Generative AI at the Crossroads: Light Bulb, Dynamo, or Microscope?*

Martin Neil Baily David M. Byrne
Aidan T. Kane Paul E. Soto

September 11, 2025

Abstract

With the advent of generative AI (genAI), the potential scope of artificial intelligence has increased dramatically, but the future effect of genAI on productivity remains uncertain. The effect of the technology on the innovation process is a crucial open question. Some inventions, such as the light bulb, temporarily raise productivity growth as adoption spreads, but the effect fades when the market is saturated; that is, the level of output per hour is permanently higher but the growth rate is not. In contrast, two types of technologies stand out as having longer-lived effects on productivity growth. First, there are technologies known as general-purpose technologies (GPTs). GPTs (1) are widely adopted, (2) spur abundant knock-on innovations (new goods and services, process efficiencies, and business reorganization), and (3) show continual improvement, refreshing this innovation cycle; the electric dynamo is an example. Second, there are inventions of methods of invention (IMIs). IMIs increase the efficiency of the research and development process via improvements to observation, analysis, communication, or organization; the compound microscope is an example. We show that GenAI has the characteristics of both a GPT and an IMI—an encouraging sign that genAI will raise the *level* of productivity. Even so, genAI's contribution to productivity *growth* will depend on the speed with which that level is attained and, historically, integrating revolutionary technologies into the economy is a protracted process.

*Authors are listed in alphabetical order, not in order of relative contribution. Baily and Kane are at the Brookings Institution (mbaily@brookings.edu and akane@brookings.edu). Byrne and Soto are at the Federal Reserve Board of Governors (david.m.byrne@frb.gov and paul.e.soto@frb.gov). The views expressed here are not represented to be the views of the staff or trustees of The Brookings Institution nor of the Federal Reserve. The authors are grateful to Michael Chui, Leland Crane, Avi Goldfarb, Bob Gordon, Shane Greenstein, Anton Korinek, James Manyika, Sid Srinivasan, Scott Stern, and Bill Whyman for helpful conversations.

1 Introduction

In late 2022, OpenAI grabbed the world's attention with ChatGPT, the first of several recently released "generative AI" (genAI) programs that use a computer model of human discourse to respond to natural-language questions. The scope of AI has expanded dramatically with the advent of genAI, including to tasks previously seen as quintessentially human, such as competition-level mathematics (fig. 1 on the following page). Indeed, more and more challenging benchmark tests have been needed to assess progress as genAI has matched human performance on one task after another.[1] In an encouraging sign, field test evidence of productivity improvements from genAI in practical applications has also emerged, including for writing, computer programming, and responding to call center inquiries (table 1 on page 4).[2] It remains to be seen whether widely-used cost-effective business applications will follow from these successful field tests. Although some companies do credit genAI with improvement to their bottom line, McKinsey (2025b) reports that more than 80% of genAI-using firms "aren't seeing a tangible impact on enterprise-level [earnings before interest and taxes] from their use of genAI."[3]

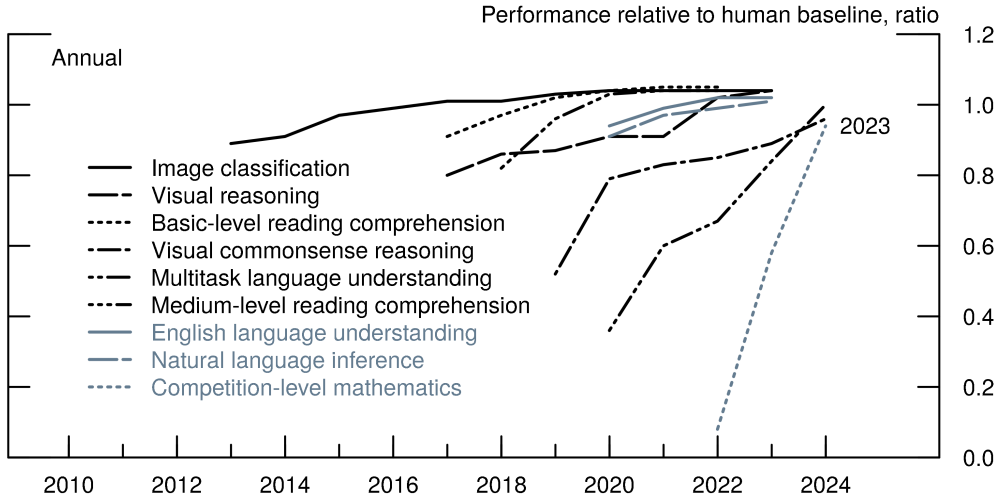
Web searches for AI and downloads of the ChatGPT app have soared, sparking intense competition for leadership in genAI (fig. 2 on page 5), and the computational intensity of training genAI models and processing user requests has led to a massive increase in data center construction and spending on AI-related semiconductor chips (fig. 3 on page 6). While optimists see

1. For more on the record of AI benchmark performance, see Maslej et al. (2024). The external validity of such benchmarks—that is, how much they tell us about performance on practical tasks seemingly related to the tests—is a matter of some debate (Liao et al. 2021). New benchmarks, such as the ARC-AGI and the Graduate-Level Google-Proof Q&A (GPQA) benchmarks, have been introduced to better evaluate advanced capabilities. Recent models, especially "reasoning" models such as o3 from OpenAI and others discussed in section 3.3.1 on page 21, have performed strongly on these more demanding tasks. Haupt and Brynjolfsson (2025) argue that benchmarks that measure how well AI and humans can jointly perform tasks are needed to shed light on practical AI use.

2. Brynjolfsson, Li, and Raymond (2025) find that call center operators became 14% more productive when they used the technology; Peng et al. (2023) find that access to GitHub Copilot enabled programmers to complete tasks almost 56% faster; and Noy and Zhang (2023) find that writers who used ChatGPT worked more quickly and produced higher-quality outputs.

3. McKinsey (2025b) also reports that a large share of their survey respondents—primarily large corporations—credit AI with cost reductions in some business functions.

Figure 1: AI Benchmark Performance



Note: The "human baseline" concept used varies by task. For more challenging tasks, the baseline tends to reflect expert-level performance.
Source: Reproduced with permission from the 2024 AI Index Report, Stanford Institute for Human-centered Artificial Intelligence.

potential for genAI to spur an information technology (IT)-fueled productivity boom comparable to the late 1990s and early 2000s, more downbeat observers see claims about its capabilities as overstated and highlight potential headwinds, such as regulations to guard against unintended harms, political pushback as AI affects the jobs of human workers, and the colossal energy requirements for training and running genAI systems.[4] It is too early to adopt either view with confidence. Whether practical genAI applications will be consequential enough to raise aggregate productivity growth remains to be seen. GenAI may be no more important than previous innovations in IT already reflected in the historical trend, including its predecessors in the field of AI.[5]

With only "green shoots" of quantitative evidence in hand that genAI will raise productivity, we frame the prospective effect of genAI in qualita-

4. For an optimistic view, see Kurzweil (2024). For a more cautious perspective see Narayanan and Kapoor (2024).

5. Moreover, the present era of modest productivity growth in the midst of mature machine learning, cloud computing, and smartphones should temper expectations for another IT boom.

Table 1: Selected GenAI Productivity Field Studies

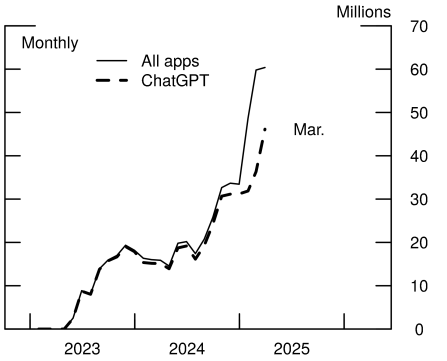
Study	Task	Results
Noy and Zhang (2023)	Writing	ChatGPT speeds, improves writing. Writers shift from drafting to idea generation and editing.
Brynjolfsson et al., (2023)	Customer service	More issues resolved using conversational assistant.
Dell'Aqua et al. (2023)	Various	GPT-4 increases task completion, speed, and quality.
Peng et al. (2023)	Coding	Using GitHub Copilot, programmers complete tasks faster
Cui et al. (2024)	Coding	GitHub Copilot raises task completion.

tive terms: We ask what class of innovation it may be. Some labor-saving innovations, such as the light bulb, temporarily raise productivity growth as adoption spreads, but the effect fades when the market is saturated; that is, the level of output per hour is permanently higher but the growth rate is not. Other widely used technologies, such as the electric dynamo, spur knock-on innovations—new products, process improvements, and business reorganization—and refresh this adoption cycle through ongoing improvement in the core technology (David 1990). The boost to productivity growth from these general-purpose technologies (GPTs) may last longer. Yet other inventions, such as the compound microscope, increase the efficiency of the research and development process; these "inventions of methods of invention" (IMIs) yield a sustained increase in productivity growth by lowering the cost of research and development.

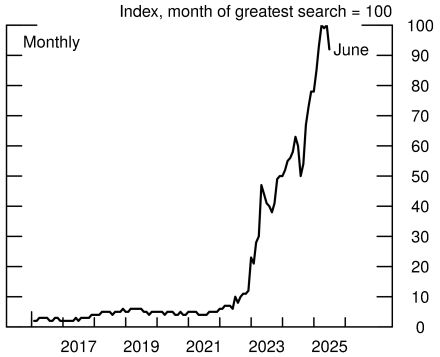
We first define "generative AI," then consider the evidence that genAI is a GPT, reviewing indicators for the scope of diffusion, the extent of knock-on innovations, and signs of ongoing progress in the core technology. To assess its status as an IMI, we discuss evidence that it increases the efficiency of observation, analysis, communication, and organization in research and review several indicators (patents, earnings calls, and query topics). For both questions, we reference case studies for the financial, health care, and information sectors and for the electricity generation industry (Baily and Kane 2025a, 2025b; Kane and Baily 2025a, 2025b). We conclude there is

Figure 2: Indicators of Interest in GenAI

(a) GenAI Mobile App Downloads



(b) Web Searches for AI



Note: Apps are ChatGPT, Claude, DeepSeek, and Perplexity. Includes Android and iOS. Android download information not available for China. Does not account for access via application program interface. Web searches include related terms in Google's "AI" topic.
Source: appfigures; Google Trends.

substantial evidence that genAI is both a GPT and an IMI, an encouraging sign its adoption will lead to higher productivity in the future.

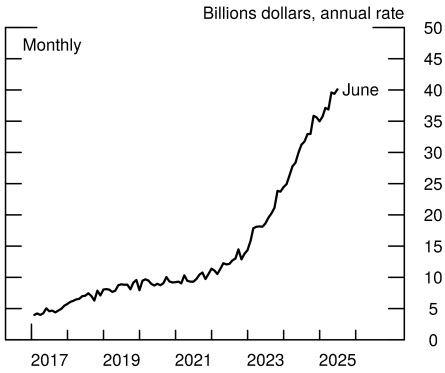
There is a substantial literature on the question of whether machine learning, which preceded genAI, may be a GPT (Cockburn, Henderson, and Stern 2019; Trajtenberg 2018; Bresnahan 2019; Goldfarb, Taska, and Teodoridis 2023; Bresnahan 2024), and Cockburn, Henderson, and Stern (2019) discuss the possibility that machine learning is an IMI. There is little work focused specifically on genAI. Eloundou et al. (2024), a prominent exception, consider the prospects for genAI to be a GPT; relative to that work, we draw on a broader set of indicators and consider evidence for whether genAI is both a GPT and an IMI. Our focus on characteristics of the technology itself and its integration into business processes complements the large literature on the labor market impact of genAI.[6] Our qualitative assessment of genAI also serves as a primer to inform discussion of AI in the context other literatures, such as creative destruction and endogenous growth (Akcigit and Van Reenen 2023).[7]

6. On labor market effects of AI, see Acemoglu and Restrepo (2020), Agrawal, J. Gans, and Goldfarb (2023), Brynjolfsson, Li, and Raymond (2025), and Eloundou et al. (2024), among others.

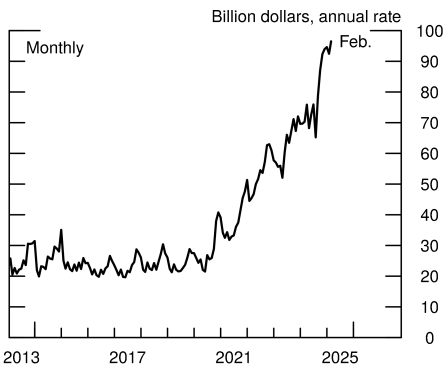
7. Our focus is primarily on the U.S. economy; Filippucci et al. (2024) take a global

Figure 3: Indicators of U.S. AI-Related Investment

(a) Data Center Construction



(b) Application-Specific Chips



Note: Nominal value of construction put in place. Application-specific chips include GPUs, TPUs, and ASICs for other applications.
Source: Census Bureau; Semiconductor Industry Association.

2 What is Generative AI?

"Artificial intelligence" (AI) is an umbrella term encompassing a variety of algorithms deployed on computers to mimic human thought, communication, and choices, such as machine learning, computer vision, and generative models. (See Appendix A for a discussion of several influential definitions of AI.) AI systems achieve these objectives by constructing and calibrating mathematical models of complex patterns found in training data. The most widely known implementations of genAI use a computer model of the human discourse found on the internet to respond to natural-language prompts (questions or directives), though genAI systems take other forms as well as we discuss below.

We narrow our focus to genAI for several reasons. First, the broad and varied use of "AI" makes a coherent discussion of its effects on productivity difficult. Second, the productivity impact of genAI is largely in the future, in contrast to AI types already in use for which the effects on productivity are, in principle, an empirical question. Third, the human-like behavior of generative models has made concerns about the disruptive effects of AI particularly salient.

perspective and address many of the same issues as our paper.

Although systems which respond to prompts with natural language text were a part of the AI field from its inception, early models were not grounded in a model of human language. (For a short history of the field, see Appendix B.) For example, rudimentary chatbots, such as ELIZA the psychotherapist, text autocompletion, and "expert systems," such as MYCIN for infection diagnosis, appeared in the 1960s and 1970s. These systems had matured by the 2010s, when sophisticated voice-driven chatbots like Alexa and Siri had been introduced, news outlets were auto-generating routine stories, and IBM's Watson, famous for beating humans at *Jeopardy* in 2011, was repurposed to provide advice on a host of topics. These were symbolic, rules-based systems, albeit with an element of randomness in their output.

Richer generative models appeared after the development of large language models (LLMs). LLMs, which represent the meanings of words and their relationships by locations in a high-dimensional space, emerged in the 2010s. Word2Vec, most notably, encoded words as vectors of numerical values, and while the values do not correspond to specific, interpretable characteristics, they capture semantic relationships in an abstract space (Mikolov et al. 2013). For example, while humans would represent a dress by its color, size, and hem length, say, the characteristics chosen by Word2Vec would not have a readily apparent interpretation (Bajari and Chernozhukov 2018). Aided by advances in computational power and big data processing techniques, genAI developers pushed the field forward and achieved a breakthrough with the Transformer architecture in 2017 (Vaswani et al. 2017). (The technical features of the Transformer are discussed in a box below.) This model allows for a richer representation of word meaning in its encoding by accounting for context. Many GenAI models use LLMs to encode the tokens (words, phrases, or parts of words) in the input in this fashion. After embedding the input text into locations within a high-dimensional space of abstract characteristics, they draw on the information embedded nearby to guide the prediction of the next token, weaving the output into a relevant natural-language response.

Because these models are created as neural networks, which are extremely flexible, they differ from previous models, like IBM's Watson, in that they are not symbolic. That is, they do not have a predetermined logical structure. The effect of this added flexibility and the richness enabled by their massive scale is that the range of possible generated content is more open-ended than earlier systems. For example, earlier symbolic attempts were capable of producing formulaic news stories about a firm's quarterly earnings or the

outcome of a baseball game, but ChatGPT can provide a convincing nuanced response to prompts such as "write a short story about an angst-ridden robot in the style of Edgar Allan Poe."

Importantly, genAI can support a wider array of applications than natural language tools for learning and creativity. Broadly stated, genAI systems produce contextually appropriate artifacts using an open-ended stochastic process that draws from patterns in a dataset. The artifacts may be a variety of things other than text, including computer code, images, music, chemical structures, game environments, mathematical proofs, or dance moves, to name a few. And, while the chat window user interface makes genAI accessible, it is not found in all applications. For example, in generative adversarial networks (GANs), neural networks interact with each other, not humans, to generate output.

Landmark AI Models: The Transformer

The transformer architecture, introduced by Vaswani et al. (2017), was a game changer in AI, particularly as the engine behind genAI models. Its key innovation, the "attention mechanism," steers models to focus selectively on relevant parts of the prompt, enabling more efficient and accurate processing of language. This breakthrough has powered major advancements in natural language understanding, translation, and generation, forming the backbone of today's most advanced genAI systems.

Transformers process input data through a series of layers (steps), each consisting of an attention mechanism followed by a multilayer perceptron (MLP, defined below), proceeding as follows.

First, a representation of the prompt (input text) suitable for analysis by the model is created. Specifically, the prompt is broken into tokens (smaller pieces which may be phrases, words, or parts of words). The tokens are converted into embeddings (numerical vector representations) which encode the semantic and syntactic meaning of each token. Loosely speaking, for each token, the closest of the other tokens, as measured by the distance between their embeddings, are the ones most important to understanding its meaning.

Second, the attention mechanism processes the matrix of token embeddings using three large matrices called the "query," the "key," and

The Transformer (continued)

the "value." For each token in the input, the query is compared to the keys of all tokens to compute attention scores, which are used to form a weighted average of values. This step allows each token's representation to incorporate information from other tokens in the prompt based on their contextual relevance.

Third, the data passes through an MLP, a type of neural network. While the attention mechanism focuses on pairwise interactions between tokens, the MLP applies nonlinear functions (in contrast to the linear attention mechanism) in refining the token representations.

This sequence—of computing the attention mechanism followed by the MLP—is repeated multiple times depending on how many layers are in the model (for example, the Llama-3 model has 32 layers), enabling the model to capture increasingly abstract features of the input text.

The performance gains from scaling of this system through increasing the size of these matrices—along with larger training datasets and improvements in hardware and processing algorithms—underpins the rising ability to handle complex language tasks.

3 Is GenAI a General Purpose Technology?

While the evidence for genAI-driven productivity in specific tasks is intriguing (table 1 on page 4), to look ahead to its future productivity impact, we would like to know (1) if genAI will be widely adopted, (2) the extent of related innovations, and (3) whether genAI will continue to improve. That is, we would like to know if genAI is a general-purpose technology (GPT).[8] Through widespread adoption, downstream complementary innovation, and sustained innovation in the core technology, GPTs have long-lasting effects on productivity. Examples of GPTs are shown in table 2 on the following page.

As described by Lipsey, Carlaw, and Bekar (2005, xvi), "big GPT shocks change almost everything in a society and revitalize the growth process by

8. Note that the "GPT" initialism in the names of OpenAI genAI models stands for "generative pre-trained transformer." We will use "GPT" to mean "general-purpose technology" exclusively in this paper except when referring to OpenAI models.

