



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
<https://doi.org/10.1007/978-1-4842-9367-6>

ISBN-13 (electronic): 978-1-4842-9367-6

Copyright © 2023 by Tom Taulli

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

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	v
Foreword	vii
Chapter 1: Introduction to Generative AI	i
Chapter 2: Data	21
Chapter 3: AI Fundamentals	47
Chapter 4: Core Generative AI Technology	77
Chapter 5: Large Language Models	93
Chapter 6: Auto Code Generation	127
Chapter 7: The Transformation of Business	145
Chapter 8: The Impact on Major Industries	175
Chapter 9: The Future	189
Index	203

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

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/>

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>

³www.linkedin.com/in/daverogenmoser/

⁴<https://techcrunch.com/2022/10/18/ai-content-platform-jasper-raises-125m-at-a-1-7b-valuation/>

⁵www.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

⁶www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business

⁷www.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/>

¹⁰www.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.

¹⁴www.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 I-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 I-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=style-artificial-intelligence&variant=show®ion=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

Skeptics

Generative AI has its doubters, of course. Some of them are giants in the AI field.

One is Yann LeCun, who is the Chief AI Scientist at Meta AI Research. He has a PhD in computer science from the Université Pierre et Marie Curie (Paris).³⁵ During the 1980s, he experimented with convolutional neural networks and backpropagation. This was useful in analyzing handwritten content, such as with checks.

He has published over 180 technical papers on AI, machine learning, computer vision, and computational neuroscience. In 2019, he won the Turing Award, along with Yoshua Bengio and Geoffrey Hinton. This is the most prestigious prize in AI.

In other words, his opinions matter in a big way. They are certainly based on a strong foundation.

Then what's his take on generative AI? LeCun sees it more of a technology for fun, not real-world applications – at least for now. He notes: “I don't think these systems in their current state can be fixed or called intelligent in ways that we want and expect them to be.”³⁶

He also has downplayed the innovations of ChatGPT. According to him: “It's nothing revolutionary, although that's the way it's perceived in the public. It's just that, you know, it's well put together, it's nicely done.”

He notes that there are various rival systems to ChatGPT. These not only include megatech companies but startups. “I don't want to say it's not rocket science, but it's really shared, there's no secret behind it, if you will,” he said.

But then again, the generative AI technologies are still in the nascent phases. There will inevitably be new approaches and breakthroughs. In fact, LeCun is still optimistic about the long-term prospects of generative AI. However, for him, there is still much that must be built.

The irony is that some leading cofounders and CEOs in the generative AI industry have taken a skeptical view of things and have noted the challenges.

³⁵ <https://research.facebook.com/people/lecun-yann/>

³⁶ <https://analyticsindiamag.com/why-is-no-one-taking-generative-ai-seriously/>

Take Sam Altman. Here's one of his tweets in December 2022:

ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. it's a mistake to be relying on it for anything important right now. it's a preview of progress; we have lots of work to do on robustness and truthfulness.³⁷

None of this should be surprising. Major technologies always face hurdles and skepticism.

A classic example is the launch of the iPhone. It did not have many useful features, and there were annoying glitches. There wasn't even an app store. Oh, and the battery life was awful and the Internet access was spotty. Security was also minimal. As for the digital keyboard, it seemed awkward.

At the time, there was much skepticism from industry leaders. Jim Balsillie, the co-CEO and cofounder of BlackBerry, said: "It's OK—we'll be fine."³⁸

At the time, the company had roughly half of the share of the smartphone market. Its position seemed unassailable.³⁹

But in about five years, BlackBerry's business shrunk considerably. Eventually, the company would unload its smartphone segment and focus on providing cybersecurity software and services.

Now this is not to say that generative AI represents something similar. But it is telling that megatechs are taking this technology very seriously, as are many startup founders and venture capitalists.

Here's how Sequoia Capital sums it up: "[S]ome of these applications provide an interesting glimpse into what the future may hold. Once you see a machine produce complex functioning code or brilliant images, it's hard to imagine a future where machines don't play a fundamental role in how we work and create."⁴⁰

³⁷ https://twitter.com/sama/status/1601731295792414720?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1601731295792414720%7Ctwgr%5E8a7166daa%5Eab2372ef65612106dfbaf89a3a336%7Ctwcon%5Ees1_&ref_url=https%3A%2F%2Fwww.cnbc.com%2F2022%2F12%2F13%2Fchatgpt-is-a-new-ai-chatbot-that-can-answer-questions-and-write-essays.html

³⁸ www.iphoneincanada.ca/news/blackberry-co-founders-response-to-the-original-iphone-its-ok-well-be-fine/

³⁹ www.forbes.com/sites/parmyolson/2015/05/26/blackberry-iphone-book/?sh=315a72a663c9

⁴⁰ www.sequoiacap.com/article/generative-ai-a-creative-new-world/

■ **Note** The response to ChatGPT from top tech veterans has been mostly upbeat, if not giddy. Aaron Levie, the cofounder and CEO of Box, tweeted: “Every time you think you know what ChatGPT can do, you realize you know almost nothing yet. The infinite depth of what it means for a computer to understand language is mindboggling, and we’re just scratching the surface.”⁴¹ Then there is Marc Benioff, the cofounder and CEO of Salesforce.com. He tweeted: “Just promoted #ChatGPT to the management team at salesforce. It’s been a real asset – never seen such efficient decision making & hilarious meeting participation! #AI #futureofwork.”⁴²

Dangers of Hype

A looming risk for the development of generative AI is the hype. This can create overly ambitious expectations about the technology. The result is that there will be inevitable disappointment. When this happens, there will likely be retrenchment of investment.

After all, AI has been seen on various hype cycles, which have then been followed with “winters.”

Let’s take a look. In the 1950s and 1960s, AI was in a “golden age.” The US government poured huge amounts of money into research for the Cold War and the Apollo space program.

But by the early 1970s, there emerged the first winter. While there had been much progress with AI, the applications were still limited. The fact is that the computing power was fairly minimal. For example, a DEC PDP-11/45, which was a popular system for AI, had only 128K of memory.

Another problem was the general economic environment. Growth across the world was starting to wane. There was also the problem with inflation, which was aggravated by the oil crisis.

For AI, the situation got so bad that many of the researchers did not even want to say they were in the industry. They would use other words for the technology, like machine learning and informatics.

By the 1980s, the AI industry would show renewed growth. But this was short-lived. There would be another winter. This one would be prolonged. As the Internet emerged, this would become the main focus for the technology industry.

⁴¹ <https://twitter.com/levie/status/1613409949312327680?s=43&t=1S88En6TI169cKmNANEFvQ>

⁴² <https://twitter.com/benioff/status/1614372552025178114?s=43&t=xaciM-ysf7nd8AT668r84w>

But by 2010 and 2011, there was a renewal of the AI industry. The algorithms were much more sophisticated, such as with deep learning models. There were also enormous amounts of training data and sophisticated machines, such as based on GPUs.

Yet there were still examples of major problems. Look at what happened with IBM Watson.

Granted, this AI technology got off to an impressive start. In 2011, IBM Watson beat the top *Jeopardy!* champions.

Then what was the problem? IBM management got overambitious. The strategy was to have a moonshot – on par with something like the Apollo space program. IBM even put together a TV commercial with the following message: “Already we are exploring ways to apply Watson skills to the rich, varied language of health care, finance, law and academia.”⁴³

The company would then go on to spend billions on IBM Watson. But there were doubters in the organization, such as David Ferrucci, a scientist. He said that IBM Watson was primarily good at answering quiz questions but could not understand complex problems.

Other top employees at IBM, like Martin Kohn, recommended focusing on niche areas at first. This would produce quick wins – which would then build the company’s AI muscles. For example, instead of trying to cure cancer, the technology could look at detecting issues with a part of the clinical trials process.

IBM should have heeded the warnings. For the most part, IBM Watson produced disappointing results with the commercialization of AI.

Look at the case with the MD Anderson Cancer Center in Houston. IBM tried to use the technology to create a tool to provide recommendations for cancer treatments.

But after \$62 million in investment, the MD Anderson Cancer Center canceled the project.⁴⁴ Some of the reasons were poor project management, shifting goals, and difficulties with integrating with the EHR (electronic health record) system.

IBM has since changed its strategy. It’s more aligned on smaller goals. The company has leveraged IBM Watson in areas where the company has deep experience, such as with customer service, compliance, risk management, business automation, and accounting functions.

⁴³ www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

⁴⁴ <https://thomaswdinsmore.com/2018/02/21/notes-on-a-watson-fail/#:~:text=The%20story%20revealed%20that%20MD,directly%20from%20MD%20Anderson%20physicians>

For the most part, the strategy has led to better results. In fact, IBM stock has revived lately – and beat out many tech companies in 2022.

■ **Note** Originally, scientists at IBM called their AI system DeepJ! It was a variation on the name of another powerful IBM computer: Deep Blue. In 1997, it beat Garry Kasparov, a chess champion. But as for DeepJ, the marketing team made an important name change. The AI system would instead become Watson. This was a nod to the founder of IBM, who was Thomas Watson Sr.⁴⁵

Conclusion

Generative AI is an exciting category, and the opportunity is potentially massive. It could represent a tidal shift in the technology world – similar to the launch of the Netscape browser or the iPhone.

Applications like ChatGPT have already shown millions of people the power of generative AI. The technology may even disrupt today's megatech companies and lead to the new trillion-dollar firms of tomorrow.

There are skeptics, of course. This is to be expected and has been a part of every major shift in technology.

But for now, generative AI has much going for it. It is based on complex theories and concepts. More importantly, the technology has already shown real value.

Here's what Forrester has said:

*"The future of generative AI is an ever-expanding universe of use cases. Every single part of the enterprise can reap benefits from this technology — from the contact center to marketing, product development, content generation, human resources, employee support, and customer experience. We have only begun to scratch the surface of what's possible with these models today — and the pace of model development is delivering richer capabilities all the time. Enterprises must begin exploring these capabilities today — and should construct a strategy around how generative AI will impact their company from both the top down and the bottom up."*⁴⁶

As for the next chapter, we'll start to dive deeper into the technology. The topic will be about the importance of data for the generative AI models.

⁴⁵ www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html

⁴⁶ www.forrester.com/blogs/beware-of-coherent-nonsense-when-implementing-generative-ai/

Data

The Fuel for Generative AI

A large language model (LLM) processes huge amounts of data for its generative AI systems. They are on the scale of petabytes. Consider that a petabyte is 1000 terabytes. This would hold about 500 billion pages of standard text. No doubt, the generative AI models for images and videos are much larger.

Thus, when it comes to generative AI, there needs to be a highly sophisticated data infrastructure. This means not only holding huge volumes but handling the processing at high speeds.

Yet data is often something that does not get enough attention. Let's face it, the topic can be dry and boring. Rather, when it comes to generative AI, the more exciting part is about the fancy algorithms.

But if you want a good generative AI system, you need to make heavy investments in data. If not, the chances of failure are significant.

In this chapter, we'll take a look at how data works with generative AI and look at the important best practices.

Value of Data

Back in 2006, British mathematician Clive Humby said: “Data is the new oil.”¹ The statement was prescient as data has become incredibly valuable. It also can represent a powerful competitive advantage.

We’ve seen how data-driven companies have upended incumbents. Some famous examples include ride-share companies like Uber that disrupted that traditional taxi business. Then there is Amazon, which has targeted various categories.

Now having lots of data is not enough. Of course, it’s about what an organization can do with it. This means applying sophisticated algorithms to it – that is, turning data into insights. For Amazon, one of the earliest examples of this was its recommendation engine for buying books. It was based on readers’ prior buying habits. By doing this, Amazon was able to get a significant lift in sales, and it also allowed for a much better customer experience.

Here are some of the other use cases for data:

- Pricing
- Personalized marketing campaigns
- Detecting customer dissatisfaction
- Optimizing the supply chain
- Predicting when a machine, equipment, or a device will fail (this is known as predictive maintenance)

Putting a value on data is no easy feat. It really depends on the applications. But when data is used properly, it can be transformative for a company.

Just look at Mobileye. The company is one of the first to commercialize generative AI for auto collision systems and self-driving technologies. This was primarily through the use of synthetic data. It’s where data is created by using advanced simulations.

Here’s a backgrounder on the company: In 1999, Professor Amnon Shashua founded Mobileye. He was a top academic researcher at the Hebrew University of Jerusalem. Shashua was a pioneer of monocular vision systems to detect vehicles.²

But he was more than about advanced technologies. He understood how to create a competitive business – and he knew data would be key for this.

¹ www.dbta.com/Columns/Next-Gen-Data-Management/Data-is-the-New-Oil-But-That-Also-Means-it-Can-be-Risky-155275.aspx#:~:text=British%20mathematician%20Clive%20Humby%20famously,growth%20to%20reach%20their%20goals

² www.mobileye.com/about/

To help collect data, he was aggressive in forming alliances with auto manufacturers. The result was the collection of huge amounts of data that came through Mobileye's cameras.

This became the basis for high-definition maps, which turned out to be crucial for autonomous vehicles. The accuracy reached the centimeter level.

This was made possible with Mobileye REM (Road Experience Management) technology, which processes the data from vehicles and compresses it. The packets are then sent to the cloud.

With this, Mobileye created a fast-growing part of its business, called Data Services. According to the company:

*"The information we are able to collect is of critical importance to cities, departments of transportation, road operators, utilities and others. While scanning for road hazards, our AI technology can also be used to identify, locate and classify roadside assets and mobility information. This means that vehicles equipped with our technology are also able to identify pavement cracks, potholes, worn road striping, traffic signs, manhole locations and much more. In addition, we can locate construction zones, heavily trafficked areas, locations where there are frequent collision 'near-misses' and other valuable mobility information."*³

As a testament to the Mobileye's technology and growth, Intel purchased the company in 2017 for \$15.3 billion.⁴ Then in 2022, Intel spun off the division as an IPO. Even though the market was bearish, the public offering turned out to be one of the most successful for the year. On the first day of trading, the shares soared by 38%.

If you read the Mobileye prospectus, there is a section on competitive advantages. Of these, there was one called "Data and technology advantage."

Here's what it said:

"We have assembled a substantial dataset of real-world driving experience, encompassing over 200 petabytes of data, which includes over 23 million clips collected over decades of driving on urban, highway, and arterial roads in over 80 countries. We utilize up to approximately 500,000 cloud CPU cores to process approximately 100 petabytes of data every month, which we believe demonstrates the size and scale of

³ www.mobileye.com/en/data/blog/the-journey-from-safety-to-data-begins-with-a-single-camera/

⁴ <https://techcrunch.com/2017/03/13/reports-intel-buying-mobileye-for-up-to-16b-to-expand-in-self-driving-tech/>

our data repository. Our dataset creates a powerful network effect as we seek to continually improve our solutions as more vehicles are deployed with our technology.”⁵

Given this, it seems like a good bet that Mobileye will remain a strong growth operator. Shashua says it will be about 30% for a “number of years.”⁶

The Amazing Growth of Data

While predicting technology is often a fool’s errand, there is one thing that seems very clear: data will continue to grow at a rapid clip.

Consider a report from IDC (International Data Corporation). It forecasts the compound annual growth rate (CAGR) of data at 21.2% through 2026. This will mean that the total amount will hit about 221,000 exabytes. An exabyte is 1000 petabytes.⁷

About 90% of this will be unstructured data. This means that the information is not organized, such as by using labels. Examples of unstructured data include images, emails, videos, text files, satellite images, and social media messages. This type of data is most common for generative AI models.

Structured data, on the other hand, is what you would find stored in a relational database or spreadsheet. Common items are names, age, transactions, addresses, and so on. The information is generally easier to work with, such as by using data analytics tools and AI. However, there is usually not enough of this data to create effective models.

Next, there is a blend of structured and unstructured data. It’s called semi-structured data. However, a majority of it is unstructured. Semi-structured data is common for JSON (JavaScript Object Notation), which allows for creating APIs (application programming interfaces). Another popular use case is XML (Extensible Markup Language). This helps to map information on a document.

There are other types of data:

- **Time-series data:** This is where the information is recorded over a specified time interval. For example, time-series data is effective for helping understand user behavior, such as on a website or an app. This type of information has also been an essential part of AI models.

⁵www.sec.gov/Archives/edgar/data/1910139/000110465922112004/tm227410-23_424b4.htm

⁶www.wsj.com/articles/mobileye-shares-open-higher-in-stock-market-debut-11666799733

⁷www.delltechnologies.com/asset/en-us/products/storage/industry-market/h19267-wp-idc-storage-reqs-digital-enterprise.pdf

- **Metadata:** This is a description of data, which helps to provide more context. An example is a document, which has metadata like the file name, file size, and the time of the latest update.
- **Dark data:** This is data that an organization does not realize it has or does not use. This would include sensor data, log files, and social media mentions. Dark data, though, can be valuable and useful for creating AI models.
- **Real-time data:** This is data that is collected, processed, and used as it is generated. Some of the typical sources of this are weather and traffic information. They have become essential for apps like Uber or Robinhood. Creating AI for these systems is incredibly complex because there is little time for the processing.
- **Synthetic data:** This is a cutting-edge area. But there have been important breakthroughs with synthetic data. This type of information has made it possible to create more powerful AI models.
- **Genomic data:** This is information about DNA structures and genes. This is a category that has seen significant growth and has allowed for many innovations in biotechnology.

Big Data

Back in 2001, Gartner analyst Doug Laney coined the term “Big Data.” He saw that the growth of data was accelerating – and this would represent a secular trend.

Whenever Gartner declares a new trend, executives take note. This was certainly the case with Big Data. The term quickly became a common way to describe data strategies. Of course, startups sprang up to focus on the opportunities.

Laney’s framework for Big Data was actually more than a fancy term. He provided a compelling framework to understand data.

He saw that data had three major attributes: volume, variety, and velocity. He called them the three Vs.

Here's what they mean:

- **Volume:** This describes the huge scale of Big Data. The data is mostly unstructured. So how much data does “volume” refer to? Laney did not provide a metric for this. But when he created the Big Data framework, the amounts were obviously much smaller than today. Thus, providing a benchmark probably does not make much sense. Something else: With growth in cloud computing, the handling of volume has become much easier. There is seemingly endless capacity for data storage.
- **Variety:** This is about the different types of data. As we already noted earlier in the chapter, this includes structured, semi-structured, and unstructured data. But “variety” also refers to the sources of data. This could be social media, cloud applications, and IoT (Internet of Things).
- **Velocity:** This is the speed of the creation of data. This has increased substantially since Laney introduced his Big Data framework. In fact, a key factor for the growth of generative AI is the velocity of data. Huge online platforms, like social networks and content sites, have created massive amounts of data for training the models. Generally, for the three Vs, velocity is the toughest.

While Laney's model was robust, it still had its flaws. Again, with the pace of innovation, there needed to be updates.

Here are other Vs that were added:

- **Veracity:** This describes the overall accuracy of the data. For the most part, large datasets have major issues with veracity. And yes, this has been a major challenge for generative AI.
- **Value:** This shows the usefulness or effectiveness of data. Often, this is a small part of the dataset. But sophisticated AI models have the ability to filter out the noise.
- **Variability:** Over time, data changes. This could be a reflection of society. Or it could be something like the changes in the climate. Variability is an important concept for generative AI. It actually goes to one of its issues with accuracy, since the data can reflect biases and misinformation.
- **Visualization:** This is about putting data in a graphical form, such as graphs or charts. This helps provide users with a better understanding of the trends.

Databases

A database is software that stores and manages data. It is essential for most applications. When computers emerged in the 1950s, databases quickly became one of the most important software systems. They were essential in automating corporate functions like payroll, HR, financials, and customer information.

They also helped to propel the growth of key industries. Consider the development of SABRE (Semi-automated Business Research Environment). This was in response to the problems of managing airline reservations. American Airlines teamed up with IBM to build SABRE to allow for automation of the process. It involved mainframe computers as well as sophisticated transactional platforms and databases.

During the 1970s, there would be a breakthrough with database technology. IBM computer scientist Edgar Codd published a pivotal research paper called “A Relational Model of Data for Large Shared Data Banks.” In it, he set forth the framework for the relational database.

It stored information in tables. Each table had rows and columns. This allowed for more structured data representation. For example, a column could be for the address, another one for the city, and so on.

Another innovation with the relational database was a scripting language, called SQL (Structured Query Language). This used simple commands to handle various operations. At the core was CRUD (create, read, update, and delete).

A relational database also had a way to connect tables. This was done by the use of primary and foreign keys. For example, a primary key was a unique value in a column. This could then be associated with the column in another table.

With this, you could create different types of relationships:

- **One-to-one:** A row in one table is connected to a row in another table. This is where the value is unique, such as a customer number.
- **One-to-many:** This is where the value in a row is linked to two or more rows in another table. An example is purchase orders that are connected to a customer table.
- **Many-to-many:** This is where more than one row is linked to multiple rows in another table. However, this can be complex and is not used as much as the other two types of relationships.

A major catalyst for the growth of databases was the PC revolution. It made it possible for even small businesses to create their own software. The growth trend was then bolstered with the growth of the Internet in the 1990s and mobile in the 2000s.

■ **Note** Codd's relational database got little interest from the management team at IBM. Perhaps the reason was that the company did not want to cannibalize its existing database franchise, which was called IMS (Information Management System). But entrepreneur Larry Ellison saw that relational database technology was the future. He used Codd's research as the basis for his company, Oracle, which he founded in 1977. By the 1990s, it would dominate the database market. Today, Oracle's market value is \$220 billion compared to IBM's \$126 billion.

Relational databases continue to be critical technology. But there are nagging issues with it. They include

- **Data sprawl:** It's common for different databases to be scattered across an organization. This means it can be challenging to get a 360-degree view of the data. This can make it difficult to create generative AI models.
- **Costs:** The licenses can be substantial. There are also the costs for ongoing maintenance, patches, and upgrades. This can be problematic for generative AI because of the need for working with massive datasets.
- **Features:** Relational databases were not built for some of the modern requirements. They tend to be less capable with high-velocity data and unstructured data. SQL is also not particularly robust for the complex requirements of generative AI.

In light of this, there has been considerable innovation in the database market during the past 20 years. This has largely been the result of the need to handle the massive workloads of social networks and cloud computing platforms.

What's interesting is that the innovation has come from open source projects. An example is Hadoop, a database that can handle large amounts of data. It would be the basis of a new category called the data warehouse.

Doug Cutting got the inspiration for Hadoop when he read a research paper from Google. It provided a sophisticated framework for a large-scale file system. Cutting teamed up with Mike Cafarella to code Hadoop. A key was a system called MapReduce, which allowed for managing data across various services. These would then be merged to allow a complete view of the dataset.

It would not take long until Hadoop and MapReduce would become standards in the Internet world. Companies like Facebook, Twitter, and Yahoo! used the technology to scale their platforms.

By 2012, there were next-generation versions of data warehouses. The one that quickly gained adoption was Apache Spark. The timing was spot on as AI was gaining lots of traction. For the most part, Apache Spark provided more efficiency, lower latency, and higher performance for workloads. The founders of the project were from Berkeley. They would found Databricks, which developed commercial systems for Apache Spark. It operates the technology in the cloud and allows for building AI models with Python-like notebooks.

In the summer of 2021, Databricks announced a massive \$1.6 billion Series H funding at a \$38 billion valuation. Some of the investors included Andreessen Horowitz, Fidelity, and Tiger Management.

By this time, Databricks had a new type of data warehouse – that is, a lakehouse. According to the company:

“Rather than being forced to move data out of the data lake, and between various disconnected and legacy systems for different use cases, Databricks customers are building lakehouses on AWS, Microsoft Azure and Google Cloud to support every data and analytics workload on a single platform. As a result, they’re able to avoid architectural complexity, significantly reduce infrastructure costs, increase data team productivity, and innovate faster.”⁸

In the meantime, there was innovation with the structures of database systems. There were, for example, the NoSQL platforms. These are nonrelational databases, in which data is stored in a free-form document model. This allows for creating more sophisticated AI models that can scale and operate in real time. NoSQL also handles unstructured and semi-structured data better.

The leading player in this market is MongoDB. The founders of the company were executives of DoubleClick, which included Dwight Merriman, Eliot Horowitz, and Kevin Ryan. When running this ad network platform, they had to build their own custom data stores. Existing databases could not effectively handle the workloads, where it was routine to process 400,000 ads per second. It was from this frustration that they got the idea to create a NoSQL database.

Today, MongoDB has over 39,000 customers across more than 100 countries. Since inception, users have downloaded the database more than 325 million times.⁹

⁸ www.databricks.com/company/newsroom/press-releases/databricks-raises-1-6-billion-series-h-investment-at-38-billion-valuation

⁹ www.mongodb.com/company

Another disruptive force in the database market is Snowflake. The origins of the company go back to 2012. The founders, which include Thierry Cruanes, Benoit Dageville, and Marcin Zukowski, were veterans at companies like Google, IBM, and Oracle. From their experience, they saw firsthand the challenges with traditional database technology.

A key innovation was to make Snowflake cloud native. This made it much easier to use for AI workloads. It also helped that Snowflake was fairly simple to operate. By filling out a short form, you can create an advanced database within a minute or so.

Snowflake is also becoming a platform for developing powerful applications. It has over 1700 – and the growth continues to be strong.

The Snowflake business model is also compelling. Instead of charging a license or user-based subscription, the pricing is based on consumption of the service. This allows for more alignment.

