subject field of scientific discipline.

Figure 29. The Race for Top-Quality Research in AI



Source: OECD Open AI

A FWCI of 1 means that the publication is cited as much as the average publication in that subject field or scientific discipline with higher and lower values showing more or less citations per paper. To identify the most promising research outputs, OECD.AI defines three categories of scientific publications based on their FWCI score: (1) Low impact: O < FWCI <= 0.5; (2) Medium impact: 0.5 < FWCI <= 1.5; and (3) High impact: FWCI > 1.5.

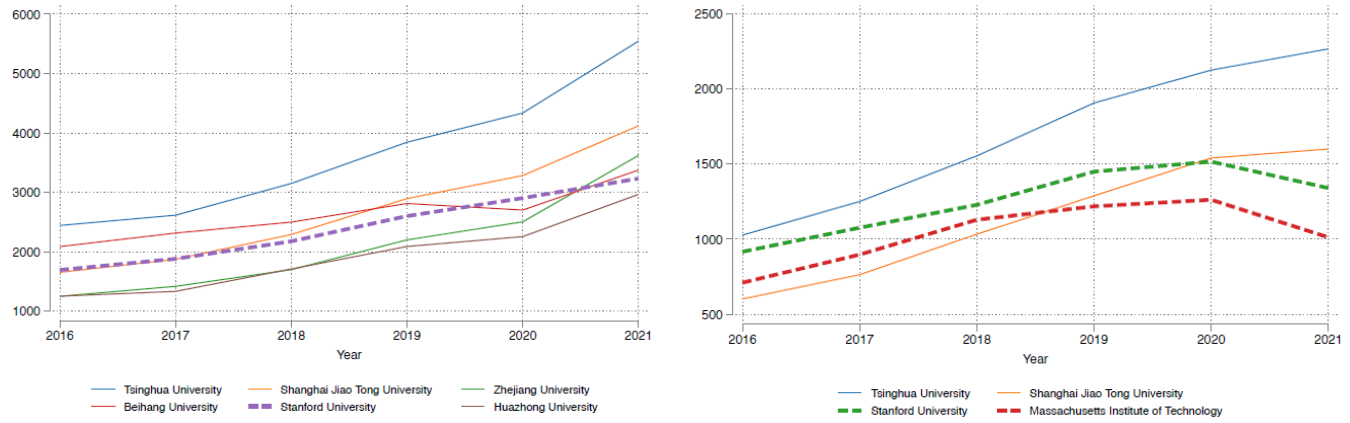
Given the lag in citation counts for the most recent publications, it is likely that the last two years (2021 and 2022) underrepresent the significance of each output's impact. Figure 29 shows that China is leading in high-impact publications since 2019, closely followed by the U.S. and the EU. This is a remarkable achievement given the intensity of the global competition for AI dominance. Even more surprising is the place of the UK and India, who take up the fourth and fifth place in high-impact research outputs. Germany follows in sixth place globally, with Canada, Italy, France, and Japan completing the Top-10 list.

Are U.S. Universities Leading the AI Race?

To continue our analysis, we look into specific universities that focus on AI research and strive to lead in the global competition. We first look at the total number of papers produced and then focus on the high-impact ones. Aligned with the country-level findings and using data from the OECD.AI database, we find that the global leader in AI outputs is Tsinghua University followed by Shanghai Jiao Tong University, Zhejiang University, and Beihang University in China, and Stanford University in the U.S. As with the top four spots, the sixth spot goes to China's Huazhong University competing the Top-6 in the world.

The high-impact research results are not as rosy for Chinese institutions. In Figure 30, we see that Tsinghua University remains the global leader in this domain, closely followed by Stanford and MIT, which have recently been surpassed (with the caveat of undercounted citations for 2021) by Shanghai Jiao Tong University. This finding clears the picture for global dominance in AI with the U.S. and China taking the top spots and their leading universities challenging each other for pole positions. There is, however, a lot more diversity within these metrics, as we depart from total and high-impact counts and look for the research domains by country.

