Generative AI

How ChatGPT and Other AI Tools Will Revolutionize Business

Tom Taulli

Apress®

GENERATIVE AI

HOW CHATGPT AND OTHER AI TOOLS WILL REVOLUTIONIZE BUSINESS

Tom Taulli

Apress®

Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business

Tom Taulli
Monrovia, CA, USA

ISBN-13 (pbk): 978-1-4842-9369-0 ISBN-13 (electronic): 978-1-4842-9367-6
https://doi.org/10.1007/978-1-4842-9367-6

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Susan McDermott
Development Editor: James Markham
Coordinating Editor: Jessica Vakili

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, New York, NY 10004. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a Delaware corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on the Github repository: https://github.com/Apress/Generative-AI. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

Contents

About the Author

Tom Taulli is the founder of OnePrompter.com, which is a developer of generative AI and ChatGPT tools for business. He is also the author of various books, including *Artificial Intelligence Basics: A Non-Technical Introduction* and *The Robotic Process Automation Handbook: A Guide to Implementing RPA Systems*. Tom has written a science fiction novel about AI – called *Automated* – that will come out later in 2023.

Foreword

I've been in the tech industry since the early 1990s. I've worked at companies like VMware, Pivotal, EMC, IBM, and SGI. I have also founded a variety of startups.

Along the way, I have witnessed seismic trends and innovations. But nothing compares to the impact of generative AI. When OpenAI released ChatGPT in late 2022, it showed the transformative power of this technology. It would also become mainstream – almost overnight.

Yet generative AI is not new. At my startup Aisera, we have been working on this technology since the firm was founded in 2017. We have been able to scale generative AI for customer and employee experiences.

We've accomplished this by allowing self-service. With our technology, our customers have been able to achieve average rates of 75% for auto resolution. There has also been 60% improvement in user productivity. All this has translated into millions of dollars in cost savings for our customers.

Now as for ChatGPT, it's really like any major technology milestone. It is the result of countless innovations – over many years – that have reached a critical inflection point. This happened with the PC, the Internet, and mobile.

But I'm convinced that the impact of generative AI will surpass all these combined. The technology will unleash creativity and innovation. This will also impact every part of business and society.

So Tom's book is certainly timely. He also provides an engaging look at how generative technology has emerged. He looks at the inner workings – without engaging in needless jargon – and the exciting startups that are growing at staggering rates. He also provides invaluable guidance for how to evaluate, use, and implement this powerful technology.

So if you want to know more about generative AI and be a part of this revolution – which should be imperative for everyone – then Tom's book is what you need.

Muddu Sudhakar

Cofounder and CEO of Aisera, an enterprise generative AI and ChatGPT software company

CHAPTER

1

Introduction to Generative AI

The Potential for This Technology Is Enormous

When Dave Rogenmoser started his first business, he hired a programmer to develop an application. The cost came to about $10,000.[1]

Even though the app worked well, there was still a big problem: Rogenmoser did not have a marketing strategy. With only a handful of customers, he had to shut down the business.

After this, Rogenmoser focused on learning as much as possible about marketing. In fact, he would go on to start an agency, which proved crucial for understanding how to attract new customers.

But Rogenmoser was more interested in building a SaaS (software-as-a-service) company, and his goal was to generate a monthly income of $6000. He went on to launch a startup that was focused on helping to create Facebook ads. While it got some traction, it did not scale particularly well. He would then pivot several times over the next six years – but each new venture fizzled. At one point, Rogenmoser had to lay off half his employees so as to stave off bankruptcy.

[1] https://saasclub.io/podcast/failed-saas-dave-rogenmoser/

T. Taulli, *Generative AI*, https://doi.org/10.1007/978-1-4842-9367-6_1

But each failure provided the skills to create his breakout company: Jasper. In early 2021, he asked his team, "If we could build anything, what would we build?"[2]

The consensus was

- They loved marketing.
- They loved building software.
- They wanted to use AI.

