Generative AI at the Crossroads: Light Bulb, Dynamo, or Microscope? === *preliminary and incomplete* ===*

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

March 20, 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. GenAI-induced productivity gains will depend on the range of tasks enhanced by the technology, the extent of adoption for those tasks, and the success of firms in integrating the adopted technology; it is too early to draw confident conclusions on those questions. Moreover, the effects of the technology on the innovation process will play a crucial role. 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. 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 are (1) widely adopted, (2) spur abundant knock-on innovations (new goods and services, process efficiencies, and business reorganization), and (3) improve continuously, refreshing this innovation cycle; the electric dynamo is an example. Second, there are inventions in the method of invention (IMIs). IMIs increase the efficiency of the research and development process, generating new ideas more quickly and cheaply; the compound microscope is an example. We show that GenAI has the characteristics of both a GPT and an IMI—an encouraging sign. Even so, for genAI to boost productivity growth, its contribution will have to outpace the past IT innovations which are baked into the trend, including machine learning.

---

* *Preliminary and incomplete. Please do not circulate without permission.* 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, Shane Greenstein, Anton Korinek, and James Manyika for helpful conversations.

1 Introduction

The scope of artificial intelligence (AI) has expanded dramatically with the advent of generative AI (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 (Maslej et al., 2024).[1] Nevertheless, the prospective effect of recent genAI developments on labor productivity remains uncertain. The external validity of 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). And, human-level performance on benchmarks does not guarantee that cost-effective practical application at scale in a business context is possible. There are only a few examples of genAI improving the bottom line in practice, in contrast to the abundance of examples available for machine learning, an AI technology that preceded genAI.

In an encouraging sign, field test evidence of productivity improvements from genAI in practical applications has emerged, including for writing, computer programming, and responding to call center inquiries (table 1 on page 5).[2] Whether these and other applications will collectively be consequential enough to raise aggregate productivity growth remains to be seen. GenAI may simply represent the latest evolutionary stage of information technology (IT)—one in a long line of innovations in hardware and software, including its predecessors in the field of AI. Because these earlier innovations are reflected in the historical trend, genAI will have to contribute more than the typical IT innovation to raise productivity growth. While optimists see potential for the return of the IT-fueled productivity boom of the late 1990s and early 2000s, more downbeat observers see some claims about its capabilities as overstated and highlight potential headwinds, such as regulation to guard against unintended harms from genAI.[3]

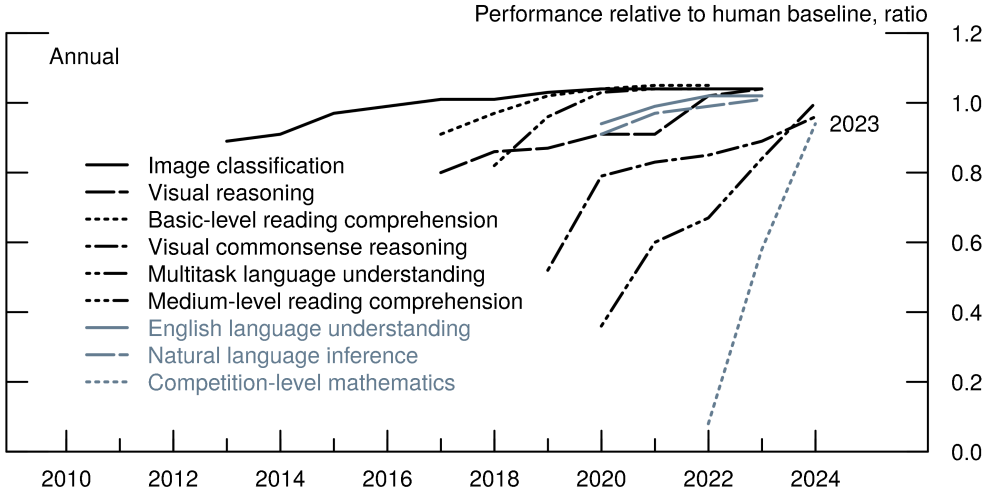
[1] The reader is encouraged to consult Maslej et al. (2024) for a far more detailed set of indicators for progress in AI.

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

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

In our view, the evidence supports optimism about the future productivity contributions from genAI. That said, caution is in order. Predictions of an IT-infused future of abundance have appeared regularly since the 1950s, but noteworthy productivity gains have only followed slowly, in tandem with time-consuming complementary changes in business organization.[4]

Figure 1: GenAI Benchmark Performance



Note: Human baseline is the peformance of [the average human?].
Source: Reproduced with permission from the 2024 AI Index Report, Stanford Institute for Human-Centered Artificial Intelligence.

One way to frame the forward-looking question of the effect of genAI on productivity is to consider 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 continually refresh this adoption cycle through ongoing improvement in the core technology. The boost to

[4]This reality was apparent from the outset of the digital age, which arguably began in the 1940s when a steady, rapid fall in the price of information processing began (Nordhaus, 2007). Martin Jr. (1960) cautioned that productivity gains would come slowly: "The data-processing system of an organization is of almost un-imaginable complexity. The introduction of a computer usually involves widespread changes in this complex system."

productivity growth for these general-purpose technologies (GPTs) persists for longer. Yet other inventions, such as the compound microscope, increase the efficiency of the research and development process; these "inventions in the method of invention" (IMIs) yield a sustained increase the productivity growth rate by lowering the cost of research and development.

We consider the evidence for genAI as a GPT and an IMI. To assess its status as a GPT, we consider indicators for the scope of diffusion, extent of knock-on innovations, and signs of ongoing progress in the core technology. To assess its status as an IMI, we review the record of genAI contributing to R&D efficiency. For both the question of genAI's status as a GPT and an IMI, we reference case studies for the financial, health care, and information sectors and for the electricity generation industry (Baily and Kane, 2025a,b; Kane and Baily, 2025a,b).

We conclude there is substantial evidence that genAI is both a GPT and an IMI, with an important caveat: The current and prospective extent of genAI diffusion is highly uncertain; adoption outside of large firms is low and evidence for the profitabililty of genAI use at large firms is slim.[5] Moreover, it is premature to say that it will boost productivity more than previously-introduced forms of IT—including earlier forms of AI like machine learning—the condition that must be met for a genAI-fueled sustained increase in productivity growth.

Before turning to the subjects of GPTs and IMIs, we sketch the evolution of AI in general and genAI in particular. That history makes clear that the field of AI is diverse, encompassing far more than just genAI, so coherent claims about the effects of "AI" on productivity must specify what kind of AI is under discussion; we focus on genAI. Moreover, the history of innovation in the field is crucial to understanding the *relative* contribution of genAI.