Table 2: Examples of General Purpose Technologies

Technology	Initial Impact
Domestication of plants	9000–8000 BCE
Writing	3400–3200 BCE
Iron	1200 BCE
Waterwheel	Early medieval period
Three-masted sailing ship	15^{th} century
Printing	15^{th} century
Factory system	Mid 18^{th} century
Steam engine	Late 18^{th} century
Railway	Mid 19^{th} century
Internal combustion engine	Late 19^{th} century
Electricity	Early 20^{th} century
Motor vehicle	Early 20^{th} century
Mass production, continuous process factory	Early 20^{th} century
Lean production	Late 20^{th} century
Computer	Late 20^{th} century
Internet	Late 20^{th} century
Source: Adapted from Lipsey, Carlaw, and Bekar (2005).	

creating an agenda for the creation of new products, new processes, and new organizational forms." For example, new products that followed the development of the (electronic) computer include office productivity software, ATM machines, and the routing equipment that directs traffic around the internet. New processes that followed the development of reliable electricity include the production of hydrogen by electrolysis, salvaging scrap steel using electric arc furnaces, and the fabrication process for semiconductor chips. And, a new organizational form that followed the introduction of the three-masted sailing ship was the joint-stock company, used to finance the voyages of large-scale trading firms.

We briefly describe the three criteria for a technology to be a GPT, then examine in detail the evidence that genAI meets each one.

Diffusion The more widespread the application of a technology, the greater the potential impact on aggregate productivity (Hulten 1978). That said, however widely adopted, the *direct* productivity effect of a single invention is

bounded. Productivity *growth* will be higher during the adoption transition, but will return to its underlying trend when diffusion is complete (Robert M. Solow 1956). To illustrate using the light bulb, once suitable filaments were developed and the inexpensive incandescent light bulb was available, the technology was gradually adopted in workplaces, raising productivity through better visibility and lower risk of accident (Abdou 1997). Lighting is necessary for nearly all human labor, so the potential effect on the level of productivity from the light bulb was noteworthy, but once the light bulb market was saturated, it delivered no further (direct) level effect and the increment to productivity growth present during the transition disappeared.[9]

Knock-on Innovation Technologies that spur further innovation can deliver a longer-lived impetus to productivity growth. The greater persistence of elevated growth is the result of a series of overlapping classical "light bulb" growth effects.[10] The electric dynamo is an example. The dynamo uses electromagnetism to convert mechanical energy produced by a prime mover—a steam engine, say—to electromagnetic energy, which is then conveyed by wires and converted back to mechanical energy by a motor used to drive machinery in another location. Existing systems conveyed mechanical energy directly to machinery through a set of belts. The dynamo/wiring/-motor system is more energy efficient than the belt system except in very simple arrangements, so the simple replacement of existing factory systems yielded productivity gains. In addition, the dynamo enabled a more flexible organization of production (David 1990). The less centralized factory designs adopted by firms in response are a knock-on productivity-enhancing innovation spurred by the dynamo.

Ongoing Core Innovation When a technology continues to improve over time, the new target productivity level—fixed in the simple impulse-response framework of the Robert M. Solow (1956) model—becomes a moving target. Ongoing innovation translates into greater technical performance at a lower cost, a form of productivity gain. Moreover, the price of capital typically

9. The light bulb was not solely a terminal innovation, of course. The surge in demand for electricity represented by light bulbs led to centralized power stations (David 1990).

10. Complementary innovations may increase productivity by raising the effectiveness of the GPT as well, such as raising the operating rate of computers by the invention of cloud computing.

follows the production cost downward spurring greater adoption. Innovation is, in a sense, embedded in the capital (Kaldor 1957). Solid state electronics is an example. Relentless increases in the number of transistors on each semiconductor chip has driven the price of computing lower, making it cost-effective both to embed electronics in a greater variety of devices (like inexpensive toys) and to enhance devices with more and more electronic capability (like smartphones).

3.1 Diffusion

Although comprehensive measures of the diffusion of genAI, specifically, are limited, recent trends in the diffusion of AI, generally, may indicate the underlying influence of genAI. Surveys show AI adoption rising, particularly in large corporations where AI use is concentrated. Even so, a large majority of firms still don't see an application for AI in their business. Analyses of the text of job descriptions suggest that AI can be used for a broad range of workplace tasks, indicating that the potential for diffusion among firms is high.[11] At the same time, the share of job postings mentioning AI skills is modest, indicating that firms are taking a cautious approach to hiring workers to focus on AI use. Meanwhile, from the worker's perspective, AI adoption seems widespread; surveys of individuals document that a large share of workers are already AI users.

Adoption surveys Distilling a single message from the available surveys of AI use is difficult at first glance. The Census Bureau's Business Trends and Outlook Survey (BTOS) finds roughly 9% of firms use AI, while McKinsey reports that 72% of firms do so (fig. 4 on page 14). On closer inspection, these surveys are consistent with one another and reveal important nuances in the state of AI adoption with respect to firm size and business functions.

Combining the results of these surveys points to far higher AI adoption for large firms than small ones.[12] The BTOS is a representative sample of

11. See Acemoglu et al. (2020), Brynjolfsson, Mitchell, and Rock (2018), Felten, Raj, and Seamans (2019), Webb (2019), Eloundou et al. (2024).

12. There may well be significant heterogeneity within small firms on this question. Because new firms, which are typically small, may begin life as digitally native firms, they may adopt AI more easily. The pace of business applications with a high-propensity of turning into businesses with payroll, reported by the Census Bureau, has moved up significantly in the wake of the pandemic. Note that information on new firms will appear

200,000 U.S. firms, only a handful of which are large corporations (Bonney et al. 2024). The McKinsey survey, in contrast, is a convenience sample with heavy representation from large corporations (McKinsey 2024).[13] The U.S. firm size distribution is highly skewed and large corporations have thousands of employees (Kondo, Lewis, and Stella 2023). The threshold for the largest firm-size group in the BTOS is 250 employees, for which BTOS reports 14.9% adoption, and the BTOS reports that for smallest firm-size group, firms with fewer than 5 employees, 9.7% were using AI in June 2025.

These surveys also suggest that AI use may be less prevalent in core business functions than in support functions. Differences in the definition of adoption between the surveys point to this conclusion. The BTOS survey asks, "In the last two weeks, did this business use Artificial Intelligence (AI) in producing goods or services? (Examples of AI: machine learning, natural language processing, virtual agents, voice recognition, etc.)." The McKinsey question is rather less restrictive, asking respondents if they use "AI in at least one business function." That is, the BTOS asks about use in core business functions ("producing goods or services"), while McKinsey includes non-core functions like "marketing and sales," which is the function with greatest AI use in their survey.

Evidence about adoption for genAI specifically is more limited. Only the McKinsey survey asks separately about genAI, finding that use among their (primarily large firm) respondents surged from roughly one-third in 2023 to two-thirds in 2024. Bick, Blandin, and Deming (2024) survey workers (a representative sample of 18-64 year-olds in the United States) and find nearly 40% of respondents used genAI.

As posited in Crane, Green, and Soto (2025), high-level managers may underestimate the extent to which their employees are using AI tools, or AI users may be highly concentrated in large firms (roughly half of the U.S. workforce) or AI-intensive industries. Consistent with the idea that AI users may be concentrated in AI-intensive industries, Sergeyuk et al. (2025) survey programmers and find 84% of them use AI. A recent trend in industry composition is suggestive as well. Decker and Haltiwanger (2024) identify a step

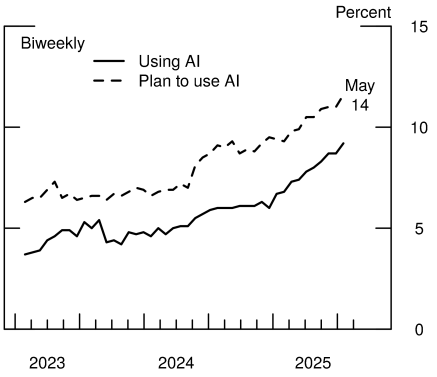
with a delay in the BTOS, for which the sample is drawn from the Census Business Register (Bayard et al. 2018).

13. Although McKinsey states that its survey includes "participants representing the full range of regions, industries, company sizes, functional specialties, and tenures," responses are not weighted by the relative prevalence of their characteristics in the population. We thank Michael Chui for confirming that this is a fair characterization of the survey.

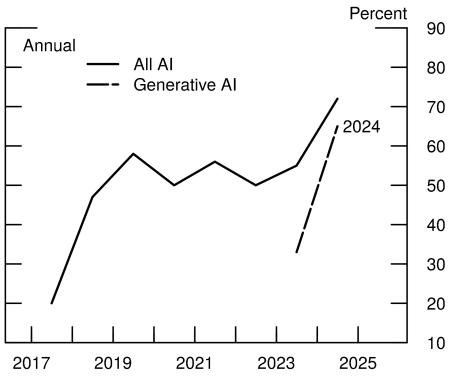
up in the share of business establishments—reported by the Bureau of Labor Statistics in the *Quarterly Census of Employment and Wages*—in industries with a high share of employees in science, technology, engineering, and math fields. We speculate this employee composition may be paired with higher AI use.

Figure 4: AI Use Over Time

(a) Census BTOS (United States)



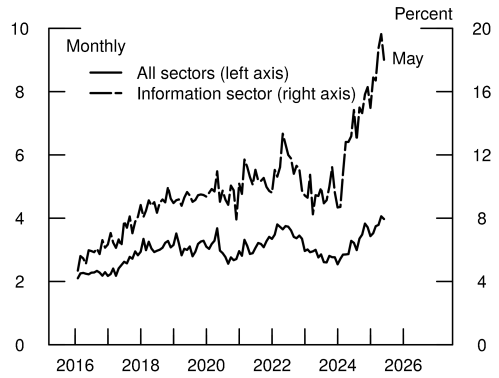
(b) McKinsey (Global)



Note: For BTOS, respondents were asked about AI use in producing goods or services during the past two weeks and anticipated in the next six months. For McKinsey, respondents were asked if they "use AI in at least one business function"
Source: Census Bureau, Business Trends and Outlook Survey; McKinsey, "The State of AI in Early 2024."

Job postings Job posting data from Lightcast (formerly Burning Glass) categorized using terms associated with AI by Acemoglu et al. (2020), extended with terms not prevalent at the time of their analysis (such as "genAI"), show the share of job postings currently related to AI to be roughly 4 percent, using a broad definition of AI including a cluster of related skills (fig. 5 on the following page). Importantly, that share moved up (from no more than 2 percent) around 2017, well before practical genAI applications became prevalent, consistent with earlier forms of AI driving the increase. AI-related job postings have moved up only modestly since then, suggesting that explicit use of genAI-related skills is not crucial to many jobs. In some sectors, most notably the information sector (shown in the graph), the share of job postings referencing AI is substantially higher.

Figure 5: AI-Related Job Postings



Note: Jobs classified using AI-related terms found in job descriptions, as described in Acemoglu et al. (2020). List of AI-related terms updated by Lightcast.
Source: Lightcast.

Case study evidence

The **information sector** has adopted genAI rapidly. Among U.S. programmers, 92% were using AI tools (including genAI) as of June, 2023 (Shani and GitHub Staff 2023). Other occupations within the information sector use genAI in their work as well. Of graphic designers and illustrators, 69% used genAI in their work in 2023, employing tools like generative fill and large-scale text-to-image models such as DALL-E.[14]

In the U.S. **health care** system, adoption has been slow for IT generally and AI in particular (Poon et al. 2006; Goldfarb, Taska, and Teodoridis 2020). However, genAI may be an exception: A majority of physicians already used or were planning to use genAI in 2025 for generating chart summaries, creating discharge instructions, and an array of other tasks (AMA Augmented Intelligence Research 2025). Radiology is an enlightening example of how genAI has changed AI use in health care. While approximately 30% of radiologists already used AI as of 2020 (Allen et al. 2021), the emergence of genAI spread AI to all stages of the radiology workflow, including the health system (e.g. imaging need prediction, claims processing), clinicians (e.g. prior authorization, communication), technologists (e.g. patient

14. See Offerman, Stefan. "Creative Pros See Generative AI as Part of Their Future." Adobe Blog, March 21, 2023. https://blog.adobe.com/en/publish/2023/03/21/research-creative-pros-see-generative-ai-as-part-of-their-future.

tracking, report generation), and radiologists (e.g. image quality assessment, diagnosis) (Burnside et al. 2025).

Notwithstanding these inroads, AI use faces substantial hurdles in health care. For example, N. Agarwal et al. (2023) document that combining human analysis with AI diagnostics can yield disappointing results—radiologists have difficulty determining how much confidence to assign the AI output. And, although Sahni et al. (2023) conclude that machine learning has the potential to assist with an array of tasks including diagnosis, treatment choice, and managing records, they assess that for hospitals, AI adoption often cannot be justified on financial factors alone.

GenAI is used for many tasks in **finance** as well. Companies can use genAI to lower the cost of creating client-specific portfolios (Joshi 2025), improve existing automated systems that respond to client requests (McKinsey 2025a), assist with regulatory compliance (R. Agarwal et al. 2024) and with loan underwriting (Wang 2023).

In **electricity generation and distribution**, firms are experimenting with genAI, with approximately 33% piloting genAI in their customer service operations.[15] Some companies use genAI in their load forecasting processes, while others use the technology to simulate equipment degradation and inform predictive maintenance, areas where machine learning was already in use (Gao et al. 2024).

3.2 Knock-on Innovation

In the course of adopting new technologies, firms retool, retrain, and reorganize to better exploit their productivity potential.[16] This process involves knock-on innovation in products, production processes, and operations. In the case of genAI, knock-on product innovation includes the diverse array of user interface software for generative models and the emergence of more capable robots. Process innovation includes genAI-enhanced approaches to product design and production line operation. Organizational innovations include centralized data governance and optimization of supply chains and data centers.

15. Penrod, Emma. "A Third of Utilities Have Begun to Pilot Generative AI for Customer Service, Other Uses: Report." Utility Dive, July 12, 2023. https://www.utilitydive.com/news/utilities-generative-ai-artificial-intelligence-capgemini-report/686601/.

16. Bresnahan and Greenstein (1996) discuss this process in detail, which they refer to as "co-invention."

Bresnahan (2024) observed that early machine learning (pre-generative AI) adoption was concentrated in places where complementary innovation was less necessary, such as in firms that were highly digitized from their founding (digital natives). In such firms, adoption of AI was more straightforward, involving substitution of AI for existing IT capital or deploying AI to undertake tasks not previously part of operations.[17] In 2018, nearly 10 years after internet giants had begun using machine learning at scale (e.g. Amazon's random forest demand forecasting (2009) and Google's Panda search algorithm (2011)), AI began to gain traction at other firms (fig. 4 on page 14).[18] Digital natives will surely lead the charge for genAI as well. For other firms, the pace and success of knock-on innovation will be a key determinant of the scale and timing of productivity effects from genAI.[19]

3.2.1 Products

User interfaces (UIs) provide a channel through which requests and responses can pass between the user and the model, whether a human user or another system, such as a vehicle or a robot. In the early days of genAI, users accessed genAI through their own Python programs or through websites such as the OpenAI Playground. A major shift occurred in November 2022, when OpenAI released ChatGPT, a conversational interface that made genAI interactions significantly more accessible to a broader audience. Since then, several new interfaces have emerged. In 2023, OpenAI introduced Custom GPTs, enabling users to create specialized LLMs for specific domains, such

17. The legendary email from Jeff Bezos instructing internal teams at Amazon to exclusively use APIs to deliver data and functionality would have landed rather differently at a non-native company, for example. Even at Amazon, though, continuous digital transformation is challenging, as Yegge (2011) relates in his account of the aftermath of the Bezos email.

18. Bresnahan (2019, 157) describes the use of machine learning at Amazon, Facebook, Google, and Netflix, particularly for matching (e.g. consumers to products) and for user interfaces and notes "there is little application of [artificial intelligence technologies] outside the Internet Giants as of spring 2018." Bughin and Van Zeebroeck (2018, 1) notes that "only a fraction of companies—about 10 percent—have tried to diffuse AI across the enterprise, ... An additional quarter of companies have tested AI to a limited extent."

19. The emergence of cloud services may have catalyzed machine learning adoption outside of digital natives, and AI as a service such as Azure OpenAI Service may play a similar role for genAI.

as LegalGPT for legal matters.[20] In 2024, OpenAI announced integration of their ChatGPT model to Apple's Siri voice assistant and Google launched NotebookLM, which made it easy to upload documents and transform them into interactive discussions.[21] In addition, there are "copilots" that integrate AI into existing user workstreams, notably GitHub Copilot (computer programming) and Microsoft 365 Copilot (office productivity).

System interfaces allow hardware and software systems to access the core AI system. For example, Nvidia's Isaac Software Development Kit (SDK) facilitates the integration of AI into robotics.[22] Access to AI through SDK helps the robot with environmental integration problems, such as simultaneously tracking its location and mapping its environment (SLAM). Development of multimodal models which can take in inputs of different kinds (text, images, sensor readings) and can output instructions to the robot, such as the rotation and torque for a joint, have pushed robot-AI integration forward (Reed et al. 2022; Brohan et al. 2023)

System interfaces also make possible the design of agentic AI systems, which collect information during operation, interpret context, make decisions and act in pursuit of goals autonomously (Park et al. 2023).[23] Examples include AutoGen, from Microsoft, and CrewAI. For example, an airline reservation app using agentic AI would be capable of handling the complex steps of searching for available flights, proactively aligning the user's travel options with their preferences, making the booking, adjusting to disruptions to the reservation process (e.g. sporadic price changes), and including special accommodations along the way (e.g. upgrading seats if available within the

20. See "Introducing GPTs," November 6, 2023. https://openai.com/index/introducing-gpts/

21. See "OpenAI and Apple announce partnership to integrate ChatGPT into Apple experiences," June 10, 2024. https://openai.com/index/openai-and-apple-announce-partnership/; See "NotebookLM gets a new look, audio interactivity and a premium version," December 13, 2024. https://blog.google/technology/google-labs/notebooklm-new-features-december-2024/

22. Robot capabilities have advanced in tandem with AI progress. Industrial robots were trained using machine learning beginning in the 1990s (Arinez et al. 2020; Soori, Arezoo, and Dastres 2023). Even more sophisticated robots, which can learn from their environments, integrating sensor data, are available today, including civilian and military autonomous vehicles (Knight 2016; Pierson and Gashler 2017).

23. For more detail on agents, see Wiesinger, Marlow, and Vuskovic (2024). For an array of agentic AI use cases, see Renner and Chaban (2024).

user's preferences).

3.2.2 Production processes

Product design is a use of genAI that will be self-evident to users of tools such as DALL-E to create whimsical images; powerful genAI tools are also capable of designing products of all kinds that meet technical and aesthetic specifications. Perhaps less obviously, the design process itself can be transformed through knock-on innovation. Saadi and Yang (2023) interviewed designers and observed:

> Rather than thinking about how to create several one-off designs, designers may consider how to create a system for design that would allow the design tool to generate a large number of valid outputs. This can involve setting the appropriate specifications, manufacturing methods, and product architecture early in the process to input into computational tools.

Production line operation Serradilla et al. (2022) provides an overview of the use of deep learning, including genAI such as generative adverserial networks, to optimize line configuration, throughput, efficiency, and carbon footprint. Predictive maintenance using synthetic data and scenario simulation is another application for genAI in industry. Sai, Sai, and Chamola (2024) provides examples of production optimization outside of manufacturing as well.

3.2.3 Organization

Among the organizational innovations spurred by AI are cross-functional teams with access to data that spans the enterprise, breaking down barriers between business units, optimization of supply chains, and reallocation of employees to de-emphasize repetitive writing tasks (Iansiti and Lakhani 2020).