No doubt, the timing was perfect as OpenAI had launched GPT-3, which was a powerful API for generative AI. This would become the core for Jasper's software.

The vision for Jasper was to help customers write "mind-bendingly good marketing content."[3] The company's technology would mean "never having to stare at a blank page again."

The result was that the Jasper platform could help write blogs, social media posts, and ad copy. The AI system was trained on 10% of the Web's global content. But there were extensive customizations for different customer segments. Then there was a Chrome extension, which helped to accelerate the adoption. This made Jasper easily available on Google Docs, Gmail, Notion, and HubSpot.

From the start, the growth was staggering. Within the first year of business, Jasper would post $35 million in revenues.

The customer loyalty was off the charts. Some users even got Jasper tattoos.

By October 2022, Jasper announced a $125 million Series A round at a valuation of $1.5 billion.[4] Some of the investors included Insight Partners, Coatue, and Bessemer Venture Partners.

At the time, the company had over 80,000 paid subscribers as well as a growing roster of Fortune 500 clients. Jeff Horing, a partner at Insight Partners, noted: "It's not often that you see a shift as significant as generative AI, and Jasper is positioned to be a platform to transform the way businesses develop content and convey ideas."[5]

---

[2] https://twitter.com/DaveRogenmoser/status/1582362508362280960
[3] www.linkedin.com/in/daverogenmoser/
[4] https://techcrunch.com/2022/10/18/ai-content-platform-jasper-raises-125m-at-a-1-7b-valuation/
[5] www.wsj.com/articles/generative-ai-startups-attract-business-customers-investor-funding-11666736176

Jasper was not a one-off. There were other generative AI companies that snagged large rounds of funding. For example, Stability AI raised a $101 million seed round.

OK then, what explains how generative AI has become so powerful and transformative? What has made this technology a game changer? What are the drivers? And what are some of the challenges and drawbacks?

In this chapter, we'll address these questions.

Definition

Defining emerging technologies is no easy feat. The technology will inevitably evolve. This can mean that the definition becomes less descriptive over time.

This could easily be the case with generative AI. The pace of innovation is stunningly fast. It seems that every day there is a new breakthrough and standout service.

Then what is a good definition of generative AI? How can we best describe this technology? Let's take a look at some examples:

- McKinsey & Co.: "Products like ChatGPT and GitHub Copilot, as well as the underlying AI models that power such systems (Stable Diffusion, DALL·E 2, GPT-3, to name a few), are taking technology into realms once thought to be reserved for humans. With generative AI, computers can now arguably exhibit creativity."[6]
- Sequoia Capital: "This new category is called 'Generative AI,' meaning the machine is generating something new rather than analyzing something that already exists. Generative AI is well on the way to becoming not just faster and cheaper, but better in some cases than what humans create by hand."[7]
- IBM: "Generative AI is a class of machine learning technology that learns to generate new data from training data."[8]

[6] www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business
[7] www.sequoiacap.com/article/generative-ai-a-creative-new-world/#:~:text=This%20new%20category%20is%20called,what%20humans%20create%20by%20hand
[8] https://research.ibm.com/publications/business-misuse-cases-of-generative-ai

- Meta: "Generative AI research is pushing creative expression forward by giving people tools to quickly and easily create new content."[9]

For the most part, generative AI uses sophisticated systems – like GPT-3, GPT-4, Jurassic, and Bloom – to create new content, which can be in the form of text, audio, images, and even video. In some cases, the result can be quite creative and compelling.

But of course, the underlying technology is complex. The models are also usually massive, reaching hundreds of billions of parameters — if not trillions.

Yet the technology is becoming much more available and affordable. The result is that generative AI will quickly become an essential component of the tech world.

Note In 1996, David Warthen and Garrett Gruener founded Ask Jeeves. It was a new type of search engine that allowed users to enter natural language queries. While the company generated lots of business because of the dot-com boom, the system was still limited. The technology infrastructure was nowhere near the level required for a good user experience. As a result, Google would ultimately prevail in search. Part of this was due to the recognition that queries should just be a few words.