Figure 30. University Research in Total and High-Impact Outputs



Source: OECD Open AI

AI Dominance in Research Sub-Domains

Robotics, Natural Language Processing (NLP), and Computer Vision

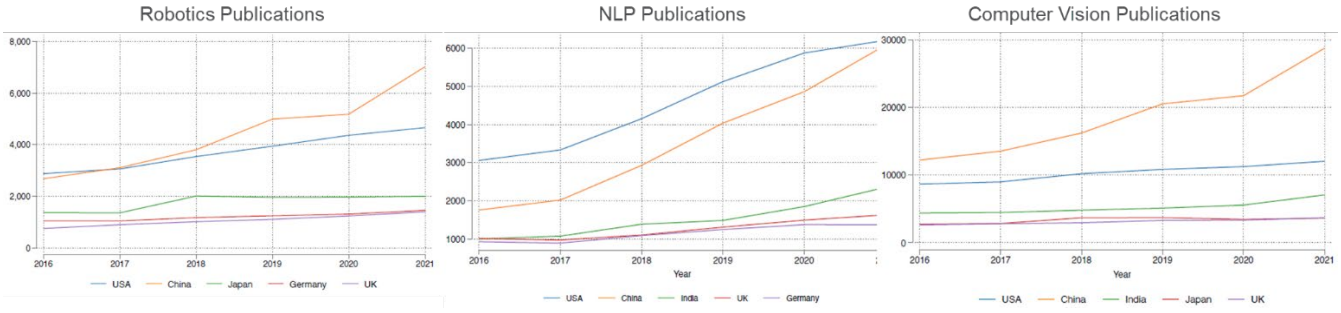
Although foundation models have emerged in the recent past and are likely to be used for Generative AI applications, each domain of AI research requires in-depth expertise on scientific methods and expert applications.

Looking at the Robotics research outputs over the past five years, we find that China has specialized further, surpassing the U.S. by a large margin. The focus on manufacturing in the Chinese economy, as China is currently the top global exporter, is linked to this shift, but it is also likely that reshoring efforts from the U.S. might reverse this trend. Japan comes third in Robotics research, followed by Germany and the UK.

The race for NLP research has been dominated by the U.S. over the past five years, although the gap between the U.S. and China is closing rapidly. NLP applications include chatbots and virtual assistants, sentiment analysis, text extraction, summarization and classification, machine translation, auto-correction applications, among others. Advances in NLP by the Big Tech firms in the U.S. along with U.S. universities, have allowed the country to maintain the top spot. Notably, India does particularly well in NLP research followed by the UK and Germany.

Computer vision is the area of AI that has produced the most publications among the three subdomains over the past five years. Computer vision was a front-runner of AI research, in terms of human-level benchmarks, long before NLP kicked in with Transformers revolutionizing the filed. Computer vision applications are used in self-driving cars for object and pedestrian detection, traffic-flow analysis, and road condition monitoring. It is also used in healthcare in CT, X-Ray and MRI analysis and the detection of various diseases. At the same time, computer vision can be used to identify and track the physical movements of individuals — a practice that raises civil liberty concerns. In this research area, the lead is clearly taken by China, and it is expanding since 2016. India is strengthening and approaching the U.S. output levels, while Japan and the UK round out the top five spots.

Figure 31. Global Race for Robotic, NLP, and Computer Vision Research



Source: OECD Open AI

AI Research Collaborations: Openness is a Strength

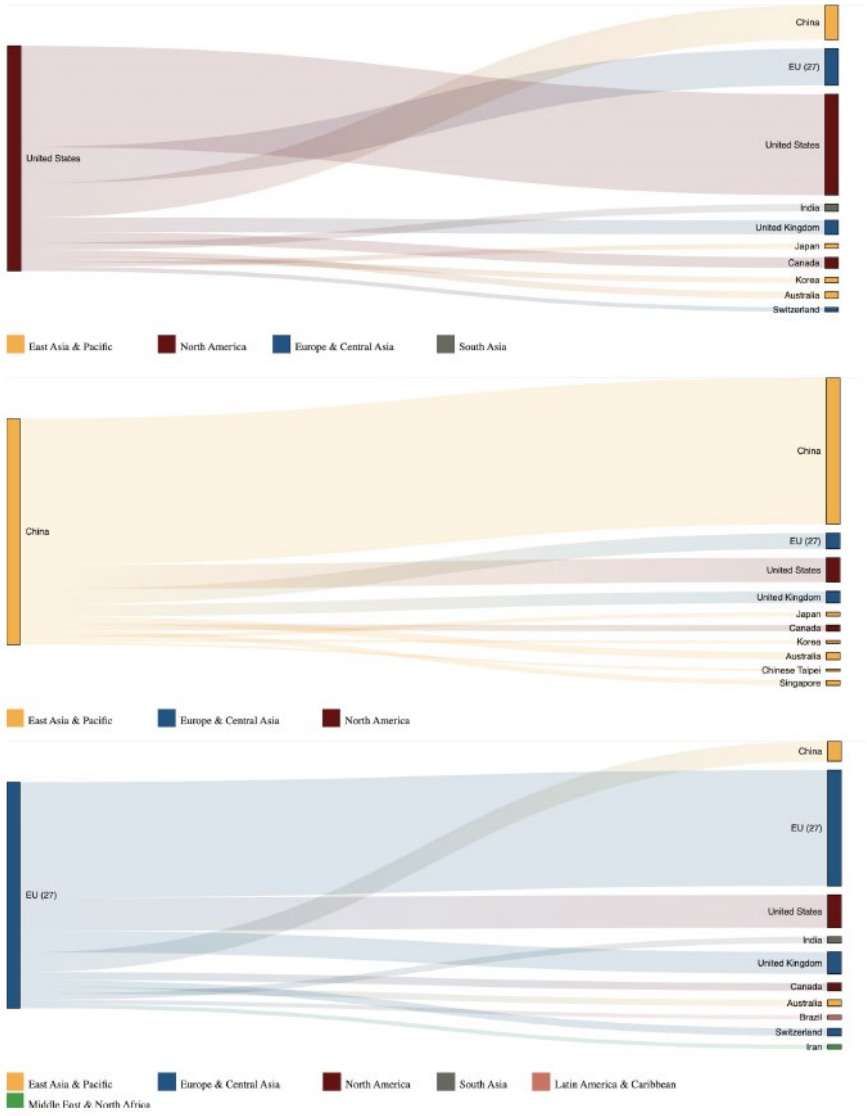
Research progresses through collaboration, specific goals, human talent, and adequate funding. The global leaders in AI have dedicated significant investments over the past decade aiming to improve their outputs across each one of these dimensions.