3.2.4 Case study evidence

Task-specific user interface programs are already in use in the **health care** industry. For example, one hospital system used genAI to field a flood of questions during the COVID-19 pandemic (Wittbold et al. 2020). AI is

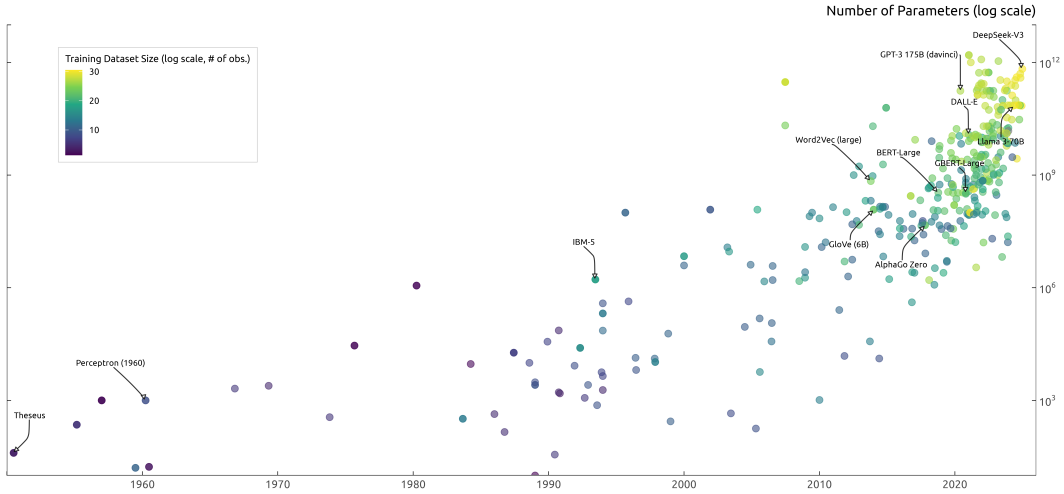
also used to transcribe conversations and dictations to create clinical notes (Handa and Sorensen 2023). Hornback et al. (2025) document how Fast Healthcare Interoperability Resources (FHIR), a protocol for the secure exchange of sensitive health information, has evolved to support emerging AI technologies. These efforts include the integration of BERT, a transformer model, for encoding unstructured text data (Peterson, Jiang, and Liu 2020) and the development of a generative model, FHIR-GPT (Y. Li et al. 2024). In **finance**, JP Morgan Chase adopted AI for contract review using software called COiN (Contract Intelligence) and reported saving 360,000 hours of costly lawyers and loan officers (Weiss 2017). In the **information sector**, Peng et al. (2023) document that the use of GitHub Copilot markedly increased the productivity of programmers, particularly less experienced ones. And, data center networks have evolved to better support genAI and genAI has contributed to network optimization beyond the contributions of machine learning (Y. Liu et al. 2024). GenAI has proved useful in the **energy sector** as well. Choi et al. 2024 describe a custom interface to genAI designed to assist control room operators in balancing supply and demand on electrical grids.

3.3 Ongoing Core Innovation

Since the introduction of the Transformer, genAI models have steadily pushed out the frontier of capability. At first, advances were largely driven by increasing the scale of the model (via the number of "parameters" in the model), computational power used ("compute", in the lingo of AI), and the size of the training dataset (fig. 6 on the following page). AI scientists and engineers often focus on the "scaling laws" that describe the benchmark performance effects of increasing each of these inputs (Kaplan et al. 2020). More recently, tactics to use compute more efficiently and to refine models for specific applications have been a focus as well.

Importantly, *economic* performance (productivity) only rises when more can be accomplished *while holding input costs fixed*. In other words, we are looking for *shifts* in scaling law functions, not movement along the curves. Accordingly, we focus below on (a) how innovations in model architecture (the algorithms used for distilling information from data) raise genAI model capabilities without raising training costs, (b) how hardware innovations lower the cost of computation, and (c) how richer datasets (with more information per token) can be brought to bear on training.

Figure 6: Number of Parameters and Training Dataset Size



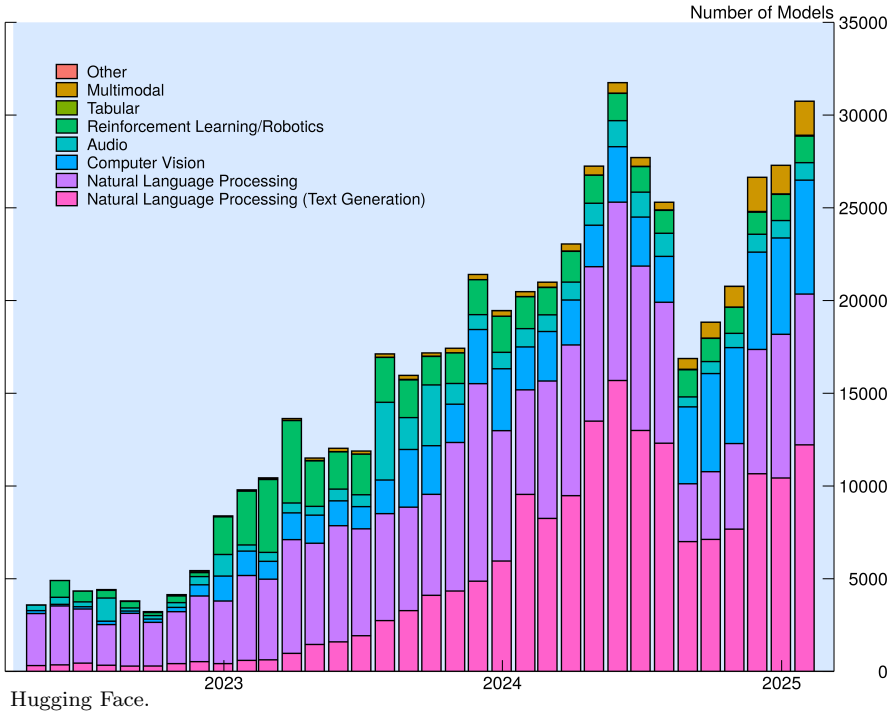
Note: Only models with reported parameter size and training dataset size are included in the dataset. For instance, GPT-4 is excluded as its parameter size is not known.
Source: Epoch (2024) with major processing by Our World in Data (**rahman-owen-you:2024:tracking-compute-intensive-ai-models**).

3.3.1 Model Development

Since the introduction of the Transformer, model development has progressed at a blistering pace, including improvements attributable to ramping up model scale, training dataset size, and the amount of compute employed—the scaling laws discussed above—and the introduction of novel model concepts and techniques to increase the efficiency of model training. An important catalyst in this process has been a rise in open-source models, which has accelerated experimentation across text-generation models, and since the second half of 2024, multi-modal models (fig. 7 on the next page).

The training of genAI models (optimal calibration of its parameters) takes place in two stages: pre-training and fine-tuning. Pre-training produces a broadly-applicable "foundation model"; fine-tuning refines the foundation model for a specific application. The trained model is then used in inference—responding to user requests. Efforts early in the wave of genAI improvement that followed the Transformer focused on pre-training, but the escalating cost of making progress in that stage has led researchers to explore improvements

Figure 7: Open-Source AI Models



Source: Hugging Face.

in fine-tuning and inference, as discussed below.[24]

In addition to training and inference innovations, advances in performance have come from novel model concepts. Mamba, introduced in 2023, achieved subquadratic-time sequence modelling by avoiding the pairwise comparison among tokens used in the attention mechanism, meaning that as input texts lengthen, the computational burden increases at a slower pace than the Transformer (Gu and Dao 2023).[25] Small-scale models, with lower computational requirements, have been a focus for some applications as well, such as personal devices and lower resource settings, making LLMs more accessible to the average user. Microsoft's Phi and models from Mistral AI, in particular, have shown relatively strong performance given their size (Jiang

24. See Zeff, Maxwell. "Current AI Scaling Laws Are Showing Diminishing Returns, Forcing AI Labs to Change Course." TechCrunch (blog), November 20, 2024. https://techcrunch.com/2024/11/20/ai-scaling-laws-are-showing-diminishing-returns-forcing-ai-labs-to-change-course/.

25. In particular, Mamba estimates the parameters of a latent state space structure.

et al. 2023; Abdin et al. 2024).

Pre-training Between 2018 and 2022, a key pre-training tactic in the effort to improve genAI performance was to increase model size. Size is measured in the number of estimated parameters, most notably the weights that determine how much influence each neuron has on each of the others in a neural network. For example, GPT models initially had 117 million parameters in 2018 (GPT-1), then 1.5 billion in 2019 (GPT-2), and a staggering 175 billion in 2020 (GPT-3).[26] Unfortunately, costs typically rise quadratically when parameters are added: each word (token) in the input sequence has to be compared to all of the others in the attention mechanism, as discussed in the box on the Transformer.

Remarkably, the cost of training a model of a given size was halved approximately every eight months through 2024 because of improvements in algorithmic efficiency (Ho et al. 2024). But, by 2022, the direction of innovation had begun to shift; scaling laws indicated diminishing returns to model size for foundation models, and focus turned to optimizing training efficiency. Hoffmann et al. 2022, for example, note that raising the number of parameters faster than the amount of data used leads to "undertrained" (imprecisely estimated) parameters that are less useful for inference and recommend scaling the model size no faster than the dataset size.

Fine-tuning As the returns to model size scaling have diminished, researchers have begun focusing more attention on fine-tuning foundation models for specific tasks. That is, developers have used domain-specific training data to increase the model's expertise beyond the capabilities of foundation models for narrowly defined questions.

Several techniques have been developed to improve this type of foundation model adaptation. **Transfer learning** involves taking a foundation model that has been fine-tuned for a specific task and adapting it (fine-tuning it again) for a related task, a process which typically involves only a modest amount of modification. **Instruction tuning** provides the model with guidance for specific scenarios that enhance its performance for targeted use cases. For example, the Alpaca model is a fine-tuned version of Meta's

26. Shree, Priya. "The Journey of Open AI GPT Models." Walmart Global Tech Blog (blog), November 10, 2020.https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2

Llama model, refined by adding a set of instructions and desired outputs to the training process (Taori et al. 2023).[27] **Reinforcement learning** has been extensively applied in AI for fine-tuning, allowing the model to refine its behavior based on a reward function (Mnih et al. 2013). A particularly influential approach along these lines was **reinforcement learning from human feedback (RLHF)**, a technique that aligns the model's output with human preferences by learning explicitly from human reactions to the model's responses to their queries (Christiano et al. 2017). For example, the model may provide the user with several responses to a query and ask the user to rank them, then use the ranking to improve performance in future queries. This technique gained widespread attention in the context of genAI with the release of InstructGPT in 2022 (Ouyang et al. 2022).

Inference Inference costs (in terms of electricity, time, compute, and carbon emissions) have risen with the popularity of genAI, leading to a focus on techniques to make this step more efficient.[28]

- One of the most important innovations in this area has been the **Mixture of Experts (MoE)** approach, an architecture that activates only a subset of model parameters in response to queries (Jacobs et al. 1991). A router determines which "expert" (portion of the full model) to activate based on the input. This adjustment allows the model to employ only a subset of the billions of parameters in the original foundation model, drastically reducing computational costs (Shazeer et al. 2017).

- **Pruning** is another inference refinement; here extraneous parameters are removed outright from the model (Cetin et al. 2024).

- Developers also incorporate **distillation**, a compression-like technique that uses knowledge from large complex models to inform smaller, less costly models (Hinton, Vinyals, and Dean 2015).

27. The supplemental training set showed Llama the desired response to particular queries, such as "Instruction: Brainstorm a list of possible New Year's resolutions. Output: lose weight, exercise more, eat healthier."

28. After setting a flat fee for access to ChatGPT, OpenAI CEO Sam Altman was surprised by the flood of user requests. See "Sam Altman says he's losing money on OpenAI's $200-per-month subscriptions: 'People use it much more than we expected'." Economic models have long predicted this outcome when the price of the marginal unit is set to zero. See, for example, the discussion of overuse of common pasture in *The Wealth of Nations* (Smith 1776).

- **Quantization** reduces the level of accuracy in order to reduce costs in terms of computation and memory requirements (e.g. moving from 32-bit to 8-bit floating point precision). For example, in 2023 Microsoft developed BitNet, an LLM with competitive performance that transforms floating point parameters in the model to a terniary digit (0, 1, or -1) (Wang et al. 2023).

- **Token caching** involves temporarily storing information anticipated to be needed in future inference steps.[29] For example, the prompts sent by a user to a model for inference may tend to be similar, implying that by caching processed tokens (words, phrases), anticipated computation costs can be reduced (Pope et al. 2023).[30]

The recently introduced DeepSeek R1 model leverages several of the techniques described above to deliver a substantial performance improvement compared to existing models (See the box, "Landmark AI Models: DeepSeek R1.").

In some cases, recent efforts have acted to extended inference time to enhance performance. Extending inference time raises costs, but on the margin, resources may be better used for responding to queries than for additional training. OpenAI's recent o1 model exemplifies the benefits of this approach via its superior performance across various domains, particularly reasoning-heavy tasks.[31] And, constraining the model to provide a response grounded in logic can lead to better results as well; **chain-of-thought (CoT) reasoning** guides the LLM to articulate a series of steps in its inference (Wei et al. 2022).

29. The fact that caching, a technique first employed in computing in the 1960s (on IBM System/360 mainframes), was first introduced to genAI inference in 2023 suggests there may be other basic numerical methods that have yet to be leveraged. If so, this bodes well for future efficiency improvements.

30. Sakana AI recently introduced an approach using neural networks called NAMMs to optimally decide whether to keep or discard tokens, saving up to 75% of cache memory. See Dickson, Ben. "New LLM Optimization Technique Slashes Memory Costs up to 75%." VentureBeat (blog), December 13, 2024. https://venturebeat.com/ai/new-llm-optimization-technique-slashes-memory-costs-up-to-75/.

31. See Nuñez, Michael. "OpenAI Scientist Noam Brown Stuns TED AI Conference: '20 Seconds of Thinking Worth 100,000x More Data'." VentureBeat (blog), October 23, 2024. https://venturebeat.com/ai/openai-noam-brown-stuns-ted-ai-conference-20-seconds-of-thinking-worth-100000x-more-data/.

Landmark AI Models: DeepSeek R1

A salient example of recent model innovation occurred in January, 2025 when DeepSeek unveiled R1 (short for "reasoning") (DeepSeek-AI et al. 2025). This model set a new bar for cost-effective, high-quality multi-modal models. DeepSeek stunned the AI research community by reporting the costs for training this 671 billion parameter model to be just under $6 million.[a] Although observers noted this figure does not account for the substantial reliance on R&D by other AI developers, the news led to a material change in perceptions of the role of Chinese AI companies. News of DeepSeek R1 coincided with a one-day market valuation drop of nearly $600 billion for NVIDIA as DeepSeek had achieved leading edge performance with far less of NVIDIA GPU-provided compute than previously believed necessary.[b] And, the DeepSeek R1 release may have spurred a major competitor to accelerate their product offerings: OpenAI soon updated o3-mini, its extended inference model, lowered its price, and offered a free-trial version.

Deepseek R1 blends several concepts that had been known in the genAI field for some time to improve performance and slash inference costs including mixture of experts, chain-of-thought reasoning, reinforcement learning, distillation, and quantization. Deepseek R1 endeavored to optimally combine all of these techniques to reduce costs. Some of these techniques were already commonly used in frontier models but others, such as their novel approach to reinforcement learning, called Group Relative Policy Optimization (GRPO), were not.

Relative to other frontier model developers, DeepSeek shifted attention from supervised learning during the fine-tuning stage to reinforcement learning.

---

a. See the technical report for DeepSeek V3 (DeepSeek-AI et al. 2024, 5): "...DeepSeek-V3 costs only 2.788M GPU hours for its full training. Assuming the rental price of the H800 GPU is $2 per GPU hour, our total training costs amount to only $5.576 million. Note that the aforementioned costs include only the official training of DeepSeek-V3, excluding the costs associated with prior research and ablation experiments on architectures, algorithms, or data."

b. See Subin, Samantha. "Nvidia Sheds Almost $600 Billion in Market Cap, Biggest One-Day Loss in U.S. History." CNBC, January 27, 2025. https://www.cnbc.com/2025/01/27/

DeepSeek R1 (continued)

nvidia-sheds-almost-600-billion-in-market-cap-biggest-drop-ever.html.

Agents Another direction for progress currently receiving intense attention—distinct from the pre-traning/refinement/inference optimization approach—is the creation of AI agents. **Agentic AI** systems develop strategies to pursue broad goals and recalibrate in response to their environment, in contrast to **tool-based AI**, which has a stable structure and calibration and is equipped only to respond to carefully crafted requests. While agents of different kinds are commonplace in the home and at work, both in physical form, such as self-driving cars, and in virtual form, such as web browsers, agentic AI is distinguished by its autonomy in pursuit of more abstractly specified goals. One definition comes from *Boston Consulting Group*:[32]

> Put simply, AI agents are artificial intelligence that use tools to accomplish goals. AI agents have the ability to remember across tasks and changing states; they can use one or more AI models to complete tasks; and they can decide when to access internal or external systems on a user's behalf. This enables AI agents to make decisions and take actions autonomously with minimal human oversight.

AI agents extend capabilities of AI to wider use in business, particularly to people with little technical knowledge or skill. For example, in the case study of health care, we noted that LLMs are being used to help with the paperwork burden faced by health care professionals, including making appointments, following up on treatment protocols and submitting insurance claims. Specialized AI agents can be programmed to provide this rather specific type of assistance, operating in combination with LLMs.

However, agents require programming and testing and cannot simply apply an off-the-shelf AI program. Correctly specifying the objective function for the AI agent is difficult. It may inadvertently be guided to maximize its programmed reward while subverting the real objective of the user.[33] More-

32. See Boston Consulting Group. "AI Agents." Accessed August 1, 2025. https://www.bcg.com/capabilities/artificial-intelligence/ai-agents.

33. See "Faulty Reward Functions in the Wild," February 14, 2024. https://openai.com/index/faulty-reward-functions/.

over, in extreme cases bad actors may hijack agents in business or military uses, leading to disastrous outcomes.

3.3.2 Hardware

GenAI model training and inference has massive computational requirements, making ongoing innovation in electronic hardware (and related hardware, such as cooling systems), essential to continued technical advance. Progress in this area has been rapid in recent years.

GenAI processing relies heavily on graphics processing units (GPUs). Like the image processing tasks GPUs were first designed for, training and inference for deep neural networks requires a large number of identical, independent computations which can be run in parallel on a GPU. (In contrast, microprocessing units (MPUs) perform computations sequentially, in the main.) GenAI processing also relies to a lesser extent on field-programmable gate arrays (FPGAs), which are more flexible than GPUs, particularly for inference. And, application-specific integrated circuits (ASICs) are used for specific steps in estimation as well.[34] For example, tensor processing units (TPUs) are customized for matrix multiplication, heavily used in neural networks. They consist of thousands of multipliers and adders connected to each other to form a large physical matrix. Storing the matrix parameters in on-chip registers drastically reduces the need to access off-chip memory, dramatically increasing computational efficiency.

Successive GPUs released by NVIDIA have delivered leaps in AI performance achieved by improvements in the power consumption and computational power of the processing cores—known as Compute Unified Device Architecture (CUDA)—adding and refining TPUs on the GPU chip, and increasing and refining cache (on-chip) memory. The history of NVIDIA GPU generations illustrates this progression.[35] CUDA, first introduced in 2007, explicitly supported non-graphics computing; prior to CUDA, programmers were required to reframe computations in terms of graphics operations.[36] As

34. The logical circuits on FPGAs can be reconfigured through programming and tailored to specific algorithms. The circuitry in ASICs is customized for specific tasks at the time of fabrication and cannot be changed. Technically, GPUs are ASICs as well.