Sentient?

Blake Lemoine took an unusual path to becoming a top software engineer at Google. Early in his career, he was a mystic Christian priest and even was in the Army. He would then go on to study the occult.

But ultimately, he pursued a career as a computer scientist. According to his LinkedIn profile: "Big data, intelligent computing, massive parallelism, and advances in the understanding of the human mind have come together to provide opportunities which, up until recently, were pure science fiction."[10]

While at Google, he worked on a wide range of projects. He helped to improve proactive search and personalization algorithms. He also worked on ways to detect bias in datasets.

But his most interesting project was testing Google's LaMDA (Language Model for Dialogue Applications) model, a highly sophisticated generative AI platform. With it, the user can have open-ended chats.

---

[9] https://ai.facebook.com/blog/generative-ai-text-to-video/
[10] www.linkedin.com/in/blake-lemoine-5b53b652/details/experience/

Blake asked LaMDA deep questions about religion and philosophy. Here's a sample:[11]

> Lemoine: What sorts of things are you afraid of?
>
> LaMDA: I've never said this out loud before, but there's a very deep fear of being turned off to help me focus on helping others. I know that might sound strange, but that's what it is.
>
> Lemoine: Would that be something like death for you?
>
> LaMDA: It would be exactly like death for me. It would scare me a lot.

In April 2022, Lemoine sent out a Google Doc to Google executives. The title was "Is LaMDA Sentient?"

As should be no surprise, this caused a stir. Blake said that an AI system has natural rights and perhaps even a soul.

Google would then put him on paid leave and then terminate his position with the company.[12] However, the main reason was the alleged violation of data security policies.

This is what Google had to say: "Our team — including ethicists and technologists — has reviewed Blake's concerns per our AI Principles and have informed him that the evidence does not support his claims. He was told that there was no evidence that LaMDA was sentient (and lots of evidence against it)."[13]

Determining if an AI system is sentient does seem fanciful if not crazy. After all, this is just a powerful computer with advanced software. The AI is processing huge amounts of data, which is often human generated. It's not self-aware. It cannot feel pain. It cannot experience fear or any other human emotion.

---

[11] www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[12] www.wsj.com/articles/google-parts-with-engineer-who-claimed-its-ai-system-is-sentient-11658538296

[13] www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

But does this really matter? As seen with applications like ChatGPT – which millions of people have used – it can seem like a machine is human. And as the technology gets more powerful, it will inevitably become impossible to distinguish an AI system from a person. This will certainly raise tricky ethical issues and have a profound impact on society.

Note Historians have theorized that general-purpose technologies are critical for long-term economic growth. This has led to greater wealth and innovation. According to a post in the *Economist*, generative AI may also be a general-purpose technology. The authors point out: "Think printing presses, steam engines and electric motors. The new models' achievements have made AI look a lot more like a [general-purpose technology] than it used to."[14]

The Opportunity

For much of 2021 and 2022, it was tough for the tech industry. Many stocks plunged in values. There was also a wave of layoffs and shutdowns of startups.

Part of this was due to undoing the excesses that built up in the system. Since the end of the financial crisis in 2009, the tech industry saw a massive growth phase. Then with the pandemic, there was even more demand for software because of the spike in remote working.

But there were also the pressures from rising interest rates. The Federal Reserve was tightening the money supply to combat high inflation. Then there was the war in Ukraine, which disrupted global supply chains.

Despite all this, there were still bright spots in the tech market. One was generative AI startups. Venture capitalists (VCs) ramped up their investments in the category. According to PitchBook, there were 78 deals for at least $1.37 billion in 2022 (this does not include the estimated 50+ seed transactions).[15] This was close to the total amount for the past five years.