2 The Evolution of Artificial Intelligence

"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

[5]For other examinations of AI and its potential to be a GPT, see, *inter alia*, Trajtenberg (2018); Bresnahan (2019); Eloundou, Manning, Mishkin and Rock (2023); Goldfarb, Taska and Teodoridis (2023); Bresnahan (2024).

Table 1: Selected GenAI Productivity Field Studies

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

models.[6] AI systems achieve these objectives by constructing mathematical models to mimic complex patterns found in training data.

The term 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).[7] The project proposal stated its premise and objectives and contained an implicit definition of AI (in italics):

> 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.*

To be clear, AI research had begun 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.[8]

We sketch subsequent developments in AI research and application below. As will be apparent, substantial theoretical and practical progress on AI

[6] Artificial intelligence is notoriously difficult to define in a way that satisfies the diverse perspectives of academia, the tech industry, and everyday users and the meaning continually evolves alongside technological advancements. U.S. legislation (15 USC 9401) contains this definition: "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."

[7] For a more comprehensive history, see Nilsson (2009), Wooldridge (2021), and Olson (2024).

[8] Indeed, Andrey Markov identified language as a use for his mathematical structures as early as 1906. From Wikipedia: "Markov chains have long been used to model natural languages since their development by Russian mathematician Andrey Markov in the early 20th century. Markov published his first paper on the topic in 1906, and analyzed the pattern of vowels and consonants in the novel Eugeny Onegin using Markov chains. Once a Markov chain is learned on a text corpus, it can then be used as a probabilistic text generator." Shai; Singer, Yoram; Tishby, Naftali (July 1, 1998). "The Hierarchical Hidden Markov Model: Analysis and Applications". Machine Learning. 32 (1): 41–62.

Table 2: Distinguishing Features of AI Models

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

preceded the explosion of attention to AI that followed the introduction of ChatGPT in 2022.

2.1 Early AI Research

Following the Dartmouth project, AI research developed models distinguished along several dimensions (table 2).

- **Symbolic AI** encoded a system of explicit rules in computer programs. For example, Newell et al. (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).[9]

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

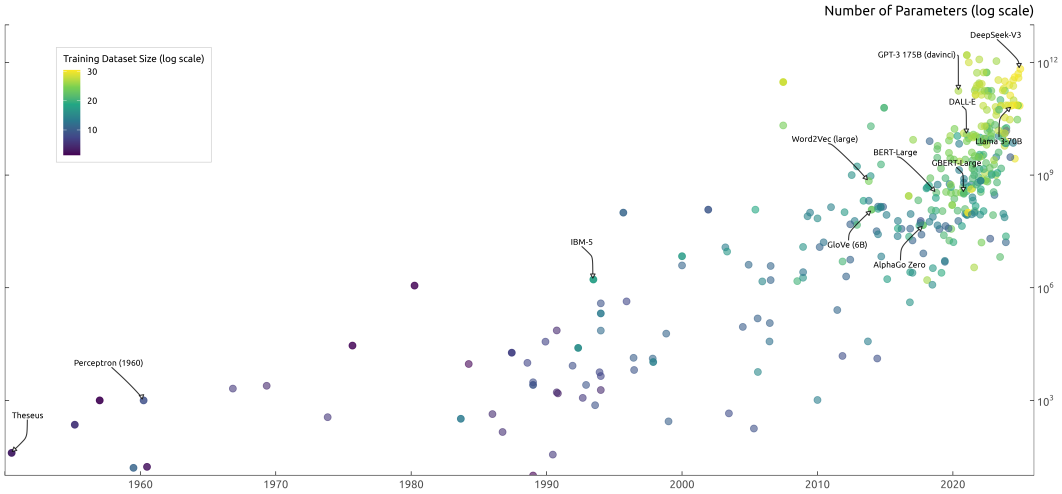
Early efforts to apply theoretical models were limited by advances in computing hardware. Greater AI system capability typically requires a larger model, where size is measured in the number of parameters (fig. 2 on the next page). Early models, like Theseus—a robotic maze-solving mouse—and the Perceptron—the 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. 2 on the following page). Computational requirements for model training and application rise with the size of the model and the size of the training dataset.

2.2 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). A key innovation behind these systems 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,

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

Figure 2: 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.

but not others. Referring to the characteristics described above, these early image systems were connectionist, deterministic, discriminative, and trained by predictive learning.

In the news industry in this period, symbolic models such as *Cyborg* at the AP 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).

Table 3: 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)	

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, Gans and Goldfarb (2018), AI practicioners 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 3). **Machine learning**, another term for this approach to AI, is credited by a wide range of companies for increasing profitability during this period (table under construction).

2.3 Generative AI

Public interest in AI surged in late 2022 with the appearance of ChatGPT, a front end to 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 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 et al., 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 **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.

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 Transformer Model" for more detail.) This advancement directly changed how machines encoded text, shifting the focus within NLP toward a deeper understanding of language.[10]

With the advent of the Transformer, genAI flourished. Innovations such as WaveNet (2016) improved raw audio and speech generation. NVIDIA's GAN (2017) and DALL-E (2022) enhanced image generation. Most notably, text generation improved at a blistering pace in this period. Especially prominent in this domain were a series of models produced by OpenAI: GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), ChatGPT (OpenAI,

[10]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)

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.

Unlike machine learning, however, few companies have reported a noteworthy impact on their bottom line from genAI (table under construction).

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 input data simultaneously, 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 text suitable for input to the model is created. Specifically, the input text is broken into smaller pieces (tokens), which may be phrases, words, or parts of words. The tokens are converted into numerical vector representations (embeddings) 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 input matrix using three large matrices called the "query," the "key," and the "value." For each token in the input, the query predicts which token will come next, picking from the choices available in the key. The value vector represents the correct answers and the distance (miss) between the prediction and the value is feedback used to refine the query.

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.

The Transformer (continued)

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 vast size of advanced transformer models, with their extensive weight matrices in the attention and MLP components, highlights their ability to handle complex language tasks and drives their scalability as demonstrated by recent breakthroughs in genAI.

3 Is GenAI a General Purpose Technology?

While the limited evidence of genAI-driven productivity in specific tasks is intriguing (table 1 on page 5), to look ahead to its 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).[11] 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 4 on the following page.