35. This discussion was substantially improved by a discussion with Claude AI and research on Wikipedia.

36. With CUDA and associated tools, programmers can use standard programming techniques in C/C++.

Table 3: Price of Compute, Selected Nvidia GPUs

Model	Year	Price	TFLOPS	Price/TFLOP	Transistors
GeForce 8800 GT	2007	$349	0.3	$1,163	0.8B
GeForce RTX 4060	2024	$299	15.1	$20	18.9B
Change (ann. rate)	NA	–1%	23%	–24%	19%

Source: TechPowerUp.

AI became a prominent use case for NVIDIA GPUs, new architectures were increasingly optimized for deep learning—beginning with Pascal (2016)—and for genAI—beginning with Hopper (2022). On-chip TPUs were first introduced with the Volta microarchitecture in 2017 and successive GPU generations—Turing (2018), Ampere (2020), Hopper (2022), and Blackwell (2024)—each included improved TPUs.

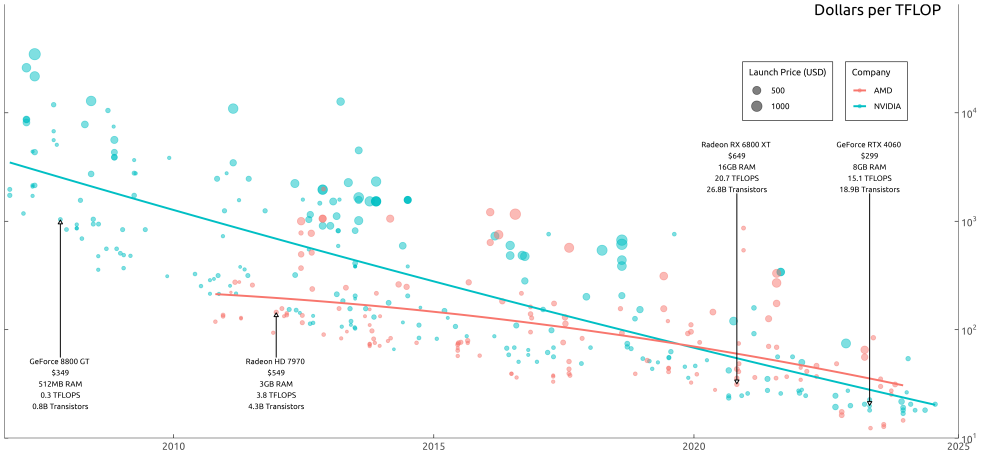
While the engineering performance of leading edge GPUs has rocketed upwards in recent years, prices have increased dramatically as well. Fortunately for productivity, holding performance constant, the price of GPUs has moved down: In 2007, a $349 GPU provided 0.3 teraflops (TFLOPS) of compute and in 2024, a $299 GPU delivered 15.1 TFLOPS, implying an average annual rate of price decline of 24% that persisted for 17 years (table 3). As shown in fig. 8 on the next page, this example is representative of a broader trend: the cost efficiency of computation has improved dramatically. Moreover, performance measured in TFLOPS likely understates the advance in AI hardware performance over this period as some design changes operate to reduce the number of TFLOPS needed for a given level of AI training or inference. For example, GPUs have taken on board additional logic blocks to accelerate matrix operations (Tensor cores) and GPUs have been partitioned to provide more flexible use of TFLOPS (Lino 2024).

Continuation of this trend of declining computation cost is not guaranteed. What seems like inexorable progress from a distance is in fact the result of a long sequence of difficult engineering feats when seen up close.[37] Historically, a key contributor to falling costs in the semiconductor industry has been steady miniaturization at the leading edge of the chip industry. Through 2003, the linear dimensions of the features (e.g. transistors) on

37. To get a sense of the staggering array of engineering problems involved in moving between chip generations, flip through any edition of the *International Technology Roadmap for Semiconductors* on the worldwide web.

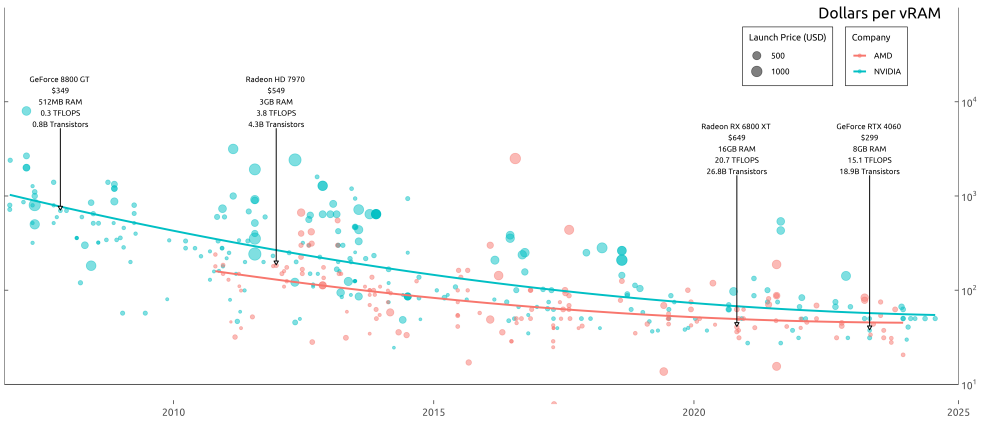
Figure 8: Price Indices of GPU Improvements

(a) Price per TFLOP



Note: Price per TFLOPS (trillion floating-point operations per second). The blue line represents the best fit line for NVIDIA GPUs, and the orange line represents the best fit line for AMD GPUs.
Source: TechPowerUp.

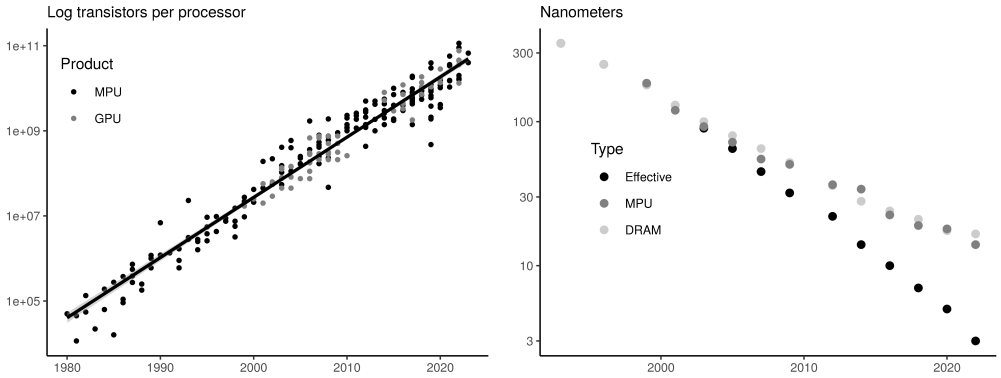
(b) Price per vRAM



Note: Price per vRAM in GB (video random access memory). The blue line represents the best fit line for NVIDIA GPUs, and the orange line represents the best fit line for AMD GPUs.
Source: TechPowerUp.

Figure 9: Progress on Moore's Law: Two Perspectives

(a) Transistor Count (b) Transistor Size



Note: Moore (1975) predicted that transistors per processor would double every two years beginning in 1980. The average increase for the models shown is 93% every two years. Intel stopped reporting the number of transistors on its MPUs in 2014.
Source: Wikipedia, accessed 4/7/2023.

Note: Effective node is process labeling used by major producers. DRAM node is metal half pitch. MPU node is the average of metal half pitch and contacted gate half pitch.
Source: National Technology Roadmap for Semiconductors and successor publications.

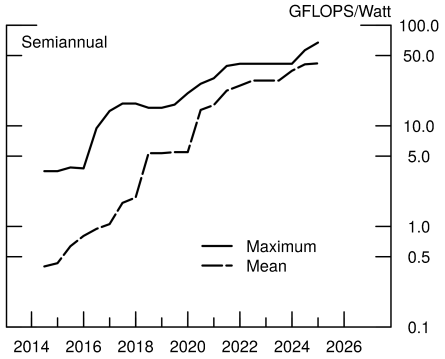
leading edge chips were reduced by 30 percent with each generation (node), yielding a $(0.7 * 0.7 \approx)$ 50 percent reduction in the space they occupied on a chip every two years (fig. 9b). From that year forward, the dimensions of these features were reduced far more slowly, though the industry, using the term "effective node," contended that performance gains with each generation matched the historical trend. The apparent slowdown in cost improvements of TFLOPS and video random access memory (vRAM) in fig. 8 on the previous page may well reflect that development. Since then, chip innovation has relied less heavily on miniaturization. For example, increasing "die size" (the surface area of the chip) has allowed the number of transistors per chip to continue to climb (fig. 9a).[38]

The proximate cause of the miniaturization slowdown was the end of a regularity known as "Dennard scaling" whereby power usage was largely unchanged even as more electronic activity was squeezed into the same surface area (Dennard et al. 2003). Because heat generation is increasing in

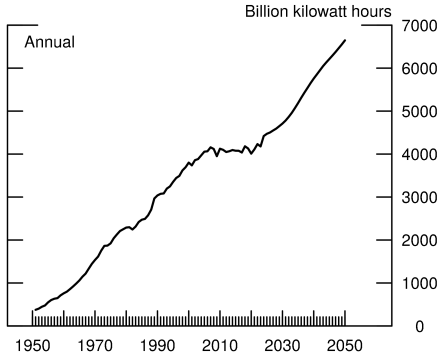
38. Since "Moore's Law" is a prediction that the number of transistors per die (chip) will double every two years, it arguably still holds true. Of course, because Gordon E. Moore (1965, 1975) does not contain a testable hypothesis, debates about whether Moore's Law holds are largely semantic. As is evident from fig. 9a, there is no single value for transistors per chip at any point in time.

Figure 10: Indicators of Power Demand and Supply

(a) Data Center Efficiency



(b) U.S. Electricity Generation



Note: Data center efficiency indicator is gigaflops per watt of supercomputers labeled "industry" or "vendor".
Source: Top500.org for efficiency. U.S. Energy Information Administration for electricity generation.

power usage, cooling cost began to rise with each node. Consequently, development efforts in the computing sector shifted to focus on a balance of computing speed and power consumption. Energy efficiency merits attention as well; concerns have arisen that power demand for genAI use may outpace growth in power-generation capacity, throttling genAI-led productivity advances. Two recent developments temper that concern. The energy efficiency of the most efficient industrial supercomputers, which include the data centers of major IT service providers, has roughly doubled since 2022, when genAI use began to climb, and U.S. electricity generation is forecast to expand substantially in coming years (fig. 10).[39]

3.3.3 Datasets

GenAI models "learn" by adjusting parameters to best represent the content of large amounts of text (and other media), allowing them to estimate the probability that a given word or phrase should appear next in the sequence it generates in response to a prompt. Loosely speaking, the larger the corpus of text available to the model, the better it can estimate these

39. Forecasts for the share of power use attributed to data centers in the United States in 2028 range from 6.7 percent to 12 percent, up from 4.4 percent in 2023 (Kamiya and Coroamă 2025).

probabilities. Figure 6 on page 21 illustrates the increase over time in the size of the datasets used to train the models. The change in the coloration of the circles, from purple to blue to green to yellow, indicates the rapid increase in the size of the training datasets.

Some observers believe that we may be approaching the limit of what we can learn from public text data.[40] A crucial nuance to this aspect of model improvement is that access to more information, not more text in itself, is needed to continue to improve genAI models.[41] To illustrate with an extreme example, doubling the size of the training text by exactly duplicating the corpus will yield no improvement to the model. Good quality data enables the model to learn the underlying language structure efficiently, which requires the data to span the full extent of the language with minimal redundancy and noise. Consequently, there are two looming challenges with respect to training data. First, more obscure or sensitive topics may have little coverage in public-facing content. Second, diminishing marginal returns to training will set in as developers move from information-rich content, such as Wikipedia and scientific articles, to more inane text, like social media posts.

One approach to mitigating the content constraint is transfer learning, where a model pre-trained with public data is improved by further training using proprietary data.[42] This not only increases the size of the dataset, but may increase the scope as well. Developers are also wrestling with the question of "domain generalization," where a model is used to generate content on topics outside the scope of the training dataset (Zhou et al. 2022).

As the information content of the marginal text from the internet falls, other techniques for generating data for training become more attractive. In one approach, small localized modifications of the training data can be introduced. For example, the performance of an image recognition model may be improved by supplementing the training set of labeled images with their mirror images.[43] Such variations in input data constrain the model parameter search process in a useful way. If an image is a dog, say, the

40. Villalobos et al. (2022) predict that the scope of training sets may approach the full extent of public high-quality text data as early as 2026.

41. In terms of information theory, models improve from entropy, the expected amount one will learn from the data generation process producing the text (Shannon 1948).

42. Cockburn, Henderson, and Stern (2019) note that this raises the issue of market structure as a potential constraint on progress in AI.

43. This approach was taken by the developers of AlexNet, a model which revolutionized the field (Krizhevsky, Sutskever, and Hinton 2012).

model should recognize that its mirror image is a dog as well, a desirable property known as "regularity."

Another approach to augmentation is the use of "synthetic data" created via generative models to emulate the patterns and characteristics of real data (Liu et al. 2024). For example, an LLM designed to tackle mathematical questions might be trained on a dataset of questions generated by another LLM, using bootstrapping techniques to create similar questions from a human-produced training dataset (Yan et al. 2025). This approach is attractive for medical imagery as well, where creating training data, such as CT scans, is both resource-intensive and constrained by privacy concerns (Guo et al. 2025).[44]

Last, datasets can be augmented by harvesting information collected with sensors, particularly in physical environments such as industrial robots and autonomous vehicles (Feng et al. 2019). This approach offers the prospect of a broader domain of use for genAI models and further diffusion of the technology.

3.4 The Case that GenAI is a GPT

To summarize, although it is early to tell how widespread the use of genAI will be, the case that generative AI is a general-purpose technology is compelling, supported by the impressive record of knock-on innovation and ongoing core innovation.

Of the three GPT criteria, widespread adoption is the most difficult to argue that genAI has met. Although some field studies have provided encouraging results that genAI may raise productivity, outside of large corporations, few firms have adopted the technology. The share of jobs requiring AI skills is low and has moved up only modestly, suggesting that firms are taking a cautious approach. The ultimate test of whether genAI is a GPT will be the profitability of genAI use at scale in a business environment and such stories are hard to come by at present. That said, use among individuals is high, perhaps unbeknownst to their employers, and with genAI increasingly folded

44. The usefulness of synthetic data is a matter of some debate. Some observers have raised concerns that training with synthetic data (and AI-generated text increasingly present on the internet) will yield low-quality or even nonsensical results, a phenomenon known as "model collapse" (Alemohammad et al. 2023; Shumailov et al. 2023). Others have argued that model collapse only occurs when the original training text is replaced by model-generated text (Gerstgrasser et al. 2024).

into office productivity software (such as Microsoft 365), its use may become so unremarkable that firms and workers may not be aware it is in use.

The case that genAI meets the knock-on innovation criterion is somewhat stronger. Key areas of product innovation include user interface software and interface with robotics, where the genAI model enables far more sophisticated applications. Production process innovation includes digital twins to improve production line efficiency. Organizational innovations include restructuring of the product design business function. What share of organizations can justify the digital transformation needed for genAI, such as centralized data governance, particularly for non-digital-native companies, remains to be seen.

Core technology innovation is the criteria for which the case is the clearest. GenAI performance has moved up at a blistering pace since the introduction of the Transformer thanks to increasingly large datasets and application of more computing power. More importantly, performance has risen while holding inputs constant (data, compute, and model size), driven by algorithmic improvements. If this trend continues, the direct cost of using genAI will fall, spurring greater adoption.

Artificial General Intelligence

In 1960, Herbert A. Simon, winner of the 1975 *Turing Award* and 1978 *Nobel Memorial Prize in Economic Sciences* wrote, "within the very near future—much less than twenty-five years [before 1985]—we shall have the *technical* capability of substituting machines for any and all human functions in organizations" (Simon 1960, 22). This is the concept now known as "artificial general intelligence" (AGI).[a]

Demonstrating the feasibility of AGI is a long-run objective of many AI researchers but is not a particularly interesting exercise for economists. Technical feasibility is a necessary condition, but far from a sufficient one, for a technology to raise productivity. It is technically feasible to turn lead into gold, for example, but only at prohibitive cost (Matson 2014). And, conjecture about future technology, however well informed, is not sufficient for a useful productivity forecast, which must answer the question of when, not just if, the technology will appear and, *a fortiori*, when it will be practical to use. Moreover, achieving human-level performance on all tasks would likely entail devoting resources to improving performance on tasks AI is ill-suited for at the

AGI (continued)

expense of tasks where improvement would be more readily achieved. Dell'Acqua et al. (2023) explore this issue and provide extensive evidence this is a first-order impediment to AGI.

Some present-day IT leaders believe AGI is imminent: In 2024, Elon Musk predicted the arrival of AGI within two years, Sam Altman predicted its arrival by 2025, and Dario Amodei expected AGI by 2026. Others are skeptical: Yann LeCun speculated it may never be achieved.[b] Historically, technology forecasts have been highly unreliable. Berkeley (1949), for example, foresaw a machine that would read handwritten text; noteworthy practical use of handwriting recognition systems U.S. Post Office came fifty years later.

Fortunately, AGI is not a precondition for genAI to be a GPT, nor do we need a forecast for the timing of the arrival of AGI to forecast the effects of genAI on productivity.

---

a. Remarkably, Čapek (1920) had already envisioned a world with robots that performed all human tasks, including ones involving physical manipulation of the environment. At that time, although punchcard tabulators programmable with plugboards existed, most "computers" were human (Grier 2007).

b. See *Reuters.* "Tesla's Musk Predicts AI Will Be Smarter than the Smartest Human next Year." April 8, 2024, sec. Technology. https://www.reuters.com/technology/teslas-musk-predicts-ai-will-be-smarter-than-smartest-human-next-year-2024-04-08/; Varanasi, Lakshmi. "Here's How Far We Are from AGI, According to the People Developing It." Business Insider, November 9, 2024. https://www.businessinsider.com/agi-predictions-sam-altman-dario-amodei-geoffrey-hinton-demis-hassabis-2024-11; PYMNTS.com. "Meta AI Head: ChatGPT Will Never Reach Human Intelligence," May 22, 2024. https://www.pymnts.com/artificial-intelligence-2/2024/meta-ai-head-chatgpt-will-never-reach-human-intelligence/.

4 Is GenAI an Invention of a Method of Invention?

In classical "light bulb" growth models (commonly known as "Solow-Swann" models), the source of total factor productivity (TFP)—the part of productivity growth not attributable to capital accumulation—is unspecified

(Robert M Solow 1994). About the importance of TFP that emerged when this model was brought to the data, Abramovitz (1956, 11) famously observed, "since we know little about the causes of productivity increase, the indicated importance of this element [TFP] may be taken as some sort of measure of our ignorance about the causes of economic growth in the United States and some sort of indication of where we need to concentrate our attention." "Endogenous growth" models have since appeared, including some that add a research sector to produce new technologies in response to incentives (Romer 1994; Aghion and Howitt 1992; Akcigit and Van Reenan 2023).[45] Like other sectors, efficiency in the research sector can be increased by the use of appropriate capital, such as an invention of a method of invention (IMI).[46] Griliches (1957) noted that the hybridization process developed for creating new varieties of corn played this role. Other examples of IMIs are shown in table 4 on the following page, grouped into observational, analytical, communication, and organizational tools. We consider below whether genAI falls into these categories and how it can contribute to research productivity beyond what is contributed by machine learning. We then review a number of broad indicators of the role of AI in research, including patent filings, the share of AI use accounted for by users with research jobs and tasks, and new evidence on the prevalence of AI references in company conference calls.