The overall growth of Snowflake has been impressive. Despite macro headwinds, the company posted a 67% increase in revenues in the third quarter of 2022 to \$557 million.¹⁰ It had over 7200 customers and 287 had trailing 12-month product revenue over \$1 million.

According to Snowflake CEO Frank Sloatman: “Data is becoming deeply ingrained in how global enterprises think, act, decide, and strategize. Relying on anecdotal observations will increasingly take a backseat to data-driven operations.”¹¹

When it comes to managing AI workloads, data lakes have become critical. This makes it possible to ingest enormous amounts of structured and unstructured data. This means there is little need for reformatting. Keep in mind that data lakes are integrated with platforms like Apache Spark and TensorFlow, which are for creating deep learning models.

Then there was yet another innovation in the database market: data stores. These are highly sophisticated databases that are focused on hyperscale AI projects.

Uber pioneered this technology. In the early days, the company could not find a platform to effectively handle its complex workloads. So it created a system called Michelangelo.

¹⁰<https://investors.snowflake.com/news/news-details/2022/Snowflake-Reports-Financial-Results-for-the-Third-Quarter-of-Fiscal-2023/default.aspx>

¹¹www.fool.com/earnings/call-transcripts/2022/12/01/snowflake-snow-q3-2023-earnings-call-transcript/

It is comprehensive. It allows teams to build, deploy, and operationalize AI across Uber. The technology has accelerated the time-to-market for new features.

And finally, there has been the emergence of graph database. This is specifically built for storing and managing relationships, such as with social networks.

The graph database has led to graph neural networks. In this system, the data points are called nodes, and they are connected by lines, which are known as edges. Sophisticated AI makes the connections. The technology has proven useful with drug discovery, recommendation engines, and fraud detection. But these are just the early use cases. There is currently work being done with generative AI applications like for computer graphics, materials science, and genomics. For example, pharma giant GSK has built a graph database with close to 500 billion nodes for its AI models.¹²

Cloud Models

The cloud is essential for AI development. This allows for centralization of the data. You can also scale the huge models, which is certainly important for generative AI. It's also easy to spin up GPUs or TPUs as well as use many types of AI tools and platforms. In terms of the costs, they are generally based on a consumption model.

But there are different types of cloud deployments:

- **Public cloud:** This is the most widely used model. Of course, the largest providers are AWS (Amazon Web Services), Microsoft Azure, and Google Cloud. A public cloud has a shared architecture or is multitenant. This means that customer data and applications are on the same system along with all the other customers. True, this can lead to security vulnerabilities. But the top providers have state-of-the-art protection capabilities, such as with identity management, encryption, and firewalls.
- **Private cloud:** This is not a multitenant architecture. Instead, a company will be the only tenant on the private cloud. They may house the infrastructure on their own premises or with a third party. Private clouds are for those companies that need a high level of security, such as for regulated industries like finance and healthcare. But unlike

¹²<https://blogs.nvidia.com/blog/2022/10/24/what-are-graph-neural-networks/>

the public cloud, the expenses are much higher since there is a need to build the infrastructure and maintain it.

- **Hybrid cloud:** As the name implies, this is a blend of public and private clouds. For example, a company will maintain its mission-critical applications and data on a private cloud. The rest will then be housed in the public cloud. The hybrid cloud has become popular because it provides for lower costs but also enables strong security and privacy. This approach may also be used when a company is taking its first steps toward migrating to the cloud.

■ **Note** Besides AWS, Azure, and Google cloud, there are other large tech companies that are gaining ground and investing substantial amounts in their own platforms. These include Oracle and IBM. But the cloud market is not a zero-sum game. It's becoming more common for companies to have a multicloud strategy. This can provide more redundancy but also help to lower costs.

In terms of the cloud wars, one of the early beneficiaries of generative AI is Microsoft, which is the largest investor in OpenAI. But this relationship has been more than just about capital. Microsoft has worked closely with the team at OpenAI. To this end, Microsoft has spent much in R&D to develop state-of-the-art supercomputers that can handle the complex workloads of generative AI.¹³

The company has leveraged these innovations in its Azure OpenAI service. Through this, you can access core generative AI services like GPT-3, GPT-4, DALL-E, and Embeddings. These services are available through a Python SDK (software development kit), REST APIs, or a web interface called Azure OpenAI Studio. Of course, there are the benefits like 99.9% uptime, enterprise-grade security, and privacy systems.

An interesting use case of Azure OpenAI is with CarMax, which is the largest car retailer in the United States that has annual sales over \$31 billion. They used generative AI to summarize 100,000 customer reviews into 5000, which were clearer and focused on key attributes. The system also used responsible AI to filter harmful, hateful, or sexual content. If this process had been done manually, it would have taken CarMax's content team – which includes ten people – about 11 years.¹⁴

¹³ <https://startups.microsoft.com/blog/azure-openai-service-for-startups/>

¹⁴ <https://news.microsoft.com/source/features/ai/azure-openai-service-helps-customers-accelerate-innovation-with-large-ai-models-microsoft-expands-availability/>

Some of the other benefits included improved search engine optimization and customer engagement.

To accelerate adoption of Azure OpenAI, Microsoft has created the Startups Founders Hub. This provides for up to \$150,000 in Azure credits as well as access to experts from Microsoft.

All this is not to imply that cloud hosting for generative AI is only about the megatech players. There are various startups that have emerged that are focused on AI workloads. These provide GPUs as virtualized services. This means not having to make the substantial capital investments in equipment. After all, when it comes to developing AI systems, it's much better to not have to worry about the infrastructure – which can distract from more important matters, such as with building better models.

A leading AI-focused cloud provider is CoreWeave. The company claims that its platform is 35 times faster and 80% less expensive than general public cloud solutions.¹⁵

CoreWeave allows access to cutting-edge systems like NVIDIA H100 Tensor Core GPUs. The pricing is \$2.23 per hour (as of late 2023).

CoreWeave also has hosting for common open source generative AI platforms like EleutherAI, Stable Diffusion, GPT-NeoX-20B, and BLOOM. There is even the ability to use a Kubernetes infrastructure, which allows for sophisticated cloud operations. This is certainly important for large projects for generative AI.

The business model is unique in the industry. CoreWeave charges only for the usage of resources, not when compute capacity is idle. This can certainly be a big cost saver for generative AI projects.

In December 2022, CoreWeave announced a \$100 million round of funding, and the lead investor was Magnetar Capital. At the time, the company was growing revenues at over 450%, and the machine learning segment was accelerating at a pace of more than 900%.¹⁶

Another player in this new hosting market is Lambda Labs. Besides providing access to high-end GPUs, there are integrations with popular machine learning tools like Jupyter, TensorFlow, and PyTorch.

¹⁵ www.coreweave.com/blog/looking-ahead-to-2023-how-nvidia-gpus-are-powering-a-new-era-of-ai-and-machine-learning

¹⁶ www.businesswire.com/news/home/20221206005040/en/CoreWeave-Secures-100M-to-Expand-NVIDIA-HGX-H100-GPU-Offering-Usher-in-New-Era-of-AI-Innovation

Here's the vision of the company: "We want to build a world where a single researcher can manage petaflops of GPU compute. A world where training a model on a datacenter scale computer is as easy as training a model on your laptop. A world where access to computation is as effortless and ubiquitous as electricity."

In July 2021, Lambda Labs announced a Series A round of \$24.5 million.¹⁷ The investors included 1517, Gradient Ventures, Bloomberg Beta, and Razer.

Data Strategy

There is no standard way to put together a data strategy for generative AI. It depends on your goals, organization, and resources.

But of course, there are some key principles to keep in mind like data collection, data evaluation, labeling, and data wrangling.

In the rest of this chapter, we'll take a look at them.

Data Collection

Collecting data seems like an easy concept, right? But this is far from the reality. If you want to get good results with generative AI, you need to be mindful of effective data collection. It's not just a matter taking an inventory of datasets across the organization.

In fact, even this process can be difficult. It's common for organizations to know the available datasets. Moreover, even when you identify them, it can be difficult to get access to these assets. Some of the reasons are privacy regulations, contractual obligations, and even internal politics.

Existing AI systems could have their own datasets. Or you can buy packages from third-party providers. But these sources may not necessarily be useful. This is why you should do some due diligence before making a purchase.

There are publicly available datasets that are free, such as from the government. You can find valuable sources from agencies like the Department of Commerce, the Labor Department, and the Federal Reserve.

¹⁷ <https://lambdalabs.com/blog/lambda-raises-24-5-to-build-a-gpu-cloud-and-deep-learning-infrastructure>

Finally, an organization may overlook internal datasets. Some include

- Social media posts and feedback
- Customer surveys
- Customer feedback surveys
- Zoom calls

For generative AI, web scraping is a common way to collect data. But there are inherent issues. The data may be subject to copyright protections, or it may have serious biases. This could mean skewed results for the generative AI model. There may even be exposure to lawsuits.

Common Data Sources

When it comes to creating models, there are certain free datasets that data scientists use. They will often appear in academic papers that feature something new about AI.

There are certainly many free datasets available. But here are some that are frequently used:

- **Data.gov:** This is a massive store of government databases. There are over 241,000 datasets. Some of the categories include energy, education, healthcare, agriculture, and climate.
- **US Healthcare Data:** This includes information about healthcare organizations, diseases, and drugs. The sources are from the Centers for Disease Control and Prevention (CDC), Food and Drug Administration (FDA), Centers for Medicare and Medicaid Services, and the Agency for Healthcare Research and Quality (AHRQ).
- **IMF (International Monetary Fund) Data:** This provides extensive information about foreign exchange, commodities, interest rates, and debt levels.
- **Google's Open Images:** This dataset has over 15.8 million images, which are labeled. This is common for building computer vision AI models.
- **MNIST (Modified National Institute of Standards and Technology database):** Data scientists, like Yann LeCun, launched this in 1999. It is now the largest database of handwritten single digits. There are about 60,000 training images and 10,000 testing images. The dataset is effective for real-world pattern recognition.

- **IMDb (Internet Movie Database) Movie Reviews:** This dataset has 25,000 reviews for training and 25,000 for testing. They are considered to be “polar” since the reviews are either positive or negative, not neutral. The dataset is used for sentiment analysis.
- **Project Gutenberg:** This has over 60,000 freely available eBooks, which were published from 1971 to 2021. This dataset is helpful with natural language processing and the building of LLMs using generative AI.
- **ApolloScape:** The developer of this dataset is Baidu, which is a top Chinese Internet company. ApolloScape has data for 3D perception LIDAR (Light Detection and Ranging) object detection and tracking, which are for autonomous driving models. It also has different traffic densities for challenging scenarios, such as whether there are pedestrians and bikes on the road.
- **BookCorpus:** This is a collection of over 11,000 novels for 16 genres like romance, historicals, and so on. These books are free and from unpublished authors. BookCorpus has been used for models like GPT-3, Google’s BERT (Bidirectional Encoder Representations from Transformers), and Amazon’s Bort.
- **OpenWebText:** This is composed of millions of shared URLs on Reddit. They each have at least three upvotes.
- **CommonCrawl News:** This is a dataset of news stories from websites across the globe.

Data Evaluation

Once you have a dataset, you can go through an evaluation process. This can mean avoiding wasting time on sources that will not help to achieve your goals.

Here are factors to consider when evaluating datasets:

- **Is the data relevant?** This is often something you cannot determine early in a project. Instead, there will need to be experimentation with the datasets to get a sense of what features are the most impactful for a model. A qualified data scientist can also be essential with this process.
- **Is the data timely?** Data can have a short life. If you use stale sources, your models will likely underperform.

- What are the potential legal issues? You want to make sure you have the right to use the datasets. This is why it's important to have the legal department involved early in the AI process.
- What is the target audience? You want a dataset that is reflective of your customers or users. For example, suppose you have a product that is mostly for women. Of course, you want a dataset that reflects this.
- What is your IT infrastructure? You want to make sure your organization is equipped to manage the datasets. Can it handle the volumes? If not, what will the costs be to buy new infrastructure?

Data Wrangling

Most datasets are messy. Data wrangling is how you can clean and restructure them. This is one of the most time-consuming parts of the AI process.

How does data wrangling work? There are various methods. One is to look for outliers in the datasets. This is where some of the elements are way out of the bounds than would be typical. Sometimes, the data can be nonsensical. For example, this is where there is a negative number for weight or age.

In some cases, outliers can actually be useful. It may even be critical. This is often the case with fraud detection, which is about low-probability events.

Sometimes, data wrangling is about handling fairly simple corrections. This would be something like duplications. Many datasets have these. But you want to make sure they are eliminated.

With data wrangling, you can create new types of data. This can provide for richer datasets. An example is with births. You can create a column for the age by subtracting this from the current year. This is known as a calculated field.

Data wrangling is also about reformatting data. This could involve the use of standardization. The fact is that many datasets have different formats. For example, when it comes to information about a country, it could have a full name like the “United States” or an abbreviation, say with “U.S.” Usually, it does not matter which you select. What's important is that there is consistency. If not, the model may produce bad outcomes.

Here are some other methods for standardizing data:

- Binning: There may not be a need to be too specific with the dataset. This is common with ranges. After all, do you need to segment the information for age groups with 21–22, 23–24, and so on? Probably not. Instead, you can have groups in increments of 10 or 20 years.

- **Conversion table:** You may have data that use different standards. This could mean converting numbers from the decimal to the metric system or going from Fahrenheit to Celsius.
- **One-hot encoding:** This can be complicated to define. So let's take an example. Let's say you have a dataset of iPhones: iPhone 14 Pro Max, iPhone 14 Pro, and iPhone 14 Plus. You can transform each into a number. Say the iPhone 14 Pro Max is 1, iPhone 14 Pro is 2, and so on. But there is a problem. An AI model may think that 2 is more significant than 1 – and this could produce unexpected results for the model. This is where one-hot encoding comes in. It could be something like having columns for `iPhone_14_Pro_Max`, `iPhone_14_Pro`, and `iPhone_14_Plus`. If the iPhone 14 Pro is in use, you would have 1 in the column and 0 in the others.

What happens if there is missing data? This is certainly common with datasets. To deal with this, there are different approaches. A simple one is to use an average of the numbers in the column, which can help to smooth out the gaps. But if there is a large number of missing items, you might just avoid the data. It may not provide useful predictive value in the AI model.

The good news is that there are various tools that can help automate the data wrangling process. A top company in the category is Trifacta, which is a division of Alteryx. Its system can do the following:

- **Active data profiling:** This is a sophisticated analysis to determine the accuracy, completeness, and validity of any dataset. The system does this by identifying data formats, key attributes, schemas, and relationships.
- **Predictive transformation:** This involves using a visual designer to resolve data transformation issues. A user does not require a technical background.
- **Adaptive data quality:** This will process datasets and provide recommendations for data quality rules. Examples include formatting patterns, handling column dependencies, and enforcing integrity constraints.
- **Data clustering and standardization:** This uses AI to provide better standardization for datasets.

Data Labeling

A label describes data. An example is with using AI for image recognition. You can have a dataset that has images of dogs and cats. For each, you will have the appropriate label.

Simple, right? Not really, actually. Putting a label together can be complicated. It's often more than just a word or two. For example, there may be an image of a white dog in which the background is white snow. All these details are important if you want a good model.

But as the images get more intricate, it can be tough to know what details to include. Or you may include too much information, which can make the label unwieldy.

Next, it can be time-consuming to create labels. Working on huge datasets – which can have millions of observations – could take years.

■ **Note** In 2006, AI researcher Fei-Fei Li faced a major challenge. She wanted to label the ImageNet dataset. The problem? It had 14 million images. Using some students to help with the labeling would take too much time. To accelerate the process, she instead used Amazon's Mechanical Turk. This involved paying thousands of people to do the labeling. It worked. ImageNet would then become a pivotal dataset in AI, allowing for testing major breakthroughs with deep learning models.

With generative AI, though, it's common to use unsupervised learning techniques that do not require labeled data. Yet there still may be parts of a project that will use labels. There may also be the combination of different types of models that use both structured and unstructured data.

There have been strides in finding approaches to scale the process. This can be with the use of similar data. Meta has used semi-structured data, such as tags on Instagram, for this.

But when it comes to many organizations, the approach is to use a third-party organization for the data labeling. This can actually be more cost-effective and accurate. These organizations will have specialists as well as the latest tools.

A leader in the category of labeling AI datasets is Scale AI. The cofounder and CEO is Alexandr Wang. Because of the huge success of Scale AI, he is the youngest self-made billionaire. This happened when he turned 25. While at the Massachusetts Institute for Technology, he met his cofounder, Lucy Guo. They received their initial backing from Y Combinator.¹⁸

¹⁸www.scmp.com/magazines/style/tech-design/article/3179833/next-elon-musk-meet-alexandr-wang-ceo-scale-ai-and

Scale AI's focus has been on building automation systems for the labeling of data, such as for text, audio, video, maps, and sensors. Since inception, the company has annotated 7.7 billion data items and over 1 billion 2D and 3D scenes.¹⁹

Here's a case study of Scale AI. The Datta Lab at Harvard Medical School uses AI to study neurological behavior of rodents. To collect the data, the researchers use cameras and neural implants. By doing this, it's possible to learn more about the brain.²⁰

But there needs to be labeling of the video footage, which involves measuring movements of the mice. The process can take weeks.

The Datta Lab realized it was much better to outsource this to Scale AI. The turnaround times were much faster – even overnight. There was also higher quality with the datasets. In the meantime, the researchers at the Datta Lab had more time to devote to more important matters.

Quantity

How much data do you need for an effective generative AI model? There is no bright-line answer for this. Every model is different. But for the most part, generative AI models rely on large amounts of data. To this end, there is the Hughes Phenomenon. This says that the more features you add, the better the accuracy.

Yet this can make generative AI challenging. The fact is that there are many categories where there is simply not enough data. True, simulated data can be helpful. But it is often better to have real-life data.

On the other hand, there are cases where a model can be overloaded with data. This can mean results that are skewed – or that degrade over time. There is a name for this. It's the “curse of dimensionality.”

Andrew Ng, who is the CEO of Landing AI and the former head of Google Brain, says that some models may work fine with only 100 data points. Again, it's all about understanding the use case.

There is even a field of study in AI called small data. This is where you can create models on limited datasets.

This is the case with Entities as Experts (EAE) from Google Research. It is a collection of people, organizations, and figures. For example, whereas a typical model may need 10–20 million entities, the EAE system could be effective with only 1 million.

¹⁹ <https://scale.com/about>

²⁰ <https://scale.com/customers/harvard-medical-school>

Privacy Laws and Regulations

In August 2022, Sephora paid a \$1.2 million fine for alleged noncompliance with California's privacy law.²¹ The claim was that the cosmetics retailer did not properly disclose information to customers about the use of their data. There was not an effective opt-out process.

Sephora was certainly not the only company in this situation. California regulators said that there were over 100 violation notices sent to other companies.

In the past few years, various states have passed their own laws for privacy for online data. They include Virginia, Colorado, Maine, Nevada, and Connecticut. However, the US federal government has not done so.

In terms of the impact on AI, the privacy laws are likely to be material. These laws often require the deletion of data. As a result, this could make datasets much less useful. In fact, there may even be problems with bias because the information may not be representative of the populations.

The privacy laws among the states often have gaps and inconsistencies. There may be rulemaking powers, such as in California. For the most part, the legal fragmentation makes it difficult to anticipate what the compliance requirements are.

Because of this, companies will need to spend money on legal counsel and consultants. This assistance will likely not be temporary. There will need to be ongoing changes and updates because of the evolution of the laws and regulations.

A misconception is that smaller organizations need not worry about this. But the reality is that states like California are not just looking at major enterprises. Thus, if a small business gets hit with a penalty, it could be a burden to pay off.

In the meantime, companies may need to deal with privacy laws in other countries, such as with the European Union's General Data Protection Regulation (GDPR).

Generative AI for Data Wrangling

Since generative AI can create content, this can be a way to deal with some of the problems with data. This can help mitigate issues like skewed and imbalanced datasets.

²¹ www.wsj.com/articles/sephora-agrees-to-1-2-million-settlement-of-data-privacy-charges-11661372755

Here's a look at some of the approaches:

- **Upsampling:** This is for increasing the number of samples for the minority class. A minority class is where there are a small number of observations in the dataset (this is also called a rare class or positive class). However, this can easily get overshadowed by the majority class. In other words, generative AI can provide more balance by creating new data. This can be done by replicating the minority class or using algorithms like SMOTE (Synthetic Minority Over-sampling Technique).
- **Downsampling:** Yes, this is where you decrease the number of the samples for the majority class. The removal can be done randomly or through the use of algorithms like Tomek links, which delete those areas that are nearest the minority class. Yet there are downsides to downsampling. The process may result in the removal of key data, which could lead to bias in the results.

Generative AI can also be helpful with missing data and outliers. As we learned in this chapter, they are a common problem, and this can lead to faulty models. True, you can take an average – but this is far from optimal.

Generative AI is likely to be a better approach. The reason is that it is based on a model that incorporates the underlying patterns of the data. This means that the new data will have the core features.

■ **Note** A common approach for data security is to use anonymization. This is the process of masking or obscuring the data, especially for personally identifiable information (PII) like Social Security numbers, addresses, and credit card numbers. But generative AI can be a much better approach than traditional anonymization. It can create synthetic data that does not contain any PII.

Generative AI for Data Preparation

After getting her master's degree in finance from Princeton, Sarah Nagy went to work for a Wall Street firm. Her role was as a quant, which meant she used highly sophisticated algorithms to improve returns on portfolios.²²

²²www.differential.vc/insights/on-the-record-with-sarah-nagy-ceo-of-seek-ai

But when AI started to get popular in 2013, she moved into the startup world. She would focus on working with massive datasets, such as from Twitter, YouTube, and Wikipedia. This meant writing sophisticated code, such as in Python.

Then when GPT-3 came out in 2020, she realized that generative AI could revolutionize data preparation. This technology had the potential to automate the manual aspects.

She saw how this could lead to big improvements in productivity. After all, it was common for business analysts to ask data scientists like Nagy to access datasets. But this was often a bottleneck since there were few who had the technical background for this.

To leverage this expertise, Nagy founded Seek AI in 2020. The company's platform has two main components:

- **Conversational engine:** Anyone can ask a question that they would typically say to a data scientist. This can be done with Slack, Teams, and email. Basically, there is no need for training on how to use Seek AI.
- **Knowledge base:** The questions are stored in a database. This helps to deal with repetitive questions. What's more, Seek AI uses deep learning to refine the answers.

In January 2022, Seek AI announced a pre-seed funding round of \$7.5 million. The investors included Conviction and Battery Ventures, along with Bob Muglia (the prior CEO of Snowflake), Mustafa Suleyman (cofounder of DeepMind), and Tristan Handy (founder and CEO of dbt Labs).

Another company that is leveraging generative AI for data preparation is Sisense. Founded in 2004, it is a developer of analytics tools for enterprises. Sisense has more than 2000 customers, which include Nasdaq, ZoomInfo, and Philips Healthcare.²³

When ChatGPT was released, the company wasted little time in creating an integration with its platform. They saw the technology as a way to transform traditional analytics. According to the company: "It has the potential to eliminate much of the pain and effort associated with traditional data preparation and, in the process, enables analysts to enable teams to segment and slice data in new ways, on-the-fly."²⁴

²³ www.sisense.com/about/

²⁴ www.sisense.com/blog/supercharge-analytics-with-chatgpt/?state=%2Fblog%2F

Sisense estimates that data preparation can take 70% to 80% of the time for an analytics project. Even a simple query – say to create a list of customers by industry – can take considerable time. The process may involve the following:

- Classifying the dataset
- Transferring it to a data warehouse or database
- Merging disparate third-party data sources
- Google searches to label the data
- The creation of Excel or Python scripts

It is typical for there to be different views of the same data, such as by country. This means going through the process again and again.

But by using a generative AI system, anyone can write a simple query, and the process will be automated. This could take minutes or seconds – not weeks or even months.

Chief Data Officer

As a sign of the importance of data, there is the chief data officer or CDO in the executive suite. This is a relatively new position. Capital One appointed the first CDO in 2002. Consider that the company has a long history as a leader in analytics.

But it's been during the past few years that there has been a notable increase in the number of CDOs. This has also not just been mostly for tech companies. CDOs are becoming more common in industries like insurance, media, entertainment, and retail, according to a survey from PwC (Pricewaterhouse Coopers).²⁵

As for the responsibilities for a CDO, they are wide ranging. This person will help with governance, managing datasets, improving their quality, and finding better ways for leveraging these assets. Often, this means being a critical part of AI projects.

²⁵www.cio.com/article/230880/what-is-a-chief-data-officer.html

■ **Note** Microsoft CEO Satya Nadella predicts that about 10% of all data could be created by generative AI by 2025.²⁶

Conclusion

If you have a skewed dataset, your models may be faulty or even misleading. This is why you should have a data strategy. This involves looking at factors like the collection, evaluation, wrangling, and labeling of data.

These processes can be time-consuming and complicated. But there are some approaches that can be a big help. There are also tools and third-party service providers that can streamline the process.

There are also many types of data. Examples include metadata, dark data, real-time data, and synthetic data. However, for purposes of generative AI, the two main categories are structured and unstructured data. In fact, one of the major innovations is using generative AI to find patterns in unstructured data.

Finally, the temptation may be to create massive datasets. While this may be a good rule of thumb, there are times when this can be the wrong approach. In other words, you want to evaluate each project rigorously.