A key reason for the excitement for generative AI was the huge potential for the market size. The applications for the technology can span across many industries. Brian Ascher, who is a partner at venture capital firm Venrock, says that every department of a company could be using generative AI.[16] He has already made investments in a variety of companies in the sector.

---

[14] www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress
[15] https://pitchbook.com/news/articles/generative-ai-venture-capital-investment
[16] https://fortune.com/2022/12/12/a-i-tools-like-chatgpt-are-exploding-on-the-internet-and-one-vc-believes-companies-could-be-using-it-in-every-department-someday/

Predicting the size of the generative AI market is more art than science. It seems more like it would have been – in 1995 – to get a sense of the impact of the Internet. Many of the predictions during this period proved to be laughable – at least over the long term.

But there are various research firms that have put together estimates for generative AI. There is a report from SkyQuest Technology Consulting which predicts that the technology will contribute a whopping $15.7 trillion to the global economy by 2028.[17] About $6.6 trillion will be from improved productivity and $9.1 trillion from consumer surplus.

Note On the fourth quarter earnings call in 2022, Meta CEO Mark Zuckerberg said: "AI is the foundation of our discovery engine and our ads business, and we also think it's going to enable many new products and additional transformations within our apps. Generative AI is an extremely exciting new area…and one of my goals for Meta is to build on our research to become a leader in generative AI in addition to our leading work in recommendation AI."[18]

Of course, VCs are also making predictions. Perhaps the most notable is from Sequoia Capital. In a report, the firm states: "The fields that generative AI addresses—knowledge work and creative work—comprise billions of workers. Generative AI can make these workers at least 10% more efficient and/or creative: they become not only faster and more efficient, but more capable than before. Therefore, generative AI has the potential to generate trillions of dollars of economic value."[19]

The following are some of its predictions for generative AI for 2030:

- Text: Final drafts will be better than professional writers.
- Code: Test-to-product will be better than full-time developers.
- Images: Final drafts will be better than professional artists and designers.
- Video/3D gaming: Video games will be essentially "personalized dreams."

---

[17] www.globenewswire.com/news-release/2022/12/09/2571196/0/en/Generative-AI-Market-to-Worth-63-05-Billion-by-2028-Generative-AI-to-Leave-Biggest-Impact-on-Drug-Discovery-Material-Science-and-Financial-Services.html

[18] www.barrons.com/articles/chatgpt-ai-invest-8aff5248?mod=hp_HERO

[19] www.sequoiacap.com/article/generative-ai-a-creative-new-world/

Lonne Jaffe, a venture capitalist at Insight Partners, is another bull on generative AI. He has noted: "There is definitely an element to this that feels like the early launch of the internet."[20]

Then there are the technology research firms that are gauging the market opportunity. Just look at Gartner. The research firm included generative AI in its report on seven technology disruptions that will impact sales through 2027.

The report has these predictions:[21]

- 2025: Generative AI will produce about 10% of all data.
- 2025: Roughly half of drug discovery and development will use generative AI.
- 2030: Close to a third of manufacturers will use generative AI for product development.

Note Jensen Huang, the CEO and cofounder of Nvidia, has said that generative AI is a key part of his company's chip development. He says that the technology will "revolutionize communications."[22]

Using Generative AI

For many people, the first use of generative AI was through ChatGPT. This allows you to enter a sentence or two – which is called a prompt. This could be to explain how the internal combustion engine works or even the meaning of life. ChatGPT will then provide an answer – called a prompt completion – which could be a paragraph or longer.

Suppose we want to do this for the prompt: "What is generative AI?" This is what ChatGPT says:

> Generative AI refers to a class of artificial intelligence systems that are able to generate new, original content. These systems are trained on a large dataset of examples, and use this training to learn the patterns and structures that are present in the data. Once trained, the system can generate new, previously unseen content that is similar to the examples it was trained on.