There are two distinct ways that AI leaders go after their dominance. From a market perspective, the total funding received by all AI-related U.S. firms sums up to $460 billion since 2012, with almost $100 billion in the first half of 2023. In the same time period, Chinese counterparts only received $220 billion since 2012 and less than one-fifth ($18 billion) of U.S. funding in the first half of 2023. In terms of direct government support, however, the situation is rather different. The Chinese government has provided more than $32.5 billion over the recent past to support its local AI industry, compared to just $12.2 billion from the U.S. government. This drastic difference can also be observed in the research collaboration patterns across the main research leaders. U.S. institutions collaborate predominantly with other domestic institutions in their home country, but they also produce research outputs with their international peers.

In China, domestic outputs in 2022 represented 55% of the total while their international collaboration outputs with the U.S. rose up to 9.1% and with the EU at 5.5%. For the EU, domestic outputs in 2022 represented 41.4% of the total while their international collaboration outputs with the U.S. rose up to 11.7% and with China at 7%. By far the most open research community is found in the U.S. with domestic outputs 32% of the total while their international collaboration outputs with the EU rose up to 11.8% and with China at 10.9%. (Figure 19.)

The repercussions of research in-breeding appear to correlate with the outcomes of each of these countries in the AI markets. Looking at the NLP field since 2020, more than 60 large language models have been developed with 39 of them originating from the U.S. (many of which have been developed by the Big Tech firms), 10 from China (again from the leading firms) and 3 from the EU (including the UK and Germany).

Figure 32. Research Collaborations Between Leading Countries



Source: OECD Open AI

Patent Trends in AI

Helen H Krause
Head of Data Science Insights
Citi Global Data Insights

Yehuda Dayan
Head of Data Science Insights
Citi Global Data Insights

Brian Yeung
Head of Data Science Insights
Citi Global Data Insights

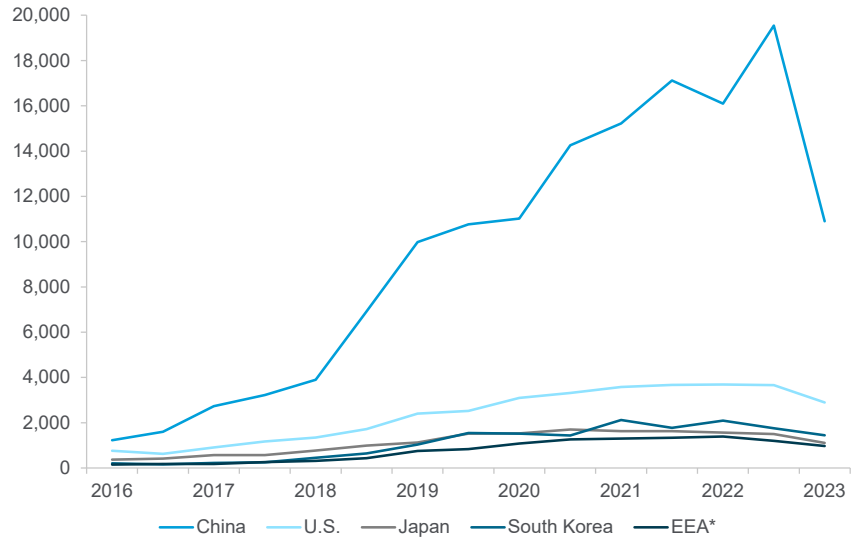
One way to investigate the global trends and growth in AI is from the perspective of investment in technological innovation. To do this, we take a bottoms-up approach, where we record the number of AI-related patent applications over time and across countries. We also investigate the impact of filtering on the quality of the individual patent applications, and compare three applicant groups: public companies, private companies, and educational institutions.

Our patent data analysis is based on a feed from Quant-IP, which provides historic patent data scraped from over 80 patent offices. Quant IP adheres to the International Patent Classification (IPC) code of each patent they store. We have made use of a set of IPC codes in our analysis to identify patents linked to AI alongside a set of terms searched in the title and abstract of an application. The curated concepts can be found in the appendix. For each patent, Quant IP attaches a proprietary patent quality rating and score derived from models of grant robustness, market potential, and citation network size.