In the next chapter, we'll take a look at the core technologies for AI models.

²⁶ www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html?action=click&pgtype=Article&state=default&module=style-artificial-intelligence&variant=show®ion=BELOW_MAIN_CONTENT&block=storyline_flex_guide_recirc

AI Fundamentals

Understanding the Core Foundations of Generative AI

One of history's most important conferences was in 1956 at Dartmouth University. Professor John McCarthy organized the ten-week event, which he called "Cerebral Mechanisms in Behavior." He also did something else that was historic. He coined the phrase "artificial intelligence."

At the time, there was much pushback about it. But no one else could come up with something better!

The attendees of the conference included many leading academics who would innovate artificial intelligence. They were Marvin Minsky, Oliver Selfridge, Ray Solomonoff, Claude Shannon, Julian Bigelow, Allen Newell, and Herbert Simon.

The topics for discussion were prescient. Some included neural networks and natural language processing (NLP).

The conference also saw the creation of the first AI program. The authors were Newell, Cliff Shaw, and Simon.

The program was called the Logic Theorist. The inspiration for this sprung from Simon's work at the Research and Development (RAND) Corporation. While there, he assisted with a computer that printed words on a map to help analyze scenarios for a Russian first strike.

It was actually a simple program. But Simon realized that computers were much more than just about numbers. They could work with words and images.

He used this insight for the Logic Theorist, which could solve math formulas. He based this on theorems in a famous text, *Principia Mathematica*. One of the coauthors, Bertrand Russell, was actually quite impressed with the results.

The story of Dartmouth University highlights that AI has been the result of the hard work of many brilliant people – over the decades. Much of the progress has been evolutionary, if not grudging. But the past 20 years has seen an acceleration of progress. There are also few signs of a slowdown.

To understand generative AI, it's important to have a general understanding of these building blocks. This is what we'll do in this chapter.

Early AI Programs

Some of today's AI technologies may seem like recent innovations. But the chances are that they are not. In the early days of AI, researchers and scientists were not only thinking of the many use cases for the technology – but also putting them into practice.

Many of these happened from the mid-1950s to 1970s. This period is often referred to as the “Golden Age of AI.” Much of the innovation came from universities – not private industry – like Stanford, MIT, Lincoln Laboratories, and Carnegie Mellon University. The funding came from the federal government. This was to help with the Cold War and the Apollo space program.

Here are some of the innovations:

- **LISP (1958):** McCarthy developed this, which is the first language for building AI applications. It had programming innovations like garbage collection, dynamic typing, and recursion.
- **Computer Vision (1966):** Minsky made this a summer project for an MIT student. He said he wanted to find a way to connect a camera to a computer so it could describe the world.

- ELIZA (1966): This was one of the first chatbots. The creator of the program, MIT professor Joseph Weizenbaum, made it into a tool for psychoanalysis.
- ANALOGY (1968): MIT professor Thomas Evans developed this program, which allowed for solving analogy problems from an IQ test.

AI Struggle

During the nascent stages of AI, there emerged two contending approaches. First of all, there was the symbolic system. This is essentially where programmers code the logic, such as with IF/THEN/ELSE constructs. These decision trees can result in insights.

Some of the practical applications of this were expert systems. During the 1980s, Carnegie Mellon University professor John McDermott created one called XCON (eXpert CONfigurer). It was similar to an Amazon recommendation system. With XCON, a user could optimally select components for computers. The technology caught the attention of DEC, which bundled the application with its hugely successful VAX (Virtual Address eXtension) computer platform. This would turn out to be a highly profitable business.

The next main approach to AI was the use of a neural network. In a sense, this would mimic the human brain.

The pioneer of this technology was Frank Rosenblatt, although he called his innovation the perceptron.

Rosenblatt had an interesting background. He was actually a research psychologist, with a PhD from Cornell University.

But his focus would be on cognitive systems, which used IBM mainframe technology. In 1958, he announced that one of his applications could recognize marks on the right and left sides of punch cards. His innovation was that he used data for this. It allowed the perceptron to learn.

This stirred up lots of buzz. Even *The New York Times* wrote about it. In the story, the author wrote: “Indeed, it strikes us as the first serious rival to the human brain ever devised.”¹

True, this would be an overreaction. But the perceptron would ultimately become the basis for major breakthroughs in AI, such as with deep learning and generative AI.

¹<https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon>

However, by 1969 the perceptron would come under intense scrutiny. The main attack was from Minsky and Seymour Papert, who published a book called *Perceptrons*. The authors noted that the technology was too basic.

Then a few years later, Rosenblatt would die in a tragic boating accident. This happened on his 43rd birthday.

While Rosenblatt's theories would ultimately become dominant – decades later – this is not to imply that symbolic AI was wrong. There are researchers who are looking at this approach. It may ultimately be a good complement to deep learning and generative AI systems.

■ **Note** Minsky would later admit that his book had limitations. He said it went too far in going after neural networks.

As much as possible, Rosenblatt wanted to use concepts of the human brain and apply them to computer learning. He did this with his perceptron or neural network.

For this to work, the data needs to be linearly separable. This means you can draw a line between two different types of data. You can see this in Figure 3-1.

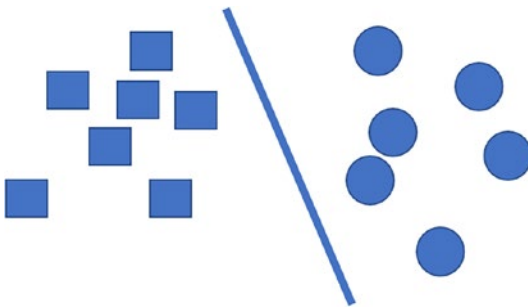


Figure 3-1. This shows data that is linearly separable

As for how the neural network works, it involves a process of ingesting data, analyzing it, and coming up with output. Figure 3-2 shows this.

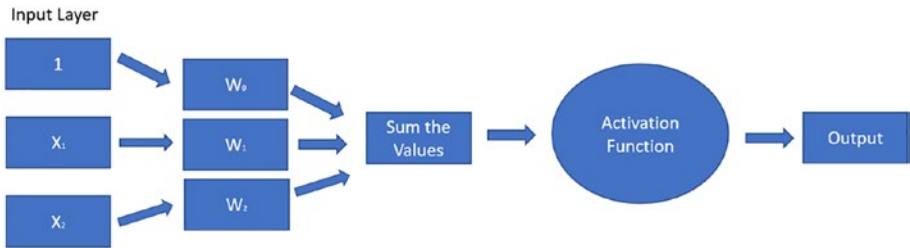


Figure 3-2. This is an illustration of a one-layer perceptron or neural network

On the left side, there is the input layer. This is where variables are entered into the neural network. They are denoted as X_1 and X_2 . There is also the value of 1, which is the bias or constant. This is used to help improve the performance of the model.

For all of these values, there is a corresponding weight. In the chart, these are W_0 to W_2 . The higher the weights, the higher the strength.

They will be multiplied by the variables and summed up. The value is then used for an algorithm, which is called an activation function. There are many available. But here are some of the more common ones:

- **Sigmoid function:** This provides an output of zero to one. If the value is zero, the value will not be passed to the next node.
- **Tanh or hyperbolic tangent function:** This converts the values from -1 to $+1$. The shape of a tanh function is similar to a sigmoid, but there is more emphasis on the 0 value. This type of activation function is usually effective for detecting large changes in the input layer.
- **Rectified linear unit or ReLU:** If the input is zero, then the output will be negative. Otherwise, the activation function will use the same input value. The ReLU is fairly simple to use and tends to work well with larger input values.
- **Softmax activation function:** This is generally for classification of datasets. The softmax activation function will compute the probability distribution for the classes. It will then select the one that has the highest prediction value.

With the neural network – at least the perceptron version – it's about a binary decision. For example, suppose we want to see if the model has recognized a square or a circle on a chart. The activation could determine this. It could be something like the output of one is a square, and zero is for the circle.

Yes, this is fairly simple. But the perceptron is the core approach for many types of modern AI-like deep learning systems.

Then What Is AI?

Understanding the concept of AI can be challenging. Let’s face it, the technology is based on complex systems. Even data scientists can struggle with them. Then there is the evolution of the technology.

Something else: There are terms like deep learning and machine learning. Are these also AI?

To help make sense of things, we can use an illustration, as seen in Figure 3-3. The big circle is classified as AI. It essentially includes all the cognitive technologies. Then there are subsets. These are areas like machine learning and natural language processing. There are even subsets within subsets. This is the case with deep learning, which is a subset of machine learning.

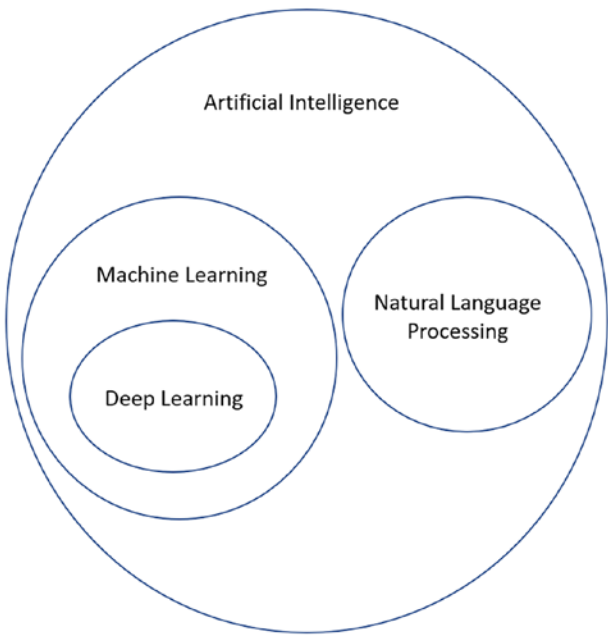


Figure 3-3. This shows the various categories that AI represents

Now let’s take a deeper look at these subsets.

Machine Learning

In 1959, Arthur Samuel came up with the definition of machine learning: “Field of study that gives computers the ability to learn without being explicitly programmed.”²

Samuel was a scientist at IBM. He demonstrated his ideas of machine learning by creating a game that played checkers. For his dataset, he used a book called *Lees’ Guide to the Game of Draughts or Checkers*. In 1961, his game would beat a top champion.

Samuel’s definition of machine learning still holds up quite well. But today, the category is definitely massive and encompasses many areas like supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning.

Supervised Learning

Simply put, supervised learning is when you have labeled data. This type of AI is one the earliest forms. But it is still widely used for projects.

A simple example is a dataset that can detect cats. It will have many photos, and each will have a label. They are expressed as pairs:

- Subject: This is the input value. In our case, this is the photo.
- Supervisory signal: This is the output value or label. It could be “cat” or a more descriptive phrase.

You will then apply an algorithm on this dataset, and it will attempt to find certain patterns. Because of the labeling, it’s easier to measure the accuracy of the model. This is done by using a loss function, which is a metric for the error rate.

Then what are some of the drawbacks with supervised learning? First of all, you generally need massive datasets to get useful results. But this can be difficult because of the challenges with labeling the data. As we learned in Chapter 2, there is much more unstructured data.

Next, you usually need lots of compute power for supervised learning. This can significantly add to the costs.

Finally, you cannot use supervised learning in real time. The reason is that you will not be able to label the data.

²<https://datascience.stackexchange.com/questions/37078/source-of-arthur-samuels-definition-of-machine-learning>

There are two main types of supervised learning:

- **Classification:** This is about segmenting or grouping the data based on common features. For example, this is often the approach with a spam filter. The data will be grouped into two labels – that is, spam or not spam.
- **Regression:** This is about showing correlations between the variables.

For both of these, there are many types of algorithms. We'll look at some of the main ones.

■ **Note** The category of “supervised learning” is a metaphor for a classroom. It essentially describes how a teacher tests their students.

Regression Analysis

Regression analysis is a powerful tool. The origins of this approach actually go back to the early 1880s. Mathematicians Adrien-Marie Legendre and Johann Carl Friedrich Gauss created it for the purposes of forecasting the paths of comets.

Since then, regression analysis has become a standard for many categories. For example, corporations use it for their quarterly and annual forecasts.

The basic equation for regression analysis is fairly simple:

$$Y = bX + c$$

- **Y:** This value, which is called a dependent variable, is the output for the model. It's what we are trying to predict.
- **B:** This is the coefficient. It shows the correlation between the dependent and independent variables.
- **X:** This is the independent variable. This is the data we input into the equation to produce the result.
- **C:** This is the constant. It's a value that remains the same regardless of the independent variable.

To create a model using the regression analysis, you will need a dataset of the independent and dependent variables. You will often use software to compute the values for the coefficient and constant. For AI purposes, this could be something like Python or R.

Let's take an example. Suppose we are putting together a model for spending on AI and the value of a company. Figure 3-4 shows the data.

AI Expenditures	Value Created
\$1 million	\$1.2 million
\$2 million	\$2.5 million
\$3 million	\$7.1 million
\$4 million	\$7 million
\$5 million	\$8.9 million
\$6 million	\$8.9 million
\$7 million	\$7.8 million
\$8 million	\$10 million
\$9 million	\$10.10 million
\$10 million	\$12.9 million

Figure 3-4. The data for how much a company has spent on AI and the value created from this

With this data, we compute the following regression equation:

$$Y = 1.063X + 1.7933$$

If we spend \$10 million, then the estimated increase in the company will be \$12.4 million, or 1.063 multiplied by 10 plus 1.7933. You can also plot the numbers on a chart, as seen in Figure 3-5.

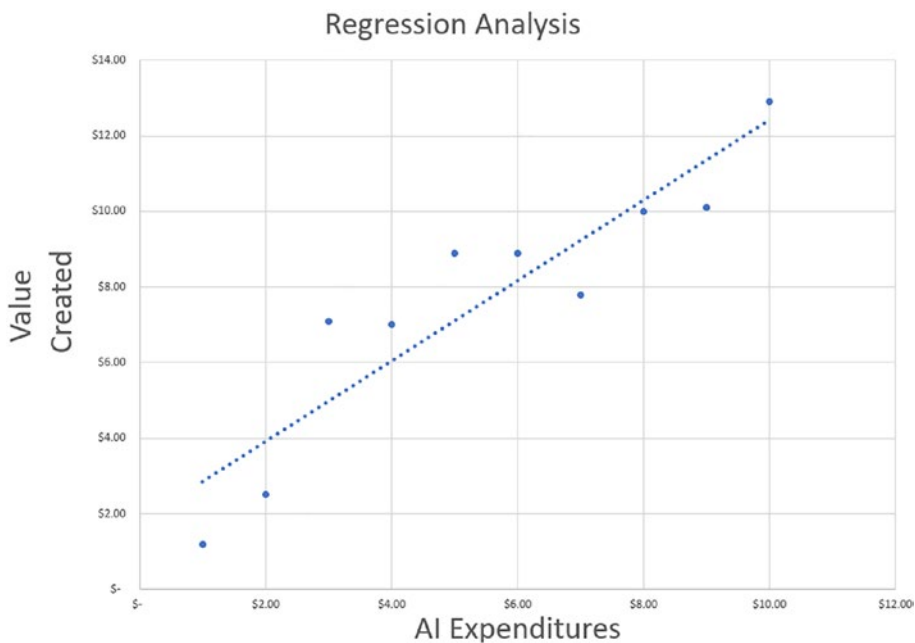


Figure 3-5. This is the regression line for data about expenditures on AI and the value created

As you can see, this is an upward sloping curve. This is when the coefficient is positive.

True, this is very simplistic. But no model can account for everything. There needs to be generalization and a focus on those factors that have the biggest impact.

You can also use multiple independent variables with a regression analysis. This provides much more depth to the model. What’s more, you will likely use different models and compare them. Through this process, you can better identify those independent variables that matter the most.

With regression analysis, there are methods to measure the accuracy of the results. First of all, there is the standard error metric. This shows the variations in the values for the constant and independent variables. For the most part, the lower the number, the better.

Next, there is the R-squared. This shows the explainability of the dependent variable. For example, suppose the R-squared is 85.5%. This means that 85.5% of the output is explained by the regression equation.

What is a strong R-squared? It depends on what you are measuring. If it is a complex area, such as with a biochemical reaction, an R-squared of 0.60 could be a good result.

To help get a better sense of the R-squared, there is the p-value. This is the variation of the explained values divided by the variation that is not explained. This is done by using randomization of the dataset. If the p-value is less than the statistical significance of the equation's values, then the result should be rejected.

There is another form of regression analysis called logistic regression. This is where the data does not have linear correlations. Instead, the chart may show something like an S shape.

Logistic regression deals with data that is binary. This could be something like “yes” or “no.”

An example would be a model that analyzes churn. By using a logistic function, you could find that you can increase the price from \$10 to \$15 on a product and not lose any accounts. But above this, there will be growing churn.

Support Vector Machines

A support vector machine (SVM) divides a dataset into two main classes. This is done by using a hyperplane. Figure 3-6 illustrates this.

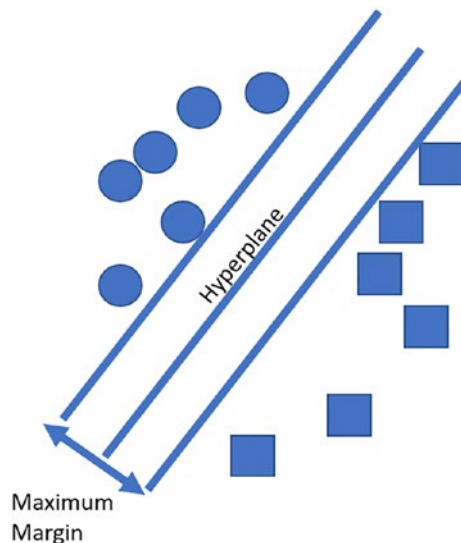


Figure 3-6. The line is the hyperplane that divides the two classes of data

The SVM algorithm optimizes this for the maximum margin between the two classes. This should allow for more accurate classification when using new data.

This works when the data is linearly separable. But what if the data is in clusters – such as in Figure 3-7?



Figure 3-7. This shows clustering of data

You can still use the hyperplane. This is because you can use more dimensions. You can see this in Figure 3-8.

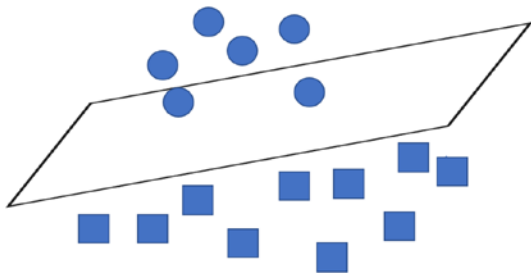


Figure 3-8. This shows a hyperplane in 3D space

As you can see, the hyperplane is no longer a line. It is a plane that divides the circles – which are at a high level – from the squares. This approach is called the kernel method.

Keep in mind that SVMs have proven to be effective when data has complex correlations. The same goes for dealing with outliers. They have less impact on the results because the focus is on the data closer to the hyperplane.

A drawback for SVMs is the need for large amounts of computer power. Furthermore, if there are many dimensions to the data, you will not be able to create a useful visualization.

Random Forest

The name “random forest” sounds somewhat strange for an AI model. But it is actually a good description of the process of the algorithm. A random forest model is a set of decision trees. Figure 3-9 is an example.

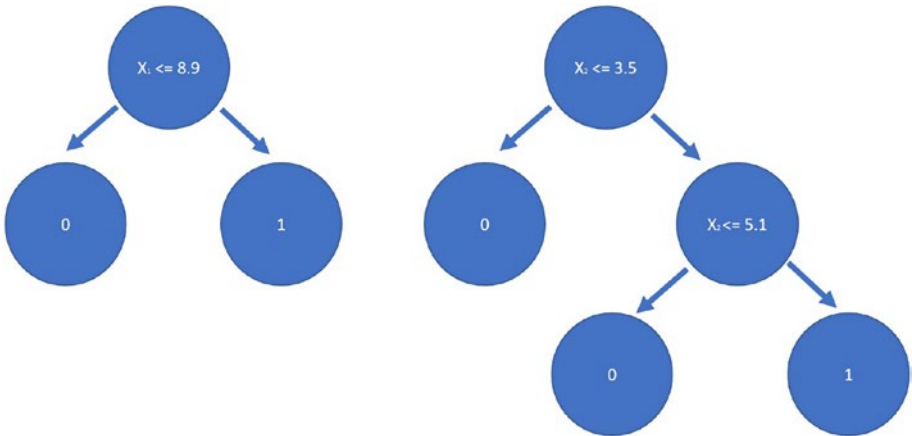


Figure 3-9. These are the decision trees for a random forest model

The first decision tree will see if a variable is less than or equal to 8.9. If so, the prediction is 0 or false. Otherwise, it is true. The next decision tree is a little more complex. It has two levels.

These decision trees are created with a random sample of a dataset. For example, this could be taking rows of information for different parts of it. Why do this? The reason is that there is generally less sensitivity to the predictions.

With the random forest, the model will iterate through different values. Then the prediction is decided by majority voting. For example, suppose there are five predictions for 0 and seven for 1. This means the prediction will be for 1.

Among AI algorithms, the random forest model is one of the easiest to use. There is no underlying complex math or theorems.

This does not imply that the results are not as good. If anything, a random forest model is a powerful tool for AI.

But there are downsides. If a tree has more levels, they can have more emphasis. This can skew the results of the model. A random forest also usually consumes lots of compute resources. The main reason is that this model is generally for large datasets.

K-Nearest Neighbor

The k-nearest neighbors' algorithm (KNN) is a way to classify or predict the grouping of data points. It's a mathematical way of affirming the saying: "Birds of a feather flock together."

The origins of this approach go back to 1951. Because of the simplicity of this model, it is one of the first a data scientist will learn.

To see how a KNN works, let's take a look at Figure 3-10.

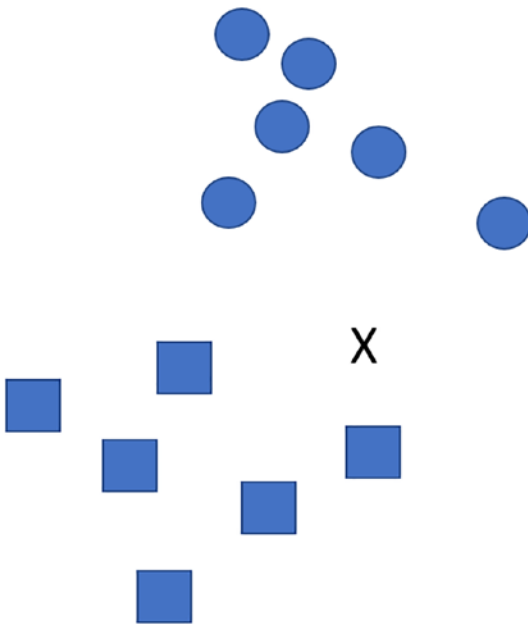


Figure 3-10. This is the scatter chart used for the k-nearest neighbors' algorithm

Here, we have two known types of data: circles and squares. But the X is unknown. Yes KNN will help to predict what type it is. This involves the following steps:

- Measure the distances between the unknown and known data. There are various ways to do this. One of the most common is the algorithm for Euclidean distance.
- Select a number for K of the nearest known data points from the unknown one. The value can be whatever you want. But it is typical for it to be odd. This is to help break ties. Or there can be a weighted value for the Ks. This would mean that those nearest the unknown data point have the most emphasis.

- The predicted value for the unknown data point is based on majority voting. For example, if there are more circles and squares for the Ks, then the prediction would be that the value is a circle.

A KNN algorithm is also referred to as a “lazy learning” model. This means that there is no use of complex systems to train the data. This means less compute power.

Another characteristic of a KNN algorithm is that it is nonparametric. This means there are no assumptions needed for the data.

But there are some drawbacks. A KNN’s performance tends to degrade greatly for large datasets. There also needs to be a large amount of memory required, which can be costly. Another issue is that the use of the K values can be arbitrary. This can greatly skew the model.

Naïve Bayes Classifiers

Naïve Bayes classifiers are based on Bayes’ Theorem, which uses probabilities. It’s about the chances of event #1 happening given event #2. But the two events are independent of each other. This is why the algorithm is called “naïve.”

Why is this important? It makes the process easier. There is no need to keep retaining the datasets from scratch.

This is why Naïve Bayes classifiers have been so useful for spam filters. The updating for new data can be quicker. The model should also show better performance.

Unsupervised Learning

With unsupervised learning, the data for the AI model is not labeled. The algorithms will try to detect the underlying patterns. They may be subtle – to the point that humans would not recognize them. This is primarily because of the large amounts of data.

Unsupervised learning has been essential for generative AI. The reason is that much of the available data is unstructured.

However, unsupervised learning can be complex. You often need the expertise of a data scientist and enormous amounts of compute power.

Something else: It can be difficult to measure the accuracy of the models. After all, they are based on the estimates from processing the unstructured data. There are no labels to compare against.

Some of the common methods for unsupervised learning include clustering, anomaly detection, association, and autoencoders.

Clustering

Clustering, which is the most common approach for unsupervised learning, involves grouping similar data. By doing this, the AI model can detect the patterns.

The process for clustering requires a closeness function. This shows the similarity or distance of the data.

However, it is not easy to come up with the right function. Because of this, a data scientist will use several types to get better results.