As described by Lipsey, Carlaw and Bekar (2005), "big GPT shocks change almost everything in a society and revitalize the growth process by creating an agenda for the creation of new products, new processes, and new organizational forms." For example, new products that followed the invention 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

[11] Note that the "GPT" acronym 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 4: Selected Examples of General Purpose Technologies

Technology	Date
Domestication of plants	9000–8000 BCE
Writing	3400–3200 BCE
Iron	1200 BCE
Waterwheel	Early medieval period
Three-masted sailing ship	15^{th} century
Steam engine	Late 18^{th} to early 19^{th} century
Factory system	Late 18^{th} to early 19^{th} century
Railway	Mid 19^{th} century
Internal combustion engine	Late 19^{th} century
Electricity	Late 19^{th} century
Motor vehicle	20^{th} century
Mass production, continuous process factory	20^{th} century
Computer	20^{th} century
Lean production	20^{th} century
Internet	20^{th} century
Biotechnology	20^{th} century
Source: Reproduced from Lipsey, Carlaw and Bekar (2005), p. 132.	

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 gain from a single invention is limited. Productivity *growth* will be higher during the adoption transition, but will return to its underlying trend when diffusion is complete. The seminal work on this classical growth effect is 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. Lighting is necessary for nearly all human labor, so the potential effect on the level of productivity from the light bulb was large, but when the light bulb market became saturated, it delivered no further (direct) pro-

ductivity gains and the increment to productivity growth present during the transition disappeared.

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. The electric dynamo is an example. Dynamos were a replacement for existing systems that used a prime mover—steam or water power—to drive a set of belts that conveyed power to multiple machines. The dynamo was a more efficient source of power, so the replacement of an existing system yielded productivity gains. In addition to that direct productivity effect, the dynamo enabled a more flexible approach to the organization of production (David, 1990). The less centralized factory designs adopted by firms in response to this flexibility are themselves a productivity-ehancing innovation spurred by the dynamo.

Ongoing Core Innovation When a technology continues to improve over time, the new productivity level—fixed in the Solow (1956)—becomes a moving target. Ongoing innovation translates into greater technical performance at a lower cost, a form of productivity gain. Moreover, the lower price of the technology's service that typically follows the production cost reduction spurs greater adoption.[12] Solid state electronics is an example. Relentless increases in the number of transistors on a semiconductor has driven the price of computing lower, making it cost-effective both to embed electronics in more and more devices (like inexpensive toys) and to adopt devices with more and more electronic capability (like smartphones).

3.1 Diffusion

Widely cited analyses of the text of job descriptions support the apparent applicability of AI for a broad range of workplace tasks and suggest that the potential for diffusion among firms is high.[13] However, surveys find that AI adoption remains concentrated in large corporations and that a large

[12] A lower price need not follow if a firm with market power over the innovation raises markups.

[13] See Acemoglu et al. (2020), Brynjolfsson et al. (2018), Felten et al. (2019), Webb (2019), Eloundou et al. (2023).

majority of firms don't see an application for AI in their business. The share of job postings mentioning AI skills is modest, indicating that firms are taking a cautious approach to AI use. However, from the worker's perspective, adoption seems widespread; surveys of individuals document that a large share of workers are already using genAI.

Adoption surveys An array of AI adoption surveys are available and distilling a single message about AI use from them is difficult. In great part, the apparent contrasts among them can be rationalized by referencing differences in the definition of adoption and the nature of the sampling used.

Results from firm-level surveys of AI adoption seem wildly different at first glance. The Census Bureau's Business Trends and Outlook Survey (BTOS) finds roughly 7% of firms are using AI, while McKinsey reports that 72% of firms do (fig. 3 on the next page). On closer inspection, these surveys are consistent with one another and reveal important nuances in the state of AI adoption.

Differences in the firm size distribution of their samples help to rationalize the apparent tension between the Census and McKinsey surveys. The BTOS is a representative sample of 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). The BTOS finds that smaller firms are significantly less likely to adopt AI: 5% of Firms with fewer than 5 employees were using AI at the time of the early-2024 survey versus 11% of firms with 250 or more employees. Importantly, because the BTOS threshold is 250 employees for the largest firm-size group, even the figures for that group are not strictly comparable to the McKinsey survey; the U.S. firm size distribution is highly skewed and large corporations typically have thousands of employees (Kondo et al., 2023).

Differences in the definition of adoption between the surveys are relevant as well. 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.)"; and, the McKinsey survey asks respondents if they "use AI in at least one business function." The McKinsey question is rather less restrictive with respect to recency and business function for responding in the affirmative than the BTOS, which asks about use in core business functions

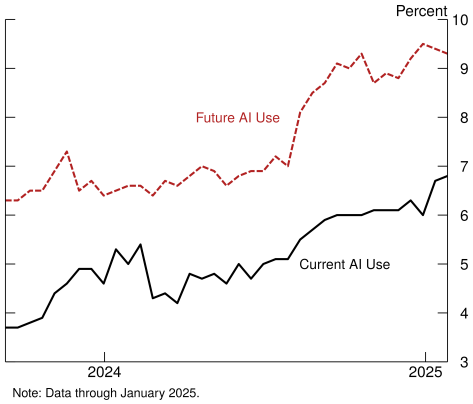
("producting goods or services") in *the last two weeks*. Indeed, in their 2025 survey, McKinsey found that the business function with the greatest use of AI was "marketing and sales."

All told, firm-level adoption surveys suggest that AI adoption is low except among large firms and that AI use may be less prevalent in the core business functions than in support functions.

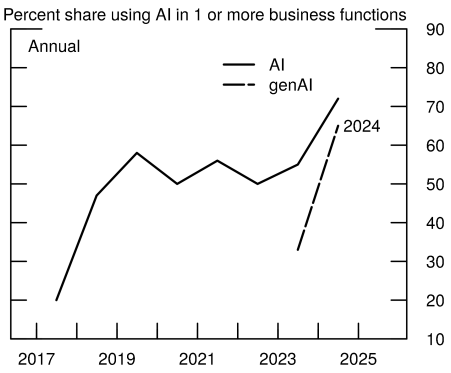
Adoption figures from user surveys suggest fairly widespread AI use. Bick, Blandin and Deming (2024) survey a representative sample of 18-64 year olds in the United States and find 40% of workers use genAI.[14] As this is a representative sample, it is most directly comparable to the BTOS and seems to indicate significantly more prevalent AI use than the Census survey. As posited in Crane, Green and Soto (2025), high-level management in firms may underestimate the extent to which their employees are using AI tools, or AI users may be highly concentrated in large firms or AI-intensive industries. Consistent with the latter idea, Sergeyuk et al. (2025) survey programmers and find 86% AI usage.

Figure 3: AI Adoption Over Time

(a) Census BTOS



(b) McKinsey



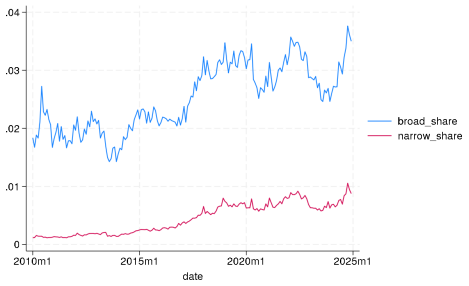
Note: For BTOS, respondents were asked about AI use in producing goods or services during the past two weeks and anticipated in the next six months. McKinsey states that "participants representing the full range of regions, industries, company sizes, functional specialties, and tenures."
Source: Census Bureau, Business Trends and Outlook Survey; McKinsey, "The State of AI in Early 2024."