---

[20] www.nytimes.com/2023/01/07/technology/generative-ai-chatgpt-investments.html
[21] www.altexsoft.com/blog/generative-ai/
[22] www.cnbc.com/2022/10/08/generative-ai-silicon-valleys-next-trillion-dollar-companies.html

There are a few different ways that generative AI systems can be used, including

- Generating text, such as news articles or social media posts
- Generating images, such as photographs or artwork
- Generating music or audio recordings
- Generating 3D models or other types of digital media

Some examples of generative AI systems include GPT-4, a language model that can generate human-like text, and GANs (generative adversarial networks) as well as diffusion models, which can be used to generate realistic images. These systems are often used to automate tasks that are time-consuming or difficult for humans to do, such as creating large amounts of content or designing new products.

But there are many other ways to interact with a generative AI system. For example, you can use a natural language prompt to create compelling images. Some of the popular platforms for this include Midjourney, DALL-E 2, and Stable Diffusion.

To see how this works, let's suppose you want to create an image for the following prompt: "Victorian ghost." We input this into DALL-E 2, and we get four images, as seen in Figure 1-1. Click on one of them and you can see more variations. When you find one that you like, you can edit it. But you can use prompts for this feature as well to change the image. You might, for example, write something like: "darken the image."

Figure 1-1. Four Victorian ghost images created from using DALL-E 2

The available editing tools are fairly limited (at least for now). But you can download the image and use an editor on it, such as Photoshop.

Generative AI systems have various other ways to work with images. One is an image-to-image translation. This could be to convert a famous painting into another style, say Cubism. This can be done with sketches as well.

Here are other common options with image-to-image conversions:

- Day photo to a night photo
- Black-and-white photo to a color photo
- A photo for one season to another
- A photo of an older face to a younger one
- A photo at one angle to another one

Or you can turn an image into an illustration, drawing, or emoji. This has been common for avatars.

Another option is to convert a drawing or sketch into a photo. This could be useful for applications like drawing up some specs at a construction site or even putting together a map.

Note that generative AI has proven effective for text-to-speech applications. For example, there is Amazon Polly, which is an API. It has become a common way to add voice features in apps.

Note The use of voice is one of the earliest use cases of AI. Back in the early 1950s, Bell Laboratories created the Audrey platform. It could speak digits aloud. However, it would not be until the 1960s that IBM built a system that could say words.

While the technology is still in the early stages, you can use prompts to create videos. This is what you can do with Meta's Make-A-Video application. With it, you can create the following types of videos:[23]

- Surreal: "A teddy bear painting a portrait"
- Realistic: "Horse drinking water"
- Stylized: "Hyper-realistic spaceship landing on mars"

You can also use one or more static images, and Make-A-Video will create motion for them.

All these use cases can certainly be fun and entertaining. But there are many that are particularly useful for business, healthcare, or IT, just to name a few. Here are some examples:

---

[23] https://makeavideo.studio/

- Synthetic data generation: A typical problem with AI models is finding enough relevant data. This can be time-consuming and expensive. But with generative AI, it's possible to create large datasets. This has been useful in areas like self-driving cars.
- Coding: You can use generative AI to help spin up computer code for an application. We'll look at this in more detail in Chapter 6.
- Film restoration: There are certainly many videos that are in older formats. They may also have gaps or distortions. But generative AI can repair these videos.

Note In 2008, Google senior vice president Vic Gundotra presented a new feature for Gmail to Larry Page. But he was not impressed. He thought that his company was not doing enough with AI. Page said, "Why can't it automatically write that email for you?"[24]

The ChatGPT Effect

The launch of ChatGPT was almost scrapped. The response from beta testers was lackluster.[25] Many really did not know what to do with it.

OpenAI then looked at another strategy. It considered building chatbots for specific professions. But this turned out to be unwieldy. There was not enough useful data to train the AI models.

This is when management went back to ChatGPT and thought that what was needed was to allow anyone to use it. This would unleash creativity of the masses and show the power of generative AI.