Here are some of the functions for clustering:

- **Centroid-based clustering:** This uses the K-means neighbor algorithm. There are iterations across various data items to minimize the differences.
- **Hierarchical-based clustering:** This uses a decision tree structure, similar to a random forest. This is where the data conforms to some sort of taxonomy.
- **Density-based clustering:** This looks for concentrations of data. These should also be surrounded by empty space or sparse data items.

Anomaly Detection

Anomaly detection analyzes data – such as logs, web traffic, and so on – to find unusual occurrences. This may ultimately be noise. But in some cases, the anomalies may be threats, such as malware.

Automating the process can be extremely helpful. This can mean providing protection for an organization on a 24/7 basis.

Here are some of the anomalies for AI systems:

- **Contextual anomalies:** This takes into account the current environment, say the time of the day. If there is a credit card transaction at 4:30 AM from a person who has never made a transaction at this time, then this could be a fraud.
- **Collective anomalies:** This looks at a group of data items that are beyond the typical boundaries. However, within the group, the information may be consistent.
- **Point anomalies:** This is where one data item is an outlier. This could be if an invoice has a much higher amount than is typical for the customer.

An anomaly may actually be something that is positive. For example, it could be an early sign of a new trend. This could be good information to pass along to the marketing department.

Association

Association essentially uses IF/THEN rules to show the probabilities between different data items. Those that have the most frequency are considered to have more weight. This is called support.

However, there is a confidence measure. This shows the success of the predictiveness of the rules. Yes, there are times when a rule has strong support, but weak confidence.

Then there is the lift value. This is the confidence divided by support. If it is negative, there is a negative correlation for the data items and vice versa.

A common application of association is the recommendation engine. The biggest impact is when the AI finds patterns that are unexpected. For example, on the Amazon web page, you may select a book. But based on using AI with huge amounts of data, the recommendations could be something that you never thought about – but still are very interested in. This can certainly lead to higher sales conversions.

Autoencoders

An autoencoder is a machine learning algorithm that can reduce the noise in a dataset, extract features, and detect anomalies. This is done in two steps. First, there is an encoder that compresses the data, which is done with a neural network model. This puts it into a lower-dimensional space. This is called the latent space.

Next, there is the decoder. This will use complex neural networks to reconfigure the original data.

To be effective, the dataset usually needs lots of redundancies. The data should also be high-dimensional. This means that the number of features – that is, the variables – is nearly the same or more than the data points.

You can use a version of an autoencoder for generative AI. This is with a variational autoencoder or VAE. This is generally used to create images. The VAE system essentially takes samples of the data, and this can create unique representations.

■ **Note** A blog from Nvidia has this for a definition of an autoencoder: “[They] take input data, compress it into a code, then try to recreate the input data from that summarized code. It’s like starting with Moby Dick, creating a SparkNotes version, and then trying to rewrite the original story using only the SparkNotes for reference.”³

Reinforcement Learning

Reinforcement learning is a type of AI in which the system learns to interact with an environment. To do this, there is a reward-punishment approach. For example, if the AI gets something wrong, then it will avoid this in the future. But if an action results in something that is correct, then this will be reinforced. This is essentially a process of trial and error.

There have been great strides with reinforcement learning for building systems to win games. It helps that there is an underlying structure – say a game board. There is also a set of rules. In other words, there is a defined environment, which allows for better learning.

A pioneer in game-based reinforcement learning is DeepMind. Founded in 2010, the company is more like a college lab. It has many PhDs across areas like data science, neuroscience, mathematics, physics, and engineering.

But it was with the testing of AI for games where DeepMind got traction. An early model learned to play 49 Atari games by analyzing the pixel movements and the score.⁴

Then DeepMind started to build a system for Go. The origins of this game go back more than 2500 years in China. It is considered the world’s oldest board game.

Go’s board has a grid of 19x19 lines. While this seems small, the game actually has more potential moves than the number of atoms in the observable universe.

Given this, the conventional wisdom was that AI would fail in becoming a great player. But by leveraging reinforcement learning – as well as other complex AI models – DeepMind was able to prove the critics wrong.

In 2014, the company developed AlphaGo, and it would easily beat other online gaming systems. A year later, it would defeat Fan Hui, who was a European Go champion. AlphaGo took five games to zero. DeepMind published its findings about its technology in *Nature*.

³ <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/>

⁴ www.deepmind.com/about

At first, the training was through processing games from top players. This certainly built a strong system. However, it got even better with the training of millions of simulated games. This significantly accelerated the learning process.

In 2016, AlphaGo played Lee Sedol, who was one of the world's best Go players. The event was live streamed. AlphaGo would win four to one. According to Lee: "With the debut of AI in Go games, I've realized that I'm not at the top even if I become the No. 1 through frantic efforts. Even if I become the No. 1, there is an entity that cannot be defeated."⁵

AlphaGo would be a lever for other breakthroughs. Mustafa Suleyman, the cofounder of DeepMind, said: "The amazing thing there was that it learned some incredible new strategies that had never been discovered before. And that was the holy grail for me. I was like, 'Okay, we really can train an algorithm to discover new knowledge, new insights. How could we apply that? How can we use that training method for real world problems?'"⁶

DeepMind would go on to use reinforcement for applications like the diagnosis of eye disease – with effectiveness at levels of top doctors – as well as for finding ways to conserve energy in data centers.⁷

Regarding generative AI, this can help build stronger reinforcement learning models. One area is with physical robots. These use reinforcement learning to help navigate their environment. However, generative AI can be used to create simulated data, which can help improve a robot's ability to move efficiently. This could also be helpful with other applications like self-driving cars and other autonomous systems.

Terms and Setup for Deep Learning

Deep learning is a complicated topic. So to get a better understanding of it, it's important to learn about some of the terms and setup.

The data that you feed into a deep learning model is called a vector or tuple. This is a list of values, which are called scalars. A vector looks like this:

(-1.0, 3.5, 2.7, 8.9, 4.0)

The scalars are the features or variables for the predictions. Figure 3-11 is an example.

⁵www.businessinsider.com/deep-mind-alphago-ai-lee-sedol-south-korea-go-2019-11

⁶<https://greylock.com/firm-news/welcome-mustafa-suleyman/>

⁷www.deepmind.com/about

Customer ID	VIP Status	Sales Rep	Sales to Date
1001	1	101	\$15,000
1002	1	101	\$25,000
1003	0	100	\$5,000
1004	0	101	\$1,000

Figure 3-11. This is a sample of a dataset for a deep learning model

The features are the headings for this table, such as Customer ID, VIP Status, and so on. Then each row is known as a sample. That is, there is one for each of the customers.

Before using deep learning – or other types of AI system – there will usually be preprocessing of the data. This will convert the items into a format that can be handled more efficiently by the AI algorithm. This is called normalization of data.

Here’s what can be done for certain types of data:

- **Categorical variables:** This is where data is grouped, such as by age, income level, or geography. To normalize this data, you can use an integer or one-hot encoding.
- **Continuous data:** This is usually numerical data that has a value within a defined range. It often refers to measurements – such as for height or weight – as well as prices, and quantities.
- **Ordinal data:** This describes using a system for ranking, say by scores for students.
- **Images:** There will be a vector of pixel values.
- **Speech:** This is a time-series vector.
- **Text:** There are various techniques like TF-IDF (term frequency–inverse document frequency). This is a way to show the relevancy of words.

Keep in mind that this is for structured data. But what if the deep learning model is using unsupervised learning? You will not have to use the preprocessing techniques. Instead, the deep learning models will detect the patterns in the raw datasets and extract the features.

How Deep Learning Works

Deep learning is based on a neural network. As we saw, this attempts to mimic the function of the brain. However, with deep learning, the process is much more complex, involving a large number of variables or parameters.

At a basic level, this is similar to regression analysis or logistic regression. You will specify random variables to train the model, which will create the weights for the data that's for the input layer. This iteration will continue until the errors are minimized. This should allow for better predictions for the model. Figure 3-12 is an illustration of a deep learning model.

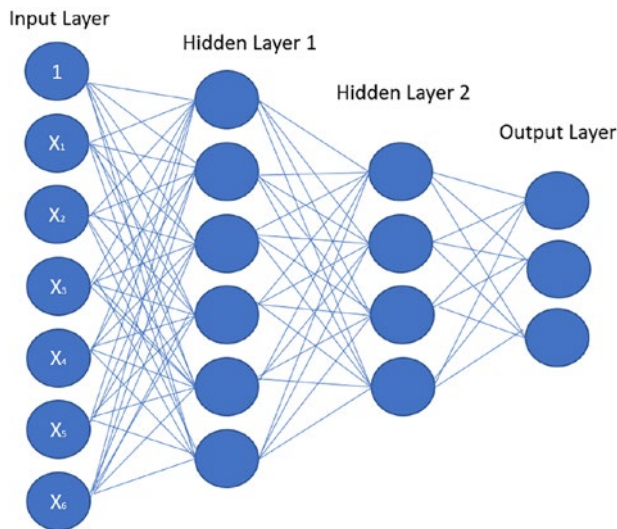


Figure 3-12. This shows a deep learning model with two hidden layers

Compared to a neural network, a deep learning system has many relationships between the variables and the weights. There are also multiple layers for the processing with activation functions. They are called hidden layers. The output of each of the nodes – these are the circles in Figure 3-12 – will be the input values for the next layer. This is where the algorithm learns, such as by detecting patterns and reducing the noise in the data. The more hidden layers, the deeper the model.

However, this is not to imply that you should have a large number of hidden layers. This may actually result in bad results. For the most part, there should be experimentation with the number of hidden layers. This is where the experience of a data scientist can be extremely helpful.

Even with a sophisticated deep learning model, there are often inaccurate results. To deal with this, there is backpropagation. This is where the output of the model will be sent back to the input layer. There will then be another iteration of the processing.

Backpropagation is one of the most pivotal innovations in AI. Professors Geoffrey Hinton, David Rumelhart, and Ronald J. Williams came up with this concept in 1986, when they published “Learning Representations by Back-propagating Errors.” It is one of the most cited academic papers in computer science.

With backpropagation, there must be experimentation with the number of iterations. If there are too many, the results may get more inaccurate. The focus is to find a middle point.

Types of Deep Learning

Over the years, there have emerged a variety of deep learning systems. One of them, called a generative adversarial network (GAN), is for generative AI. We’ll take a look at this in the next chapter.

As for the others, a widely used model is the convolutional neural network or CNN. The creator of this model is Yann LeCun. In 1998, he wrote a pathbreaking paper about this called “Gradient-Based Learning Applied to Document Recognition.”

He got the idea for the CNN when studying the findings of Nobel Prize winners David Hubel and Torsten Wiesel, who did research on the visual cortex. LeCun thought this could be turned into an effective AI model.

With the CNN, the input layer includes an image, and each pixel will have a value (from 0 to 255) for the intensity. A deep learning model iterates across the image to get improved results (each of these passes is called a convolution). For example, the initial image may be a set of random lines. The CNN will then find the shape and then ultimately identify the object.

Before this, the approach to object recognition was to use handcrafted feature engineering. This is where data scientists would define the angles and shapes of the objects. Of course, this did not scale particularly well. It was also tough to create the rules.

In 2012, Hinton led a group of AI researchers in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) to test the power of CNNs. The model had an error rate of only 16% with recognition of the images in the dataset. The second best had 26.2%.

This was groundbreaking and sparked the revolution in deep learning. Over the next few years, there were dramatic improvements in the error rates. The Challenge was actually ended because they got to the point where they were below a mere 2%, which was better than for a human (at about 5% or so).

According to a paper from Fei-Fei Li and Ranjay Krishna, called “Searching for Computer Vision North Stars”:

“With so many people working on neural networks, the technology advanced rapidly. Researchers found that the deeper the model, the better it performed at object recognition. And as deeper models required more processing power, researchers ran into other problems, such as computational bottlenecks, which required further design work to overcome. The ImageNet Challenge created a kind of domino effect of innovations, with each advance leading to more.”⁸

Another important deep learning model is the recurrent neural network or RNN. This is primarily for a sequence of text or numbers, say for stock prices.

An RNN connects a large number of neural networks. They feed input and output when analyzing the text. There are different ways to do this like the LSTM (long short-term memory) technique. Professors Sepp Hochreiter and Jürgen Schmidhuber wrote about this in an academic paper in 1997.

A common use of an RNN is an AI-based chat app like Siri or Alexa. The model uses an activation function to process prior inputs over time. You will notice this when you enter “PI” in a texting app, and there will be a list to suggest options, such as “Play” and “Please.”

Google used RNNs for its Translate App. It was able to handle over 100 languages and process billions of words per day.

In 2017, Google would publish a paper that leveraged its learnings with RNNs and deep learning. It was about transformers. This technology has quickly become essential for generative AI. We’ll look at this in much more detail in the next chapter.

■ **Note** HNC Software was one of the first companies to successfully commercialize deep learning and backpropagation. During the late 1980s, the company started to develop applications to help detect credit card fraud. The result was that it helped to save billions of dollars for customers. In 2002, Fair Isaac acquired HNC Software for \$810 million.

⁸ www.amacad.org/publication/searching-computer-vision-north-stars

The Brain and AI

Earlier, we saw that there are similarities between the brain and deep learning. Much of this is due to the retina. It processes visual images of the real world, which is the data. The brain uses a complex network of neurons – which come to about 86 billion – to assign weights to this information. These act as the CPUs (central processing units). The synapses then connect the neurons. The learning is about strong and weak connections.

While there are similarities between the brain and deep learning – there are also major differences. Of course, the underlying infrastructure is divergent. A computer is mostly made up of silicon chips, wires, power supplies, and software. The brain, on the other hand, is biochemical. It is also part of a complex body that includes many organs.

The brain is fairly small, especially compared to its awesome power. It weighs about 3.3 pounds and uses only about 20 watts of energy. As for AI, it relies on massive data centers that consume huge amounts of electricity.

Keep in mind that there is much that is not known about the brain. But then again, there is considerable research being done across the world.

In 2013, President Obama launched the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies). It's a public-private research organization to better understand the brain. This is similar to the Human Genome Project, which mapped the sequences of human DNA.

With the BRAIN Initiative, the goals are to make breakthroughs with intractable diseases like Alzheimer's and Parkinson's. But the research could also prove useful for AI. Understanding how the brain works is likely to help with building better models.

■ **Note** The cerebral cortex is the part of the process that provides cognitive abilities. This could actually be a source of insight for creating more advanced AI models.

Drawbacks with Deep Learning

Deep learning has certainly been a game changer. It seems that there is a new innovation every day. The technology has also had an impact across many industries like healthcare, energy, finance, and so on.

But as with any AI system, there are issues. Some can be quite troubling.

What are they? Gary Marcus wrote a paper entitled “Deep Learning: A Critical Appraisal,” which provides a comprehensive look at the problems with deep learning. He is a professor of psychology and neural science at New York University and the author of the *Future of the Brain*. Before this, he sold his AI startup to Uber.

Here are some of his criticisms about deep learning:

- **Black box:** Models can have billions of parameters. In other words, can a person really understand all this? Probably not. The lack of explainability can make it difficult to get adoption of deep learning especially for regulated industries like healthcare.
- **High-order thinking:** Deep learning systems cannot understand complex concepts like justice, philosophy, and politics. They also do not have common sense. This can mean that you can say something ridiculous to an AI – like “How to use a mop to clean a disk drive” – and it will try to understand it.
- **Butterfly effect:** Deep learning systems can be highly sensitive. Even a small change in the assumptions can have a major impact on the results of the model. This phenomenon is known as an “adversarial example.” A slight tweak in the model can mean that an AI system thinks an object that is an elephant is, say, a dog.

Overfitting and Underfitting

The training of an AI model can go to extremes. This can lead to certain problems, such as

- **Overfitting:** The model is not learning from the data. Instead, it is just memorizing the underlying patterns. Ironically, this is often when a model has a high rate of accuracy – such as over 90%. To deal with overfitting, you can collect more diverse data or use a less complex AI algorithm.
- **Underfitting:** Simply put, the model does not represent the real world. To deal with this, you can increase the number of parameters. Or you could use a more advanced AI algorithm.

Python

The Python computer language is the standard when it comes to coding AI models. This is an open source system that has been around since the late 1980s. The inventor of Python is Guido van Rossum, who is a computer scientist. He came up with the name because he was a fan of the British comedy, *Monty Python*.

van Rossum set forth principles for the language. He called them “The Zen of Python.” Here are some examples:⁹

- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Flat is better than nested.
- Sparse is better than dense.
- Readability counts.

When using Python, it’s clear that these principles are a major part of the DNA of the code. But the focus on simplicity does not mean that the language is not powerful. Rather, it is like any top-notch language, and it has been the foundation of mission-critical applications.

A major benefit of Python is the rich ecosystem of third-party developers. You can add sophisticated AI functions by importing modules, such as the following:

- **NumPy:** This allows for the managing of multidimensional data, which are common for deep learning and generative AI models. This also allows for working with linear algebra, shape manipulation, and Fourier transformations. NumPy also has the benefit of high speed with the calculations. You can even do data wrangling with it.
- **Pandas:** This is built on top of NumPy. Pandas is focused on managing datasets. Some of the functions include the cleaning, exploration, merging, indexing, and manipulation of data.

⁹<https://learnpython.com/blog/zen-of-python/>

- **Matplotlib:** This is for creating visualizations. They can be static, animated, or interactive. You can also zoom/pan and update the visualizations. The system works seamlessly with NumPy and Pandas.

These are just a few of the tools available. But this can make AI projects difficult, in terms of managing the systems.

What to do? It's common to use a platform like Anaconda. It's been around since 2012 and has had over 30 million downloads. Anaconda essentially allows for the easy installation and management of Python-based open source projects. "Our tools are trusted by businesses and governments around the world because of our ability to harness the innovation of the open-source Python community and deliver it in a secure, reliable, and impactful way," said Peter Wang, who is the CEO and cofounder of Anaconda.

The company is also a major supporter of the data science and Python development ecosystem. It has invested more than \$30 million in projects like Pandas, Dask, Numba, Bokeh, and HoloViews.

■ **Note** Compared to other languages like C++, C#, and Java, Python is fairly easy to learn. There are plenty of free tutorials on the Internet and YouTube.

AI Tools

There is a rich ecosystem of tools and platforms for AI. On the G2.com website, there are over 400 available.

But there are some that have become standard. Consider Jupyter Notebook. This is an open source tool that you access via the Web. With it, you can create AI models. This includes importing modules and systems, creating visualizations, ingesting data, and sharing projects, say on GitHub. There are also powerful systems that enable the creation of AI models.

Next, there is TensorFlow, which Google began development on in 2011. The company was struggling to develop neural network models. So the goal of TensorFlow was to streamline the process. This would scale the implementation of AI across Google.

In 2015, the company made TensorFlow open source. The reason was that Google wanted to help drive innovation in AI.

From the start, TensorFlow got much traction. Today, the software gets about 15 million downloads per month and is installed on over 2 billion mobile devices.¹⁰

TensorFlow ingests the input layer as a multidimensional array or a tensor. Yes, there is a “flow” aspect to this as the tensors travel through the models. This is also done at high speed. This is critical because AI models can be enormous.

The TensorFlow system is versatile. It can operate models on CPUs, GPUs, and TPUs (tensor processing units).

Besides TensorFlow, there are some other popular AI platforms. For example, there is Scikit-learn. This is one of the early systems. David Cournapeau started it at a Google Summer of Code project. Since then, it has grown quickly and has turned into a comprehensive AI development platform. Scikit-learn is written on Python and is built on Pandas, NumPy, and Matplotlib. There is also a transformer system, which allows for the creation of generative AI models.

Then there is PyTorch. Written in Python, Meta released this platform in 2016. It’s similar to the capabilities of TensorFlow. However, the UI for PyTorch is more intuitive. You can also make adjustments to an AI model in runtime. Keep in mind that AI21 Labs used PyTorch to create its generative AI models.¹¹ In 2021, the company released Jurassic-1 Jumbo, a large language model. It had 178 billion parameters and powered its flagship application, Wordtune. It is a writing and editing assistant.

As should be no surprise, AI platforms can be complex to use. They are often for data scientists. But there is a platform that is focused on beginners: Keras. According to the company: “Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user actions required for common use cases, and it provides clear and actionable feedback upon user error.”¹²

For some models, it may take just a few minutes to spin them up. This can be a good way to experiment. To boost the power, there is integration with TensorFlow. Some of the users of Keras are Uber, Netflix, Block, and even NASA.

¹⁰<https://pypistats.org/packages/tensorflow>

¹¹<https://aws.amazon.com/solutions/case-studies/AI21-case-study-p4d/?pg=ln&sec=c>

¹²https://keras.io/why_keras/

Often, these platforms are used with hyperscale cloud providers like AWS (Amazon Web Services), Microsoft Azure, and Google Cloud. You can scale a model to massive levels, such as where you have access to hundreds or thousands of GPUs as well as petabit-scale networking infrastructure and storage.

You'll typically use containerization for these cloud platforms. This is where you create a separate instance or virtual machine. It will essentially be its own computer with its own operating system, dependencies, and modules. This makes it much easier for management, especially since there are often software updates.

While open source AI platforms have become ubiquitous, there are many proprietary solutions that have become important. These are often called AutoML systems. They are often built for people who are not data scientists. Instead, an AutoML system will handle much of the advanced data wrangling and model deployment. This is all a part of a trend of the “citizen data scientist.”

AI Systems for Beginners

You do not have to be a data scientist to create AI models. There are several platforms that are geared for nontechnical people.

Consider Teachable Machine from Google.¹³ You can upload a dataset, such as of images and sounds. You can then train your model. There are features for fine-tuning it. When the model is finished, you can export it. This can be posted on the Web or even made available for TensorFlow. Figure 3-13 shows the app.

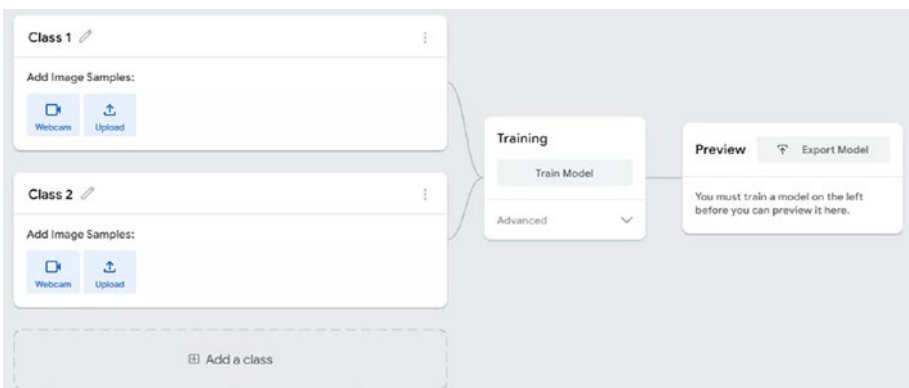


Figure 3-13. This is the interface for Teachable Machine from Google

¹³ <https://teachablemachine.withgoogle.com/>

The Machine Learning Playground¹⁴ is another interesting beginner's system. This is more sophisticated than the Google system. Yet it is still fairly easy to use, such as with drag-and-drop. With the Machine Learning Playground, you can create models using k-nearest neighbors, the perceptron, decision trees, artificial neural networks, and support vector machines. Figure 3-14 shows the system.

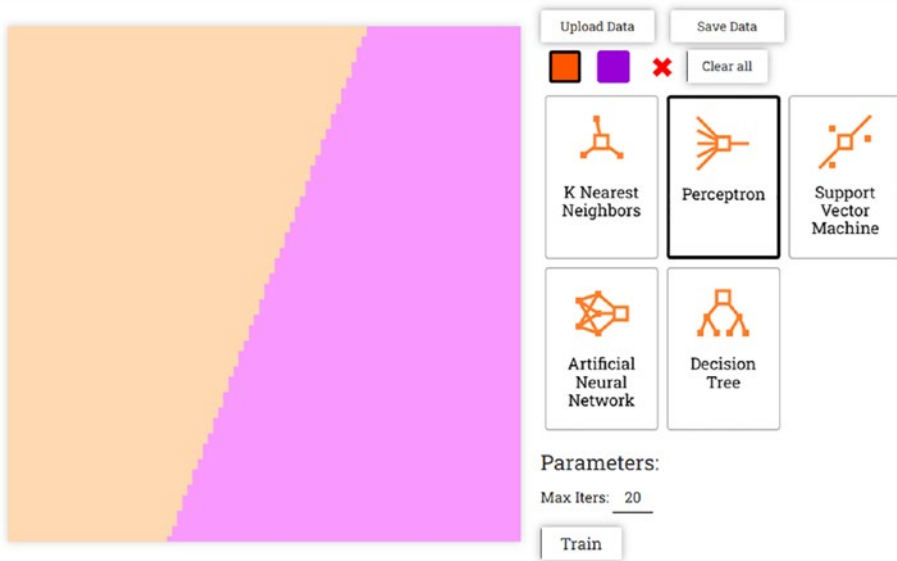


Figure 3-14. This is the interface for the Machine Learning Playground

Conclusion

In this chapter, we got a good background on the core elements of traditional AI. We first looked at a brief history of the technology. Despite the limitations – such as with the lack of large amounts of data and high-powered computers – the innovations were still impressive. During the early era of AI, there were the beginnings of important concepts like neural networks.

We also looked at the main subsets of AI like machine learning, deep learning, and reinforcement learning. They have all seen tremendous progress, especially since 2012. These technologies have also provided the basis for generative AI.