[14] Our focus is primarily on the United States, but we note that Humlum and Vestergaard (2024) survey Danish workers from 11 industries and find 36% uptake of ChatGPT.

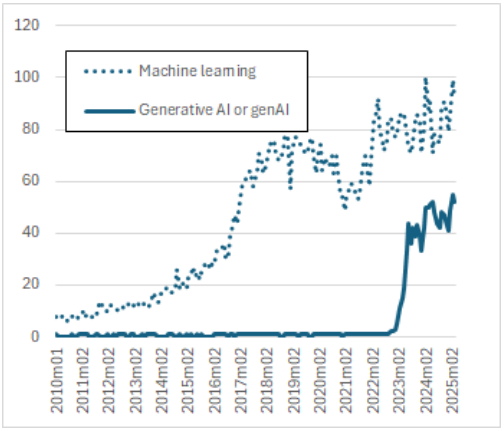
Job postings Job postings referencing AI and related technologies rose in the late 2010s along with the surge in interest in machine-learning and large-language models. Lightcast (formerly Burning Glass) job posting data using search terms associated with AI by Acemoglu, Autor, Hazell and Restrepo (2020) and extended with terms not prevalent at the time of their analysis (such as "genAI") bound the share of job posting currently related to AI between roughly 1 percent (narrow definition) and roughly 4 percent (broad definition including a cluster of related skills). Importantly, that share moved up (from no more than 2 percent) well before genAI became prevalent. In fact, the series show a jump in postings around the time that "machine learning" surged in prevalence in Google searches. AI-related job postings have moved up modestly since then, roughly corresponding with the rise in prevalence of "generative AI" and "genAI" in Google searches, though the apparent impact of genAI is much smaller than the apparent impact of the earlier wave of AI employing machine learning for discriminative and predictive tasks.

Figure 4: AI-Related Job Postings and Google Search Activity

(a) Job Postings



(b) Search



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

Case study evidence

Sahni et al. (2023), in their overview of research on AI use in **healthcare**, conclude that AI has the potential to assist with an array of tasks including diagnosis, treatment choice, and managing records. A majority of physicians already use or are planning to use genAI for generating chart summaries, creating discharge instructions, and an array of other tasks (Research, 2025). Other AI technologies are in use in some medical fields as well: Machine learning was used by 30% of radiologists as of 2020 (Allen et al., 2021). Since then, 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 tracking, report generation), and radiologists (e.g. image quality assessment, diagnosis) (Burnside et al., 2025). Notwithstanding these inroads, AI use faces substantial hurdles to adoption in healthcare. For example, 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, Sahni et al. (2023) conclude that for hospitals, AI adoption often cannot justified on financial factors alone.

The **information sector** has also adopted genAI rapidly. Among U.S. programmers, 92% were using AI tools as of June, 2023 (Shani and GitHub Staff, 2023). Other occupations within the information sector use genAI in their work as well. 69% of graphic designers/illustrators use genAI in their work, employing genAI tools like generative fill and large-scale text-to-image models such as DALL-E. 75% of customer service leaders say they are using some form of AI technology in their daily tasks. GenAI is very beneficial for customer service representatives' productivity, which, by one study, improves number of issues resolved per hour by 14%, on average.

In **finance**, there has been an explosion of interest in using AI, as evidenced by very rapid growth in the number of articles on the topic (Baloo et al. 2024) The impact of AI on productivity is harder to pin down. Much of the research on AI in finance involves developing new algorithmic trading models designed to get ahead of other traders, largely a zero-sum effort (OECD 2021). There are more productive efforts, such as the following. In portfolio management, finance companies can, at low cost, create client-specific portfolios to match individual preferences (BlackRock Advisor Center, 2025). AI is good at detecting anomalous patterns, which assists in fraud detection and prevention (Infosys BPM Analytics, 2025). Financial companies answer many customer queries and can cut costs by using AI to improve

the existing automated response systems (McKinsey Company, 2025). Regulatory compliance and reporting is a major activity for banks and AI is being developed to assist this activity (ABA Banking Journal, 2024). AI is also being applied to predict loan failure rates, improving loan underwriting and credit scoring, although with the risk of introducing bias (Wang 2022). Other AI efforts include data analytics and insights (Hamberg, 2024), and back-office operations and automation (Prakash, 2024). Some of these functions were in use prior to the development of genAI but are being improved as the technology is developed.

In **electricity generation and distribution**, machine learning has been used for many years to assist with matching energy supply, or the amount of electricity produced by generators, with energy demands of industrial, commercial, and residential users. Load forecasting utilizes machine learning to predict how energy demand will change over a short-term or long-term period. In addition, predictive maintanence and vegetation management utilize machine learning to predict what parts of power plants are most likely to fail and where power lines are likely to be affected by vegetation. These capabilities have expanded with genAI, as it allows utilities to use more data in the prediction processes. Utilities are now increasingly experimenting with genAI, with approximately 33% piloting genAI in their operations. Some companies have already begun using genAI in their load forecasting processes, while others have used the technology to simulate equipment degradation.

3.2 Knock-on Innovation

In the course of adopting new technologies, firms retool, retrain, and reorganize to better exploit their productivity potential. This process involves knock-on innovation in products, production processes, and operations. In the case of genAI, related products include 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 changes include centralized data governance and optimization of supply chains and data centers.

3.2.1 Products

User interfaces (UIs) provide a channel through which requests and responses can pass between the model and the user, 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 conversation-style "chatbot" that made AI 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 as as LegalGPT for legal matters. Advances in audio-to-text further enhanced AI-user interactions. In the summer of 2024, OpenAI announced integration of their ChatGPT model to Apple's Siri voice assistant. Google launched their NotebookLM in September 2024, which made it easy to upload documents and transform them into interactive discussions, pushing the boundaries for AI-generated human-like conversations. In addition, there are "copilots" that integrate AI into existing user workstreams, notably GitHub Co-pilot (computer programming) and Microsoft 365 (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.[15] Access to AI through SDK helps the robot with environmental integration problems, such as simultaneously tracking its location and mapping its environment (SLAM).