AI Technological Innovation: Continued Growth with Large Quality Impact on Global View

The ongoing growth or technological investment in AI-related innovation continues, with China dominating the landscape with over 30,000 applications over the last 12 months, which in absolute terms is more than all other countries combined.

Figure 33. General AI Patent Semi-Annual Count by Country or Economic Area



* In this study, we include the UK and Switzerland in the EEA block of countries.
Source: Quant IP, Citi Global Data Insights

As further detailed in Figure 34, the dominance of China in terms of IP production compared to its closest peers is growing. Over the 2016-23 period, China has increased the number of semi-annual patent applications more than seven-fold, while the U.S. and Japan increased by 3.5x and 2x, respectively. When looking further in individual countries, South Korea and especially Sweden, stand out in terms of their growth rates of AI innovation. Saudi Arabia is also noteworthy, although this growth is from a small baseline in 2016.

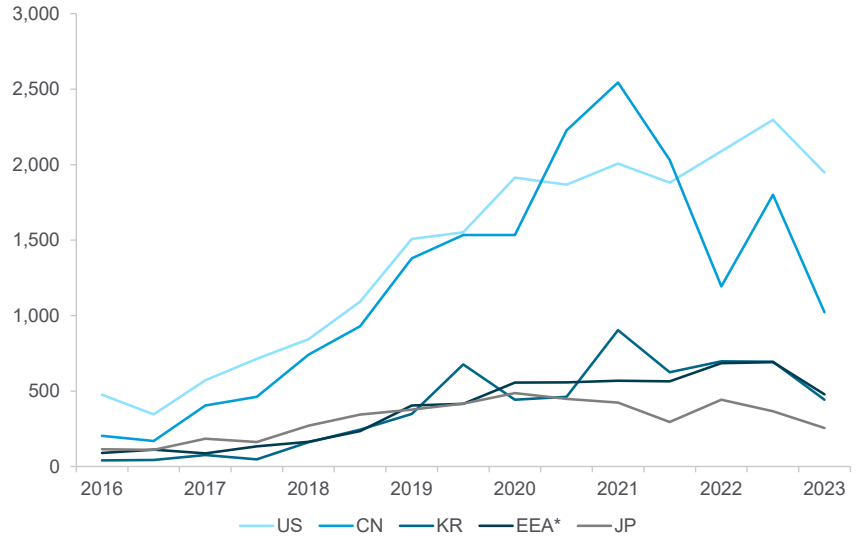
Figure 34. Country-Specific Count and Growth of General AI Patent Applications, 2016 vs. 2023

Country	Total	Growth	Country	Total	Growth
China	144,516	7.36	Canada	1,017	3.46
United States	35,385	3.46	France	955	7.86
Japan	17,012	2.08	Netherlands	854	4.41
South Korea	16,684	8.03	Hong Kong	811	4.73
EEA	11,667	5.99	Sweden	683	24.82
Germany	5,473	6.75	Finland	604	5.70
Taiwan	2,820	5.34	India	480	11.54
Switzerland	1,127	8.94	Israel	434	3.66
United Kingdom	1,049	1.21	Saudi Arabia	334	38.25

Source: Quant IP, Citi Global Data Insights

A cross-country comparison of the total number of patents is, however, misleading as not all patents are equal in terms of their quality, broader impact, and economic value. In Figure 35 below, we repeat our trend analysis, this time filtering on the top 25% of AI-related patents ranked on their quality index. Filtering on quality removes more than 80% of the Chinese patents, compared to roughly 40% of the U.S. patents. This adjustment creates a more nuanced and crowded country landscape where the U.S. and China are following a similar trajectory.

Figure 35. General AI Patent Semi-Annual Count Filtered on Top Quartile Quality



* In this study, we include the UK and Switzerland in the EEA block of countries.
Source: Quant IP, Citi Global Data Insights

Drilling down into the country level and calculating the overall growth over the period, South Korea is the standout performer, with a 10-fold increase in high quality patents over the period. The high-quality production rate of the European Economic Area (EEA), driven by Germany, Switzerland, and France, is also noteworthy.

Figure 36. Country-Specific Count and Growth of General AI Patent Applications, 2016 vs. 2023 Filtered on Top Quartile Quality