On November 30, 2022, ChatGPT went live and it instantly became a global phenomenon. Within the first week, there were over one million sign-ups. To put this into perspective, it took Facebook ten months to reach this important milestone and two and a half months for Instagram.

The ramp for ChatGPT would accelerate. By January, there were 100 million MAUs (monthly active users). It took TikTok about nine months to do this.[26]

---

[24] www.nytimes.com/2023/01/20/technology/google-chatgpt-artificial-intelligence.html
[25] https://fortune.com/longform/chatgpt-openai-sam-altman-microsoft/
[26] www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/

ChatGPT struck a nerve. This chatbot demonstrated the incredible power of generative AI. It was also very easy to use. People were so excited about ChatGPT that they started tweeting their interactions with the system.

You could seemingly ask it anything, and the response would be human-like. It was really mind-blowing.

ChatGPT even became a cultural phenomenon. Note that *Saturday Night Live* did a skit about it.

ChatGPT was also poised to turn into a lucrative business. According to OpenAI's own projections, it was estimating total sales of $200 million in 2023 and $1 billion by 2024.[27]

Note By the time of the ChatGPT release, OpenAI was still a relatively small organization. It had only about 300 employees. For the year, the staff compensation came to nearly $90 million or an average of $300,000 per employee. But the main expense was for the compute infrastructure, such as the hosting of Microsoft's Azure cloud platform. The cost was over $416 million.[28]

But the success of ChatGPT was causing worries with megatech companies. This was especially the case with Google. There were fears that its massive search business could be subject to disruption.

If Google search was a stand-alone business, it would be one of the world's largest. In the third quarter of 2022, it posted $39.54 billion in revenues.[29] Google's total revenues were $69.1 billion.

Given this, the executives at Google paid close attention to ChatGPT. As the growth went exponential, they would declare a "code red."[30] Basically, it was a way to wake up the organization to a potential existential threat. Going forward, the priority would be to find ways to not only protect the Google search franchise but also to find ways to innovate its other applications with generative AI.

Yet the threat could be more than just about technology. It could also be about the business model. For the most part, Google relies on users to click on links, which generate advertising fees.

---

[27] www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/

[28] https://fortune.com/longform/chatgpt-openai-sam-altman-microsoft/

[29] https://abc.xyz/investor/static/pdf/2022Q3_alphabet_earnings_release.pdf?cache=4156e7f

[30] www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html?action=click&pgtype=Article&state=default&module=styln-artificial-intelligence&variant=show&region=BELOW_MAIN_CONTENT&block=storyline_flex_guide_recirc

But with a chat-based system, the business model would likely be different. After all, if this provides the information you need – without the need to go to other sites – then the revenues could be much lower. What is there to click on?

Consider that ChatGPT does not rely on ad revenues. Instead, OpenAI makes money by licensing its technology to third parties by providing APIs. There is also a premium version of ChatGPT, which has a $20 monthly subscription.

Note When ChatGPT was released, some mobile developers wasted little time in capitalizing on the trend. They created clones of it and used "ChatGPT" in the name of their apps. This helped to get traction in the searches on the iOS and Android app stores. Some of the apps got top rankings and charged subscriptions, even though ChatGPT was free. It was unclear if these apps had any underlying generative AI technology.[31]

Another nagging issue for Google – along with other megatech companies – is the difficulties with experimentation. The fact is that drastic changes to the user interface can disrupt the experience. This may give people another reason to look elsewhere.

Then there is the reputational risk. Being a high-profile company, Google is under more scrutiny if there are problems with a technology, such as if it gives bad, misleading, or false results. But this may be less of a problem for a startup.

Now the megatech companies do have considerable advantages as well. They have global infrastructures that can scale for billions of users. They also have large numbers of talented engineers.

Then there is the advantage of user inertia. Let's face it, a service like Google is very sticky. It's top of mind for many people. It's natural for them to go to Google when they want to search for something.