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).[16] Examples include AutoGen, from Microsoft, and CrewAI.[17] 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 preferences with the best travel options, making the booking, adjusting to disruptions to the reservation process (e.g. sporadic price increases), and including special accommodations along the way (e.g. upgrading seats if available within the

[15]Robot capabilities have advanced in tandem with AI progress. Theseus, a symbolic AI robot mouse that navigated mazes was developed in the 1950s (Wallace, 1952). Industrial robots were trained using machine learning beginning in the 1990s Arinez et al. (2020); Soori et al. (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).

[16]For more detail on agents, see Wiesinger et al. (2024).

[17]For an array of agentic AI use cases, see Hall (2024).

user's preferences).

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) 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. Choi et al. (2024) describe a custom interface to genAI designed to assist control room operators in balancing supply and demand on electrical grids.

3.2.2 Production processes

under development

3.2.3 Organization

under development

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 in the scale of the model, computational power used ("compute", in the lingo of AI), and the size of the training dataset. 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.

3.3.1 Model Development

Since the introduction of the Transformer, model development has progressed at a blistering pace, including ramping up model scale, dataset size, and compute employed—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. 2 on page 9).

The training of genAI models (optimal calibration of its parameters) takes place 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 in fine-tuning and inference, as discussed below.[18]

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).[19] 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 et al., 2023; Abdin et al., 2024).

VIllalobos

[18] "AI scaling laws are showing diminishing returns forcing AI labs to change course" Techcrunch, 11/20/2024/

[19] In particular, Mamba estimates the parameters of a latent state space structure.

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). 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 above.

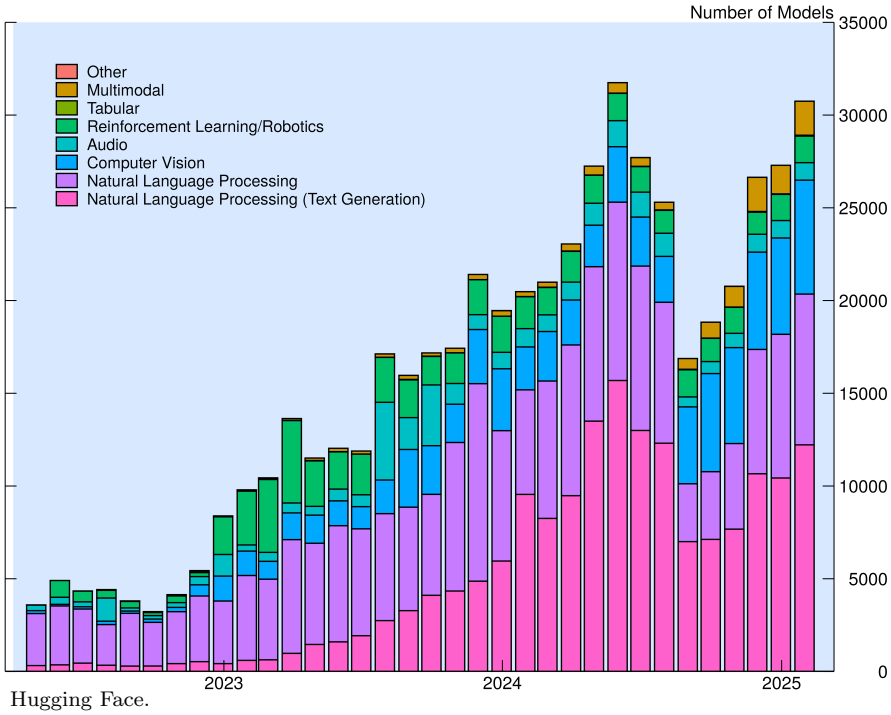
Remarkably, the cost of training a model of a given size was halved approximately every eight months through 2024 (Ho et al., 2024). But, by 2022, the direction of innovation had begun to shift; scaling laws indicated diminishing returns to model size for foundation models, and focus turned to optimizing 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 (the number of tokens).

Fine-tuning As the returns to 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 Llama model, refined by adding a set of instructions and desired outputs to the training process (Taori et al., 2023).[20] **Reinforcement learning** has been extensively applied in AI for fine-tuning, allowing the model to refine

[20]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."

Figure 5: Open-Source AI Models



Source: Hugging Face.

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)** (Christiano et al., 2017), a technique that aligns the model's output with human preferences by learning explicitly from human feedback. 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). This innovation directed fine-tuning efforts towards a framework that made genAI models more controllable and aligned with human interests.

Inference Because of the high volume of inference requests—leading to variable costs in terms of electricity, time, compute, and carbon emissions—

techniques to make this step more efficient have arisen.[21] 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, Jordan, Nowlan and Hinton, 1991). A router determines which "expert" (smaller 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). **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. 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).[22] The recently introduced DeepSeek R1 model leverages several of the techniques described above to deliver a substantial performance improvement compared to existing models (See "Landmark Models: DeepSeek AI").

In some cases, recent efforts have acted to extended inference time to enhance performance. **Chain-of-thought (CoT) reasoning** guides the LLM to articulate a series of steps in its inference, constraining the model to provide a response grounded in logic (Wei, Wang, Schuurmans, Bosma, Xia, Chi, Le, Zhou et al., 2022). Extending inference time raises costs, but on the margin, resources may be better used for responding to queries than for

[21] After setting a flat fee for access to ChatGPT, OpenAI CEO Sam Altman remarked on *X*, "we are currently losing money on openai pro subscriptions! people use it much more than we expected!" (**?**) With the price of the marginal query set to zero, model operators have to confront the possibility that revenue will not cover costs.

[22] 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 https://venturebeat.com/ai/new-llm-optimization-technique-slashes-memory-costs-up-to-75/).

additional training. OpenAI's recent o1 model exemplifies the benefits of this approach via its superior performance across various domains, particularly reasoning-heavy tasks.[23]

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, 2025).[a] 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 $6 million.[b] 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 million for NVIDIA as DeepSeek had achieved leading edge performance with far less of NVIDIA GPU-provided compute than previously believed necessary. 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.[c]

[a] R1 was not the first model from DeepSeek. The company had previously released three major versions: V1 (67B parameters) in November 2023, V2 (236B) in May, 2024, and V3 (671B)—the foundation model for R1—in December, 2024.

[23] OpenAI scientist Noam Brown stuns TED AI Conference: '20 seconds of thinking worth 100,000x more data'

DeepSeek R1 (continued)

[b]See the technical report for DeepSeek V3 (DeepSeek, 2024): "...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."

[c]DeepSeek had opted to initially build a model, dubbed R1-Zero, using purely reinforcement learning. This model was highly effective but suffered from notable setbacks, predominantly language mixing and poor readability.

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 **?**:

> 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.