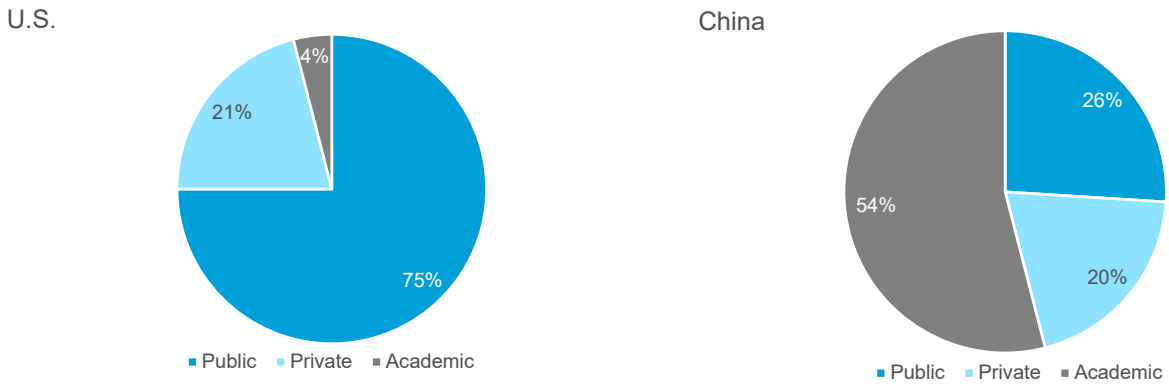
Country	Growth	Total	Country	Growth	Total
USA	3.5	21,107	Switzerland	6.7	756
China	4.2	18,178	Taiwan	3.4	656
South Korea	10.5	5,909	Canada	4.1	510
EEA*	5.4	5,742	France	7.8	492
Japan	1.6	4,698	United Kingdom	1.0	476
Germany	5.6	2,485	Netherlands	2.6	442

Source: Quant IP, Citi Global Data Insights

AI Technological Innovation: Public Companies Drive U.S. Growth, Universities Drive China's

The three main groups of entities that apply for IP production are (1) public companies, (2) private companies, and (3) academic institutions. When comparing China to the U.S., in terms of distribution of high-quality patent applications across these three groups, it is clear the underlying drivers of innovation differ between the two countries.

Figure 37. Country-Specific Count and Growth of General AI Patent Applications, 2016 vs. 2023, Filtered on Top Quartile Quality



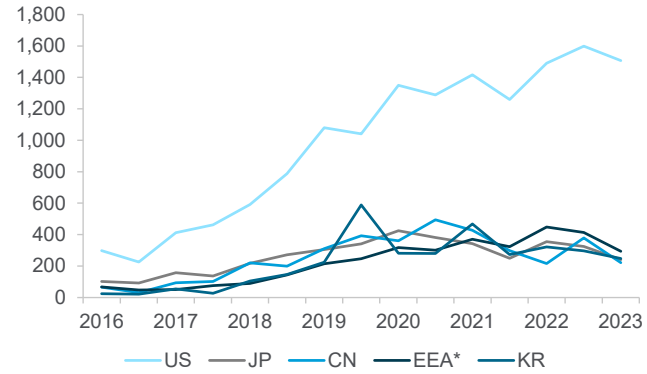
Source: Quant IP, Citi Global Data Insights

Figure 37 shows that U.S. public companies account for 75% of applications while private companies account for just 21%. Together, they account for more than 95% of the entire innovation pipeline. On the other hand, in China, the main engine of innovation is academic institutions, accounting for 54% of the innovation pipeline. Private and public commercial entities split the remaining almost equally.

Given the different drivers, it is instructive to look separately at these three entity groups. Similar to our earlier analysis, we continue with filtering on the high-quality patent universe, using the same top 25% filter.

In Figure 38 and Figure 39 below, the dominance of U.S. public companies is clear, with the total number of U.S. patents roughly equal to all the patents of its largest peers combined for the same period. It is also interesting to note the low growth of Japanese public company patents over this period, as well as the high growth of South Korean and EEA public company patents.

Figure 38. General AI Patent Count of Public Companies Filtered on Top Quartile Quality



* In this study, we include the UK and Switzerland in the EEA block of countries.
Source: Quant IP, Citi Global Data Insights

Figure 39. Country-Specific Public Company Counts and Growth of General AI Patent Applications, 2016 vs. 2023

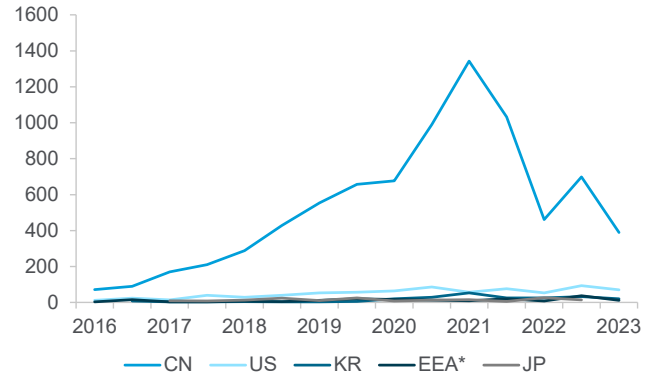
Country	Total	Growth
USA	14,807	3.91
Japan	3,938	1.62
China	3,811	3.28
EEA	3,401	6.00
South Korea	3,359	7.67

Source: Quant IP, Citi Global Data Insights

Not shown in Figure 39 are the countries driving the growth in EEA, notably Germany, with a five-fold increase, and Sweden with a 57-fold increase. India and Saudi Arabia also showed rapid growth, albeit from a small baseline.