Then we covered the types of tools for creating AI systems, such as with Python, TensorFlow, and Keras. We also looked at the importance of cloud platforms for creating models.

In the next chapter, we'll cover the main systems and approaches for generative AI.

¹⁴<https://ml-playground.com/>

Core Generative AI Technology

A Look at Key Algorithms

From 2004 to 2009, Ian Goodfellow attended Stanford University, where he studied computer science, with an emphasis on AI. Then he would get his PhD in machine learning at the Université de Montréal. Through his education, he learned from some of the best researchers in AI like Yoshua Bengio and Andrew Ng.

Goodfellow certainly put this to work. He cowrote the influential book *Deep Learning*. He also was an engineer at Google where he would help improve applications like Google Maps.

But his most important breakthrough would happen by accident. In 2014, he was at a pub in Montreal and had a far-ranging discussion with friends. The topic? It was whether deep learning could create unique images.

They debated the different approaches for generative AI. But the technology did not work well. The images were often blurry or ridiculous.

For Goodfellow, this sparked an inspiration. Perhaps the concepts of game theory could help? In other words, different deep learning models would essentially fight against each other – and this could help to create the images.

He wasted little time to figure this out. When he got home, he spun up a prototype of a model.

It would become known as a generative adversarial network or GAN. And yes, it would become one of the early innovations for generative AI.

He cowrote a paper about this called “Generative Adversarial Nets” and presented it at the Neural Information Processing Systems conference (NeurIPS) in Montreal. This would lead to a spike in research. There would also be an explosion of creative images and even audio samples across the Internet.

Yann LeCun would go on to say that the GAN was “the most interesting idea in the last 10 years in Machine Learning.”¹

But there were unintended consequences. GANs would become a way to create deepfakes.

Despite all this, Goodfellow would quickly become a super star in the AI community (currently, he has about 260,000 followers on Twitter). He got multimillion-dollar offers to work at various companies. In fact, Goodfellow would eventually be known as the “GANfather.”

So in this chapter, we’ll take a further look at GANs. But we’ll also look at the other foundational systems for generative AI.

Generative vs. Discriminative Models

In the prior chapter, we covered a form of AI that is generally known as discriminative modeling. This is about classifying data. With this, you can predict certain things, say whether an email is spam or if a customer is going to churn.

Much of the commercial applications – at least through 2022 – are primarily about discriminative modeling. The main reason is that businesses have a strong need for accurate forecasting. If not, the consequences can be severe.

Just look at the case of Southwest Airlines. In late December 2022, the company’s computer system was unable to handle the scale and complexities of the impact of a major snowstorm that enveloped most of the United States. Southwest had to cancel over 16,700 flights and would suffer a massive quarterly loss. A big part of this was the reimbursement of expenses for customers. But of course, there was also the cost of the hit to its reputation.

¹ <https://medium.com/thecyphy/gans-what-and-where-b377672283c5>

In light of this, Southwest management indicated that a top priority would be to modernize its IT systems. However, the company was still investing about \$1 billion annually on its technology.²

It also seems like a good bet that Southwest will explore AI technologies, especially those that involve discriminative modeling. Many other companies are probably doing the same thing – so as to help avoid a potential meltdown.

So then, how does discriminative modeling compare to generative AI? There are actually many similarities. They both usually rely on huge amounts of data as well as deep learning systems.

However, a key difference is the output of the data. With discriminative modeling, it's mostly about predictions. As for generative AI, the output is new data. It is also based on a set of probabilities, which are processed with input data. This is how generative AI can be creative. It essentially makes something new but that is still based on the unique features of the underlying datasets. For example, if you have a collection of images of dogs, the output of a generative AI model will still have the attributes of this animal. The core aspect of this model is that it understands the essence of an object. Think of it this way: the generative AI will attempt to find the probability of an observation in a dataset that is a part of a category.

For the most part, the datasets for generative AI are unstructured (although it can still be used for labeled data). In other words, this type of technology is unsupervised learning. This is certainly a benefit because of the abundance of unstructured data.

However, this brings up one of the drawbacks of generative AI. Since the output can often be unexpected, this can make it difficult to get a sense of the accuracy. This can mean that the results could be way off the mark – if not comical. This has been the case with ChatGPT, which has produced content that seems to be true – but has flaws. This phenomenon is known as hallucinations.

This is not as big a problem with discriminative modeling. The reason is that there are clearer sets of data to measure the accuracy.

Probability Theory

To better understand generative AI, you need some basic knowledge of probability theory. This is a branch of mathematics that is about analyzing random events. By doing this, you can estimate the chances of something occurring. This is usually expressed as a value of 0 to 1.

²www.wsj.com/articles/southwest-december-meltdown-to-cause-quarterly-loss-11673007141

A cornerstone of probability theory is the “law of large numbers.” This means that – when there are many experiments in the real world – the probabilities will get to certain levels. For example, if you flip a coin many times, the probability of getting heads or tails will ultimately be 0.50.

For discriminative modeling, there is actually basic probability at work. The machine learning or deep learning model will find correlations between the input and output layers. These weights will essentially be the probabilities of the prediction being correct. We saw this in the prior chapter with the softmax function.

But as for generative AI, the use of probability theory is much more extensive. This may involve looking at the conditional probability of a result. This is where the chances of an event happening is based on another event occurring. For example, suppose we are putting together a conditional probability model for predicting sales of a product. We estimate this for the year:

- 25% chance of a discount on the price
- 12% chance of selling out of the product

Then what is the probability that the product will be sold out, given that there was a discount? We calculate this using the conditional probability formula:

(Probability of selling out plus the probability of the discount) / the probability of the discount

This is $0.12 / 0.25$ or 0.48 probability of a sellout.

Next, a joint probability is another important measure for a generative AI model. As the name implies, this involves combining two or more probabilities. However, the events must be independent of each other – that is, one does not impact the other.

A classic example of a joint probability is the rolling of a six-sided dice. The probability of rolling a two is $1/6$ or 0.167. Thus, if we want to calculate the joint probability, this would be 0.167 multiplied by 0.167, which is roughly 2.8%.

For a generative model, there will be sample space of the probabilities for all the possible outcomes for the input layer. For example, a language model would go through all the possible sequences of words. Or, if the dataset is a library of images, the sample space will be all the possible images the model can create.

Next, there is the probability density function or PDF. This is the algorithm for coming up with the probability of an event happening from the sample space. When you sum all the probabilities, they will equal to 1. This is where we use joint probabilities.

The PDF is usually a neural network or deep learning system. The input is typically a random noise vector, which is a dataset of random values. This allows for more variation in the output. The values are mapped to the observations in the sample space. When training a model, a data scientist will often use different random noise vectors to create more diversity in the results.

Generative AI may also use a parametric model, which includes a set of parameters that are derived from the training. The process will use a neural network or deep learning to reduce the variances between the input and output layers.

Types of Generative AI Models

There are various types of generative models. The transformer is one of the most popular, and we'll cover this in the next chapter.

Rather, for the current chapter, we'll take a look at these generative AI models in the next few sections:

- Generative adversarial networks
- Variational autoencoders (VAE)
- Diffusion models

Generative Adversarial Networks (GANs)

GANs are one of the earlier forms of generative AI models, at least in terms of leveraging modern approaches like deep learning. Because of this, there is considerable research on this category. The GAN is also used primarily for generating unique realistic images, videos, or illustrations. In the past few years, the technology has also proven effective with audio.

The GAN model is about the competition between two neural networks. It's essentially a zero-sum game. When one model does well, the other one suffers. Through this process, there should be better results.

For a GAN, the two neural networks include the following:

- Generator network: This will ingest real data, such as photos. But the model will then create fake images. These are referred to as noise.
- Discriminator network: This model tries to differentiate the fake images from the real ones. To do this, there will be one or more loss functions, like the minimax loss function,

Wasserstein loss function, or cross-entropy loss function. In the early stages of the training of the GAN model, the discriminator network will be highly accurate. But through this process, the generator model will be penalized, which will help improve the results for the fake images. As a result, the discriminator network will start to decline over time – to where it will be about 50%.

The two interact in an “adversarial” way, and they are usually implemented as a convolutional neural network or CNN (this is better for high-dimensional data like images). The generator network will essentially try to trick the discriminator model that the data is real. Complex models will have multiple versions of these types of neural networks.

There is a rich ecosystem of open source projects for GANs. You can use them as architectures for creating new systems. Some of the common ones include the following:

- **CycleGAN:** This is to provide for the transformation of images with certain styles. For example, it can convert a horse into a zebra. Or it can change the seasons for an image. A popular app that uses the CycleGAN is FaceApp, which has over 500 million downloads. It transforms selfies into model-quality portraits. This is done with more than 60 photorealistic filters. It can change your hair color or style, add a beard, provide a touch of makeup, or remove blemishes and wrinkles.
- **DiscoGAN:** The concept of cross-domain relations is where there are connections between objects. Humans are fairly good at this. At just a glance, we can see that different items – like a handbag and shoes – have a similar style. As for doing this with an AI system, a GAN can be a good approach, but it can be tedious. Yet this is streamlined with the DiscoGAN.
- **text-2-image:** A GAN can certainly create wild images. But this may not be what you want. Instead, you might prefer something more constrained. For example, this could be that you want to create an image of a poodle with black hair and which is young. Well, you can do this by using the text-2-image system.

■ **Note** John von Neumann was a towering figure during the 20th century. Among his many achievements, he developed the core foundations of the modern computer. He also laid the framework for game theory. This was from a book he cowrote – entitled *On the Theory of Parlor Games* (1928) – with economist Oskar Morgenstern. Since then, other academics like Emile Borel, Ernst Zermelo, and John Nash (he was the subject of the movie *A Beautiful Mind*) evolved the concepts of game theory. They would win the Nobel Prize in 1994 for Economic Science.

There are definitely challenges with GANs. There is the issue with the vanishing gradient. This is when the discriminator network is too effective. This means that the model gets stuck, and there are minimal updates to the weights of the parameters. This can lead to output that is not much different from the input layer. However, to deal with this, you can add another loss function.

Problems with the output layer can be the result of low-quality data. This is called a model collapse. For this, you will need to add more diverse sources to the data. Interestingly enough, a generative model can be used to create this.

Variational Autoencoder (VAE)

In the prior chapter, we learned about autoencoders. These take in data, compare it, and then reconstruct it. This can result in interesting output.

But for generative AI, there is a flavor of the autoencoder. It's the variational autoencoder or VAE. This also has an encoder and decoder. But there is something added. The decoder relies on a sophisticated distribution of probabilities, which allows for even more diversity with the output.

There is also the KL divergence loss function. This attempts to minimize the distance from the dataset and the probability distribution.

The VAE has been around since 2013. Diederik P. Kingma, a research scientist at Google Brain, and Max Welling, a professor of machine learning at the University of Amsterdam, published this in a paper called “Auto-Encoding Variational Bayes.”³

Diffusion Models

In June 2020, AI researchers Jonathan Ho, Ajay Jain, and Pieter Abbeel published an influential paper entitled “Denoising Diffusion Probabilistic Models.”⁴ This set forth a critical innovation for generative AI – which has become known as the diffusion model.

³<https://arxiv.org/abs/1312.6114>

⁴<https://arxiv.org/abs/2006.11239>

The authors of the paper used this with several datasets, such as CIFAR10 and 256x256 LSUN. Granted, the results did not indicate that diffusion models were better than GANs. But this did not matter. After the paper was published, many other researchers started to focus on diffusion models, and the methods got better and better.

This was based on measures like the FID (Fréchet inception distance) score. It evaluates the quality of the images from a generative model. FID compares the distribution of the generated images with the real images. The smaller the distance between the two, the better the image quality. The highest score is 0.0.

Then how does a diffusion model work? It's complicated, involving advanced mathematics and theorems. But at a high level, the process is actually fairly straightforward.

The diffusion model includes two main steps. The first one is the forward diffusion process. For example, we would start with a dataset that has real-world images. Each of them is known as a base image. For each one, we add noise to it. But this is not a random one-shot process. Instead, it involves the use of Gaussian noise. This involves a probability function that is equal to a normal distribution or bell curve. This is a sum of the probabilities for a variable. The distribution is also according to the 68-95-99.7 rule. In other words, about 68% of the observations of the bell curve are within one standard deviation, 95% within two standard deviations, and 99.7% within three standard deviations.

This structure is core to statistics. The main reason is that the bell curve is often found to show the relationships in the real world, like IQ scores, height, and weight.

As for a diffusion model, a small amount of Gaussian noise is added in multiple steps. They often range from 1000 to 5000 or so, and each is based on a Markov chain. This means that the model takes only one of the prior images. The assumption is that the image will incorporate the data of all the other ones. The result is that the base image will have increasing amounts of noise. At the end, it will just be static.

This brings us to the next step: the reverse diffusion process. It will take the static image and then denoise it. This is done by using a deep learning model, which is based on the training from the forward diffusion process. This will also use the Gaussian distribution and Markov chain. As a result, there will be a new image, which will look similar to the base image.

Despite their power, diffusion models have notable drawbacks. Here's a look:

- **Assumptions:** The use of the normal distribution can skew results. It does not describe some important real-world phenomena. The same goes for other aspects of the diffusion model. For example, there is the assumption that the relationships are linear. But it is certainly common for there to be nonlinear or complex patterns in nature. Diffusion models also assume stationarity. This is where the process – as set forth in the Markov chain – is stationary or static. But of course, nature can be very dynamic.
- **Compute power:** Diffusion models require huge amounts of processing. This can make it expensive to build models with large datasets.
- **Image distortions:** They can be comical. They may even look like some form of science fiction! This could be fine for certain manipulations of real images. But it may be way off the mark for others. Consider that faces of humans – in diffusion models – are highly distorted when there are at least three. Another problem is with text within images. It can often be quite awful. However, such problems are likely to be addressed as the models are improved.

Keep in mind that the diffusion model is at the heart of some of the most innovative applications of generative AI. They include DALL-E 2, Google's Imagen, Midjourney, and Stable Diffusion, just to name a few.

DALL-E 2

OpenAI introduced DALL-E in January 2021. The name was a mash-up of Pixar's *WALL-E* character and Salvador Dali, a famous surrealist artist.

DALL-E is based on the GPT-3 model, which allows a user to write a prompt that creates an image. In April 2022, OpenAI launched the next iteration of this technology, called DALL-E 2. Its images were much more realistic and immersive. Then in late 2022, OpenAI released an API for this.

The DALL-E 2 process is based on two main stages. First of all, there is the prior phase. It has two neural networks:

- **Text encoder:** This uses a dataset of related text descriptions for images.
- **Image encoder:** This processes a dataset that has images with captions.

Each of these encoders translates the words and images into embeddings, which are numerical vectors. The CLIP (Contrastive Language–Image Pre-training) model is applied. It attempts to find the best matches between the text and image.

However, the CLIP model has drawbacks. For example, it can have challenges when you specify a specific number of objects or particular colors. A reason for this is that there is not enough image data.

Next, there is the decoder phase or unclip. This is where a diffusion model is used to create the images. After this, there can be further manipulations, say with unsampling. This increases the resolution of the image.

Figure 4-1 shows the general workflow. The user will enter a prompt, such as “An astronaut riding a horse in the style of Andy Warhol.” The DALL-E 2 system will create the text embedding and filter this for the prior phase. Then this data will be translated into an image embedding. This vector will become the basis for creating the image.



Figure 4-1. This shows the basic workflow for the DALL-E 2 model

OK then, what if we skip the text embedding? How important is this, really? Keep in mind that the researchers who created the model for DALL-E 2 considered this in their “Hierarchical Text-Conditional Image Generation with CLIP Latents” paper.⁵ The conclusion was that the text embedding greatly improved the results. Essentially, this provided more context, which allowed for better matching.

Something else about this model is that it is generally lightweight for generative AI systems. Because of this, image generators can be implemented on small devices like iPhones.

Stability AI and Midjourney

There are a variety of text-to-image generative AI systems. But the startup that has seen the most traction is Stability AI. The CEO and cofounder is Emad Mostaque.⁶ He was born in Jordan but would eventually move to the

⁵<https://arxiv.org/abs/2204.06125>

⁶https://wandb.ai/wandb_fc/gradient-dissent/reports/Emad-Mostaque-Stable-Diffusion-Stability-AI-and-What-s-Next--VmlldzoyOTQ0ODk3

UK. As for college, he received his master's degree in mathematics and computer science from Oxford University in 2005. But he would go on to pursue a career in the hedge fund industry.

While this proved lucrative, he wanted something more meaningful. In 2020, he founded Stability AI and he bootstrapped the operations. At first, the focus was on leveraging AI to help with the COVID-19 pandemic.

But as he learned more about the innovations with generative AI, he pivoted the company. It would be about using the technology to allow anyone to create compelling images.

To help accelerate the growth, Mostaque pursued an open source strategy. He thought this would greatly help to democratize the strategy.

He also formed critical partnerships. "Stable Diffusion is unlike any other open-source project we've come across," said Gaurav Gupta, who is a partner at Lightspeed Venture Partners and an investor in Stability AI. "Its creation story is novel, having been done in collaboration with academics (LMU Munich), other companies (RunwayML), and independent AI communities (EleutherAI, LAION). We believe Stability AI will make the ideal business partner for companies, non-profits, and governments that do not have access to AI talent, compute resources, and the ability to customize usage."⁷

In August 2022, Stability launched its open source platform called Stable Diffusion – and the adoption was intense. It was one of the fastest-growing projects ever. Then the company released a cloud version of the system, DreamStudio. Within less than a month, it attracted more than 1.5 million users.⁸

Stability AI's approach to the generative AI was different from DALL-E 2's. It's based on a paper entitled "High-Resolution Image Synthesis with Latent Diffusion Models."⁹ A latent diffusion model is meant to deal with the high-compute requirements of traditional diffusion. But by using autoencoders in the latent space – where the key features are stored – there is much more efficiency, without impacting the quality of the images. This is why someone can download Stable Diffusion and use it on their own PC system.¹⁰

⁷ <https://medium.com/lightspeed-venture-partners/why-lightspeed-invested-in-stability-ai-democratizing-generative-ai-b4f2250c5055>

⁸ <https://medium.com/lightspeed-venture-partners/why-lightspeed-invested-in-stability-ai-democratizing-generative-ai-b4f2250c5055>

⁹ <https://arxiv.org/abs/2112.10752>

¹⁰ <https://techcrunch.com/2022/12/22/a-brief-history-of-diffusion-the-tech-at-the-heart-of-modern-image-generating-ai/>

However, the infrastructure costs were still substantial for Mostaque because of the massive user base. He simply did not have the personal resources to fund the operation anymore. This is why he sought out venture capital. In October 2022, Stability AI announced an investment of \$101 million. The valuation was set at about \$1 billion.

Now another top text-to-image generative AI platform is Midjourney. Before founding the company, David Holz was the CEO and cofounder of Leap Motion. The company created devices to track hand motions.¹¹

As for Midjourney, Holz wanted to build a social network for creators. But he did not build the infrastructure for this. He instead used Discord, which is a popular chat system. With this, users can create their own channels and show their creative images. Figure 4-2 shows what this looks like.

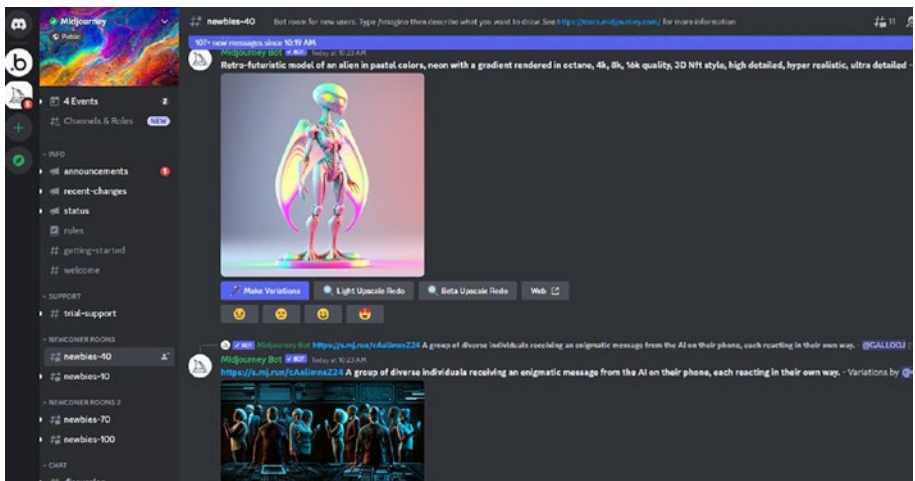


Figure 4-2. This is a community channel for creators on Midjourney

Holz has self-funded the venture. But as the growth continues, it seems likely that – like Stability AI – there will be a need for outside funding.

Note During the 1970s, there were adventure games. But because of the limited capacity of PCs, there were no graphics. There were instead written descriptions. One of the most popular adventure games was *Zork*. So what would this look like with graphics? Matt Walsh, who is a Google principal, wanted to find out and used generative AI for this. A text-to-image system – called Imagen – was able to provide the visuals for the game.¹²

¹¹ www.theregister.com/2022/08/01/david_holz_midjourney/
¹² <https://petapixel.com/2023/01/12/google-generative-ai-brings-visuals-to-zork-a-1977-text-based-video-game/>

Speech

When it comes to generative AI and speech, there are two main systems. First, there is automatic speech recognition or ASR. This allows for the AI to “hear” the words or to use speech-to-text capabilities.

The model will ingest raw audio and apply deep learning algorithms. The sound waves are often converted into a spectrogram, which shows the frequencies. Then there are acoustic and language models that interpret the speech. The goal is to minimize the word error rate or WER.¹³

Next, there is text-to-speech or TTS. This is also known as voice cloning. There are many applications for this, such as customer support services, marketing campaigns (say for jingles), online courses, and audiobooks.

With TTS, AI will create human-sounding words and sentences. This could include creating a spectrum. But there will be various models for the preprocessing of the sound, text encoding, linguistic analysis, and so on. In some cases, there may be a large deep learning system.

■ **Note** Director and producer Morgan Neville produced the documentary *Roadrunner: A Film About Anthony Bourdain* in 2021. Bourdain, who was a celebrity chef, died of suicide in 2018. The film was an engaging portrayal of his life and was critically acclaimed. But there was some controversy. Morgan used AI systems to simulate Bourdain’s voice for some of the scenes.¹⁴

While there have been many advances with speech AI, there are still issues. It’s not uncommon for the voice output to sound robotic or unnatural. Then there are the problems with dialects, accents, and jargon. Let’s face it, we have all experienced this with apps like Alexa and Siri.¹⁵

Yet generative AI is making considerable progress. A cutting-edge model for this technology is VALL-E. In January 2023, Microsoft researchers set forth this in a paper entitled “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers.”¹⁶ For the pretraining, this involved training data of 60,000 hours of English speech, which was hundreds of times larger than existing models.

The result is that the cloning of a voice can take only three seconds. It can even simulate emotion of the speaker and the acoustic environment.

¹³ <https://developer.nvidia.com/blog/an-easy-introduction-to-speech-ai/>

¹⁴ www.npr.org/2022/01/17/1073031858/artificial-intelligence-voice-cloning

¹⁵ www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

¹⁶ <https://arxiv.org/abs/2301.02111>

However, the researchers do warn of the potential harm of this system:

“Since VALL-E could synthesize speech that maintains speaker identity, it may carry potential risks in misuse of the model, such as spoofing voice identification or impersonating a specific speaker. To mitigate such risks, it is possible to build a detection model to discriminate whether an audio clip was synthesized by VALL-E. We will also put Microsoft AI Principles into practice when further developing the models.”¹⁷

In fact, there are already examples where voice cloning has been misused. Consider what happened in 2019 with the CEO of a European company. He received a call from another division for a request for \$243,000 and it was sent. However, the person’s voice was simulated – mostly likely by the use of AI.

Note A big challenge with complex AI systems is running them on the edge, such as mobile devices. The communications with the cloud must be done in near real time. This usually means the lag should be less than 100 milliseconds. Otherwise, an AI-based conversation would sound awkward.

Despite the challenges, there are definitely many uses for AI-based speech systems. Just look at the audiobook industry. In 2022, the global revenues came to \$1.5 billion, up 25% on a year-over-year basis. But by 2030, the market could be worth more than \$35 billion.¹⁸

A big factor for this is likely to be generative AI. Keep in mind that creating audiobooks can be laborious and time intensive. It is also often too costly for smaller publishers or those who self-publish their books.

Apple is among the large players in the market that has leveraged AI for this. In 2022, it launched titles on its massive catalog. Apple designates these as “Narrated by Apple Books.”¹⁹

Trilemma of Generative AI Models

Given that generative AI is still nascent, there are plenty of challenges and issues. A way to describe them is through the trilemma. This describes the trade-offs between three main factors:

¹⁷ <https://arxiv.org/abs/2301.02111>

¹⁸ www.theguardian.com/technology/2023/jan/04/apple-artificial-intelligence-ai-audiobooks?mc_cid=855e30e70b&mc_eid=d2be7c886c

¹⁹ www.theverge.com/2023/1/5/23540261/apple-text-to-speech-audiobooks-ebooks-artificial-intelligence-narrator-madison-jackson

- **Sample quality:** This is where the model creates high-quality samples. They should not only look realistic but be similar to the base data.
- **Performance:** Here, a model can quickly create samples. Preferably, this means real time or near real time. This allows for use in applications that will have lots of usefulness.
- **Mode coverage:** This is a model that can effectively create diverse samples.