(Google case studies) uses the example of booking travel, exploring the programming required. While managing excutive travel needs was formerly a secretarial task, most executives now manage their own travel, using the internet to locate flights, a time-consuming and often frustrating activity. An AI agent, programmed to incorporate the travel policy of the company and the preferences of the traveler, will search airline websites to gather the latest flight information, present the user with a potential flight plan for approval—produced with a large language model—and book the flight.

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 healthcare, we noted that LLMs are being used to help with the paperwork burden faced by healthcare 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, they require programming and testing and cannot simply apply an off-the-shelf AI program. (OpenAI,2016) highlights the problem of correctly specifying the objective function for the AI agent, which may inadvertently be guided to maximize its programmed reward while subverting the real objective of the user. Moreover, in extreme cases bad actors may hijack agents in business or military uses, leading to disastrous outcomes.

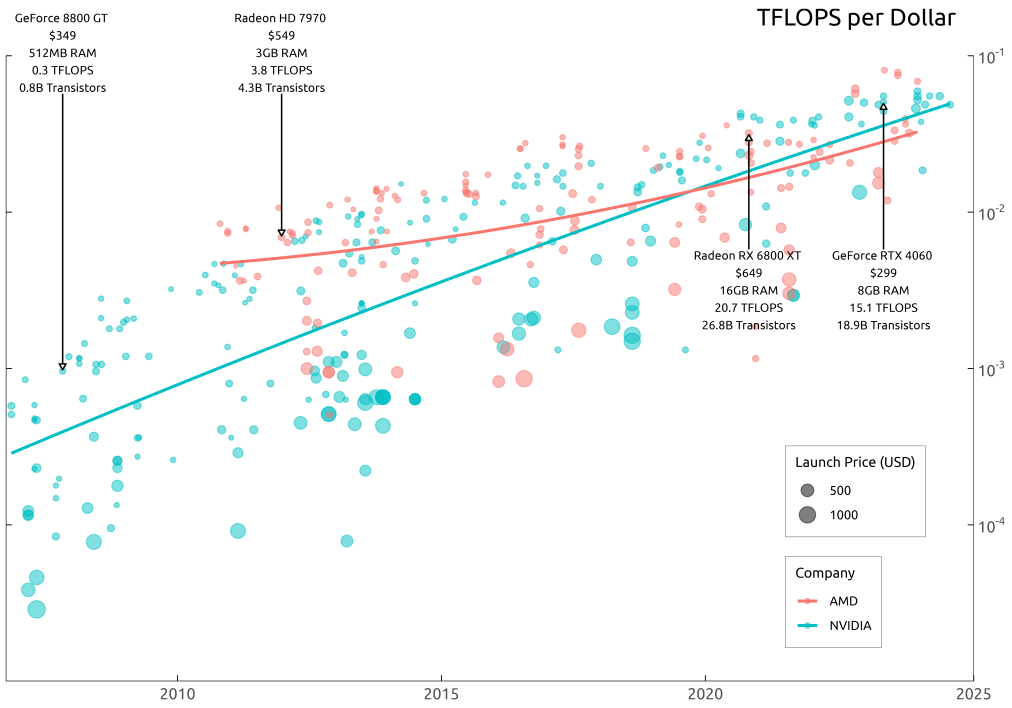
3.3.2 Hardware

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 as well.

GenAI processing relies heavily on graphics processing units (GPUs). Like the image processing tasks GPUs were 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 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.[24] For example, tensor processing units (TPUs) are customized for matrix multiplication, heavily used in neural networks. They consisting 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.

[24]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.

Figure 6: TFLOPS per Dollar



Note: This figure shows the number of TFLOPS (trillion floating-point operations per second) per dollar for GPUs over time. Each circle represents a GPU, with the corresponding year of release on the X-axis. The Y-axis, on a logarithmic scale, shows the number of TFLOPS divided by the cost of the GPU upon release. Orange dots denote GPUs released by AMD, while blue dots denote those from NVIDIA. The size of each dot represents the launch price of the GPU. 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.

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.[25] CUDA, first introduced in 2007, explicitly supported non-graphics computing; prior to CUDA, programmers

[25]This discussion was substantially supported by a discussion with Claude GPT and research on Wikipedia.

Table 5: Price of GPU Compute

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

Source: TechPowerUp.

were required to reframe computations in terms of graphics operations.[26] As 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 annual rate of price decline of 24% that persisted for 17 years (table 5). As shown in fig. 6 on the preceding 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 flexibile 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.[27]

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

[27] 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.

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 leading edge chips were reduced by 30 percent with each generation (node), yielding a $(0.7 * 0.7 =)$ 50 percent reduction in the space they occupied on a chip every two years (fig. 8b on page 34). 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 stepdown in the growth of TFLOPS per dollar in fig. 6 on page 30 around that year 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. 8a on page 34).[28]

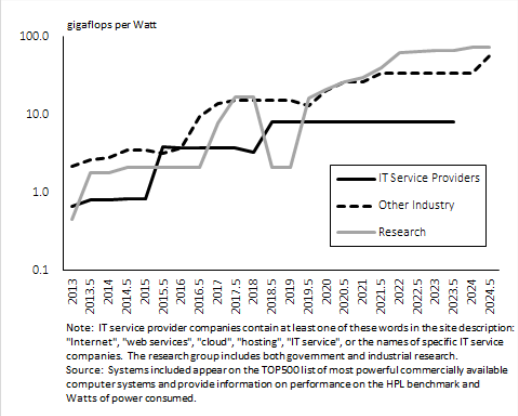
The proximal cause of the miniaturization slowdown was the end of a regularity known as "Dennard scaling" whereby power usage was largely unchanged even as more electronic activity was squeezed into the same surface area. Because heat generation is increasing in power usage, cooling cost began to rise with each node. Consequently, development efforts in the computing sector shifted away from miniaturization and toward power consumption. The energy efficiency at the frontier of research computing has risen roughly 100-fold since 2013 (fig. 7 on the next page). The looming problem of prospective growth in genAI use outpacing growth in power-generation capacity makes energy-efficiency gains a crucial determinant of the productivity impact of genAI.

3.3.3 Datasets

GenAI models learn from reviewing large amounts of text (and other media), allowing them to estimate the probability that a given word or phrase should appear next in a sequence. Loosely speaking, the larger the corpus of text available to the model, the better it can estimate these probabilities. Figure 2 on page 9 illustrates the increase over time in the size of the datasets

[28]Since "Moore's Law" is a prediction that the number of transistors per die will double every two years, it arguably still holds true. Of course, because Moore (1965, 1975) does not contain a testable hypothesis, debates about whether Moore's Law holds are largely semantic. As is evident from fig. 8a on page 34, there is no single value for transistors per chip at any point in time.

Figure 7: Performance of the Most Energy-Efficient Supercomputers



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.