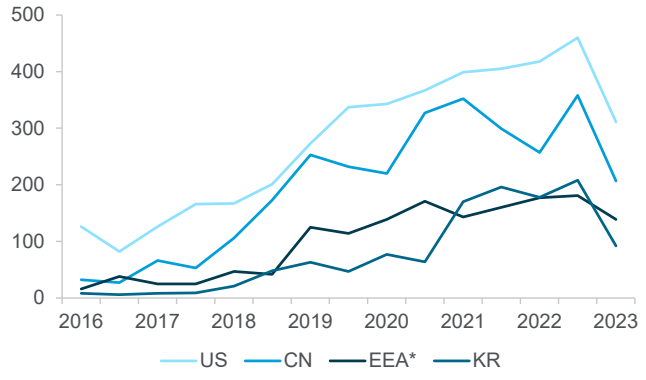
Moving to the second applicant group, the dominance of Chinese academic institutions in patent applications is significant. Figure 40 shows that at its peak in 2021, China applied for ten times more top-quality applications than all other countries combined. This is also reflected in individual institutions where no non-Chinese universities reach the top 20 institutions. Still, the rate of growth for Chinese universities has been slowing significantly over the past two years.

Figure 40. General AI Patent Count of Academic Institutions Filtered on Top Quartile Quality



* In this study, we include the UK and Switzerland in the EEA block of countries.
Source: Quant IP, Citi Global Data Insights

Figure 41. General AI Patent Count of Private Companies Filtered on Top Quartile Quality



* In this study, we include the UK and Switzerland in the EEA block of countries.
Source: Quant IP, Citi Global Data Insights

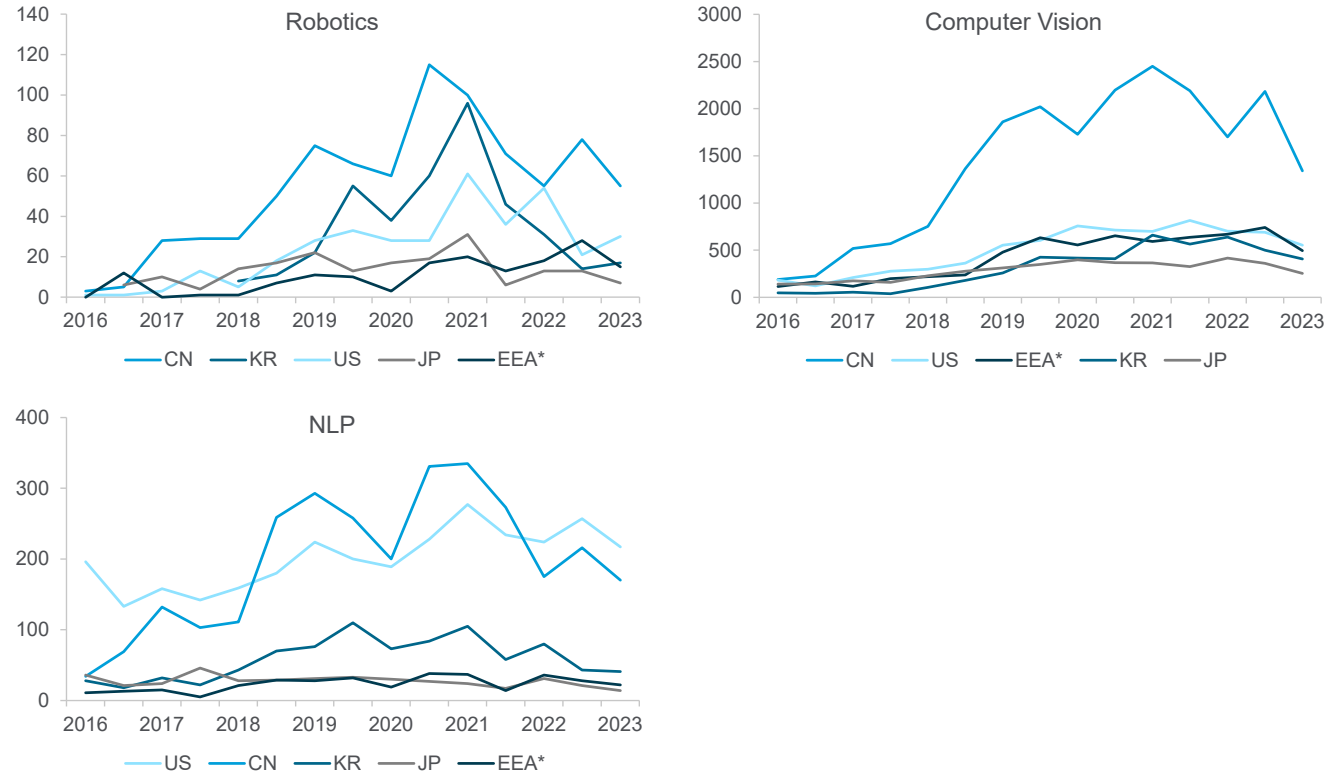
Chinese and U.S. private companies lead the pack, however, while China's growth rate has stalled over the last two years, The U.S. and EEA are growing steadily. The latter's growth is driven mainly by Germany, which over the entire period, accounts for over 50% of the private company applications of the EEA block, but over 75% in the past two years. It is also notable the jump South Korea had in private company patent production in 2020, leading to the country being at parity with the EEA block in absolute numbers.