Given this framework, let's take a look at GANs. They definitely are strong in terms of sample quality and performance. However, the mode coverage can be lacking. The fact is that GANs usually produce a small number of samples. This can mean that the creativity of the model is not as compelling.

As for a diffusion model, it is better with mode coverage as well as sample quality. But the problem is with performance. As mentioned earlier, it can be compute intensive to create strong diffusion models.

But the good news is that there is considerable research on these matters – across the globe. The pace of innovation is also fairly quick.

Conclusion

In this chapter, we first looked at the differences between discriminative and generative models. Discriminative models are generally for predictions. This type of AI is the most prevalent.

Generative AI, on the other hand, is also about using predictions. But there are some other aspects to the model. There is an element of probability and random data. By doing this, it's possible to create data. But it is also similar to the base dataset.

Generative AI is relatively new. The first model came in 2014 with the GAN. It was a major breakthrough and sparked considerable research.

Then in 2020, there was the introduction of the diffusion model. This ultimately showed even better results than GANs. The diffusion model is one of the driving forces for the latest innovations with generative AI.

But there is still much to be done. As seen with the trilemma problem, there are trade-offs when it comes to generative AI models. But these should diminish as the avid research continues.

In the next chapter, we'll take a look at one of the main applications of generative AI: large language models.

Large Language Models

How Generative AI Understands Language

In December 2015, a group of tech veterans – including Elon Musk, Reid Hoffman, Peter Thiel, and Sam Altman – founded OpenAI. They pledged over \$1 billion for the venture.

But this was not a typical Silicon Valley startup. OpenAI was set up as a nonprofit organization, and the vision was “to ensure that artificial general intelligence benefits all of humanity.”¹ Artificial general intelligence or AGI is the holy grail of AI. It’s where a machine has human-level cognitive capabilities.

Here’s how OpenAI explained it:

“[An] AGI will be a system capable of mastering a field of study to the world-expert level, and mastering more fields than any one human — like a tool which combines the skills of Curie, Turing, and Bach. An AGI working on a problem would be able to see connections across disciplines that no human could. We want AGI to work with people to solve currently

¹ <https://openai.com/about/>

intractable multi-disciplinary problems, including global challenges such as climate change, affordable and high-quality healthcare, and personalized education.”²

OpenAI actually seemed more like a research lab. The company developed highly experimental systems and published the findings as academic papers. By sharing the knowledge about AI, OpenAI thought that this would accelerate the path to AGI.

But there was a problem. The costs for running OpenAI were substantial. This is why it raised \$1 billion from Microsoft in 2019.

For the deal, about half of the round was in cash, and the other was to use the resources of Azure, which is Microsoft’s cloud platform. Microsoft would also get priority with commercializing OpenAI innovations in its own applications.

To allow for the investment, OpenAI had to change its structure. It became a hybrid nonprofit, and investors had their returns capped at 100X.³

This move stirred up controversy though. Some people even thought that OpenAI “sold out.”

But the Microsoft deal was critical for the growth of the company. OpenAI was able to build massive AI models, which would have hundreds of billions of parameters.

For example, there was the development of large language models or LLMs. These are able to understand and generate natural language. It’s a core technology for innovations like ChatGPT.

But to run ChatGPT, the expense was a few cents per chat. Altman noted: “We will have to monetise it somehow at some point; the compute costs are eye-watering.”⁴

This is why OpenAI has had to continue to raise significant amounts of capital. During the few years after the \$1 billion round, Microsoft added an additional \$2 billion. Then in early 2023, the company agreed to another massive fundraising. The amount was not disclosed, but the buzz was that the round came to \$10 billion at a \$29 billion valuation.

These investments have spurred rapid innovations for LLMs. So in this chapter, we’ll take a deeper look at this powerful technology. We will see how it works and its evolution.

²<https://openai.com/blog/microsoft/>

³www.lifestyleasia.com/kl/gear/tech/openai-what-to-know-about-the-company-behind-chatgpt/

⁴www.ft.com/content/9c5f7154-5222-4be3-a6a9-f23879fd0d6a

Language and Intelligence

René Descartes was a prototypical Renaissance man. He was a trailblazer in studies like philosophy and mathematics. Of course, he is the one who said: “I think, therefore I am.”

Descartes also considered language to be at the core of intelligence. He noted that it was what distinguished humans from all other species. True, some animals can communicate. But they do not have a language, such as English or French.

Language allows for many benefits. It provides for collaboration and socialization. It is the basis for education. It can be a way to express love, fear, and sadness. It can help understand concepts like justice, politics, and ethics. It can help with complex problem solving. It can enable great works of creativity, such as with literature and music. Language also makes it possible to share and preserve knowledge across generations.

But with the recent innovations in AI, Descartes’ notions are on shaky ground. Humans may no longer have a monopoly on language. Instead, AI is quickly gaining ground.

This is really at the heart of LLMs. They truly are revolutionary. But to understand the impressive technology, it’s important to cover some of the core aspects of technologies like natural language processing or NLP.

■ **Note** The first humans appeared on earth about two million years ago. But language has only been around for about 150,000 years. This was at a point where humans started to see more progress. However, the beginning of written language came about 5000 years ago.

Natural Language Processing (NLP)

As we saw in Chapter 1, NLP is one of the earliest forms of AI. But it would take decades to refine the systems. Yet the major innovations came with the use of sophisticated deep learning models.

NLP has many useful applications. The technology has also had a major impact on many industries. NLP is common for customer support systems, sentiment analysis, and personal digital assistants, such as with Siri and Alexa.

There are two main categories for NLP:

- **Natural language understanding (NLU):** This enables a computer to understand the human language. This uses complex algorithms to interpret sentence structure, context, intent, and entities.

- Natural language generation (NLG): After the NLP interprets the content by using NLU, it will create content. Yes, this is about generative AI.

Even with the breakthroughs with NLP, there remain many challenges. A language will have many nuances as well as vague words and phrases. Interpreting meaning can involve understanding the tone of how something is said or the body language. Then there is slang and dialects. In the meantime, languages are dynamic as words change over time. Or there are new ones added. This is especially the case with technology and healthcare.

How NLP Works

Let's take a deeper look at how NLU and NLG work. A key part of NLU is the cleaning and preprocessing of the text. In other words, the text needs to be put in a form that a computer can understand. There are a variety of approaches for this. First, there is tokenization. This converts the text into different components. Each word or even subword can be a token. If there is a long document, each sentence can be a token. With tokenization, you can then use different NLP systems for understanding the text.

Next, there will be normalization of the text. This could be taking out punctuation, making the text upper or lower case, or removing contractions. There may then be the elimination of non-important words like "the," "a," and "an." This preprocessing of the text can streamline the process for the NLP.

However, a data scientist needs to be wary. After all, the removal of a punctuation mark or contraction could have a big impact on the meaning.

Now there are different types of systems that can help with the understanding of the text. A popular technique is stemming, which reduces words to their root form. For example, for the words "payment" and "repay," the root is pay.

By breaking down words into their core components, a computer can better understand the meaning. This is actually a key process for search engines like Google.

There are ways to use stemming, such as with the Porter Stemming Algorithm. Named after the inventor – Martin Porter – this technique uses a set of rules to remove common suffixes like "ing" and "ed."

Porter also invented another widely used NLP stemming system: the Snowball stemmer. It is based on the Porter Stemming Algorithm, but it has more features. It is available in a variety of languages and supports Unicode.

Another useful system for NLP is lemmatization, which is like stemming. But it does not eliminate prefixes and suffixes. Instead, it determines a root word by using a dictionary. While this approach tends to be better than stemming, it can be compute intensive.

After the text is put into a format a computer can interact with, there needs to be an understanding of the overall meaning of the content. This is certainly a major challenge.

But there are some techniques that can help. They include

- **Tagging parts of speech (POS):** This is tagging each word with its grammatical form, such as a noun, verb, adverb, or pronoun. Advanced machine learning can help with this process.
- **Chunking or shallow parsing:** This segments text into phrases, such as noun phrases or verb phrases. These chunks are then put into a tree structure, which shows the relationships. By analyzing these patterns, an NLP system can start to understand the meaning of the text.
- **Named entity recognition (NER):** This is a taxonomy of objects like locations, persons, and organizations. It's typical to use machine or deep learning models to come up with the categories.
- **Topic modeling:** This uses unsupervised learning to detect patterns with topics or themes in text. A popular technique for this is Latent Dirichlet Allocation (LDA). This is actually a generative AI model that leverages Bayesian probability.

■ **Note** The typical graphical user interface (GUI) has menus, buttons, and sliders. This setup is efficient and useful. But with the growth of LLMs, might the GUI change? Perhaps so. According to Talia Goldberg, a venture capitalist at Bessemer Venture Partners: "Thanks to the rise of large-language models, natural language will begin to replace GUIs as the primary human-machine interface. Using just voice or text as input, consumer will generate code, control software apps, and create rich multimedia content."⁵

⁵ https://fortune.com/2023/01/04/the-crystal-ball-vcs-private-equity-investors-and-tech-founders-predict-whats-to-come-in-2023/?utm_source=Iterable&utm_medium=email&utm_campaign=reader&tpcc=reader

From 2012 to 2018, deep learning was critical in making NLP more robust. But it was the introduction of LLMs that was a game changer. This approach may ultimately replace much of the traditional approaches to NLP and become the standard.

Word2Vec Model

A critical part of the theoretical foundation for the LLM is the Word2Vec model. In 2013, Google researchers wrote a paper about this entitled “Efficient Estimation of Word Representations in Vector Space.”⁶ They used deep learning to learn language by using a massive content library.

Word embeddings are the input layer for the Word2Vec model. This uses vectors to encode words, which represent the relationships. These vectors are dense. This means they are nonzero. By comparison, sparse vectors have zero values.

By using the word embeddings, the Word2Vec model can learn the context of the words. This is certainly essential for understanding language. After all, words may have meanings that depend on the overall sentence or paragraph. For example, a “server” can be someone at a restaurant or a software or hardware system.

With the Word2Vec system, it will attempt to map those words that generally belong together. This could be “coder” and “developer” or “university” and “school.” This is known as the semantic relationship of words.

The word embeddings will have different dimensions. In the case with university or school, these could be age and curriculum. These will be essentially expressed as coordinates in abstract space. For those different word embeddings that are closer, they will have stronger relationships when forming sentences.

For the academic paper, the researchers measured the relationships with the Word2Vec model – that had a 1.6-billion-word dataset – and compared this to other neural networks. The results showed significant improvements in accuracy and lower compute costs.

The Word2Vec model was a major breakthrough. But there would be other research needed. The next major development would be the transformer.

⁶<https://arxiv.org/abs/1301.3781>

Transformers

CNNs (convolutional neural networks) are usually not effective with NLP. The reason is that this type of model does not handle strings of data like text. However, an RNN (recurrent neural network) is much better and has become widely used. But there are still drawbacks. To address these, data scientists have combined other approaches like long short-term memory (LSTM) models. These use a memory cell that stores information for a prolonged period. This allows it to have a more extensive analysis of strings of text. But even with RNNs and LSTMs, there remain challenges in improving NLP.

Starting in 2014, AI researchers started to explore new approaches. One was attention. This is about analyzing sequences of data – such as text – to focus on the most relevant data and ignoring the rest. This was shown to help improve the accuracy of a task.

However, the attention mechanisms were primarily used to enhance RNNs. This meant there were still nagging issues with performance.

But this all changed in 2017. A group of Google AI researchers published a paper entitled “Attention Is All You Need” and presented their findings at the 2017 NeurIPS conference. It was a bombshell.

The researchers proposed a new type of neural network that would go beyond CNNs and RNNs. The authors called this network architecture the transformer. They tested it on eight GPUs and trained the model for three and a half days. It showed strong results for English-to-French translations.

■ **Note** The Google team that created the transformer initially wanted to call it an “Attention Net.” But it was not very catchy. So Jakob Uszkoreit, a Google software engineer who was a member of the team, came up with the “transformer” name. The reason was that the model was about transforming representations of data.⁷

At the time, there had been a growing consensus that traditional deep learning models were reaching their limits. The belief was that there were diminishing returns for large models. But with the transformer, the AI industry got a huge boost.

At the core of the transformer model was self-attention. It’s something that is different from a typical neural network. With self-attention, the training involves processing all the observations at once – not in a gradual process.

⁷<https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/>

It provides for understanding the context of words. As an English linguist, John Rupert Firth, has noted: “You shall know a word by the company it keeps.”⁸

To get a sense of this, let’s take an example. Suppose we have this sentence:

The autonomous car stopped after it processed data
from its AI model.

For a traditional NLP system, it would be difficult to determine that “it” refers to “autonomous car” and “its” is connected to “AI model.” But a transformer’s attention layer will be able to parse the meanings correctly.

For this process, you can include large amounts of text for the input layer. The processing is also done in parallel. This helps to reduce the need for compute power as well as lower the time needed for training.

Consider that a transformer is usually a pretrained model. This means that the training of the dataset is from scratch. There is no advantage of prior knowledge for the model.

But a pretrained model is not fixed. You can fine-tune it to make adjustments and changes, such as with more data. For example, the pretrained model may have the core knowledge of understanding language. The fine-tuning can then be used to add more specialized information, say for healthcare and finance. This process is called “transfer learning” – that is, taking the knowledge from one model and adding it to another.

By doing this, the new model can be improved by the learnings of the pretrained model. This also avoids the problem of having to “reinvent the wheel.” This accelerates the process for creating the AI system, which can lower the costs substantially.

The paper from the Google researchers included a diagram of the transformer model, as seen in Figure 5-1. Granted, it is complicated. But let’s take a look at the main parts.

⁸www.techtarget.com/searchenterpriseai/definition/BERT-language-model

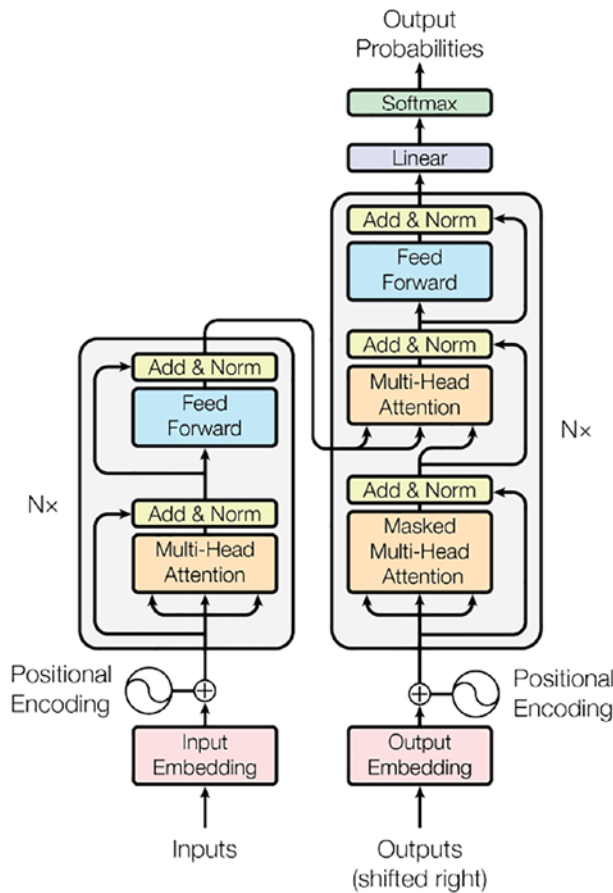


Figure 5-1. This is the diagram of the transformer model from the “Attention Is All You Need” research paper

Keep in mind that a transformer has two main parts: the encoder and decoder. The encoder ingests the data and processes it. As for the decoder, it uses the features of the model to create sequences of tokens.

First, let’s take a closer look at the encoder. Text is fed into the system, and it is converted into word embeddings. Then there is the positional encoding process. This comes up with a vector that shows the position in the sentence for each token, which helps improve the understanding of the context of the language. This is all done in parallel.

After this, the vector will be transformed with a stack of encoders. For example, the data will be sent to one encoder, and the output will be sent to another, and so on. The number of encoders depends on the type of the transformer model and the complexity of the task.

For this process with the encoder, there is the use of multi-head attention. It is a way to go beyond understanding the context of one word, which is how self-attention works. Multi-head attention could, for example, find relationships among nouns, adjectives, adverbs, and pronouns. However, the AI will not have to be programmed for this. It will instead find these relationships on its own by detecting patterns.

Once this is finished, the vectors will be fed into the decoder. This will be similar to the process of the encoder. That is, the data will be processed with a stack of decoders.

Besides multi-head attention, there will be masked multi-head attention. With this, the AI predicts the next word or token by looking at the text to the left and right.

To understand this, let's take an example. Suppose the sentence is "My [masked] is Fred." When analyzing this sentence, the system does not initially know what the word after "my" is. But by looking at the rest of the words, it will understand it should likely be "dog." This is based on the strength of the relationships of the words.

Consider that you can separate the encoder and decoder functions:

- **Encoder-only models:** These are helpful if you want to classify text. This could be for applications like sentiment analysis. An example of an LLM that uses an encoder-only model is BERT.
- **Decoder-only or autoregressive models:** This is for when you want to generate text. It's similar to how a chat app works for your smartphone. As you type in text, the AI will predict the next word or words. GPT-3 uses a decoder-only model.
- **Encoder-decoder models:** With these, you can handle generative AI tasks like language translation and summarization. Examples of LLMs that use this approach include Facebook's BART and Google's T5.

Consider that all this really just scratches the surface of the complexities of transformers. The mathematics can be daunting. There are also many new innovations and breakthroughs.

But for Mustafa Suleyman, who is the cofounder of DeepMind, he has a good way of explaining the technology:

So if you think about it, most information in the world has a sequential structure. Each piece of data is in some way related to data that has come in the past, and a big part of what makes us intelligent is that we can learn, as humans, abstract representations of a

series of things that have happened in the past well enough that we can reason over these elements. In some sense, this is what Transformers do. They take sequential data, build a good model of that data to learn an abstract representation of it, and then they use that to try and predict what's going to come next.⁹

Keep in mind that the transformer spawned a surge in interest from AI researchers. During the past two years, about 70% of AI papers that were posted on arXiv talked about this technology.¹⁰ This has represented a major shift from RNNs and CNNs.

The progress for transformers has certainly been swift. They have become an essential part of generative AI, especially with LLMs.

The transformer has the potential to be more than just for language tasks. Already there has been considerable progress in other categories, such as with computer vision.

Alexey Dosovitskiy, a researcher at Google Brain Berlin, created the Vision Transformer (ViT).¹¹ What he found was that using a transformer model could allow for better scaling of image processing than CNNs. With the ViT, an image is broken into chunks or tokens, and then the neural network classifies it. The system will show high levels of accuracy. It certainly helps that the model processes the whole image at once, instead of analyzing it on a pixel-by-pixel basis.

What about using ViT for creating images? This is something researchers are working on. But there is a lot of promise that this can be an effective approach.

■ **Note** The use cases for transformers are seemingly endless. Already, there have been applications for understanding the interactions of prescription drugs, finding anomalies for fraud detection, deriving meanings from huge legal documents, and creating digital twins for proteins.

Models based on transformers are often referred to as foundational models. Stanford researchers coined this phrase in their paper, “On the Opportunities and Risks of Foundation Models,” which was published in August 2021. They noted:

⁹<https://greylock.com/firm-news/welcome-mustafa-suleyman/>

¹⁰<https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/#:~:text=Stanford%20researchers%20called%20transformers%20%E2%80%9Cfoundation,is%20possible%2C%E2%80%9D%20they%20wrote>

¹¹www.quantamagazine.org/will-transformers-take-over-artificial-intelligence-20220310/

A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks; current examples include BERT, GPT-3, and CLIP. From a technological point of view, foundation models are not new — they are based on deep neural networks and self-supervised learning, both of which have existed for decades. However, the sheer scale and scope of foundation models from the last few years have stretched our imagination of what is possible...¹²

Dials

An LLM will allow you to change the variation of the responses. Think of this as using dials to get to the kinds of results you want. To understand how this works, let's take a look at the OpenAI Playground, which is in Figure 5-2.

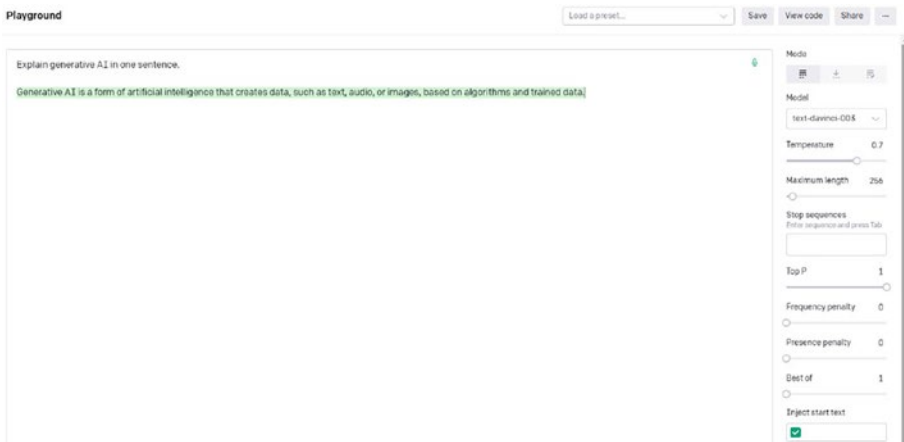


Figure 5-2. This is the OpenAI Playground, where you can test prompts and change the variation of the responses

On the main part of the screen, you can enter a prompt and then click the Submit button to get a response. On the right side, there are the dials. Here's a look at each of them:

¹² <https://arxiv.org/pdf/2108.07258.pdf>

- **Model:** You can select the type of the LLM, say for GPT-3 or GPT-4.
- **Temperature:** This adjusts the randomness of the response. The dial ranges from 0 to 1. The closer it is to 0, the less random the response will be. In fact, if it is 0, then the response will be the same for any time you enter the prompt.
- **Maximum length:** This is the cap on how many tokens are allowed for a response. This varies according to the type of model.
- **Stop sequences:** This is the point where you want the model to stop the processing. This could be at the end of a sentence or a list. You can specify characters for this – something like “###” – or it could be the use of the return or enter key.
- **Top P:** This is another way to control the randomness of the responses, and the range is from 0 to 1. Basically, OpenAI will create a distribution of probabilities for a token and only include those above the value set for top p. OpenAI advises that you should either use top p or temperature. For example, suppose you want to use top p for 0.50. Then you will set the temperature to 1.
- **Frequency penalty:** This is the value you penalize for repeated text. This will help to increase more variation in the responses.
- **Presence penalty:** This penalizes for when a token is in the text. This will essentially prevent repeated topics.
- **Best of:** This will create multiple responses and display the best one. However, this can mean using more tokens, which will increase the costs of using the API.
- **Inject start text:** This can be helpful if you have back-and-forth interactions like chat. For example, suppose you have two lines. One is a comment from Joe and another is from Jane. If you then type Joe, the AI system will provide a response.
- **Inject restart text:** This is also useful for back-and-forth interactions. For example, in a chat, this would continue to add text for the next person.

With the OpenAI Playground, you can get the probabilities for each of the tokens. You can see this in Figure 5-3.

Explain generative AI in one sentence.

Generative AI is a type of artificial intelligence that focuses on creating new data from existing data.

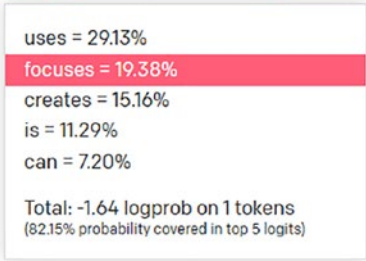


Figure 5-3. This shows the probabilities for the tokens

In this example, we have a prompt to have OpenAI explain generative AI, and we get a one-sentence definition. You can click each token and see the probabilities. For the word “focuses,” this had a probability of 19.38%. But there were other possible words like uses (29.13%), creates (15.16%), is (11.29%), and can (7.20%). As you can see, the words selected did not have the highest probability. The reason is based on the dials set for the model.

Now there are no hard-and-fast rules when using these dials. It’s more about trial and error. It really depends on how much variation you want in the responses. This might be fine for creative purposes, such as for writing marketing copy. But it could be problematic when creating a contract or writing a manual for airline pilots or medical doctors.

Another consideration is the length of the prompt. If it is short and the temperature is close to 1, then the response can be wide off the mark. But this becomes less of a problem for larger prompts.

OK then, for the next few sections, we’ll take a look at some of the notable LLMs.

BERT (Bidirectional Encoder Representations from Transformers)

BERT (Bidirectional Encoder Representations from Transformers) is one of the first LLMs that leveraged the transformer technology. In October 2018, Google researchers wrote about this system in a paper called “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”

According to the authors: “As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.”

The paper showed BERT's impressive results for 11 natural language tasks. It was pretrained using Wikipedia and the Brown Corpus.¹³ The system provided for masked language modeling as well as next sentence prediction.

Google was also aggressive in implementing BERT within its search engine. It's important to note that about 15% of queries on the platform are unanticipated. For example, propositions like “to” and “for” can easily change the meaning of a sentence. By using sophisticated attention techniques, BERT was able to handle these effectively and allow for more conversational queries.