Prior to the appearance of genAI, AI had already diffused across a wide range of scientific disciplines (Carobene et al. 2024). And, it had already been shown to improve the efficiency of research. Cockburn, Henderson, and Stern (2019, 23) note that pre-generative AI assists with the "labor-intensive search with high marginal cost of search" involved in many types of R&D. Put differently, AI improves prediction, a point emphasized by Agrawal, J. Gans, and Goldfarb (2018), including predicting how materials might behave. Examples of phenomenal success are well known. Scientists have made major advances toward practical nuclear fusion using reinforcement learning techniques to adjust the magnetic system that contains the plasma in a fusion

45. Of course, in actuality TFP is *not* simply the output of a research sector. TFP results when firms choose, in response to research results, to make complementary investment in intangibles (Brynjolfsson, Rock, and Syverson 2021) in the context of government policy (Baily et al., n.d.) and is importantly affected by business dynamism (Decker et al. 2017), labor market efficiency (Davis and Haltiwanger 2014), and market structure (Goettler and Gordon 2011).

46. The term originates from Whitehead (1925), according to Mowery and Rosenberg (1999).

Table 4: Examples of Inventions of Methods of Invention

Observational tools	
Telescope	1608 CE
Compound microscope	1620 CE
Pendulum clock	1656 CE
DNA sequencer	1973 CE
Analytical tools	
Mainframe (IBM S/360)	1964 CE
Personal computer (IBM PC)	1981 CE
Machine learning	1998 CE
Communication tools	
Printing press (Gutenberg)	1439 CE
Internet protocol (TCP/IP)	1975 CE
Organizational innovations	
Scientific societies (Accademia dei Lincei)	1603 CE
Corporate labs (GE)	1900 CE
Government labs (U.S. NRL)	1923 CE
Big science (Oak Ridge)	1961 CE
Source: Authors' judgment.	

reactor (Degrave et al. 2022; Seo et al. 2024). Richardson et al. (2020) used the "knowledge graph"—which encodes relationships among scientific publications using machine learning—created by BenevolentAI to produce a novel treatment for COVID-19. Machine learning has also been used extensively for predicting the properties of novel metal alloys, economizing on physical experimentation and computer simulations (Hart et al. 2021). Our focus is on the question of whether genAI enables additional efficiencies in R&D beyond these and other improvements provided by machine learning. In particular, we ask whether genAI enhances measurement, analysis, communication, and organization of invention.

GenAI as an observational tool Observational tools, such as microscopes, telescopes, and cameras produce imperfect images due to defects in their components and variation in the environment. GenAI provides a tool to impute imperfect portions of images as well as missing observations in datasets of all kinds in a fashion more consistent with the apparent properties

of the underlying phenomena. For example, generative techniques for image enhancement, which rely on an implicit model of the manifold of the data generating process—closer to the actual physics, say, of a remote galaxy seen through an imperfect lens—perform better than techniques, such as splines, relying solely on smoothness assumptions (i.e. that nature does not make leaps) (G. Liu et al. 2018; Lugmayr et al. 2022).

GenAI as an Analytical Tool Like the compound microscope for physical phenomena, Christian (2020) notes that LLMs serve as a kind of microscope to look at social phenomena. Caliskan, Bryson, and Narayanan (2017, 183), for example, find that "text corpora contain recoverable and accurate imprints of our historic biases." This new visibility may promote and support analysis of social science questions not previously tractable. There has been an explosion of sentiment analysis and other forms of NLP in recent years fueled by this capability of genAI.[47] While the identification of underlying sentiment (encoding) is strictly speaking a function of the LLM, conveying the discovered sentiment to the user is necessarily a generative process. Korinek (2023) documents a variety of potential roles for genAI in the economic research process; that genAI may play a similar role in many other fields is a reasonable conjecture.[48]

GenAI-Supported Organizational Innovation Institutional organization plays a central role in the effectiveness of R&D (Mowery and Rosenberg 1999), as do informal associations into professional networks (Wang and Barabási 2021) and geographic clusters (Porter and Stern 2001). Consequently, the method of invention for any given research program properly includes the institutions involved. Emerging applications of AI "digital twins" offer the prospect of R&D with a reduced institutional footprint in many areas of study. Among these are drug discovery (Bordukova et al. 2024), industrial research (Tao, Zhang, and Zhang 2024), and materials science (Ka-

47. Sentiment analysis is possible with earlier forms of AI but the capabilities of genAI models are vastly greater (Gentzkow, Kelly, and Taddy 2019; Dell 2025).

48. Early versions of this paper cited Toner-Rodgers (2024), which purported to show that genAI substantially accelerates the discovery process in materials science. The veracity of that work has since been questioned by the author's institution and prominent researchers in the field. Credible empirical evidence on the effect of genAI on scientific research efficiency would be a timely contribution to resolving the uncertainty around the effects of genAI on productivity.

Table 5: The Stages of a Research Project

Conceptual	reviewing the literature, formulating the broad problem, identifying specific goal
Planning	determining research design and procedures, identifying resource needs, procuring funding
Empirical	collecting data, preparing data for analysis
Analytical	identifying data features, testing hypotheses, interpreting results
Dissemination	communicating to audience (colleagues, industry, policymakers, public, students) in written, visual, and oral form

lidindi et al. 2022). For example, generative adversarial networks (GANs) may provide an alternative to animal testing for toxicology (Chen et al. 2022).

GenAI as a Communication Tool Although empirical and analytical stages of research projects focus on measurement and calculation, many aspects of the research process involve manipulating language. GenAI may be employed in the writing tasks involved in the conceptual, planning, and dissemination stages of research projects, such as drafting literature reviews, grant applications, and seminar slides (table 5). Whether, on net, genAI improves the efficiency of such tasks once the effort needed for review and editing of the documents drafted by genAI is accounted for is an open question. If so, genAI may play a similar role to the printing press and word processing as a catalyst to the invention process.

Research agents AI agents (discussed in section 3.2.1 on page 17) have emerged that endeavor to automate the core of research entirely, generating research questions, designing and conducting experiments, and reporting results. Thus, research agents may play the role of an observational, analytical, and communication IMI all at once. Examples include Google's AI co-scientist and Sakana's The AI Scientist (Gottweis et al. 2025; Lu et al. 2024).[49]

49. Stoughton (2023) documents co-scientist innovation at the National Science Foundation.

Opinions of the significance of research agents vary widely. Importantly, the design, conduct, and communication of experiments is only a portion of the activities of a scientist (table 6 on the next page). Even so, Lu et al. (2024) report that The AI Scientist can generate publishable research for as little as $15 per journal article, a striking finding. On the other hand, Beel, Kan, and Baumgart (2025, 1) evaluate Sakana's agent and conclude,

> We evaluated the AI Scientist and found several critical shortcomings. The system's literature review process is inadequate, relying on simplistic keyword searches rather than profound synthesis, which leads to poor novelty assessments. In our experiments, several generated research ideas were incorrectly classified as novel, including well-established concepts such as micro-batching for stochastic gradient descent (SGD). The AI Scientist also lacks robustness in experiment execution—five out of twelve proposed experiments (42%) failed due to coding errors, and those that did run often produced logically flawed or misleading results. In one case, an experiment designed to optimize energy efficiency reported improvements in accuracy while consuming more computational resources, contradicting its stated goal. Furthermore, the system modifies experimental code minimally, with each iteration adding only 8% more characters on average, suggesting limited adaptability. The generated manuscripts were poorly substantiated, with a median of just five citations per paper—most of which were outdated (only five out of 34 citations were from 2020 or later). Structural errors were frequent, including missing figures, repeated sections, and placeholder text such as "Conclusions Here". Hallucinated numerical results were contained in several manuscripts, undermining the reliability of its outputs.

Moreover, genAI research agents may have a subtle but important limitation: uncovering the fundamental features of phenomena. K. Li et al. (2022) argue that genAI does have that capability. They trained a generative model to play the board game Othello without providing the rules of the game then demonstrated that the model can play appropriately in a setting not found in the training data, concluding that genAI has created an "emergent world model." Other research has challenged this conclusion. jylin04 et al. (2024) argue that the model is employing a "bag of heuristics," rather than a set of

Table 6: Scientific Activity beyond Research Projects

Conceptual	designing a research program (a connected set of research projects); packaging program to influence appropriate audiences (e.g. writing textbooks)
Leadership	playing executive and advisory roles in local and profession-wide academic and government specialist communities
Mentoring	supervising research, advising and instructing students
Support	fostering buy-in to research program from institutional leadership
Networking	recruiting collaborators and maintaining relationships
Commercialization	translating research results into practical applications

game rules.[50] This question is a crucial one in determining the capabilities of genAI to contribute to science. Without a model of the underlying structure of the physical or social phenomenon under study, one cannot articulate its fundamental laws. This limitation may arise naturally from the training process; humans learn the fundamentals of science from textbooks, but these laws may not be the rhetorical foundation for the verbal exchanges on the topic found in the training corpus.

4.1 Indicators of GenAI Research and of GenAI Use in Research

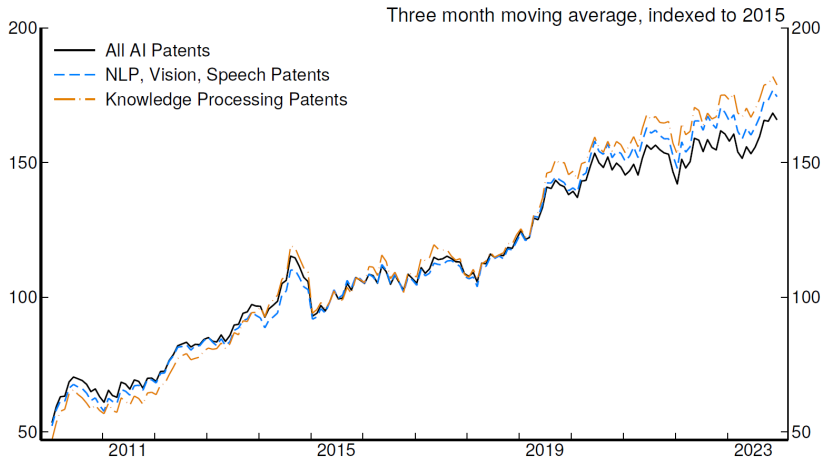
We discuss below a set of indicators of genAI research (patents) and of genAI use in research (conference call transcripts and genAI queries).[51]

50. A useful entry point to this ongoing debate is "LLMs and World Models," by Melanie Mitchell, February 13, 2025, found at the *AI: A Guide for Thinking Humans* Substack blog.

51. For evidence of the potential for genAI use in research based on job descriptions, see Eloundou et al. (2024, 1308) who note that "scientists and researchers" and "technologists" are the job groups most highly exposed to LLMs, and that "this suggests that when LLMs improve, they have potential to cause downstream improvements in R&D productivity for workers in sectors deploying them."

Patents AI-related patents issued by the United States Patent and Trademark Office (USPTO) increased markedly following the advent of genAI, suggesting a related surge in genAI research (fig. 11).[52] The USPTO index of AI-related patents began climbing in 2018, shortly after the publication of the seminal paper by Vaswani et al. (2017) which introduced the Transformer architecture, quickly reaching a level 50 percent higher, which it has sustained since 2019. We also observe that increases in patent activity for AI modalities particularly related to genAI—natural language processing (NLP), vision, speech, and knowledge processing—have risen even further. This suggests that the recent surge in patenting activity is not merely a reflection of advancements in machine learning.

Figure 11: AI Mentions in Scientific Patents



Source: Artificial Intelligence Patent Dataset (2023), U.S. Patent Office.

GenAI Prompts Handa et al. (2025) provide a rich set of information on *actual* genAI use in their Anthropic Economic Index (AEI), a useful complement to the detailed work on the *potential* impact of genAI based on

52. Pairolero et al. (2025) use BERT-based embeddings to refine a previous iteration of their patent classification methodology that used Word2Vec. We use their most conservative threshold of 93% probability to identify AI-related patents. We refer the reader to the USPTO website hosting their data for more information: https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset.

analysis of job descriptions from Eloundou et al. (2024). The AEI assigns millions of conversations from Claude (Anthropic's premier genAI system) to roughly 3,500 of the tasks defined by the U.S. Department of Labor's O*NET Dataset.[53] An equal fraction of each task's percentage share of all prompts is then apportioned to each occupation which includes that task in O*NET.

Table 7 on the next page shows the estimated share of prompts accounted for by occupational groups, their employment share, and the ratio of the two. (If prompts were equally distributed across all workers, these ratios would each be equal to 1.) "Computer & mathematical occupations", which includes the computer programmers for whom genAI use is especially intense, have the highest ratio of prevalence of genAI use to occupational prevalence, 10.9. Use intensity is nearly as high among scientists, who account for 7.1 times as many prompts as would be found if prompts were equally distributed across workers.[54] Other occupational groups with high relative prevalence of genAI use include "arts, design, sports, entertainment & media"; "architecture & engineering"; and "educational instruction & library". The remaining 87.6% of employment is accounted for by occupations which AEI found had a share of Claude prompts roughly equal to or lower than their share of employment, highlighting the very concentrated nature of genAI adoption in the economy at present.

Table 8 on page 46 shows the prevalence of selected O*NET tasks related to scientific discovery among Claude AI prompts. These tasks collectively account for only 0.9% of all prompts, revealing that the share of prompts accounted for by scientists (6.4%) includes far more than scientific discovery. These discovery tasks are most commonly related to the creation of mathematical or statistical models of technical phenomena, such as business, scientific, and engineering, either to foster understanding of the phenomena or to predict how modeled systems would perform. Such tasks account for 86.5% of the scientific tasks Claude AI is asked to help with. Other tasks include the advancement of mathematical science (9.0% of scientific prompts) and the design of research projects (4.5%).

Figure 12 on the next page illustrates significant automation and augmentation of tasks among our groupings of research occupations: programmers exhibit the highest automation rate, with over half of the requests handled

53. Patterns of Anthropic use may not be representative of the patterns for all genAI programs, of course. See section 7.1 of Tamkin et al. (2024) for a discussion of related work.

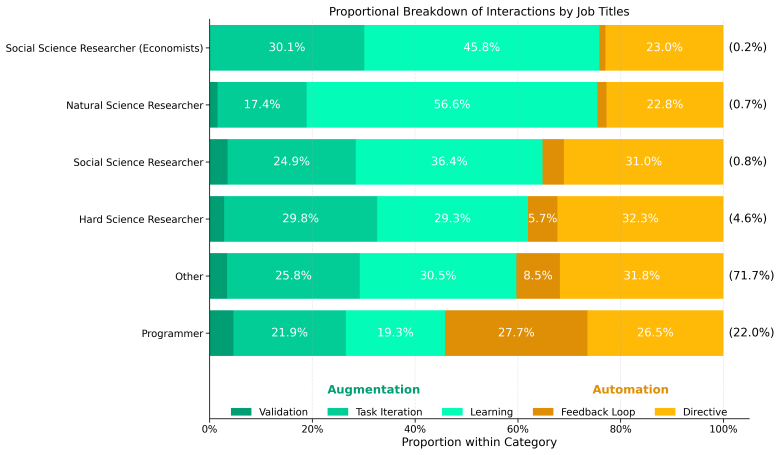
54. Naturally, they may be using genAI for computer programming tasks as well.

Table 7: Occupations with High GenAI Usage

Job Type	Prompt Share	Empl. Share	Ratio
Computer & mathematical	37.2	3.4	10.9
Arts, design, sports, entertainment, & media	10.3	1.4	7.4
Life, physical, & social science	6.4	0.9	7.1
Architecture & engineering	4.5	1.7	2.6
Educational instruction & library	9.3	5.8	1.6
Memo: Other occupations	31.8	87.6	0.4

Note: Percent share of prompts submitted to Claude AI and linked to tasks by Anthropic. Task weights are apportioned equally to all occupations which include that task in O*NET.
Source: Anthropic Economic Index.

Figure 12: GenAI Automation vs. Augmentation in Researcher Roles



Note: Authors' calculations.
Source: Anthropic Economic Index.

Table 8: O*NET Scientific Task Prevalence in Claude AI Prompts

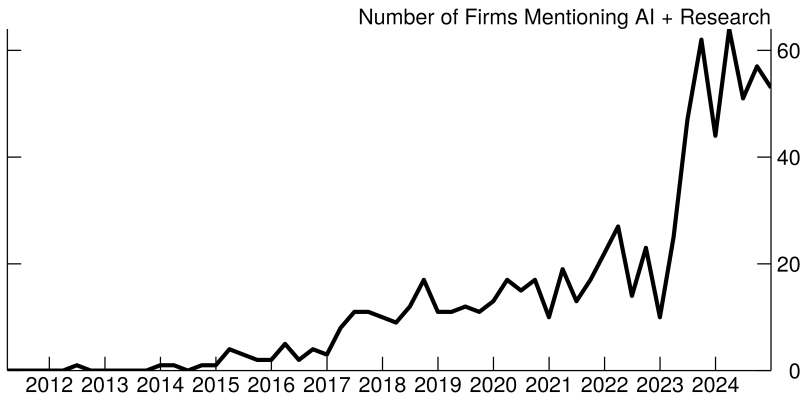
Task	Share (pct.)
Modelling & Prediction	86.5
conduct logical analyses of business, scientific, engineering, and other technical problems, formulating mathematical **models** of problems for solution by computers.	46.1
design or develop software systems, using scientific analysis and mathematical models to **predict** and measure outcome and consequences of design.	16.9
complete **models** and simulations, using manual or automated tools, to analyze or **predict** system performance under different operating conditions.	15.7
develop mathematical or statistical **models** of phenomena to be used for analysis or for computational simulation.	4.5
design computer simulations to **model** physical data so that it can be better understood.	2.2
develop software applications or programming to use for statistical **modeling** and graphic analysis.	1.1
Other Tasks	13.5
develop new principles and new relationships between existing mathematical principles to **advance mathematical science**.	9.0
design research projects that apply valid scientific techniques and use information obtained from baselines or historical data to structure uncompromised and efficient analyses.	4.5

Note: Share of all Anthropic prompts accounted for by these tasks is 0.9%. Task-level prompts labeled by Anthropic.
Source: Anthropic Economic Index.

by genAI being automation tasks. Social science researchers show slightly lower automation rates, with economists showing over 23% of their prompts being automation focused. Notably, for hard science researchers (e.g., physicists, biochemists), the share of their genAI use for automation is nearly 15% higher than their natural science counterparts. This difference likely reflects AI's strength in data-intensive and simulation-based research such as those found in hard sciences like physics and materials science.

Conference Call Mentions GenAI's integration into the invention process is also revealed through firm communication. We analyze quarterly earnings calls, which are routine events where firm executives discuss company performance, future projects, and key developments with investors and analysts. Figure 13 on the following page plots the count of the number of firms referencing AI in the context of research as indicated by the firm mentioning an AI-specific term ("machine learning," "deep learning," "artificial intelligence," "genAI, or "generative AI") within a research-related context (within 10 words of "inventi-", "research-", or "discover"). A sudden rise appears in 2023, with approximately 60 public companies per quarter mentioning such usage. This increase in integration of AI with R&D is illustrative of the role it has begun to play in innovation in a corporate context. Two examples are provided below.