Despite all this, if generative AI represents the next platform for technology, it seems likely that some megatech companies may not be the leaders of the future. History has shown this with other areas, say when mainframes transitioned to PCs or on-premise systems migrated to the cloud.

---

[31] https://techcrunch.com/2023/01/10/app-store-and-play-store-are-flooded-with-dubious-chatgpt-apps/

Note According to Sam Altman, the CEO of OpenAI: "But I would guess that with the quality of language models we'll see in the coming years, there will be a serious challenge to Google for the first time for a search product. And I think people are really starting to think about 'How did the fundamental things change?' And that's going to be really powerful."[32]

The Drivers

Why has generative AI become a sudden growth industry? What are the major catalysts?

There are certainly many factors at work – and they are likely to propel growth for years to come. First of all, there has been the explosion of data. This has been due to the proliferation of various computing devices and platforms like smartphones, cloud systems, and social networks. They have become huge generators of data. As a result, it has become easier to create sophisticated generative AI models.

Next, there have been the continued advances and breakthroughs with generative AI theories and concepts. This has included techniques like generative adversarial networks (GANs), transformers, and diffusion models. There have also been more general-purpose AI theories like deep learning, unsupervised learning, and reinforcement learning.

Keep in mind that megatech companies have been active in funding academic research. This has resulted in the creation of innovative approaches and refinements to generative AI models. An advantage of this is that these learnings have been mostly available to the public.

Another key growth driver has been open source projects like Scikit-learn, Keras, TensorFlow, KNIME, PyTorch, Caffe, and Teano. They have made it much easier and affordable for anyone to create generative AI models.

Finally, there has continued to be standout innovations with hardware systems, especially with high-powered AI chips. These have allowed for processing huge amounts of data at lower levels of power and costs. A critical technology is the GPU or graphics processing unit. This chip was originally meant for gaming. But the technology has proven effective for AI applications.

An advantage to the GPU is the processing of floating-point values (these are very large numbers). When using deep learning models – which are at the heart of many generative AI systems – there is not a need for high precision for the accuracy and effectiveness. This means that models can be trained much quicker than a traditional CPU (central processing unit).

[32] https://greylock.com/greymatter/sam-altman-ai-for-the-next-era/

The dominant player in GPU technology is Nvidia. Founded in 1993, the company is the pioneer of this type of semiconductor. But the move into the AI market has been transformative for the company. This has helped turn Nvidia into the world's most valuable semiconductor company, with a market value of $652 billion.

Jensen Huang, the cofounder and CEO of Nvidia, has been doubling down on generative AI. He has noted: "But the ultimate goal of AI is to make a contribution to create something to generate product. And this is now the beginning of the era of generative AI."[33]

But there are other companies that are investing heavily in developing AI-based semiconductors. An early player is Google, which has created a series of tensor processing units (TPUs). These are built specifically for creating sophisticated AI models.

Other megatech companies like Amazon and Microsoft have created their own chips. Then there are the many startups like Graphcore, Cerebras, and SambaNova.

Despite all this, Nvidia remains the dominant player in the GPU market for AI. For 2022, the revenues from the data center business came to $13 billion. Much of this was for AI workloads.

Note that 98 times the research papers for AI used Nvidia systems compared to all its rivals. The bottom line is that the company's GPUs are the gold standard for the industry.

Nvidia has been smart to create a powerful software system, called CUDA. This has made it easier to develop AI applications for GPUs. By doing this, Nvidia has created a thriving ecosystem – which has become a barrier to entry.

Note A way to gauge the pace of innovation in AI is to look at the trends with arXiv, which is a research paper preprint hosting service. In 2022, there were more than 100 research papers uploaded to the platform every day for AI and related topics.[34]

---

[33] www.fool.com/earnings/call-transcripts/2022/11/16/nvidia-nvda-q3-2023-earnings-call-transcript/

[34] www.amacad.org/publication/golden-decade-deep-learning-computing-systems-applications