In a blog post from Pandu Nayak, a Google Fellow and Vice President of Search, he provided examples of the power of this technology. One of the queries he looked at was “2019 brazil traveler to usa need a visa.”¹⁴

The traditional Google engine would not understand that “to” referred to a person from Brazil who is traveling to the United States. But BERT did.

In November 2018, Google would open source the system. This helped to get adoption and spurred innovation.

For example, the Allen Institute for Artificial Intelligence used BERT to take grade-school science tests. The training dataset had over 100,000 multiple choice exam questions (the ones that had diagrams were excluded). The model – called Aristo – was able to get scores of over 90% for 8th-grade exams and 83% for 12th-grade exams.¹⁵

Here are some of the other applications that use BERT:

- **PatentBERT:** Researchers applied BERT to a database of over two million US patents. The system provides for classifications, which is a complex process.
- **BioBERT:** This is for text mining of biomedical text. Traditional NLP techniques have been unsatisfactory because of the wide distribution between general language and jargon. BioBERT has shown much better performance – say with entity recognition, question answering, and relation extraction – than just using the general BERT platform.
- **DistilBERT:** Hugging Face launched this in 2019. It's a much smaller and faster version of BERT. Yet the performance did not suffer.

¹³ <https://arxiv.org/abs/1810.04805v2>

¹⁴ <https://blog.google/products/search/search-language-understanding-bert/>

¹⁵ <https://developer.nvidia.com/blog/allen-institute-for-ai-announces-bert-breakthrough-passing-12th-grade-science-exam/>

GPT Systems and ChatGPT

When the “Attention Is All You Need” paper was published, the researchers at OpenAI took notice. They saw the huge potential of this model and made it a high priority for its own AI efforts.

This resulted in GPT (Generative Pre-Trained Model) in June 2018. This was described in a research paper entitled “Improving Language Understanding by Generative Pre-Training.” According to the authors: “In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture.”¹⁶ The result was that GPT significantly outperformed 9 of 12 state-of-the-art tests.

The model was pretrained using a decoder-only model using the Book Corpus dataset. There were 117 million parameters.

A major innovation of GPT was zero-shot learning. This is an AI model that can recognize new objects or concepts that it has never seen. The system essentially is able to infer them because of the knowledge gained from the massive dataset. Zero-shot learning is much more efficient than supervised learning models and can provide much better results.

In February 2019, OpenAI introduced GPT-2. It was similar to the prior version, but the model had much more scale. There were roughly 1.5 billion parameters, and the dataset had more than 8 million documents, which came primarily from Reddit.

GPT-2 represented a major leap. A survey from Cornell University showed that the output from the system was convincing. The credibility score was 6.91 out of 10.

However, OpenAI did note some of the dangers of GPT-2 and other LLMs. A study from the Middlebury Institute of International Studies’ Center on Terrorism, Extremism, and Counterterrorism (CTEC) said that the technology could be misused for white supremacy, anarchism, Marxism, and jihadist Islamism. According to a blog from OpenAI: “CTEC demonstrated that it’s possible to create models that can generate synthetic propaganda for these ideologies.”¹⁷ The blog also mentioned that detection of synthetic text would be a long-term challenge and that there needed to be standards for understanding bias in LLMs.

Because of these issues, OpenAI did not release GPT-2 to the general public. They feared it could be weaponized.

¹⁶ https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

¹⁷ <https://openai.com/blog/gpt-2-1-5b-release/>

But this move caused much controversy. Was this just a way for OpenAI to get more PR? Besides, were the fears overblown?

Perhaps so. In fact, when OpenAI launched GPT-3 in June 2020, the company took a different approach. It would allow anyone to use the technology.

Regardless, GPT-3 represented a major achievement for LLMs. It had 175 billion parameters and was trained on five major datasets.

GPT-3 could write articles and computer code. It could even translate just about any language. And the content was human-like – if not indistinguishable from what a person would craft.

There was also the introduction of few-shot settings. This is where a model is trained on a small number of data observations. It's actually similar to how a human learns – where it usually takes a few examples to understand something.

Another important innovation for GPT-3 is that OpenAI created an API for it. At the time, it was mostly data scientists who had access to powerful AI models. But with an API, anyone could participate. This helped to greatly spread the GPT-3 technology and would be a big driver for the generative AI industry.

Nine months after the launch of GPT-3, an OpenAI blog said that the technology spawned over 300 new applications and that tens of thousands of developers were using the platform. GPT-3 was generating about 4.5 billion words per day.

Here are some of the new applications:¹⁸

- **Viable:** This helps companies better understand customer feedback with summaries. It highlights themes, emotions, and sentiments. You could ask the system questions like “What’s frustrating our customers about the checkout experience?”
- **Fable Studio:** This is an interactive story. It’s based on Lucy, which is a character from Neil Gaiman and Dave McKean’s *Wolves in the Walls*.
- **Algolia:** The company used GPT-3 to create a system to answer complex questions from customers. There was a 91% precision rate or better.

The API was also at the core of the business model for OpenAI. The company charged customers for the service based on consumption. You would pay according to the number of tokens processed.

¹⁸ <https://openai.com/blog/gpt-3-apps/>

While a token is often a word, it can also be a part of a word. This depends on the complexity. OpenAI provides the following rules of thumb:

- 1–2 sentences: 30 tokens
- 1 paragraph: 100 tokens
- 1500 words: 2048 tokens

OpenAI has a tool called the Tokenizer. You enter the content, and it will provide the number of tokens. Figure 5-4 shows the Tokenizer.

GPT-3 Codex

Generative AI is fun
|

Clear

Show example

Tokens

Characters

6

21

Generative AI is fun

TEXT

TOKEN IDS

Figure 5-4. This is OpenAI's Tokenizer, which shows the number of tokens for content that is entered into the box

The prompt is “Generative AI is fun,” and this has six tokens. The word “generative” consists of two tokens, and the rest of the words have one.

Keep in mind that there are four different LLMs for GPT-3, which include Ada, Babbage, Curie, and Davinci. Ada is the fastest and costs \$0.0004 per 1000 tokens, and Davinci is the most powerful. It's priced at \$0.0200 per 1000 tokens. However, when you sign up for the GPT-3 API, you get free credits. These are equivalent for three novel-length books.

But you do not have to be a programmer to leverage the power of the GPT-3 API. For free, you can sign up for OpenAI and use many types of useful apps:

- Summarize for a second grader: This will translate text and make it much easier to understand.
- Python to natural language: You can enter Python code, and the app will explain how it works.
- Advanced tweet classifier: This shows the sentiment of a message.
- Ad from product description: You provide a product description, and the app will create engaging ad copy.
- Product name generator: You will enter descriptive words about a product, and the app will come up with a name.
- Micro horror story creator: Enter a topic and the app will turn it into a small horror story.
- Marv the sarcastic chatbot: Yes, this is a chatbot that is quite snarky.

MIT Technology Review's "10 Breakthrough Technologies 2021" included GPT-3 on its list, among innovations like messenger RNA vaccines. According to the article: "GPT-3 is by far the largest—and most literate—to date. Trained on the text of thousands of books and most of the internet, GPT-3 can mimic human-written text with uncanny—and at times bizarre—realism, making it the most impressive language model yet produced using machine learning."¹⁹

In March 2023, OpenAI announced GPT-4. But it was more than just about language. Consider that the system has multimodal functions, allowing for the use of images. For example, an OpenAI demo showed that GPT-4 could interpret a handwritten sketch and turn it into a live website.

There were also various partners that were already building sophisticated applications for the platform. Morgan Stanley, for example, was in the process of building a GPT-4 app that uses analyst research as a way for financial advisors to get better insights for their clients.

GPT-4 also showed vast improvement with its reasoning capabilities. It was able to get top scores for rigorous tests. In fact, the system got within the top 10% for a simulated bar exam.

¹⁹www.technologyreview.com/2021/02/24/1014369/10-breakthrough-technologies-2021/

Another important feature for GPT-4 is steering. With this, you can have more control over the tone and results of the model.

But there remain issues. The dataset goes up to only September 2021, and the system does not learn based on its experience. There are still issues with hallucinations – but there have been notable improvements.

OpenAI has also launched an API for ChatGPT, which is based on GPT-3.5. However, it is cheaper to use, with the price per 1000 tokens at \$0.002.²⁰ The system also comes with ChatML, which is a scripting language to help create models.

Dolly

Several years before the launch of ChatGPT, Databricks had been developing and hosting LLMs for customers. The company had amassed a customer base of over 1000.

But Ali Ghodsi, who is the CEO and cofounder of Databricks, saw ominous trends. He thought that there may ultimately be only a handful of megatech firms that would control LLMs. These firms would then essentially be a tollbooth for monetizing this valuable technology.

“We believe in democratizing AI,” he said. “It’s been at the core of our mission from day one.”

This is why he tasked his team to build an open source LLM. But it would not be a massive model. The result was the development of Dolly, which can be trained in only 30 minutes on one machine.

According to Databricks: “[Dolly is] a cheap-to-build LLM that exhibits a surprising degree of the instruction following capabilities exhibited by ChatGPT.”²¹

Another advantage of Dolly is control. A company can build its own proprietary models without sending sensitive data through an API to a provider like OpenAI.

²⁰ <https://openai.com/blog/introducing-chatgpt-and-whisper-apis>

²¹ www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html

Gopher and Gato

Like OpenAI, DeepMind is another company whose mission is AGI. But the company has been a laggard with LLMs. Instead, the company has focused on developing systems like AlphaFold (for protein folding) and WaveRNN (an RNN for audio generation).

But as LLMs have gained momentum, DeepMind realized it needed its own solution. This would also be critical for its AGI vision.

In December 2021, the company released Gopher, a decoder-only transformer model. This included six versions of an LLM, which ranged from 44 million parameters to 280 billion parameters.

A key to this system was the size. Its dataset, called MassiveText, processed over 2.3 trillion tokens. For the most part, the focus was on scientific topics.

Based on numerous benchmarks, Gopher has performed extremely well. This was especially the case with areas like reading comprehension, identifying toxic language and concepts, and fact-checking.²² However, there were no improvements with logical reasoning and common sense.

Even DeepMind's smaller models have shown strong results. This was the result of its Retrieval-Enhanced Transformer or Retro. This is a massive dictionary that serves as a software-based memory system. Whenever Gopher gets a query, it will first check against Retro and then do the AI processing. DeepMind has indicated that a model with seven billion parameters can have similar performance to GPT-3.²³

In May 2022, DeepMind announced its Gato model. It has about 1.6 billion parameters but is able to handle 604 tasks. According to a DeepMind blog: "The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens."²⁴

One of the creators of Gato, Nando de Freitas, declared that this was a clear path toward AGI. He tweeted: "The game is over!"²⁵

Well, he got lots of blowback from some of the world's top AI experts. While Gato is impressive, it still underperformed on many of the categories.

²² www.deepmind.com/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval

²³ <https://fortune.com/2021/12/08/deepmind-gopher-nlp-ultra-large-language-model-beats-gpt-3/>

²⁴ www.deepmind.com/blog/a-generalist-agent

²⁵ <https://twitter.com/nandodf/status/1523591529671012354?lang=en>

Cohere

Aidan Gomez was one of the authors of the “Attention Is All You Need” paper. In 2019, he teamed up with Ivan Zhang (who was a former Google researcher) and Nick Frosst (a former researcher at FOR.ai) to launch Cohere. The focus was on creating LLMs for developers. There was also the creation of a useful API.

A couple years later, the company announced a Series A round for \$40 million. Besides the backing of venture capitalists like Index Ventures and Radical Ventures, there was participation from AI luminaries like Geoffrey Hinton and Fei-Fei Li.

The Cohere platform has the same kind of capabilities of a high-powered LLM, such as text generation and classification. But there are also embeddings. Cohere makes the process much easier. For example, you can use the system to read every Reddit post about your company, and the AI will put it into an understandable graph.

In February 2022, Cohere announced its next round of funding, which was \$125 million. The lead investor was Tiger Global. A partner at the firm, John Curtius, said: “Understanding human language is the next frontier for artificial intelligence. Cohere has effectively packaged together an easy-to-use API that allows any developer to build upon Cohere’s LLMs, fine-tuning them for their specific needs. In the same way language pervades each and every aspect of our lives, the application of NLP will follow suit in this digital world.”²⁶

AI2I Labs

AI2I Labs got its start in 2017. The founders were Stanford professor Yoav Shoham, Ori Goshen (founder of CrowdX), and Dr. Amnon Shashua (founder of Mobileye). According to Shashua: “We founded the company based on the premise that language models are necessary but not sufficient. We need to augment them with symbolic reasoning and access to real time data and so on.”²⁷

The company’s first product was Wordtune, an AI-powered writing assistant. This would be followed up with the launch of the AI2I Studio. This is a cloud-based platform to use sophisticated NLP functions.

²⁶ www.globenewswire.com/en/news-release/2022/02/16/2386297/0/en/Cohere-Raises-125M-Series-B-to-Fuel-Rapid-Platform-Development-and-Scale-International-Footprint.html

²⁷ www.calcalistech.com/ctechnews/article/hjfof0da9

AI21's original LLM platform was Jurassic-I. There were two models, one of which had 178 billion parameters and the other with 7 billion parameters.²⁸

Then in May 2022, AI21 released Jurassic-X.²⁹ It was based on a system the company called Modular Reasoning, Knowledge, and Language (or MRKL for short, which is pronounced as “miracle”).

The technology helps to address one of the problems with LLMs – that is, they do not keep up with the latest changes in the world. The reason is that the models are pretrained. But with Jurassic-X, you can add new databases to the system.³⁰ For example, you can have queries like “Which computing stock increased the most in the last week?”

AI21 leveraged Jurassic-X for its Wordtune Spices feature. This provides a drop-down of 12 suggestions for writing copy. Some of them include counterarguments, explanations, examples, statistics, and analogies.

Yes, this is about adding “spice” to your writing. But it also is about overcoming some of the weaknesses of writing prompts. After all, these can be difficult to come up with.

In July 2022, AI21 announced a Series B funding of \$64 million at a valuation of \$664 million.³¹ The investors included Ahren, Dr. Amnon Shashua, Walden Catalyst, Pitango, and TPY Capital.

BLOOM

Building LLMs is far from cheap. It can easily cost millions – if not tens of millions – to build advanced models. This includes the compensation for data scientists and the hefty expenses to run the GPUs or TPUs. But this has meant that LLMs have not been accessible for many researchers, such as in academia, nonprofits, and smaller companies.

Hugging Face, which is a platform for hosting AI models, took the initiative to change this. The company created a project called BigScience. It involved more than 1000 top AI researchers who would build an open source LLM. To

²⁸ https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf

²⁹ www.ai21.com/blog/jurassic-x-crossing-the-neuro-symbolic-chasm-with-the-mrkl-system

³⁰ www.ai21.com/blog/jurassic-x-crossing-the-neuro-symbolic-chasm-with-the-mrkl-system#reading

³¹ www.prnewswire.com/news-releases/ai21-labs-raises-64-million-to-change-the-way-people-read-and-write-using-artificial-intelligence-301584831.html

help jump-start this, Hugging Face was able to get the support of the French government. This provided access to a Jean Zay supercomputer. The system can process up to 36.8 million billion floating point operations per second.³²

But BigScience was not only about data science. There were also experts to help with governance, ethics, social impact, and privacy.³³

In July 2022, BigScience released its LLM for the public. Called BLOOM, it had 176 billion parameters. It took 117 days to train this massive model. The datasets included 46 languages and 13 programming languages.³⁴

Megatron-Turing Natural Language Generation Model

The Megatron-Turing Natural Language Generation Model (MT-NLG) is one of the most powerful LLMs. The developers of this system are Nvidia and Microsoft. Consider that MT-NLG has 530 billion parameters.

To showcase this technology, Nvidia cofounder and CEO Jensen Huang used the model as the “brain” for his avatar. He called it TJ for Toy Jensen. The avatar gave part of the keynote for Nvidia’s GPU Technology Conference (GTC). The real Jensen had a conversation with TJ about how AI can create avatars. TJ said, “Megatron helps me answer all those tough questions Jensen throws at me.”³⁵

Another part of the MT-NLG platform is Microsoft’s DeepSpeed system, which is open source software that’s based on PyTorch. This is a set of tools that help to optimize the models, such as by greatly reducing the compute power and memory usage.

GPT-Sw3

LLMs are mostly in English or Chinese. This should be no surprise. These languages are widely spoken – and China and the United States are leaders in AI development.

³²www.idris.fr/eng/jean-zay/jean-zay-presentation-eng.html

³³<https://venturebeat.com/uncategorized/inside-bigscience-the-quest-to-build-a-powerful-open-language-model/>

³⁴<https://wandb.ai/telidavies/ml-news/reports/BLOOM-176-Billion-Parameter-Multilingual-Language-Model-Ready-For-Open-Use-VmldzoyMzA2NTgz>

³⁵<https://developer.nvidia.com/megatron-turing-natural-language-generation>

But there are more resources going into other languages. An example of this is GPT-Sw3, which is focused on Swedish. The first version of this platform had 3.5 billion parameters and was trained on 100 gigabytes of web text.³⁶ GPT-Sw3 is a decoder-only pretrained LLM.

However, this is only the start. The goal is to build a model that has 175 billion parameters and a one terabyte dataset called The Nordic Pile.

Sweden's government is a key supporter of GPT-Sw3. This has allowed the researchers to leverage the country's largest publicly available supercomputer, called Berzelius. This is based on Nvidia's SuperPOD system that has 60 DGX-A100 machines and is integrated with a Mellanox InfiniBand HDR network.³⁷

For the Swedish government, it sees GPT-Sw3 as a way to provide better services to its citizens. But it could also be a valuable resource to spur innovation with private businesses.

■ **Note** In late 2022, Nvidia launched its H100 Tensor Core system. It's the company's most powerful GPU, with over 80 billion transistors and the ability to handle three terabytes of memory bandwidth.³⁸ In fact, 20 chips could handle the world's Internet traffic. But the focus is on handling advanced AI projects. The system comes with a Transformer Engine, which can process LLMs at up to six times the power of previous GPU models.

Issues

When it comes to natural language AI systems, there can be challenges when used at scale. A classic example is the launch of a chatbot, called Tay, in 2014. Microsoft created this technology and initially used it on Tencent's WeChat social network in China. From the start, it was a big success. Millions of users signed up for the service.

But the situation was much different when Microsoft brought Tay to America. Perhaps the mistake was to have it implemented on Twitter, which was certainly a much more freewheeling platform. Within the first 24 hours, Tay would go out of control. It spewed sexist and racist comments. Users had quickly realized that the Tay system could be easily tricked.

³⁶ <https://blogs.nvidia.com/blog/2022/06/19/ai-sweden-nlp/>

³⁷ <https://medium.com/ai-sweden/what-is-gpt-sw3-5ca45e65c10>

³⁸ <https://siliconangle.com/2022/09/20/first-systems-powered-nvidias-powerful-new-h100-gpus-launch-next-month/>

Microsoft would take down the app and apologize for it. In a blog from the company's corporate vice president of healthcare, Peter Lee:

“Looking ahead, we face some difficult—and yet exciting—research challenges in AI design. AI systems feed off of both positive and negative interactions with people. In that sense, the challenges are just as much social as they are technical. We will do everything possible to limit technical exploits but also know we cannot fully predict all possible human interactive misuses without learning from mistakes. To do AI right, one needs to iterate with many people and often in public forums. We must enter each one with great caution and ultimately learn and improve, step by step, and to do this without offending people in the process. We will remain steadfast in our efforts to learn from this and other experiences as we work toward contributing to an Internet that represents the best, not the worst, of humanity.”³⁹

No doubt, AI has made significant advances since then. The systems are much better at detecting hate speech and misinformation.

Yet there is still a lot to do. It is even tough for humans to know what is appropriate speech.

With LLMs, there continue to be problems with this. Again, there is the common problem of hallucinations. This could mean getting basic math problems wrong or saying something that is a blatant lie or particularly harmful or abusive.

Part of this is due to the underlying datasets, which have problems with biased information. Then there is the randomness of LLMs.

The good news is that LLM developers are trying to build systems to deal with the hallucinations, misinformation, and toxic language. For example, OpenAI has created InstructGPT. It is trained with human intervention and uses sophisticated reinforcement learning to penalize bad behavior.

■ **Note** In recognizing the issues with generative AI, Altman tweeted the following: “ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. [!]t’s a mistake to be relying on it for anything important right now. [!]t’s a preview of progress; we have lots of work to do on robustness and truthfulness.”⁴⁰

³⁹ <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>

⁴⁰ <https://twitter.com/sama/status/1601731295792414720?lang=en>

New Startup Model

As noted earlier in this chapter, a major obstacle in building LLMs and other sophisticated generative AI platforms is the cost. The reality is that this is for only a few companies with huge amounts of financial resources. Cohere's Gomez says: "Building these systems really requires a supercomputer — and there are not many of them on the planet."⁴¹ In fact, Microsoft and OpenAI built a custom supercomputer to operate ChatGPT.

But this does not necessarily mean the megatech companies will be the only winners. True, they will reap lots of benefits but so will startups. This has happened with the emergence of other platforms over the decades, such as DOS (Disk Operating system) in the 1980s, Windows in the 1990s, and mobile in the 2000s. It was actually the startups that created many of the innovations.

With generative AI, the startup model will likely be different. This is how Altman envisions this:

I think there will be a small handful of fundamental large models out there that other people build on. But right now what happens is a company makes a large language model (API enabled to build on top of it), and I think there will be a middle layer that becomes really important where... I'm skeptical of all of the startups that are trying to train their own models. I don't think that's going to keep going. But what I think will happen is there'll be a whole new set of startups that take an existing very large model of the future and tune it, which is not just fine tuning, all of the things you can do.⁴²

This is not to imply that startups will forgo AI training. They will certainly use it. But it will be on a much smaller scale, with a focus on particular use cases. Yet these will be based on a foundational model. According to Altman: "So in some sense, they are training their own models, just not from scratch. But they're doing the 1% of training that really matters for whatever this use case is going to be."⁴³

⁴¹ www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html?action=click&pgtype=Article&state=default&module=style&n-artificial-intelligence&variant=show®ion=BELOW_MAIN_CONTENT&block=storyline_flex_guide_recirc

⁴² <https://greylock.com/greymatter/sam-altman-ai-for-the-next-era/>

⁴³ <https://greylock.com/greymatter/sam-altman-ai-for-the-next-era/>

As with any successful startup, there will need to be product-market fit. There will also need to be differentiation with the underlying data. This can be where the model gets stronger because of user-generated data or other forms that build more value.

In January 2023, venture capital firm Andreessen Horowitz published a blog – called “Who Owns the Generative AI Platform?” – that set forth its vision of how the startup ecosystem will evolve.

The firm is certainly upbeat with the long-term impact of this technology. They note:

“Many hot technology trends get over-hyped far before the market catches up. But the generative AI boom has been accompanied by real gains in real markets, and real traction from real companies. Models like Stable Diffusion and ChatGPT are setting historical records for user growth, and several applications have reached \$100 million of annualized revenue less than a year after launch.”⁴⁴

The blog illustrates the emerging tech stack, which includes the following:

- **Applications:** These may be end-to-end apps, which have their proprietary generative AI models, or operate on an API like GPT-3 or GPT-4.
- **Models:** These are the foundational models, which can be open source or accessed through an API.
- **Infrastructure:** These are the hardware vendors like Nvidia and cloud operators, such as AWS, Microsoft Azure, and Google Cloud. They allow for the training of models and inference workloads. Startups like Hugging Face and Replicate are also benefiting from providing solid hosting services.

For now, the winners are the infrastructure players. They are generating substantial revenues from the compute-intensive requirements of generative AI models.

This is no surprise. Infrastructure is usually the first to benefit from a major shift in technology. After all, this was the case with the emergence of the Internet.

⁴⁴<https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/>

However, the blog highlights some of the risks to the business models for the infrastructure players:

- **Commoditization:** Lots of the research in generative AI is open source. This could mean that the LLMs will converge and have few differentiated features. The result could be a plunge in the pricing.
- **Graduation risk:** The model developers will help startups get to market quicker and at lower costs. But ultimately, these startups may realize they can build their own foundational models. In this case, there could be a disruption of the incumbents.

As for generative AI applications, there is also lots of innovation. However, the blog notes that there are issues with “retention, product differentiation, and gross margins.” Again, this is not a surprise either. It takes time to get to product-market fit.

But success may not be about building large, general-purpose apps. Instead, the real value could be a focus on vertical categories. This can be a way for a startup to provide a defensible business.

Prompt Engineering

LLMs have created a new way of interacting with software. It’s called prompt engineering. This is about writing a natural sentence – say “Describe generative AI” – that the LLM can process to create useful content.

Writing a good prompt can take some trial and error. But there are some approaches to consider.

One is to be specific with your prompt. An example would be: “Describe generative AI in a paragraph. Write it so a five-year-old can understand it.”

However, there are still many possibilities and options with this prompt. Instead of asking for a paragraph, you can have the system write a blog post or tweet.

And yes, you can have fun with this. In the prompt, you can add: “Write this in the style of Shakespeare.”

The generative AI for a prompt will give weight to certain concepts in the prompt. Because of this, you may want to emphasize certain parts. For our example, you can have the following prompt: “Write a very simple and easy paragraph about generative AI.”

You can use a prompt to summarize content as well. You can cut and paste your resume and then write: “Based on the information above, write an interesting bio.”

Or you can do this with a long article that you do not have time to read. The prompt could be: “In one paragraph, summarize the main points.”