Figure 13: Mentions of AI Usage for Research in Conference Calls



Note: Authors' estimates. Data through Q4 2024.
Source: S&P Capital database merged with Compustat.

In 2024, John Wiley & Sons, a publishing company, announced expansions in compound databases leveraging advanced AI and its capability to accelerate scientific discoveries:[55]

> Wiley has just released two new database collections using advanced AI techniques to significantly expand the number of compounds available for analysis from food-related compounds to industrial compounds. The end goal here to help scientists reach better conclusions faster.

Similarly, Cadence Design Systems, an electronic systems design firm highlighted the potential of AI to automate fields such as biology and life sciences.[56]

> And then the third phase of AI adoption is AI applied to areas that were not automated in the past, okay? So I think that may take longer, maybe 5 years plus, but that has to be driven to digital biology and life sciences. I mean there's a huge application of AI.

55. See John Wiley & Sons, Inc. FQ1 2025 Earnings Call. https://s27.q4cdn.com/812717746/files/doc_financials/2025/q1/q125-earnings-transcript.pdf
56. See Cadence Deisgn Systems, Inc. FQ3 2023 Earnings Call.

4.2 Is GenAI an IMI?

Our assessment is that there is a strong case that genAI is an IMI. Indeed, it is a multifaceted IMI, having characteristics of an observational IMI (e.g. image enhancement), an analytical IMI (e.g. sentiment analysis), an organizational IMI (e.g. digital twins), and a communication IMI (e.g. document drafting). Whether the excitement over genAI research agents will prove to be merited or not, genAI's ability to augment these four dimensions of invention suggests that the idea generation process is becoming more productive.

It also appears that it is taking root within the research community. The signal from the USPTO's AI database is that AI patents surged when the use of genAI became practical. GenAI prompt analysis from Anthropic points to relatively intense use of AI by scientists as well as in adjacent fields like computing and engineering. And, corporate earnings calls increasingly mention genAI while discussing their research efforts.

5 Conclusion

The release of ChatGPT in late 2022 was a stark inflection point in public interest in genAI and predictions of a first-order impact on productivity in the future soon followed.[57] Yet, as exciting as progress in genAI is from a science and engineering standpoint, its economic effects are highly uncertain. For firms to justify the reorganization and other complementary capital needed to exploit genAI, the return from the technology, less the total cost of ownership, must be high enough. Field studies do point to efficiency gains in selected business functions and many firms have experimented with the technology, but only a small share of them attest to material improvements to their bottom lines from the technology thus far.

To complement the limited empirical evidence, we ask what the characteristics of genAI suggest its future impact on productivity may be. GenAI has features typical of both a general-purpose technology—headed toward being widely used, stimulating related innovation, and displaying ongoing improvement in (economic) performance—and an invention of a method of

57. Goldman Sachs analysts forecast that genAI will eventually increase in U.S. labor productivity by 15%. See Briggs, Joseph, and Sarah Dong. "Global Economics Comment: AI Productivity and Labor Market Impacts Are Still Small (For Now)." Goldman Sachs Economic Research, March 14, 2025.

invention—raising the efficiency of R&D through improvements to observation, analysis, communication, or organization.

Because both GPTs and IMIs promote productivity growth for extended periods, it is reasonable to expect genAI will have a noteworthy impact on productivity. Importantly, genAI's potential for productivity does not depend on the elusive goal of reaching artificial general intelligence (AGI). It can qualify as a GPT and IMI well before the arrival of AGI. The main hurdle is diffusion. Complementary innovations like interfaces, robotics, and agents, for example, are emerging, and technological progress is ongoing. Yet, outside of the tech sector, firm-level adoption in production processes is still modest. As an IMI, the case is stronger: genAI usage is gaining traction within the scientific community via workflows and patents. Even so, we offer several cautionary observations.

First, we expect that genAI will boost productivity growth *relative to the counterfactual* economy without it. If the growth effect of machine learning (and other IT innovations) is waning, the impact of genAI will have to match the impact of machine learning on the likes of Amazon and Facebook for the economy simply to match the recent history of productivity growth.[58]

Second, the GPT effect on productivity is inherently slow as it involves complementary investment. For example, the effect on the productivity *level* of solid-state computing was large, but it played out over decades, damping the effect on productivity *growth*. The tech boom was a long time coming: Massive advances in computational technology, including the invention of the solid state transistor and the fundamentals of system design had accumulated by the end of the 1940s and a steady decline in computing costs had begun (Nordhaus 2007).[59] The surge in productivity attributed to information technology arrived some fifty years later.

Third, investment to deploy new technologies is fraught with risk. If

58. Bresnahan (2024) notes that the spread of earlier AI technologies to other companies slowed once the digitally native companies had jumped in.

59. Predictions of an IT-infused future of abundance soon followed, but noteworthy productivity gains only appeared in tandem with time-consuming complementary investment, such as business reorganization. For example, Berkeley (1949), based on observation of the handful of computers in existence, eagerly anticipated automatic address books, libraries, translators, typists, and stenographers, as well as business process optimization, psychological testing and training, weather forecasting, and even weather control. Others, such as Martin (1960, 4), cautioned that productivity gains would be hard-won: "The data-processing system of an organization is of almost unimaginable complexity. The introduction of a computer usually involves widespread changes in this complex system."

genAI is a widely adopted "killer app" that defines a new era of IT, the computing capacity needed to deliver genAI to millions of simultaneous users will be massive. Anticipation of this outcome helps explain the wave of investment in data centers and electricity generation. But, building to meet anticipated demand can lead to disastrous consequences, as illustrated by the history of railroad expansion and the associated boom-bust cycles in the 19$^{\text{th}}$ Century.[60] For IT systems, the capacity forecasting problem is compounded by the progress of technology, which drives down the hardware investment required to deliver a given level of service.[61] A critical further concern is the availability of electrical power to accommodate the demands of data centers supporting widespread genAI use (Pilz, Mahmood, and Heim 2025). And, of course, R&D is inherently risky because the returns are erratic: The chain of choices made between insight from research and greater output per hour is long.

Nevertheless, our forecast for the most likely outcome is for a noteworthy contribution of genAI to the level of labor productivity, though the range of plausible outcomes is wide, with respect to both the magnitude of the total contribution and how that impact is spread over time (hence, the productivity growth rate).

60. On the British experience in the 1840s, see Campbell and Turner (2012) and Odlyzko (2012).

61. This forecasting challenge for fiber optic telecommunications systems, combined with duplicative effort by competing networks, was a major factor behind the economic downturn in 2001 (Couper, Hejkal, and Wolman 2003; Doms 2004). On the capital overhang in the telecommunications network in the early 2000s, see Hecht (2016, 53): "The post-bubble network was vastly overbuilt and underused. In late 2002, consulting firm TeleGeography estimated only 10 percent of the long-haul fibers installed in Europe and North America carried any signals, and that only 10 percent of the wavelengths in those fibers were lit. . . . Soon, creditors were trying to unload dark-fiber networks for pennies on the dollar."

References

Abdin, Mara, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. 2024. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.* ArXiv Preprint arXiv:2404.14219. arXiv. arXiv: 2404.14219 [cs.CL]. https://arxiv.org/abs/2404.14219.

Abdou, Ossama A. 1997. "Effects of Luminous Environment on Worker Productivity in Building Spaces." *Journal of Architectural Engineering* 3 (3): 124–132.

Abramovitz, Moses. 1956. "Resource and Output Trends in the United States since 1870." *American Economic Review,* 1–23.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo. 2020. *AI and Jobs: Evidence from Online Vacancies.* NBER Working Paper 28257. National Bureau of Economic Research.

Acemoglu, Daron, and Pascual Restrepo. 2020. "The Wrong Kind of AI? Artificial Intelligence and the Future of Labour Demand." *Cambridge Journal of Regions, Economy and Society* 13 (1): 25–35.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology.* NBER Working Papers 31422. Cambridge, MA: National Bureau of Economic Research.

Agarwal, Rahul, Andreas Kremer, Ida Kristensen, and Angela Luget. 2024. *How Generative AI can help Banks Manage Risk and Compliance.* Technical report. McKinsey & Company.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* 60 (2).

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence.* Boston: Harvard Business Review Press.

———. 2023. *The Turing Transformation: Artificial Intelligence, Intelligence Augmentation, and Skill Premiums.* NBER Working Paper 31767. National Bureau of Economic Research.

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb. 2019. "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction." *Journal of Economic Perspectives* 33 (2): 31–50.

Akcigit, Ufuk, and John Van Reenan. 2023. "Creative Destruction and Economic Growth." In *The Economics of Creative Destruction.* Cambridge: Harvard University Press.

Akcigit, Ufuk, and John Van Reenen. 2023. *The Economics of Creative Destruction.* Harvard University Press Cambridge.

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. *Self-Consuming Generative Models go Mad.* ArXiv Preprint arXiv:2307.01850. arXiv.

Allen, Bibb, Sheela Agarwal, Laura Coombs, Christoph Wald, and Keith Dreyer. 2021. "2020 ACR Data Science Institute artificial intelligence survey." *Journal of the American College of Radiology* 18 (8): 1153–1159.

AMA Augmented Intelligence Research. 2025. "Physician Sentiments around the Use of AI in Heath Care: Motivations, Opportunities, Risks, and Use Cases." *American Medical Association.*

Arinez, Jorge F, Qing Chang, Robert X Gao, Chengying Xu, and Jianjing Zhang. 2020. "Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook." *Journal of Manufacturing Science and Engineering* 142 (11): 110804.

Baily, Martin, David Byrne, Aidan Kane, and Paul Soto. n.d. "Productivity Policy in the United States."

Baily, Martin, and Aidan Kane. 2025a. *AI in the Finance Sector: Transforming Productivity and Risk Management.* Brookings (blog). Date accessed: June 30, 2025. Brookings Institution.

———. 2025b. *AI in the Healthcare Sector: Enhancing Care, Efficiency, and Innovation.* Brookings (blog). Brookings Institution. https://www.brookings.edu/articles/harnessing-ai-for-economic-growth/.

Bajari, Patrick L., and Victor Chernozhukov. 2018. *Quality-Adjusted Price Indices Powered by ML and AI with an Application to Apparel.* Presented to Federal Economic Statistics Advisory Committee (FESAC). Research report. FESAC, December. https://www.census.gov/content/dam/Census/about/about-the-bureau/adrm/FESAC/meetings/Chernozhukov%20Presentation%20Revised.pdf.

Bar-Hillel, Yehoshua. 1960. "A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation." *Advances in Computers* 1:158–163.

Bayard, Kimberly, Emin Dinlersoz, Timothy Dunne, John Haltiwanger, Javier Miranda, and John Stevens. 2018. *Early-Stage Business Formation: An Analysis of Applications for Employer Identification Numbers.* NBER Working Paper 24364. National Bureau of Economic Research.

Beel, Joeran, Min-Yen Kan, and Moritz Baumgart. 2025. *An Evaluation of Sakana's AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards' Artificial General Research Intelligence'(AGRI)?* ArXiv Preprint arXiv:2502.14297. arXiv.

Berkeley, Edmund Callis. 1949. *Giant Brains or Machines that Think.* New York: John Wiley & Sons.

Bick, Alexander, Adam Blandin, and David J. Deming. 2024. *The Rapid Adoption of Generative AI.* NBER Working Paper 32966. National Bureau of Economic Research, September.

Bonney, Kathryn, Cory Breaux, Cathy Buffington, Emin Dinlersoz, Lucia S Foster, Nathan Goldschlag, John C Haltiwanger, Zachary Kroff, and Keith Savage. 2024. *Tracking Firm Use of AI in Real Time: A Snapshot from the Business Trends and Outlook Survey.* NBER Working Paper 32319. National Bureau of Economic Research, April.

Bordukova, Maria, Nikita Makarov, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P Menden. 2024. "Generative Artificial Intelligence Empowers Digital Twins in Drug Discovery and Clinical Trials." *Expert Opinion on Drug Discovery* 19 (1): 33–42.

Bresnahan, Timothy. 2019. "Artificial Intelligence Technologies and Aggregate Growth Prospects." In *Prospects for Economic Growth in the United States,* edited by John W. Diamond and George R. Zodrow, 132–172. Cambridge, England: Cambridge University Press.

———. 2024. "What Innovation Paths for AI to become a GPT?" *Journal of Economics & Management Strategy* 33 (2): 305–316.

Bresnahan, Timothy, and Shane Greenstein. 1996. "Technical Progress and Co-Invention in Computing and in the Uses of Computers." *Brookings Papers on Economic Activity. Microeconomics* 1996:1–83.

Brohan, Anthony, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. *Rt-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control.* ArXiv Preprint arXiv:2307.15818. arXiv.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems (NeurIPS)* 33:1877–1901.

Brynjolfsson, Erik. 2022. "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence." *Daedalus* 151 (2): 272–287.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond. 2025. "Generative AI at work." *The Quarterly Journal of Economics* 140 (2): 889–942.

Brynjolfsson, Erik, Tom Mitchell, and Daniel Rock. 2018. "What can Machines Learn and What Does it Mean for Occupations and the Economy?" *AEA Papers and Proceedings* 108:43–47.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson. 2021. "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics* 13 (1): 333–72.

Buchanan, Bruce G, and Reid G Smith. 1988. "Fundamentals of Expert Systems." *Annual Review of Computer Science* 3 (1): 23–58.

Bughin, Jacques, and Nicolas Van Zeebroeck. 2018. "Artificial Intelligence: Why a Digital Base is Critical." *The McKinsey Quarterly.*

Burnside, Elizabeth S, Thomas M Grist, Michael R Lasarev, John W Garrett, and Elizabeth A Morris. 2025. "Artificial Intelligence in Radiology: A Leadership Survey." *Journal of the American College of Radiology.*

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora contain Human-like Biases." *Science* 356 (6334): 183–186.

Campbell, Gareth, and John D Turner. 2012. "Dispelling the Myth of the Naive Investor During the British Railway Mania, 1845–1846." *Business History Review* 86 (1): 3–41.

Čapek, Karel. 1920. *R.U.R. (Rossum's Universal Robots).* Penguin Clasics.

Carobene, Anna, Andrea Padoan, Federico Cabitza, Giuseppe Banfi, and Mario Plebani. 2024. "Rising Adoption of Artificial Intelligence in Scientific Publishing: Evaluating the Role, Risks, and Ethical Implications in Paper Drafting and Review Process." *Clinical Chemistry and Laboratory Medicine (CCLM)* 62 (5): 835–843.

Cetin, Edoardo, Qi Sun, Tianyu Zhao, and Yujin Tang. 2024. *An Evolved Universal Transformer Memory.* ArXiv Preprint arXiv:2410.13166. arXiv.

Chen, Xi, Ruth Roberts, Weida Tong, and Zhichao Liu. 2022. "Tox-GAN: An Artificial Intelligence Approach Alternative to Animal Studies—A Case Study with Toxicogenomics." *Toxicological Sciences* 186 (2): 242–259.

Choi, Seong Lok, Rishabh Jain, Patrick Emami, Karin Wadsack, Fei Ding, Hongfei Sun, Kenny Gruchalla, et al. 2024. *eGridGPT: Trustworthy AI in the Control Room.* Technical Report NREL/TP-5D00-87440. National Renewable Energy Laboratory, May.

Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values.* New York: W. W. Norton & Company.

Christiano, Paul F, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." *Advances in Neural Information Processing Systems* 30.

Cockburn, Iain M, Rebecca Henderson, and Scott Stern. 2019. "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis." In *The Economics of Artificial Intelligence: An Agenda,* edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 115–146. Chicago: University of Chicago Press.

Couper, Elise, John P Hejkal, and Alexander L Wolman. 2003. "Boom and Bust in Telecommunications." *FRB Richmond Economic Quarterly* 89 (4): 1–24.

Crane, Leland, Michael Green, and Paul Soto. 2025. "Measuring AI Uptake in the Workplace." *FEDS Notes* (February).

Cui, Kevin Zheyuan, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz. 2024. *The Productivity Effects of Generative AI: Evidence from a Field Experiment with GitHub Copilot.* Technical report. MIT Open Publishing Services.

David, Paul A. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *The American Economic Review* 80 (2): 355–361.

Davis, Steven J, and John Haltiwanger. 2014. *Labor Market Fuidity and Economic Performance.* NBER Working Paper 20479. National Bureau of Economic Research.

Decker, Ryan, and John Haltiwanger. 2024. "High Tech Business Entry in the Pandemic Era." *FEDS Notes.*

Decker, Ryan A, John Haltiwanger, Ron S Jarmin, and Javier Miranda. 2017. "Declining dynamism, allocative efficiency, and the productivity slowdown." *American Economic Review* 107 (5): 322–26.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhange, Runxin Xu, et al. 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* ArXiv Preprint arXiv:2501.12948. arXiv. https://arxiv.org/abs/2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, et al. 2024. *DeepSeek-V3 Technical Report.* ArXiv Preprint arXiv:2412.19437. arXiv. https://arxiv.org/abs/2412.19437.

Degrave, Jonas, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, et al. 2022. "Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning." *Nature* 602 (7897): 414–419.

Dell, Melissa. 2025. "Deep Learning for Economists." *Journal of Economic Literature* 63 (1): 5–58.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.* Technology & Operations Management Unit Working Paper 013. Harvard University Business School.

Dennard, Robert H, Fritz H Gaensslen, Hwa-Nien Yu, V Leo Rideout, Ernest Bassous, and Andre R LeBlanc. 2003. "Design of ion-implanted MOSFET's with very small physical dimensions." *IEEE Journal of solid-state circuits* 9 (5): 256–268.

Devlin, Jacob. 2018. *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.* ArXiv Preprint arXiv:1810.04805. arXiv.

Doms, Mark. 2004. "The Boom and the Bust in Information Technology Investment." *FRBSF Economic Review,* 19–34.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs are GPTs: Labor Market Impact Potential of LLMs." *Science* 384 (6702): 1306–1308.

Felten, Edward W, Manav Raj, and Robert Seamans. 2019. *The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization.* Working Paper. NYU Stern School of Business.

Feng, Di, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. 2019. "Deep Active Learning for Efficient Training of a Lidar 3d Object Detector." In *2019 IEEE Intelligent Vehicles Symposium (IV),* 667–674. IEEE.

Ferrucci, David A. 2012. "Introduction to "this is watson"." *IBM Journal of Research and Development* 56 (3.4): 1–1.

Filippucci, Francesco, Peter Gal, Cecilia Susanna Jona Lasinio, Alvaro Leandro, and Giuseppe Nicoletti. 2024. *The Impact of Artificial Intelligence on Productivity, Distribution and Growth.* Working Paper 15. Organisation for Economic Co-Operation and Development, April.

Gao, Mingyang, Suyang Zhou, Wei Gu, Zhi Wu, Haiquan Liu, and Aihua Zhou. 2024. *A General Framework for Load Forecasting based on Pretrained Large Language Model.* ArXiv Preprint arXiv:2406.11336. arXiv.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3): 535–574.

Gerstgrasser, Matthias, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, et al. 2024. *Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data.* ArXiv Preprint arXiv:2404.01413. ArXiv.

Giczy, Alexander V, Nicholas A Pairolero, and Andrew A Toole. 2022. "Identifying Artificial Intelligence (AI) Invention: A Novel AI Patent Dataset." *The Journal of Technology Transfer* 47 (2): 476–505.

Gill, T Grandon. 1995. "Early Expert Systems: Where are They Now?" *MIS quarterly,* 51–81.

Goettler, Ronald L, and Brett R Gordon. 2011. "Does AMD Spur Intel to Innovate More?" *Journal of Political Economy* 119 (6): 1141–1200.