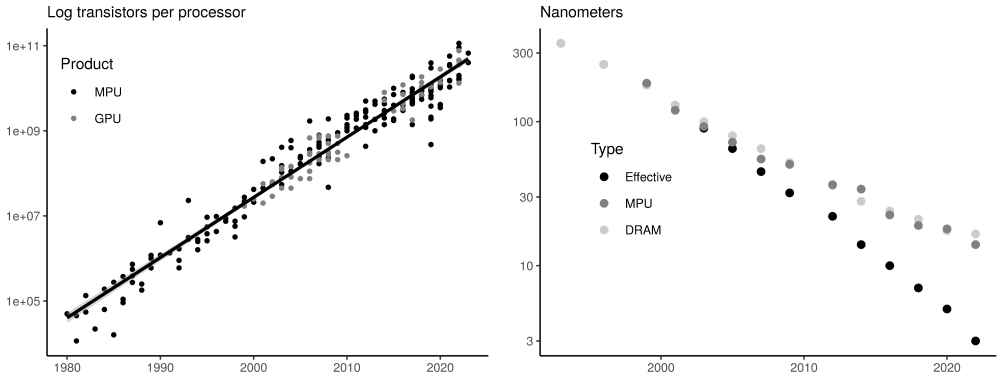
A crucial nuance to this direction for model improvement is that models improve with access to more information, not more text, per se. To illustrate with an extreme example, doubling the size of the training text by duplicating the corpus will yield no improvement to the model. Quality data enables 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. Some observers believe the availability of high-quality text data may be approaching its limit. Villalobos et al. (2022) predict high-quality text data may become scarce as early as 2026. One approach to mitigating this constraint is to use public data (such as internet sources) in pre-training and include proprietary data in the model refinement stage. Even so, in principle proprietary sources can be exhausted as well.

One potential solution is the use of "synthetic data", that is, data created via generative models.[29] Synthetic data can overcome data limitations in various domains—text, imagery, video—by emulating the patterns and characteristics of real data. It is easily scalable, can be tailored to specific requirements, and can be anonymized to hide personal information (Liu, Wei,

[29] An early example illustrates this approach. Image recognition systems were improved when the training set was increased by including the mirror images of the images initially collected. The mirror image contains some new information, but less than contributed by the original image.

Figure 8: 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.

Liu, Si, Zhang, Rao, Zheng, Peng, Yang, Zhou et al., 2024). For example, an LLM designed to tackle mathematical questions might be trained on a dataset of artificially-created or synthetic ("fake") questions. These questions might be generated by another LLM, using bootstrapping techniques to produce and rephrase similar questions from a human-produced training dataset but with varying degrees of complexity (MetaMath, Yu et al. 2025). Another example lies in medical imagery, where creating training data for cases such as CT scans is both resource-intensive and constrained by privacy concerns. Synthetic image generation of artificial yet realistic medical images provides a cost-effective and safe way to produce high-quality CT scans for use in training models that could be applied for diagnoses, monitoring or treatment (MAISI, Gu et al. 2024).

Since the nascent stages of genAI research, a prevailing hypothesis has been that models trained on synthetic data perform poorly compared to models trained on authentic data (Alemohammad, Casco-Rodriguez, Luzi, Humayun, Babaei, LeJeune, Siahkoohi and Baraniuk, 2023). This phenomenon, known as model collapse (Shumailov, Shumaylov, Zhao, Gal, Papernot and Anderson, 2023), has led researchers to believe that using synthetic data for training could lead to an overly narrow range of outputs due to a loss in the ability to remember underlying data distributions. Gerstgrasser et al. 2024

showed that model collapse can be avoided by using accumulating human-generated and machine-generated data. While model collapse continues to be a topic of ongoing debate, genAI models trained on the extant internet will progressively encounter synthetic data as generated data populates the web. Understanding the impact this has on future AI models that are trained with this data will be crucial to determine if synthetic data can serve as a panacea to the limitations of high-quality, human-generated data.

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.

GenAI innovation has proceeded at a blistering pace since the introduction of the Transformer.[30] While genAI performance has soared through the use of massive datasets and tremendous comoputational 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.

Knock-on innovation has been impressive as well. Key areas include interface with the user (through user interface software) and interface with robotics, where the genAI model enables far more sophisticated applications.

The weakest leg of the three-legged GPT stool is adoption. Although some field studies have provided encouraging results, adoption of genAI remains quite low, with the exception of large corporations and the share of jobs requiring AI skills 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.

As an aside, we note that the concept of "artificial general intelligence" (AGI)—genAI that can perform a wide array of tasks at a human level and function in novel environments—is quite different from the GPT idea. Demonstrating the feasibility of AGI is a primary long-run objective of many AI researchers but is not particularly interesting to economists. Technical

[30]Of course, the Transformer was the culmination of decades of AI research. As explained in section 2 on page 4, genAI research dates back to the beginnings of AI in the 1960s and present-day genAI would not be possible without the Word2Vec model.

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). Fortunately, AGI is not a precondition for genAI to be a GPT.

4 Is GenAI an "Invention in the Method of Invention"?

Under development

Another type of technology that is a source of sustained productivity growth is "inventions in the method of invention" (IMIs). Such technologies increase the efficiency of the research and development process. The compound microscope, which promotes deeper understanding of physical phenomena, is an example. Examples of technologies that play this role are shown in table 6 on the following page including obervational, analytical, communication, and organizational tools. These innovations hold out the possibility of permanently raising the growth rate of productivity because they raise the efficiency of the idea development and diffusion process. This channel for endogenous productivity growth is the foundation for many models that avoid the transitory nature of economic growth in classical models (Romer, 1994).

5 Conclusion

The release of ChatGPT in late 2022 was a stark inflection point in public and academic interest in genAI and expectations of a massive impact on productivity are abundant.

We argue that the characteristics of genAI and the record of its applications so far provide support for an optimistic outlook for its contribution to productivity. In particular, it appears that genAI is both a general-purpose technology—headed toward being widely used, stimulating related innovation, and advancing in (economic) performance—and an invention in the method of invention, raising the efficiency of R&D. Both GPTs and IMIs promote productivity growth for extended periods.

As exciting as technological progress in genAI is from a science and engineering standpoint, economic progress requires that the return from the

Table 6: Examples of Inventions in the Method 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
Generative AI	2017 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.

technology, less the total cost of ownership, be high enough to merit the requisite investment in reorganizing firms to exploit the innovation. Whether this is true for genAI remains to be seen. There are reports of efficiency gains, but few firms attest to material improvements to their bottom lines from the technology. And, use of genAI in small firms appears minimal.