AI Technological Innovation: Sub Themes

To compliment the academic publication analysis, we ran an analogous investigation on three AI sub-themes: (1) robotics, (2) computer vision, and (3) natural language processing, in terms of patent applications. As before, we chose to control our comparison on patent quality and filter out the bottom 50% of applicants globally. We display the three subthemes in Figure 42.

China dominates the AI computer vision race and is clearly investing heavily in the space. We note that 65% of patents for this sub-theme are from academic institutions. In Robotics, China and South Korea lead, while in the NLP space, the U.S., although being overtaken for three years by China, still leads in innovation applications in this field. The lead is significantly bigger when looking at the top quartile of patents in terms of quality.

Figure 42. Country-Specific Count and Growth of General AI Patent Application, 2016 vs. 2023, Filtered on Top Quartile Quality



* In this study, we include the UK and Switzerland in the EEA block of countries.
Source: Quant IP, Citi Global Data Insights

Glossary of Key Terms

- **Algorithm:** A set of instructions that tells a computer what to do to solve a problem or make a decision.
- **Artificial intelligence:** Computer systems that are able to have some form of intelligence to be able to replicate human expertise.
- **Artificial neural networks (ANN):** Method of developing artificial intelligence by modeling the brain, where a large number (often billions) of nodes are connected together.
- **Chatbot:** AI-powered software that is designed to engage in human-like conversations via text or voice interfaces. It leverages NLP and machine learning techniques.
- **Data bias:** Results that are biased or inaccurate based on the training data, particularly if it is built on the gender, racial, and myriad other biases of the internet and society more broadly.
- **Deep learning:** A subfield of machine learning that employs deep neural networks to learn from large amounts of data and make decisions without being explicitly programmed.
- **Embeddings:** A way to encode text as numbers, which also captures the semantics (meaning) of the text that is encoded. As a result, two words with similar embeddings have similar meaning.
- **Foundational model:** Models that are trained on a broad dataset via training cycles. This process can be expensive and time-consuming.
- **Generative AI:** Artificial Intelligence systems that generate or create new content, such as images, text, or audio, that closely resembles human-generated content.
- **Generative adversarial networks (GANs):** A class of deep learning models consisting of two components, a generator, and a discriminator. The generator generates new content, while the discriminator tries to distinguish between real and generated content.
- **Gradient descent:** An optimization algorithm for machine learning models and neural networks in order to reduce the loss function.
- **Grounding:** A process of anchoring AI systems in real-world experiences, knowledge, or data to improve the AI's understanding of the world, so it can effectively interpret and respond to user inputs, queries, and tasks. This makes AI systems become more context-aware, allowing them to provide better, more relatable, and relevant responses or actions.
- **Hallucination:** This refers to high-confidence responses that are fictional and not grounded in the training data. While this may be advantageous in some creative applications such as art generation, it can be a significant challenge in the broader applications for Generative AI, particularly in areas such as copywriting or code generation.

- **Inference:** The process of applying a trained machine learning model to new, unseen data to generate predictions, classifications, or other desired outputs, for example deriving what an object in a picture is or answering a question based on its knowledge.
- **Large language models (LLMs):** AI models that generate or predict sequences of words or sentences based on a given input. These are used in various Generative AI applications including chatbots and text generation.
- **Loss function:** The difference between the predicted output of a machine learning model and the actual output. The aim is the minimize the loss function using optimization algorithms, e.g., gradient descent.
- **Machine learning:** A subset of AI that enables computers to learn from and make predictions or decisions based on data, rather than being explicitly programmed for specific tasks.
- **Multi-modal:** Models that are able to handle mixed media inputs, such as text and images.
- **Natural language processing (NLP):** A subfield of AI and linguistics that helps computers understand, interpret, and generate human languages, like turning spoken words into text or answering questions in a chatbot.
- **Overfitting:** When a machine learning model learns the training data too well, including noise and random changes, rather than just the main pattern.
- **Parameters:** Numerical values that define a large language model's structure and behavior.
- **Supervised learning:** A machine learning approach where models are trained using labeled data and learn to map inputs to outputs by minimizing the difference between predicted and actual targets.
- **Tokenization/tokens:** The process by which neutral networks are encoded into numbers to feed passages of text into the model and perform the equivalent decode for the output.
- **Unsupervised learning:** A machine learning approach where models are trained using only input data, without corresponding target outputs. The aim is to discover patterns, structures, or relationships within the data on its own, without any prior knowledge or guidance.