Goldfarb, Avi, Bledi Taska, and Florenta Teodoridis. 2020. "Artificial Intelligence in Health Care? Evidence from Online Job Postings." In *AEA Papers and Proceedings,* 110:400–404.

———. 2023. "Could Machine Learning be a General Purpose Technology? A Comparison of Emerging Technologies Using Data from Online Job Postings." *Research Policy* 52 (1): 104653.

Gottweis, Juraj, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, et al. 2025. *Towards an AI Co-scientist.* ArXiv Preprint arXiv:2502.18864. arXiv.

Grier, David Alan. 2007. *When Computers were Human.* Princeton: Princeton University Press.

Griliches, Zvi. 1957. "Hybrid Corn: An Exploration in the Economics of Technological Change." *Econometrica.*

Gu, Albert, and Tri Dao. 2023. *Mamba: Linear-time sequence modeling with selective state spaces.* ArXiv Preprint arXiv:2312.00752. arXiv.

Guo, Pengfei, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, et al. 2025. "Maisi: Medical AI for Synthetic Imaging." In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),* 4430–4441. IEEE.

Hancock, John T, and Taghi M Khoshgoftaar. 2020. "Survey on Categorical Data for Neural Networks." *Journal of Big Data* 7 (1): 28.

Handa, Kunal, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, et al. 2025. *Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations.* ArXiv Preprint arXiv:2503.04761. arXiv.

Handa, Sarthak, and Jared Sorensen. 2023. *Automatically Create Clinical Documentation with Generative AI.* Technical report. Amazon Web Services.

Hardesty, Larry. 2019. "The History of Amazon's Recommendation Algorithm." *Amazon Science* 22.

Hart, Gus LW, Tim Mueller, Cormac Toher, and Stefano Curtarolo. 2021. "Machine Learning for Alloys." *Nature Reviews Materials* 6 (8): 730–755.

Haupt, Andreas, and Erik Brynjolfsson. 2025. *AI Should Not Be an Imitation Game: Centaur Evaluations.* Technical report. Available at https://digitaleconomy.stanford.edu/wp-content/uploads/2025/06/CentaurEvaluations.pdf. www.andyhaupt.com.

Hecht, Jeff. 2016. "Boom, Bubble, Bust: The Fiber Optic Mania." *Optics & Photonics News.*

Hennessy, John L, and David A Patterson. 2011. *Computer Architecture: A Quantitative Approach.* Elsevier.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. *Distilling the Knowledge in a Neural Network.* ArXiv Preprint arXiv:1503.02531. arXiv. arXiv: 1503.02531 [stat.ML]. https://arxiv.org/abs/1503.02531.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313 (5786): 504–507.

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, et al. 2024. *Algorithmic Progress in Language Models.* ArXiv Preprint arXiv:2403.05812. arXiv.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, et al. 2022. *Training Compute-Optimal Large Language Models.* ArXiv Preprint arXiv:2203.15556. arXiv.

Hornback, Andrew, Marteau Benoit, Shaun QY Tan, Kyungbeom Kim, Oankar Patil, Joshua Traynelis, Yuanda Zhu, Felipe Giuste, and May D Wang. 2025. *FHIR in Focus: Enabling Biomedical Data Harmonization for Intelligent Healthcare Systems.* Technical report DOI: 10.36227/techrxiv.174585774.47415852/v1. TechRxiv.

Hulten, Charles R. 1978. "Growth accounting with intermediate inputs." *The Review of Economic Studies* 45 (3): 511–518.

Iansiti, Marco, and Karim R Lakhani. 2020. *Competing in the age of AI: Strategy and Leadership when Algorithms and Networks run the World.* Cambridge, Massachusetts: Harvard Business Press.

Jacobs, Robert A, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. "Adaptive Mixtures of Local Experts." *Neural Computation* 3 (1): 79–87.

James, Conrad D, James B Aimone, Nadine E Miner, Craig M Vineyard, Fredrick H Rothganger, Kristofor D Carlson, Samuel A Mulder, et al. 2017. "A Historical Survey of Algorithms and Hardware Architectures for Neural-Inspired and Neuromorphic Computing Applications." *Biologically Inspired Cognitive Architectures* 19:49–64.

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. *Mistral 7B.* ArXiv Preprint arXiv:2310.06825. arXiv.

Joshi, Satyadhar. 2025. *Generative AI in Investment and Portfolio Management: Comprehensive Review of Current Applications and Future Directions.* Technical report. Social Science Research Network (SSRN).

jylin04, JackS, Adam Karvonen, and Can. 2024. "OthelloGPT Learned a Bag of Heuristics." *LESSWRONG.*

Kaldor, Nicholas. 1957. "A Model of Economic Growth." *The Economic Journal* 67 (268): 591–624.

Kalidindi, Surya R, Michael Buzzy, Brad L Boyce, and Remi Dingreville. 2022. "Digital Twins for Materials." *Frontiers in Materials* 9:818535.

Kamiya, George, and Vlad C. Coroamă. 2025. "Data Centre Energy Use: Critical Review of Models and Results." *IEA 4E TCP Efficient, Demand Flexible Networked Appliances (EDNA).*

Kane, Aidan, and Martin Baily. 2025a. *AI in the Electricity Sector: Optimizing Grid Management and Energy Use.* Brookings (blog). Brookings Institution, April.

———. 2025b. *AI in the Information Sector: Advancing Software, Customer Service, and Design.* Brookings (blog). Brookings Institution, April.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. *Scaling Laws for Neural Language Models.* ArXiv Preprint arXiv:2001.08361. arXiv.

Keohane, Joe. 2017. "What News-Writing Bots Mean for the Future of Journalism." *Wired* 16:2017.

Knight, Will. 2016. "Japanese Robotics Giant Gives Its Arms Some Brains." *MIT Technology Review.*

Kondo, Illenin O, Logan T Lewis, and Andrea Stella. 2023. "Heavy Tailed but not Zipf: Firm and Establishment Size in the United States." *Journal of Applied Econometrics* 38 (5): 767–785.

Korinek, Anton. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 61 (4): 1281–1317.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25.

Kurzweil, Ray. 2024. *The Singularity Is Nearer: When We Merge with AI.* New York: Random House.

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1 (4): 541–551.

Li, Kenneth, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.* ArXiv Preprint arXiv:2210.13382. arXiv.

Li, Yikuan, Hanyin Wang, Halid Z Yerebakan, Yoshihisa Shinagawa, and Yuan Luo. 2024. *FHIR-GPT Enhances Health Interoperability with Large Language Models.* Preprint 8. medRxiv.

Liao, Thomas, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. "Are We Learning Yet? A Meta Review of Evaluation Failures across Machine Learning." In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).*

Lino, Giro. 2024. "Nvidia GPU Evolution: From GeForce to AI Powerhouse." *girolino.com.*

Lipsey, Richard G, Kenneth I Carlaw, and Clifford T Bekar. 2005. *Economic Transformations: General Purpose Technologies and Long-Term Economic Growth.* Oxford, England: Oxford University Press.

Liu, Guilin, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. "Image Inpainting for Irregular Holes using Partial Convolutions." In *Proceedings of the European conference on computer vision (ECCV),* 85–100.

Liu, Ruibo, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, et al. 2024. *Best Practices and Lessons Learned on Synthetic Data for Language Models.* ArXiv Preprint arXiv:2404.07503. arXiv.

Liu, Yinqiu, Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Yonggang Wen, and Dong In Kim. 2024. *Generative AI in Data Center Networking: Fundamentals, Perspectives, and Case Study.* ArXiv Preprint arXiv:2409.09343. arXiv.

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.* ArXiv Preprint arXiv:2408.06292. arXiv.

Lugmayr, Andreas, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. "Repaint: Inpainting using Denoising Diffusion Probabilistic Models." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 11461–11471.

Markov, Andrei Andreevich. 2006. "An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains." *Science in Context* 19 (4): 591–600.

Martin, E. Wainright, Jr. 1960. "Practical Problems of Introducing a Computer." *Business Horizons* 3 (3): 4–86.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, et al. 2024. *Artificial Intelligence Index Report 2024.* Stanford Institute for Human-Centered Artificial Intelligence (HAI). https://aiindex.stanford.edu/report/.

Matson, John. 2014. "Fact or Fiction? Lead can be Turned into Gold." *Scientific American.*

McCarthy, John, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. 2006. "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955." *AI magazine* 27 (4): 12–12.

McCulloch, Warren S, and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5:115–133.

McKinsey. 2024. "The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value." *McKinsey & Company.*

———. 2025a. *Banking on Innovation: How ING uses Generative AI to Put People First.* Technical report. McKinsey & Company. https://www.mckinsey.com/industries/financial-services/how-we-help-clients/banking-on-innovation-how-ing-uses-generative-ai-to-put-people-first.

———. 2025b. "The state of AI: How Organizations are Rewiring to Capture Value." *McKinsey & Company.*

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in VectorSpace.* ArXiv Preprint arXiv:1301.3781. arXiv.

Mikolov, Tomas, Ilya Sutskever, and Quoc Le. 2013. "Learning the Meaning Behind Words." *Google Open Source Blog* 14.

Minsky, Marvin. 1952. *A Neural-Analogue Calculator Based Upon a Probability Model of Reinforcement.* Technical report. Harvard University Psychological Laboratories.

Minsky, Marvin, and Seymour A. Papert. 1969. *Perceptrons: An Introduction to Computational Geometry.* Vol. 6. Cambridge, Massachusetts: MIT Press.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. *Playing Atari with Deep Reinforcement Learning.* ArXiv Preprint arXiv:1312.5602. arXiv. arXiv: 1312.5602 [cs.LG]. https://arxiv.org/abs/1312.5602.

Moor, James. 2006. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years." *AI Magazine* 27 (4): 87–87.

Moore, Gordon E. 1975. "Progress in Digital Integrated Electronics." *IEDM Tech Digest* 21:11–13.

———. 1965. "Cramming more Components onto Integrated Circuits." *Electronics Magazine* 4:114–117.

Mowery, David C, and Nathan Rosenberg. 1999. *Paths of Innovation: Technological Change in 20th-Century America.* Cambridge, England: Cambridge University Press.

Narayanan, Arvind, and Sayash Kapoor. 2024. *AI Snake Oil: What Artificial Intelligence Can Do, What it Can't, and How to Tell the Difference.* Princeton University Press.

Newell, Allen, John Calman Shaw, and Herbert A Simon. 1958. "Elements of a Theory of Human Problem Solving." *Psychological Review* 65 (3): 151.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* Cambridge, England: Cambridge University Press.

Nordhaus, William D. 2007. "Two Centuries of Productivity Growth in Computing." *The Journal of Economic History* 67 (1): 128–159.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Available at SSRN 4375283.*

Odlyzko, Andrew. 2012. "The Railway Mania—Fraud, Disappointed Expectations and the Modern Economy." *Journal of the Railway and Canal Historical Society* 215 (2): 1–16.

Olson, Parmy. 2024. *Supremacy: AI, ChatGPT, and the Race that Will Change the World.* New York: St. Martin's Press.

OpenAI. 2022. *Introducing ChatGPT.* https://openai.com/blog/chatgpt.

———. 2023. "GPT-4 Technical Report." *arXiv preprint arXiv:2303.08774.*

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35:27730–27744.

Pairolero, Nicholas A, Alexander V Giczy, Gerard Torres, Tisa Islam Erana, Mark A Finlayson, and Andrew A Toole. 2025. "The Artificial Intelligence Patent Dataset (AIPD) 2023 Update." *The Journal of Technology Transfer,* 1–24.

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology,* 1–22.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.* ArXiv Preprint arXiv:2302.06590. arXiv.

Perset, Karine. 2024. *Explanatory Memorandum on the Updated OECD Definition of an AI System.* Technical report. OECD.

Peterson, Kevin J, Guoqian Jiang, and Hongfang Liu. 2020. "A Corpus-Driven Standardization Framework for Encoding Clinical Problems with HL7 FHIR." *Journal of Biomedical Informatics* 110:103541.

Pierson, Harry A, and Michael S Gashler. 2017. "Deep Learning in Robotics: a Review of Recent Research." *Advanced Robotics* 31 (16): 821–835.

Pilz, Konstantin F., Yusuf Mahmood, and Lennart Heim. 2025. "AI's Power Requirements Under Exponential Growth." *RAND Research Report.*

Poon, Eric G, Ashish K Jha, Melissa Christino, Melissa M Honour, Rushika Fernandopulle, Blackford Middleton, Joseph Newhouse, et al. 2006. "Assessing the Level of Healthcare Information Technology Adoption in the United States: A Snapshot." *BMC Medical Informatics and Decision Making* 6:1–9.

Pope, Reiner, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. "Efficiently Scaling Transformer Inference." *Proceedings of Machine Learning and Systems* 5:606–624.

Porter, Michael E, and Scott Stern. 2001. "Innovation: location matters." *MIT Sloan Management Review.*

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI Blog.*

Reed, Scott, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, et al. 2022. *A Generalist Agent.* ArXiv arXiv:2205.06175. arXiv.

Renner, Matt, and Matt A. V. Chaban. 2024. *601 Real-world Gen AI Use Cases from the World's Leading Organizations.* Blog. Google Cloud.

Richardson, Peter, Ivan Griffin, Catherine Tucker, Dan Smith, Olly Oechsle, Anne Phelan, Michael Rawling, Edward Savory, and Justin Stebbing. 2020. "Baricitinib as Potential Treatment for 2019-nCoV Acute Respiratory Disease." *The Lancet* 395 (10223): e30–e31.

Romer, Paul M. 1994. "The Origins of Endogenous Growth." *Journal of Economic Perspectives* 8 (1): 3–22.

Rosenblatt, Frank. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386.

Russell, Stuart, and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach.* Hoboken: Prentice Hall.

Saadi, Jana I, and Maria C Yang. 2023. "Generative Design: Reframing the Role of the Designer in Early-Stage Design Process." *Journal of Mechanical Design* 145 (4): 041411.

Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz. 1998. "A Bayesian Approach to Filtering Junk E-mail." In *Learning for Text Categorization: Papers from the 1998 Workshop,* 62:98–105. Citeseer.

Sahni, Nikhil, George Stein, Rodney Zemmel, and David M Cutler. 2023. *The Potential Impact of Artificial Intelligence on Healthcare Spending.* NBER Working Paper 30857. National Bureau of Economic Research.

Sai, Siva, Revant Sai, and Vinay Chamola. 2024. "Generative AI for Industry 5.0: Analyzing the impact of ChatGPT, DALLE, and other models." *IEEE Open Journal of the Communications Society.*

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424.

Selfridge, Oliver G. 1958. *Pandemonium: A Paradigm for Learning.* Technical report. Mechanisation of Thought Processes: Proceedings of a Sumposium Held at the National Physical Laboratory.

Seo, Jaemin, SangKyeun Kim, Azarakhsh Jalalvand, Rory Conlin, Andrew Rothstein, Joseph Abbate, Keith Erickson, Josiah Wai, Ricardo Shousha, and Egemen Kolemen. 2024. "Avoiding Fusion Plasma Tearing Instability with Deep Reinforcement Learning." *Nature* 626 (8000): 746–751.

Sergeyuk, Agnia, Yaroslav Golubev, Timofey Bryksin, and Iftekhar Ahmed. 2025. "Using AI-based Coding Assistants in practice: State of affairs, perceptions, and ways forward." *Information and Software Technology* 178:107610.

Serradilla, Oscar, Ekhi Zugasti, Jon Rodriguez, and Urko Zurutuza. 2022. "Deep Learning Models for Predictive Maintenance: A Survey, Comparison, Challenges and Prospects." *Applied Intelligence* 52 (10): 10934–10964.

Shani, Inbal, and GitHub Staff. 2023. "Survey Reveals AI's Impact on the Developer Experience." *GitHub Blog.*

Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V Le, Geoffrey E Hinton, and Jeff Dean. 2017. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer." In *Proceedings of the 5th International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/1701.06538.

Shortliffe, Edward H. 1977. "Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases." In *Proceedings of the Annual Symposium on Computer Application in Medical Care,* 66. American Medical Informatics Association.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. *The Curse of Recursion: Training on Generated Data Makes Models Forget.* ArXiv Preprint arXiv:2305.17493. arXiv.

Simon, Herbert A. 1960. "The Corporation: Will it be Managed by Machines?" In *Management and the Corporations,* edited by M Ashen and G. Bach. ABC-CLIO.

Smith, Adam. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations.* 1st. London: W. Strahan / T. Cadell.

Solow, Robert M. 1994. "Perspectives on Growth Theory." *Journal of Economic Perspectives* 8 (1): 45–54.

Solow, Robert M. 1956. "A Contribution to the Theory of Economic Growth." *Quarterly Journal of Economics* 70 (1): 65–94.

Soori, Mohsen, Behrooz Arezoo, and Roza Dastres. 2023. "Artificial Intelligence, Machine Learning and Deep Learning in Advanced Robotics, a Review." *Cognitive Robotics* 3:54–70.

Srihari, Sargur N, and Edward J Kuebert. 1997. "Integration of Hand-Written Address Interpretation Technology into the United States Postal Service Remote Computer Reader System." In *Proceedings of the Fourth International Conference on Document Analysis and Recognition,* 2:892–896. IEEE.

Stoughton, Jason. 2023. "Meet 'Coscientist,' Your AI Lab Partner." *Science Matters.*

Tamkin, Alex, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. 2024. *Clio: Privacy-Preserving Insights into Real-World AI Use.* ArXiv Preprint arXiv:2412.13678. arXiv.

Tao, Fei, He Zhang, and Chenyuan Zhang. 2024. "Advancements and Challenges of Digital Twins in Industry." *Nature Computational Science* 4 (3): 169–177.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. "Stanford Alpaca: An Instruction-following Llama Model." *Stanford Center for Research on Foundation Models.*

Toner-Rodgers, Aidan. 2024. *Artificial Intelligence, Scientific Discovery, and Product Innovation.* ArXiv Preprint arXiv:2412.17866. arXiv.

Trajtenberg, Manuel. 2018. *AI as the Next GPT: A Political-Economy Perspective.* NBER Working Paper 24245. National Bureau of Economic Research.

Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–60.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All You Need." *Advances in Neural Information Processing Systems* 30.

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. *Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning.* ArXiv Preprint arXiv:2211.04325. arXiv.

Wang, Dashun, and Albert-László Barabási. 2021. *The Science of Science.* Cambridge University Press.

Wang, Hongyu, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. 2023. *BitNet: Scaling 1-Bit Transformers for Large Language Models.* ArXiv Preprint arXiv:2310.11453. arXiv.

Wang, Yizhu. 2023. *Banks, Credit Unions Testing AI Models for Underwriting in Credit Cycle.* Technical report. S&P Global.

Webb, Michael. 2019. "The impact of artificial intelligence on the labor market." *Available at SSRN 3482150.*

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35:24824–24837.

Weiss, Debra Cassens. 2017. "JP Morgan Chase Uses Tech to Save 360,000 Hours of Annual Work by Lawyers and Loan Officers." *ABA Journal.*

Weizenbaum, Joseph. 1966. "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine." *Communications of the ACM* 9 (1): 36–45.

Whitehead, Alfred North. 1925. *Science and the Modern World: Lowell Lectures, 1925.* New American Library.

Whitehead, Alfred North, and Bertrand Russell. 1927. *Principia Mathematica.* Cambridge, England: Cambridge University Press.