Moreover, we caution that genAI will only boost productivity *growth* if its contribution outpaces other forms of IT already present in the underlying trend. Indeed, many firms attest that an earlier form of AI itself—machine learning—has already made a contribution to their productivity and the contribution from genAI will have to exceed that of machine learning. We believe a balanced forecast for productivity, *with appreciable risk on both sides*, include a contribution from genAI similar to that of machine learning, implying no acceleration of productivity growth from generative AI.

References

Abdin, Marah, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl et al., "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

Acemoglu, Daron, David Autor, Jonathon Hazell, and Pascual Restrepo, "AI and jobs: Evidence from online vacancies," Technical Report, National Bureau of Economic Research 2020.

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," 2023.

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

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

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

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

Baily, Martin and Aidan Kane, "AI in the Finance Sector Case Study," Technical Report, Brookings Institution 2025.

_ **and** _ , "AI in the Healthcare Sector Case Study," Technical Report, Brookings Institution 2025.

Bick, Alexander, Adam Blandin, and David J. Deming, "The Rapid Adoption of Generative AI," *NBER Working Paper Series*, 2024.

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

Bresnahan, Timothy, "Artificial Intelligence Technologies and Aggregate Growth Prospects," *Prospects for Economic Growth in the United States*, 2019, pp. 132–172.

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

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

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond, "Generative AI at work," Technical Report, National Bureau of Economic Research 2023.

_ **, Tom Mitchell, and Daniel Rock**, "What can Machines Learn and What Does it Mean for Occupations and the Economy?," in "AEA Papers and Proceedings," Vol. 108 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2018, pp. 43–47.

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

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

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

Choi, Seong Lok, Rishabh Jain, Patrick Emami, Karin Wadsack, Fei Ding, Hongfei Sun, Kenny Gruchalla, Junho Hong, Hongming Zhang, Ziangqi Zhu, and Benjamin Kroposki, "eGridGPT: Trustworthy AI in the Control Room," Technical Report, National Renewable Energy Laboratory 2024.

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

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

David, Paul A, "The dynamo and the computer: an historical perspective on the modern productivity paradox," *The American Economic Review*, 1990, *80* (2), 355–361.

DeepSeek, "DeepSeek-V3 Technical Report," *arXiv preprint arXiv:2412.19437*, 2024.

DeepSeek-AI, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," *arXiv preprint arXiv:2501.12948*, 2025.

Devlin, Jacob, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," *arXiv preprint arXiv:2303.10130*, 2023.

Felten, Edward W, Manav Raj, and Robert Seamans, "The occupational impact of artificial intelligence: Labor, skills, and polarization," *NYU Stern School of Business*, 2019.

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

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

Goldfarb, Avi, Bledi Taska, and Florenta Teodoridis, "Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings," *Research Policy*, 2023, *52* (1), 104653.

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

Hall, Brian, "321 Real-world Gen AI Use cases from the World's Leading Organizations," Technical Report, Google 2024.

Hancock, John T and Taghi M Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of big data*, 2020, *7* (1), 28.

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

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

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," 2015.

Ho, Anson, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla, "Algorithmic Progress in Language Models," *arXiv preprint arXiv:2403.05812*, 2024.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark et al., "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

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

Humlum, Anders and Emilie Vestergaard, "The Adoption of ChatGPT," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2024, (2024-50).

Jacobs, Robert A, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton, "Adaptive Mixtures of Local Experts," *Neural Computation*, 1991, *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, Timothy J Draelos, Aleksandra Faust, Matthew J Marinella et al., "A Historical Survey of Algorithms and Hardware Architectures for Neural-Inspired and Neuromorphic Computing Applications," *Biologically Inspired Cognitive Architectures*, 2017, *19*, 49–64.

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier et al., "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.

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

Kane, Aidan and Martin Baily, "AI in the Electricity Sector Case Study," Technical Report, Brookings Institution 2025.

_ and _, "AI in the Information Sector Case Study," Technical Report, Brookings Institution 2025.

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

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

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

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

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

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

Liao, Thomas, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt, "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)" 2021.

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

Lipsey, Richard G, Kenneth I Carlaw, and Clifford T Bekar, *Economic transformations: general purpose technologies and long-term economic growth*, OUP Oxford, 2005.

Liu, Ruibo, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou et al., "Best practices and lessons learned on synthetic data for language models," *arXiv preprint arXiv:2404.07503*, 2024.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark, "Artificial Intelligence Index Report 2024," 2024.

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

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

McKinsey, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," *McKinsey & Company*, 2024.

Mikolov, Tomas, Ilya Sutskever, and Quoc Le, "Learning the meaning behind words," *Google Open Source Blog*, 2013, *14*.

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

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller, "Playing Atari with Deep Reinforcement Learning," 2013.

Moore, Gordon E., "Cramming more components onto integrated circuits," *Electronics Magazine*, 1965, *4*, 114–117.

Moore, Gordon E, "Progress in digital integrated electronics," in "Electron devices meeting," Vol. 21 Washington, DC 1975, pp. 11–13.

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

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

Nilsson, Nils J, *The quest for artificial intelligence*, Cambridge University Press, 2009.

Nordhaus, William D, "Two centuries of productivity growth in computing," *The Journal of Economic History*, 2007, *67* (1), 128–159.

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

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

OpenAI, "Introducing ChatGPT," 2022.

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

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

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

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

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

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

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

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

Romer, Paul M, "The origins of endogenous growth," *Journal of Economic perspectives*, 1994, *8* (1), 3–22.

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

Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz, "A Bayesian approach to filtering junk e-mail," in "Learning

for Text Categorization: Papers from the 1998 workshop," Vol. 62 Citeseer 1998, pp. 98–105.

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

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

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

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

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

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

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

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

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

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

Srihari, Sargur N and Edward J Kuebert, "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," Vol. 2 IEEE 1997, pp. 892–896.

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

Trajtenberg, Manuel, "AI as the Next GPT: A Political-Economy Perspective," Technical Report, National Bureau of Economic Research 2018.

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

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

Villalobos, Pablo, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho, "Will we run out of data? an analysis of the limits of scaling datasets in machine learning," *arXiv preprint arXiv:2211.04325*, 2022.

Wallace, Richard A, "The Maze Solving Computer," in "Proceedings of the 1952 ACM National Meeting (Pittsburgh)" 1952, pp. 119–125.

Wang, Hongyu, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei, "Bitnet: Scaling 1-bit transformers for large language models," *arXiv preprint arXiv:2310.11453*, 2023.

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

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

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

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

Whitehead, Alfred North and Bertrand Russell, *Principia Mathematica*, Vol. 2, Cambridge University Press, 1927.

Wiesinger, Julia, Patrick Marlow, and Vladimir Vuskovic, "Agents," Technical Report, Google 2024.

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

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