If you are visually impaired and would like to speak to a Citi representative regarding the details of the graphics in this document, please call USA 1-888-800-5008 (TTY: 711), from outside the US +1-210-677-3788

IMPORTANT DISCLOSURES

This communication has been prepared by Citigroup Global Markets Inc. and is distributed by or through its locally authorised affiliates (collectively, the "Firm") [E6GYB6412478]. This communication is not intended to constitute "research" as that term is defined by applicable regulations. Unless otherwise indicated, any reference to a research report or research recommendation is not intended to represent the whole report and is not in itself considered a recommendation or research report. The views expressed by each author herein are his/ her personal views and do not necessarily reflect the views of his/ her employer or any affiliated entity or the other authors, may differ from the views of other personnel at such entities, and may change without notice.

You should assume the following: The Firm may be the issuer of, or may trade as principal in, the financial instruments referred to in this communication or other related financial instruments. The author of this communication may have discussed the information contained herein with others within the Firm and the author and such other Firm personnel may have already acted on the basis of this information (including by trading for the Firm's proprietary accounts or communicating the information contained herein to other customers of the Firm). The Firm performs or seeks to perform investment banking and other services for the issuer of any such financial instruments. The Firm, the Firm's personnel (including those with whom the author may have consulted in the preparation of this communication), and other customers of the Firm may be long or short the financial instruments referred to herein, may have acquired such positions at prices and market conditions that are no longer available, and may have interests different or adverse to your interests.

This communication is provided for information and discussion purposes only. It does not constitute an offer or solicitation to purchase or sell any financial instruments. The information contained in this communication is based on generally available information and, although obtained from sources believed by the Firm to be reliable, its accuracy and completeness is not guaranteed. Certain personnel or business areas of the Firm may have access to or have acquired material non-public information that may have an impact (positive or negative) on the information contained herein, but that is not available to or known by the author of this communication.

The Firm shall have no liability to the user or to third parties, for the quality, accuracy, timeliness, continued availability or completeness of the data nor for any special, direct, indirect, incidental or consequential loss or damage which may be sustained because of the use of the information in this communication or otherwise arising in connection with this communication, provided that this exclusion of liability shall not exclude or limit any liability under any law or regulation applicable to the Firm that may not be excluded or restricted.

The provision of information is not based on your individual circumstances and should not be relied upon as an assessment of suitability for you of a particular product or transaction. Even if we possess information as to your objectives in relation to any transaction, series of transactions or trading strategy, this will not be deemed sufficient for any assessment of suitability for you of any transaction, series of transactions or trading strategy.

The Firm is not acting as your advisor, fiduciary or agent and is not managing your account. The information herein does not constitute investment advice and the Firm makes no recommendation as to the suitability of any of the products or transactions mentioned. Any trading or investment decisions you take are in reliance on your own analysis and judgment and/or that of your advisors and not in reliance on us. Therefore, prior to entering into any transaction, you should determine, without reliance on the Firm, the economic risks or merits, as well as the legal, tax and accounting characteristics and consequences of the transaction and that you are able to assume these risks.

Financial instruments denominated in a foreign currency are subject to exchange rate fluctuations, which may have an adverse effect on the price or value of an investment in such products. Investments in financial instruments carry significant risk, including the possible loss of the principal amount invested. Investors should obtain advice from their own tax, financial, legal and other advisors, and only make investment decisions on the basis of the investor's own objectives, experience and resources.

This communication is not intended to forecast or predict future events. Past performance is not a guarantee or indication of future results. Any prices provided herein (other than those that are identified as being historical) are indicative only and do not represent firm quotes as to either price or size. You should contact your local representative directly if you are interested in buying or selling any financial instrument, or pursuing any trading strategy, mentioned herein. No liability is accepted by the Firm for any loss (whether direct, indirect or consequential) that may arise from any use of the information contained herein or derived herefrom.

Although the Firm is affiliated with Citibank, N.A. (together with its subsidiaries and branches worldwide, "Citibank"), you should be aware that none of the other financial instruments mentioned in this communication (unless expressly stated otherwise) are (i) insured by the Federal Deposit Insurance Corporation or any other governmental authority, or (ii) deposits or other obligations of, or guaranteed by, Citibank or any other insured depository institution. This communication contains data compilations, writings and information that are proprietary to the Firm and protected under copyright and other intellectual property laws, and may not be redistributed or otherwise transmitted by you to any other person for any purpose.

IRS Circular 230 Disclosure: Citi and its employees are not in the business of providing, and do not provide, tax or legal advice to any taxpayer outside of Citi. Any statements in this Communication to tax matters were not intended or written to be used, and cannot be used or relied upon, by any taxpayer for the purpose of avoiding tax penalties. Any such taxpayer should seek advice based on the taxpayer's particular circumstances from an independent tax advisor.