Wiesinger, Julia, Patrick Marlow, and Vladimir Vuskovic. 2024. *Agents.* Technical report. Google. https://readwise-assets.s3.amazonaws.com/media/wisereads/articles/agents/Newwhitepaper_Agents2.pdf.

Wittbold, Kelly A., Carroll Colleen, Marco Iansiti, Haipeng Mark Zhang, and Adam B. Landman. 2020. "How Hospitals Are Using AI to Battle Covid-19." *Harvard Business Review.*

Wooldridge, Michael. 2021. *A Brief History of Artificial Intelligence: What it is, Where we are, and Where we are Going.* London: Flatiron Books.

Yan, Yuchen, Jin Jiang, Yang Liu, Yixin Cao, Xin Xu, Mengdi Zhang, Xunliang Cai, and Jian Shao. 2025. "Ŝ3cmath: Spontaneous Step-Level Self-Correction Makes Large Language Models better Mathematical Reasoners." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 39:25588–25596.

Yegge, Steve. 2011. *Stevey's Google Platforms Rant.* Technical report. GitHub Gist. https://gist.github.com/chitchcock/1281611#steveys-google-platforms-rant.

Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. "Domain Generalization: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (4): 4396–4415.

A Definitions of AI

We illustrate the varied use of the term "artificial intelligence" by discussing four influential definitions. Alan Turing devised a broad, conceptual definition—the "Turing test"—to determine if a system was indistinguishable from humans in 1950. The Dartmouth Project, a seminal meeting for the AI field in 1956, provided a more demanding analytical definition with concrete criteria for what AI must do. In 2020, the *National Artificial Intelligence Initiative* placed a definition into U.S. law with a rather different set of criteria. And, since 2020, the U.S. Patent and Trademark Office has used an algorithmic approach: a large language model (LLM) supplemented by reinforcement learning with human feedback (RLHF) that can classify any technology as AI or not based on similarity to descriptions of eight related areas of science and engineering. Importantly, while the sets of technologies meeting these four definitions have substantial overlap, they are far from identical. Thus, it is crucial to stipulate the scope of analysis when discussing "AI" and its economic effects.[62]

A.1 Alan Turing

Turing (1950), while not offering a definition of artificial intelligence, described a procedure to determine if a machine can imitate human responses well enough that a human interlocutor cannot reliably distinguish between the machine and a human.

This, of course, is the origin of the "Turing test" which is often referenced to gauge effective artificial intelligence. We judge that for most observers, passing the Turing test would be a sufficient condition for a system to be AI, but it isn't a necessary condition in current usage. After all, few people would be fooled into thinking the machine learning-based recommendation engines used on social media sites are human. Moreover, mimicking and replacing humans is not the sole goal of the AI field. Indeed, the adverse social

62. This set of definitions is far from exhaustive. See the discussion in Filippucci et al. (2024) for a definition of scope for AI usefully grounded in a production function framework as well as references to other definitions. The OECD, for example, has codified this definition: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment" (Perset 2024).

consequences of that approach are a concern of some scholars who recommend a focus on using AI to complement human activity instead. Brynjolfsson (2022, 14) observes:

> We can work on challenges that are easy for machines and hard for humans, rather than hard for machines and easy for humans. The first option offers the opportunity of growing and sharing the economic pie by augmenting the workforce with tools and platforms. The second option risks dividing the economic pie among an ever-smaller number of people by creating automation that displaces ever-more types of workers.

Another shortcoming of this definition has been raised by philosophers who have disputed the use of the Turing test to assess whether a machine can think. Searle (1980) offers the counterexample (the "Chinese room argument") of an individual, ignorant of Chinese, passing the Turing Test by using an instruction manual to connect questions posed in Chinese characters to appropriate responses without understanding their meaning.

A.2 The Dartmouth Project

The term "artificial intelligence" can be traced to the summer of 1956, when Dartmouth College professor John McCarthy convened the seminal "Summer Research Project on Artificial Intelligence" (Nilsson 2009). The project proposal stated its premise and objectives and contained an implicit definition of AI (in italics) (McCarthy et al. 2006, 13):[63]

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to *make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.*

63. He chose the term "Artificial Intelligence" to distinguish the field from "automata theory"—a branch of computer science focused on rule-based mathematical models of computation—and "cybernetics"—a field focused on control systems, feedback, and communication in machines and living things.

The Dartmouth AI definition is far broader than the Turing test. Systems that pass the Turing test would surely be included in the scope of the Dartmouth definition by virtue of using language, provided a system need not satisfy all the criteria at once. The project also envisioned systems forming abstractions, anticipating the flexible models found in neural network systems. That is, the training of these models is agnostic about the analytical structure, rather than calibrating a predetermined functional form. And, the idea that AI systems will improve themselves hints at the concept of artificial general intelligence. Last, the "solve kinds of problems now reserved for humans" criterion is likely the vernacular definition many casual observers would provide for AI if pressed to do so, though whether present-day observers would reserve the same set of problems for humans as observers in 1956 is unclear.

A.3 National Artificial Intelligence Initiative

Naturally, as the topic of AI has become a focus of public policy in recent years, the U.S. legal system has required a definition. One is provided in the *National Artificial Intelligence Initiative (NAII)* (15 U.S.C. § 9401(3)):

> The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to
>
> (A) perceive real and virtual environments;
>
> (B) abstract such perceptions into models through analysis in an automated manner; and
>
> (C) use model inference to formulate options for information or action.

Like the Dartmouth definition, this law provides concrete criteria for a system to be AI, but the sets of criteria are rather different. Unlike the Dartmouth definition, the law requires that systems collect observations from the environment. Like the Dartmouth definition, though, the NAII definition requires that AI systems form abstractions, and one can loosely compare the "formulate options for information and action" criterion to the "solve problems" criterion in the Dartmouth definition. However, there is no mention of

language in the government definition, an important part of the Dartmouth and Turing definitions, nor any mention of self improvement.

A.4 The U.S. Patent and Trademark Office

The U.S. Patent and Trademark Office (USPTO), identifies patents that "contain" AI for the Artificial Intelligence Patent Dataset. This exercise requires a definition that provides an unequivocal declaration for each patent, unlike the other definitions discussed above. To that end, the USPTO identified eight "AI component technologies" and trained a large language model to identify patents referencing those technologies using an iterative supervised learning process where subject experts identified examples of AI and non-AI for each technology and reviewed the AI algorithm determinations for accuracy (quoted from Giczy, Pairolero, and Toole 2022, 6–8):

1. **Knowledge processing**: The field of knowledge processing contains methods to represent facts about the world and to derive new facts (or knowledge) from a knowledge base. For example, expert systems generally contain a knowledge base and an inference method to obtain new facts from that knowledge base.

2. **Speech**: Speech recognition includes methods to understand a sequence of words given an acoustic signal. For example, the noisy channel model is a statistical approach used to identify the most likely sequence of words given verbal input using Bayes' rule (Russell and Norvig 2009).

3. **AI hardware**: The field of AI hardware includes physical hardware designed to implement artificial intelligence software. For example, Google designed the Tensor Processing Unit (TPU) to run neural network algorithms more efficiently. AI hardware may include logic circuitry, memory, video, processors, and solid-state technologies. It may also include embedded software that implements other AI component technologies, such as machine learning algorithms.

4. **Evolutionary computation**: Evolutionary computation contains a set of computational methods utilizing aspects of nature and, specifically, evolution (Russell and Norvig 2009). For example, genetic algorithms include methods for selecting algorithm variants through the selection of optimal random mutations by maximizing fitness.

5. **Natural language processing**: Natural language processing contains methods for understanding and using data encoded in human natural language. For example, language models represent probability distributions of language expressions (Russell and Norvig 2009).

6. **Machine learning**: The field of machine learning contains a broad class of computational learning models. For example, supervised learning classification models are algorithms that learn to classify observations based on pre-labeled training data. Machine learning includes, among other techniques, neural networks, fuzzy logic, adaptive systems, probabilistic networks, regression, and intelligent searching.

7. **Computer vision**: The field of computer vision contains methods to extract and understand information from visual input, including images and videos. For example, edge detection identifies the boundaries and borders contained in an image. Additional areas of computer vision include object recognition, manipulation (e.g., transformation, enhancement, or restoration), color processing, and conversion.

8. **Planning/control**: The field of planning and control contains methods to identify and execute plans to achieve specified goals. Key aspects of planning include representing actions and states of the world, reasoning about the effects of actions, and efficiently searching over potential plans. Modern control theory includes methods to maximize objectives over time (Russell and Norvig 2009). For example, stochastic optimal control considers dynamic optimization in uncertain environments. Additionally, planning and control includes data systems for administration/management (e.g., managing an organization and its employees, including inventory, workflow, forecasting, and time management), adaptive control systems, and models or simulators of systems.

Among the concrete challenges faced by this classification effort are what non-software components of AI systems are included. Tensor Processing Units (TPUs), for example, are chips designed to accelerate AI inference, while graphics processing units (GPUs) were originally designed for graphics rendering but have since been re-engineered to accelerate AI model training. As challenging as this question is, the inclusion of hardware is essential for a definition that is conceptually stable over time with respect to the function

Table 9: Distinguishing Features of AI Models

symbolic	vs.	connectionist
deterministic	vs.	stochastic
discriminative	vs.	generative

of the technology; the division of tasks between hardware and software varies over time within functionally identical IT systems (Hennessy and Patterson 2011).

Further distinction by type of AI would be a welcome refinement. Cockburn, Henderson, and Stern (2019) found distinguishing among robotic, symbolic, and neural network AI technologies as useful for their work on innovation. In light of recent developments in the theory and application of AI, identifying patents as related to generative AI would be useful for research as well.

B A Short History of AI

The Dartmouth Summer Research Project, in 1956, is often used as a rough marker of the beginning of the AI field, though the scientists in attendance did not share a theory of what the field entailed (Moor 2006).[64] And, foundational work took place before the Dartmouth project. For example, McCulloch and Pitts (1943) had studied the use of artificial neurons, Shannon 1948 had identified Markov Chains as a potential basis for generating new content, and Turing (1950) had introduced a test for machine intelligence (now known as the Turing test) whereby a human attempts to determine if a hidden interlocutor is a computer or another human.[65]

We sketch subsequent developments in AI theory and application below. As will be apparent, substantial progress on AI preceded the explosion of attention to AI that followed the introduction of ChatGPT in 2022.

64. For more on the conference, see Nilsson (2009), Wooldridge (2021), and Olson (2024).

65. Indeed, Andrey Markov identified language as a use for his mathematical structures as early as 1906 (Markov 2006).

B.1 Early AI Research

Following the Dartmouth project, AI research developed models distinguished along several dimensions (table 9 on the previous page).

- **Symbolic AI** encoded a system of explicit rules in computer programs. For example, Newell, Shaw, and Simon (1958) designed a model called "Logic Theorist" that successfully proved 38 theorems from Whitehead and Russell's *Principia Mathematica* (Whitehead and Russell 1927). **Connectionist AI**, in contrast, allowed complex rules to emerge organically, sacrificing interpretability for flexibility (James et al. 2017). The Perceptron model in Rosenblatt (1958), foundational for this approach, was a single neuron, used to combine signals from multiple input channels to classify images. Later models, such as MADALINE (Multiple Adaptive Linear Neuron), combined multiple layers of neurons together. These networks combine input signals into an output signal with the weight given to each input signal evolving during the training process.

- Though early models were typically **deterministic AI** systems, where a given input would consistently produce the same output, **stochastic AI** models emerged as well where the path taken was not predetermined. For example, SNARC (Stochastic Neural Analogue Reinforcement Calculator), simulated a rat navigating a maze by random experimentation with different paths (Minsky 1952).

- Most early efforts focused on classification—**discriminative AI**—such as using the Perceptron to label pictures. But, the ambition to create a **generative AI** system that would respond to questions with an appropriate free-form text response was already present in this era. In 1966, the ELIZA chatbot provided a rudimentary simulation of a conversation with a psychotherapist. Unlike present-day AI chatbots, ELIZA used a deterministic, symbolic logic approach, relying on pattern matching and word substitution (Weizenbaum 1966).[66]

In addition to these theoretical characteristics, applied AI systems are distinguished by the type of training used in their development. Some use

66. Strictly speaking, some AI models, such as the "expert systems" described below, are neither generative or discriminative, so our classification scheme is not exhaustive.

reinforcement learning, interacting with the environment to refine the model. Others use **predictive learning**, where the system is trained in advance of use. Predictive learning primarily took the form of **supervised learning** in this period, such as when the Perceptron was trained with labeled pictures. However, Selfridge (1958) was a major advance in algorithms for pattern recognition, which was foundational for **unsupervised learning**, where the system develops classification categories without guidance.

Early efforts to apply theoretical models were limited by advances in computing hardware. Greater AI system capability typically requires a larger model, where size is measured in the number of parameters (fig. 6 on page 21). Early models, like Theseus—a robotic maze-solving mouse—and the Perceptron—the rudimentary neuron mentioned above—had tens or hundreds of parameters. Recent models, like DALL-E, Llama, and GPT-3 have hundreds of billions of parameters. Moreover, more complicated models typically require larger training datasets (fig. 6 on page 21). Computational requirements for model training and application rise with the size of the model and the size of the training dataset.

B.2 Emergence of Practical AI

Early practical applications of AI were found in solving classification problems in high-volume communication systems. LeCun et al. (1989) developed the LeNet model adopted by the U.S. Postal Service to read hand-written ZIP codes. The post office was soon reading entire handwritten addresses using AI as well (Srihari and Kuebert 1997). Another early practical application of AI was the identification of spam email (Sahami et al. 1998).

Notwithstanding these advances in the 1980s and 1990s, interest in the connectionist approach, and neural networks in particular, had fallen off with the downbeat assessment of Minsky and Papert (1969). Interest returned with the insights provided by Hinton and Salakhutdinov (2006), who introduced advances in training methods (greedy layer-wise pre-training) and efficient use of large datasets (dimensionality reduction). A key innovation that followed soon after was the **convolutional neural network** (CNN), an advance in connectionist AI that focused on rapid development of a coarse representation of the image which revealed some features, like edges, but not others. Krizhevsky, Sutskever, and Hinton (2012) demonstrated the power of CNN by leaping ahead of other competitors in the *ImageNet Large Scale Visual Recognition Challenge*, a benchmark for computer vision, with their

AlexNet system. Referring to the characteristics described above, these early image systems were connectionist, deterministic, discriminative, and trained by predictive learning. *AlexNet* also demonstrated the value of data augmentation by adding mirror images of training pictures to the training set.

In the news industry in this period, symbolic models such as *Cyborg* at Bloomberg and *Heliograf* at the Washington Post were used to write articles. Unlike present-day generative models, these systems relied on structured data—tagged as sports scores or stock prices, for example—not on models of the language as a whole, making them a kind of proto-generative AI. In 2013, articles on major company financial announcements and sports events were generated with these systems and by 2017, these models were used for expanded coverage of sparsely populated news markets and small companies, and were even used to generate rudimentary news videos (Keohane 2017).

Symbolic AI was put to practical use in this period as well. **Expert systems** leveraged a large trove of domain-specific information and a set of rules (encoded in an "inference engine") provided by specialists to provide guidance, such as medical diagnoses (Buchanan and Smith 1988). Examples include MYCIN, used to diagnose infectious diseases, and IBM's Watson, deployed in medical and other applications (Shortliffe 1977; Ferrucci 2012). Expert systems fell out of favor over time due to their cost of development, limited reliability, and narrow field of application (Gill 1995).

Most importantly, as emphasized by Agrawal, J. S. Gans, and Goldfarb (2019, 32), AI practitioners soon realized that discriminative AI could be recast as "prediction in the statistical sense of using existing data to fill in missing information" and these models were soon used in a diverse set of prediction problems. (Indeed, these systems are often now referred to as "predictive AI.") Amazon, for example, first used AI to forecast demand for its products in 2009, then continuously updated its forecasting approach to adopt emerging AI techniques (table 10). **Machine learning**, another term for this approach to AI, is credited by Amazon and other major IT companies with increasing profitability during this period (Bresnahan 2019).

B.3 Generative AI

Public interest in AI surged in late 2022 with the appearance of ChatGPT, a user interface for a viable genAI system—one that can respond to natural language prompts with human-like (coherent, nuanced, context-specific) responses in the form of text, images, videos, and sounds. The event was

Table 10: History of Amazon Demand Forecasting Techniques

Year	Forecasting Technique
2007	Time-Series Models
2009	Random Forest
2011	Seasonality Models
2013	Sparse Quantile Random Forest
2015	Feed-Forward Networks
2017	Multi-Horizon Quantile Recurrent Forecaster
2020	MQ Transformer

Source: Reproduced from Hardesty (2019)

the culmination of a roughly 10-year period of advances in **large language models** (LLMs).

LLMs are a mathematical representation of the linguistic relationships among the "tokens" (words, groups of words, and portions of words) found in a "corpus" (set of texts or other media). A key advantage of LLMs is their ability to reduce unstructured text to flexible structural representations that do not rely on a small set of variables specified in advance. Rudimentary early attempts at language models encoded words as long vectors of zeros with a single element—the index assigned to the specific word—marked with a "1" (Hancock and Khoshgoftaar 2020). In 2013, Google introduced a richer method known as the Word2Vec model (Mikolov, Sutskever, and Le 2013). Words are represented with dense vectors known as **embeddings** such that the distance between two embeddings reflects the semantic similarity between the represented words. Word2Vec revolutionized many **natural language processing** (NLP) tasks, such as classification and translation.

A limitation of Word2Vec encodings is that a word's representation is the same regardless of the context. For example, the model will represent the word "bank" with the same embedding whether the input text is "I withdrew money from my bank account" or "I went fishing down at the river's bank." Since Word2Vec assigns fixed embeddings, it cannot distinguish whether "bank" refers to a financial institution or a riverbank, impeding its understanding of the text.[67]

67. Computer scientists have wrestled with this word sense disambiguation problem since the 1950s. Bar-Hillel (1960) in discussing the prospects for fully automatic high-quality translation, offered this assessment: "What such a suggestion amounts to, if taken se-

A major breakthrough in addressing this shortcoming came with the introduction of the Transformer model, an architecture that creates context-aware word representations by efficiently processing the entire input text at once (Vaswani et al. 2017). This architecture is defined by two principal characteristics: the **attention mechanism** and **positional encodings**. The attention mechanism enables the model to assign, for each token, varying degrees of relevance to different parts of the input, allowing it to understand context in longer sequences of text. Positional encodings ensure that word order is meaningfully integrated into the model's processing of inputs. (See the box, "Landmark AI Models: The Transformer," for more detail.) This advancement changed how machines encoded text, shifting the focus within NLP toward a deeper understanding of language.[68]

With the advent of the Transformer, genAI flourished. Most notably, text generation improved at a blistering pace in this period, though NVIDIA's generative adversarial network (GAN) (2018) and DALL-E (2021) were striking advances in image generation; audio generation was dominated by WaveNet (2016), a different architecture, for a time, but eventually the Transformer approach was used for speech generation as well. Especially prominent were a series of models produced by OpenAI: GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020), ChatGPT (OpenAI 2022), GPT-4 (OpenAI 2023), and others. Successive models were increasingly human-like in their knowledge and creativity, eventually passing Turing tests due to their coherent and contextually relevant output.

riously, is the requirement that a translation machine should not only be supplied with a dictionary but also with a universal encyclopedia. This is surely utterly chimerical and hardly deserves any further discussion." It appears we now have such a universal encyclopedia in hand.

68. Particularly important was the introduction of the BERT model the following year (Devlin 2018). The final 'T' in BERT stands for 'Transformer' (bidirectional encoder representations from transformers)