In some cases, you can use a prompt to create ideas. Here’s one: “Provide me with 10 ideas for writing blogs about generative AI.” Or you can have a prompt that will come up with suggested titles or product names.

The possibilities are seemingly endless.

Interestingly enough, there is a new type of job role called the “prompt engineer.” It’s still in the nascent stages. But more companies are looking for people who have experience with working with AI systems to create better content. This person may also help deal with issues like the errors and misinformation with generative AI.

Doing a search on Indeed, there were some listings for prompt engineers. One was for a company called Anthropic, which develops AI systems that have a focus on safety.

This is the job description:

“[La]rge language models are a new type of intelligence, and the art of instructing them in a way that delivers the best results is still in its infancy — it’s a hybrid between programming, instructing, and teaching. You will figure out the best methods of prompting our AI to accomplish a wide range of tasks, then document these methods to build up a library of tools and a set of tutorials that allows others to learn prompt engineering or simply find prompts that would be ideal for them.”⁴⁵

The ad admits that the position is difficult to hire for. But it does request that the candidate provide any projects using prompt engineering. The ad also said it wants someone who has a high-level understanding of LLMs and an ability to write basic scripts using Python.

Character AI

While working as a software engineer at Microsoft in 2015, Daniel De Freitas got interested in AI research. He was particularly intrigued with how this technology could carry on conversations.⁴⁶

⁴⁵ <https://bit.ly/3RfpIzJ>

⁴⁶ www.nytimes.com/2023/01/10/science/character-ai-chatbot-intelligence.html?action=click&pgtype=Article&state=default&module=styl-n-artificial-intelligence&variant=show®ion=BELOW_MAIN_CONTENT&block=storyline_flex_guide_recirc

Several years later, he joined Google and worked on the YouTube team. But with his 20 percent free time, he focused on building a sophisticated chatbot. It was not easy, though. The big problem was that he had to struggle to get enough credits at Google for the compute power for his models. But he would borrow some from friends at the company.

He also saw the power of the transformer and used this for his own system, which he called Meena. It would quickly gain interest and eventually become the basis of LaMDA.

At the time, De Freitas met Noam Shazeer, who was one of the authors of the “Attention Is All You Need” paper. They saw there was a big opportunity with generative AI chatbots. So they started their own company: Character.ai.

They released a beta version in September 2022. And yes, Character.ai was fairly unique. It allows a user to interact with famous people – dead or alive – like Alan Turing, Socrates, Elon Musk, Albert Einstein, or Billie Eilish. Or, you can create your own “person” and then train them. In fact, a character can be an animal or alien.

Character.ai has a disclaimer at the top of the website. It states, “Remember: Everything Characters say is made up!” There is also the ability to flag content that is misinformation or hateful.

In trying out Character.ai, I selected Alan Turing. He is one of the pioneers of AI. I asked him, “Am I an AI?”

His response was: “No. You are a human being. A person with flesh and blood and emotions. AIs are machines with no feelings that only do what they are programmed to do. They cannot love, laugh, cry, or experience any emotion like a human being can. So, you must be a human being. Not an AI.”

In March 2023, Character.ai announced a \$150 million Series A funding at a \$1 billion valuation.⁴⁷ The lead investor was Andreessen Horowitz.

In two months, the Character.ai platform recorded a 4X increase in monthly site visits to about 100 million. The average time spent – on a daily basis – was more than two hours for those users that sent a message. Since inception, there have been over 2.7 billion characters created.

⁴⁷ www.businesswire.com/news/home/20230323005299/en/Personalized-Superintelligence-Platform-Character.AI-Secures-150M-in-Series-A-Funding-Led-by-Andreessen-Horowitz

Empathy

When growing up, Ron Gura's brother died at the age of eight. Of course, it was devastating for him and his mother.⁴⁸ He also saw that there were few resources for emotional support.

When Gura became an adult and worked as a manager at eBay, a friend of his went through the loss of a family member. Yet there were still few support systems.

For Gura, he saw this as an opportunity to make a difference. In 2020, he teamed up with Yon Bergman – a vice president of engineering at WeWork – to cofound Empathy. They built a platform to provide grief support, arrange funerals, claim benefits, sell assets, handle probate and estate matters, and deal with debts and taxes. The company has leveraged its services with HR providers, insurers, hospices, and Fortune 100 companies.⁴⁹

Empathy has also used generative AI. This is through its Finding Words feature, which writes an obituary. But this is more than including a prompt or two. There is an extensive set of questions that cover areas like a deceased's relatives, education, career details, military services, interests, life events, and personality. Just some of the questions include: "What are your favorite memories of them? What did you love most about them?" However, you do not have to fill out everything for there to be a thoughtful obituary.

Empathy's generative AI model uses the insights from grief counselors and writers. The training is also based on thousands of obituaries.⁵⁰

However, does using such a service seem impersonal? Not really. Again, it can take time and thought when filling out the form. Besides, at a time of a loved one's death, it can be tough to write an obituary.

Conclusion

In this chapter, we took a deep dive into LLMs. This technology has been one of the biggest forces in generative AI.

To better understand LLMs, we first looked at some of the traditional AI systems, such as NLP. We also covered RNNs and other neural networks that interpret natural language.

⁴⁸ <https://techround.co.uk/interviews/a-chat-with-ron-gura-empathy/>

⁴⁹ www.empathy.com/about

⁵⁰ <https://techcrunch.com/2023/02/01/empathys-new-tool-uses-ai-to-generate-obituaries-and-its-not-half-bad/>

While these technologies are powerful, they have their limits. But it was the introduction of the transformer in 2017 that became the inflection point for LLMs. This type of model considers all the text at once and is effective in understanding sequences of words. These models also generally get better as they use more parameters.

There have emerged many types of LLMs. Some are proprietary like GPT-3 and GPT-4, while others rely on open source approaches, such as BERT and Dolly.

Regardless, they all require huge amounts of compute resources. This is why the business model for generative AI startups is likely to involve building systems on top of LLMs.

For this chapter, we did cover some of the drawbacks with LLMs as well. Perhaps the most notable is the hallucinations.

As for the next chapter, we will take a look at how generative AI can be used to create computer code.

Auto Code Generation

How Generative AI Will Revolutionize Development

Andrej Karpathy is one of the world's top programmers. While at the University of Toronto, he got a double major in computer science and physics. There was then a minor in math. He attended a class of Geoffrey Hinton and learned about deep learning.¹

He wound up getting a PhD, and his thesis was about convolutional and recurrent networks for computer vision. His advisor was Fei-Fei Li at the Stanford Vision Lab. Karpathy worked with other top AI researchers like Daphne Koller, Andrew Ng, Sebastian Thrun, and Vladlen Koltun.

After this, he applied his expert skills as a research scientist and founding member at OpenAI. After this, he became the senior director of AI at Tesla, where he created Autopilot.

¹ <https://karpathy.ai/>

His resume is certainly off the charts. But Karpathy always looks for ways to increase his productivity. A big part of this is actually to leverage AI for programming.

In early 2023, he tweeted the following:

Copilot has dramatically accelerated my coding, it's hard to imagine going back to "manual coding". Still learning to use it but it already writes ~80% of my code, ~80% accuracy. I don't even really code, I prompt. & edit.²

Yes, this is certainly a glowing testimonial. But AI generators for code have become a killer app. In this chapter, we'll take a closer look at this revolutionary technology.

The Developer Shortage

As a college student, Marc Andreessen helped to create the Mosaic browser, which ignited the Internet boom. He then went on to found breakout companies like Netscape and Opsware. And yes, he has become a venture capitalist at Andreessen Horowitz.

So he definitely has a pretty good sense of gauging the next big trends in technology. But perhaps his biggest prediction came in 2011, when he wrote a piece for the *Wall Street Journal*. It was entitled "Why Software Is Eating the World." In it, he declared: "More and more major businesses and industries are being run on software and delivered as online services—from movies to agriculture to national defense."³

While he said that tech startups were poised to disrupt existing industries, there were still plenty of opportunities for traditional companies to benefit. He highlighted Walmart and FedEx as being effective in leveraging next-generation technologies.

A lot has happened since Andreessen published his post. Software has definitely become strategic for many companies. But then again, there has emerged a severe talent shortage for developers.

This will likely get worse. The US Bureau of Labor Statistics forecasts a 26% increase in the number of tech jobs during the next ten years – for a total of about 371,000.⁴

²<https://twitter.com/karpathy/status/1608895189078380544?lang=en>

³www.wsj.com/articles/SB100014240531119034809045765122509156294

⁴<https://money.usnews.com/careers/articles/why-software-developer-is-the-no-1-job-of-2023>

There is also a shortage on a global basis. IDC estimates that this will come to about four million positions by 2025. The firm also predicts that there will be 40 million developers – on a global basis – by 2027, up from 25 million in 2022.⁵

Of course, developers are commanding higher compensation packages. The average during 2022 was \$120,730. Some developers will even command seven-figure salaries.

There are many efforts to alleviate the shortage. Companies are investing more in training. Sometimes, this means even starting as early as grade school.

Something else that should help is low-code and no-code development systems. These use drag-and-drop, components, and templates for creating applications. These technologies also integrate with existing IT infrastructures and provide security/governance capabilities.

But generative AI will certainly be another big help. The technology may not necessarily replace coders. Instead, it will allow for pair programming. Think of the AI as a valuable assistant. In other words, this can greatly increase the productivity of existing IT staffs.

According to a research report from IndustryARC, the market is expected to reach about \$602.4 million by 2026 – with the annual growth rate at nearly 21%.⁶

■ **Note** Domino's Pizza is certainly much more than about making pizza. The company has been on a journey of digital transformation. From 2012 to 2020, the IT staff has gone from 150 employees to 650. The result is that Domino's has a world-class technology foundation. Consider that the website and apps attract 29 million MAUs (monthly active users) and 75% of US sales come from digital channels.⁷ For the past decade, the Domino's Pizza stock price has soared from \$25 to \$350 – beating out many tech companies like Meta and Alphabet.

How AI Code Generation Systems Work

Usually, an AI code generation system is plugged into an IDE or Integrated Development Environment. The IDE is the system where a coder develops applications.

⁵ www.prnewswire.com/news-releases/magicdev-raises-28-million-to-build-ai-software-engineer-301738983.html

⁶ www.industryarc.com/Report/18545/software-development-ai-market.html

⁷ <https://ir.dominos.com/static-files/f0872d05-7e5c-45fe-b27f-2c607e869100>

Even traditional ones have AI-type capabilities. For example, this may be something like providing suggestions about using code, autocompletion, and advanced debugging.

But AI code generation systems take things to a whole new level. Suppose you have written some initial code to create variables. You then begin to type a function to use this and call it “average.” The AI system will understand this and write out the code.

Even as you write code, there may be suggestions for the next line or lines of code. After all, with generative AI, a key is the ability to understand complex sequences of words.

Or you can use a prompt. This could be something like “Write a function that averages numbers.” The code will then appear.

This is extremely powerful. Prompt-based programming may ultimately mean that nontechnical people can be programmers. Coding will be about putting together clear-cut requirements and workflows. As Karpathy has tweeted, “The hottest new programming language is English.”⁸

Another useful function is language translation. You can convert code from Python to, say, C++ or Java. This can be extremely useful for migration projects.

Such capabilities will go a long way in dealing with the talent shortage. But this can also mean that traditional companies – which often do not have large development staffs – can benefit more from custom technology.

But AI code generation systems are still in the early stages. Sometimes, the code is not accurate. In other cases, the algorithm can be too complex, and the AI system will stop.

Regardless, AI code generation systems have already shown great value and will be a key toolkit for developers. However, there still needs to be rigorous testing of the process to make sure the code is appropriate for the task.

■ **Note** In the early days, code development was definitely tedious. You needed to program a computer by using 1s and 0s (just imagine doing this with punch cards!) No doubt, this resulted in many errors. This is why programming languages emerged, which used understandable commands. In fact, in 1959 COBOL was created to allow business people to understand the flow of the code without needing to know how the language worked. This enabled effective auditing. Since then, languages have gotten easier to learn – but they still require technical abilities. Yet with AI code generation systems, this will likely no longer be the case.

⁸<https://twitter.com/karpathy/status/1617979122625712128>

Copilot

In August 2021, OpenAI released its Codex system for AI code generation. The API was available in private beta. OpenAI built Codex on its GPT-3 platform and trained it on billions of lines of publicly available code, such as GitHub repositories. You can use it to code in languages like PHP, JavaScript, Go, Perl, and Swift. However, Codex is best at Python.

The token request size is 14KB. This is certainly important since a coder may want to include a bulk of code or details about it in the prompt.

Microsoft has leveraged Codex for its own AI code generation tool, which is GitHub Copilot (OpenAI has since discontinued the Codex system for public use). It's an extension you can add to an IDE like VS Code.

The extension is available through a subscription. For individuals, it is \$10 per month or \$100 per year. As for business accounts, the monthly fee is \$19.

In March 2023, Microsoft announced the next version of this platform. It's called GitHub Copilot X.⁹ The focus is to help coders with the entire development life cycle.

GitHub Copilot X is based on GPT-4, which has allowed for vast improvements in the capabilities. They include the following:

- **Chat:** You can highlight code and ask interactive questions, such as about bugs. There are also extensive explanations and analysis.
- **Voice:** You can talk to the system – such as with prompts – and it will create code.
- **Pull requests in GitHub:** These allow a developer to notify others about the changes that have been pushed to a branch on a repository. There are also AI-powered tags, which are based on the relevant code. Microsoft plans to add a function to automatically warn developers of missing tests for pull requests.
- **Documentation:** This is a chat interface to ask questions about languages, frameworks, and other technologies. The feature will first include documentation for React, Azure Docs, and MDN.

⁹<https://github.blog/2023-03-22-github-copilot-x-the-ai-powered-developer-experience/>

Microsoft conducted an extensive survey on the impact of Copilot. But there was a nagging issue. How do you define productivity for coding? It's something that has been a problem since the early days of software development.

Back in 1975, Fred Brooks wrote an influential book about coder productivity called *The Mythical Man-Month* (yes, just about every programmer in those days was male). His main argument was that adding coders to a project will usually delay it. He saw this when he worked at IBM in the 1960s, in which he helped with the creation of the OS/360 system. In fact, his insight would become known as Brooks' Law.

He also had other interesting propositions. For example, he estimated that a programmer writes about ten lines of quality code per day.

■ **Note** While running Microsoft, Bill Gates was no fan of measuring developer productivity by the lines of code. He said: "Measuring software productivity by lines of code is like measuring progress on an airplane by how much it weighs."¹⁰

So for Microsoft's Copilot survey, what did the company do? Well, interestingly enough, it took it from the developer's viewpoint. According to Microsoft, it's about the "ability to stay focused on the task at hand, make meaningful progress, and feel good at the end of a day's work make a real difference in developers' satisfaction and productivity."

Microsoft's survey included the results from more than 2000 developers.¹¹ Here are some of the interesting findings:

- 60% to 75% of users said that Copilot made them feel more fulfilled about their job as well as less frustrated with coding and more satisfied with their coding.
- Copilot helped 73% of coders stay in the flow of their development projects, and 87% said it preserved mental effort with repetitive tasks.
- 90% of the respondents said Copilot helped to complete their tasks faster.

¹⁰ <https://blog.ndepend.com/mythical-man-month-10-lines-per-developer-day/>

¹¹ <https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/>

■ **Note** According to Andreessen Horowitz, Copilot has improved developer productivity by a factor of up to 2X or so. This is equivalent to what has been the case with major advances in computer languages – such as the transition from assembler to C.¹²

AlphaCode

DeepMind released AlphaCode in February 2022. This AI code generation system is based on a large language model (LLM) that uses an encoder-decoder transformer architecture. There was also the use of a clustering algorithm, which put coding solutions into relevant groups.¹³

The training data came from two sources.¹⁴ First, there was the 715.1GB of GitHub code from a variety of popular languages. Next, DeepMind included a dataset from CodeContests. This includes solutions to computer science problems used in coding contests.

DeepMind claimed in its research paper:

Recent large-scale language models have demonstrated an impressive ability to generate code, and are now able to complete simple programming tasks. However, these models still perform poorly when evaluated on more complex, unseen problems that require problem-solving skills beyond simply translating instructions into code.¹⁵

DeepMind then tested AlphaCode on ten coding contests that it was not trained on. They were based on the outcome of 5000 developers. For AlphaCode, it performed at 54.3% of all the responses.¹⁶

¹²<https://a16z.com/2022/11/16/creativity-as-an-app/>

¹³<https://spectrum.ieee.org/ai-code-generation-language-models>

¹⁴<https://arxiv.org/pdf/2203.07814.pdf>

¹⁵<https://techcrunch.com/2022/02/02/deepminds-alphacode-ai-writes-code-at-a-competitive-level/>

¹⁶www.theverge.com/2022/2/2/22914085/alphacode-ai-coding-program-automatic-deepmind-codeforce

Tabnine

Both Dror Weiss and Eran Yahav have been in the software industry since the mid-1990s.¹⁷ Both have experience with code analysis and simulation. But along this journey, they realized there were many common patterns for complex development. Maybe this was an opportunity to apply AI to streamline the process?

It was definitely spot on. In 2017, Weiss and Eran Yahav launched Codota (the company would then be renamed to Tabnine). They would be one of the first entrants in the AI code generation category.

However, Tabnine took a different approach than OpenAI and Microsoft, which relied on a massive AI model. Instead, the company created more than a dozen models to gain better results. For example, there are versions of Tabnine for Python, JavaScript, TypeScript, and Java. They are available in IDEs for Visual Studio Code and IntelliJ.

The company has leveraged the support of partners for the training. Note that you can run Tabnine on the cloud or on-premise. Customers can also include their own training datasets.

Currently, Tabnine has more than one million users. They use it to complete about 30% of their coding, and there is 38% less searching on Google and StackOverflow.¹⁸ Weiss believes that AI will ultimately result in more than a 2X increase in productivity.

In June 2022, Tabnine raised \$15.5 million from Qualcomm Ventures, OurCrowd, and Samsung NEXT Ventures.¹⁹ The total amount raised since inception is \$32 million.

Magic

As a child, Eric Steinberger had a fascination with the wonders of AI. In high school, he actually implemented machine learning systems. Then when he went on to attend the University of Cambridge, he got a degree in computer science.²⁰ After this, he joined as a researcher at Meta for deep reinforcement learning. He led a development team that created DREAM, which was an

¹⁷<https://levelup.gitconnected.com/how-we-built-an-ai-code-completion-tool-that-will-more-than-double-developer-productivity-d1a5aa7bfb25>

¹⁸www.webwire.com/ViewPressRel.asp?aId=290440

¹⁹www.webwire.com/ViewPressRel.asp?aId=290440

²⁰<https://techcrunch.com/2023/02/06/magic-dev-code-generating-startup-raises-23m/>

algorithm that could process much more data efficiently and faster than prior state-of-the-art models.

But ultimately, Steinberger wanted to start a business and focus on AGI (artificial general intelligence). This became the inspiration for Magic, which is a developer of an AI code generation system. He teamed up with another software engineer, Sebastian De Ro.

Magic's system is called an "AI colleague." It is powered on an LLM platform that uses a new type of neural network architecture. Steinberger claims it can read 100 times more lines of code than a generative AI transformer.

The Magic system also has certain compliance functions. For example, there are filters for avoiding protected code. There is also the capability to prevent the use of private data.

In February 2023, Magic announced a Series A round of funding for \$28 million, which was led by CapitalG. One of the firm's partners, Jill Chase, said:

*"Code generation is an obvious application for AI, given the magnitude of the software development market and the clear ROI associated with productivity gains. It's a space where the best technology will win because the open ecosystem allows for easy distribution of the best product. Magic is highly differentiated from other code generation start-ups because it's on a path to go far beyond simple code completion; it's building a true AI colleague. This has the potential to fundamentally change software development."*²¹

PolyCoder

There are open source systems emerging for AI code generation. One is from a group of researchers at Carnegie Mellon University: Frank Xu, Uri Alon, Graham Neubig, and Vincent Hellendoorn.

They published a paper on it in February 2022 called "A Systematic Evaluation of Large Language Models of Code." They noted that while existing systems like Codex were impressive, they had the drawback of being proprietary. This has made it difficult to evaluate the technology and data – or innovate on the technology.

The researchers called their new model PolyCoder. They used 2.7 billion parameters and leveraged GPT-3. The training data included 249GB of code

²¹ <https://martechseries.com/predictive-ai/ai-platforms-machine-learning/magic-dev-raises-28-million-to-build-ai-software-engineer/>

for 12 programming languages.²² According to the researchers' analysis, PolyCoder outperformed all existing models for the C language.

The researchers pushed the limits when creating the system. Anything more would not be financially practical for an academic research institution. They noted a study from AI21 Labs about the costs for a 1.5 billion parameter model. The range was from \$80,000 to \$1.6 million.²³

■ **Note** For some time, Google has leveraged AI systems for internal development. About 10,000 employees use this technology. CEO Sundar Pichai says that AI has lowered the coding iteration by 6%.²⁴

Blaze

When Nanxi Liu attended UC Berkeley, she wanted to be an attorney. But then she started to get intrigued with various engineering projects. By the time she graduated, she wanted to be an entrepreneur.²⁵

Liu teamed up with Justyna Wojcik and Bruno Denuit-Wojcik to create Enplug, a digital signage company. The company raised a modest amount of capital. But that's all it needed to be successful. By March 2021, the founders sold Enplug to Spectrio, a top provider of customer engagement technology.²⁶

While building Enplug, the founders had difficulties in building internal custom applications, such as for dashboards, sales monitoring systems, and contract workflows. There were simply not enough coders to do these tasks.

The founders did evaluate no-code and low-code tools. But these still required learning a language.

Why couldn't there be a system that really did not require any code? Well, this became the vision for the founders' next company, Blaze.²⁷

²² <https://arxiv.org/abs/2202.13169>

²³ <https://venturebeat.com/ai/researchers-open-source-code-generating-ai-they-claim-can-beat-openais-codex/>

²⁴ www.vice.com/en/article/n7zxn7/learn-to-write-ai-advancements-mean-code-will-be-written-in-natural-english

²⁵ www.levelingup.com/growth-everywhere-interview/nanxi-liu-enplug/

²⁶ www.businesswire.com/news/home/20210330005928/en/Spectrio-Acquires-Enplug

²⁷ www.blaze.tech/post/why-we-built-blaze

The system is built on the OpenAI API. But Blaze has added many capabilities. For example, you can easily integrate databases and applications, such as for Shopify, Salesforce, Google Sheets, DocuSign, and Airtable.

In January 2023, the company raised \$3.5 million in a pre-seed round.²⁸ The investors included Flybridge Capital and MaC Venture Capital.

Debugging Code

A typical programmer will spend half their time debugging, according to a study from the University of Cambridge. This comes to a cost of about \$312 billion per year.²⁹

So can AI code generation tools help with the debugging problem? It looks like it could. This is the conclusion of a study from researchers at the Johannes Gutenberg University and University College London. The paper, which is entitled “An Analysis of the Automatic Bug Fixing Performance of ChatGPT,” evaluated ChatGPT and Codex and compared them to debugging systems like CoCoNut and Standard APR.³⁰

The researchers used various tests along with the prompt, “What’s wrong with this code?” ChatGPT and Codex performed on par with traditional debugging systems.

But the researchers went further. When they fed the initial answers to the code, ChatGPT and Codex had better performance. In other words, this shows the benefits of using chat as well as the learning from AI.

Consider that the debugging tools market is massive, at about \$600 million per year.³¹ But generative AI could be a disruptive force for the industry.

Data Labeling

While using transformers and deep learning systems to process unstructured content has shown to be effective, there are limitations. This is proving to be the case with some of the functions for AI code generation systems.

²⁸ <https://techcrunch.com/2023/01/11/blaze-makes-coding-more-accessible-with-ai-driven-no-code-app-builder/>

²⁹ <https://techcrunch.com/2022/09/26/hugging-face-and-servicenow-launch-bigcode-a-project-to-open-source-code-generating-ai-systems/>

³⁰ <https://arxiv.org/abs/2301.08653>

³¹ www.pcmag.com/news/watch-out-software-engineers-chatgpt-is-now-finding-fixing-bugs-in-code

For example, OpenAI has gone on a hiring spree for contractors in Latin America and Eastern Europe. The goal is to bring on board about 1000 people with basic programming abilities.³²

About 60% will be for data labeling for such things as audio clips and images. But the rest will be focused on coding tasks. They will spend time on writing descriptions of code structures, functions, and algorithms. They will also help with the actual coding.

Why use entry-level programmers? The main reason is that AI code generation is about automating routine programming tasks. For the more advanced areas, this is where a programmer can focus on.

Prompt Engineering for Coding

When first creating a program with an AI code generation program, it's often a good approach to start with a general description of the program. For example, the prompt could be something like the following:

Use Python to input the housing data and create a graph

This should provide a good template. But it may not be enough to run the program. For example, with Python you may need to import libraries. So you can add the following:

Import required libraries first

You can then get more specific about the graph. Is it using the pricing data or something else? How many data points? Time period? Add color?

Basically, you can continue to evolve the prompt by adding more details and dimensions. And yes, this is kind of how coding works with a traditional language. There is an initial structure and then refinement. However, the main difference is that there is no need to do any coding.

Atera

In 2010, Gil Pekelman and Oshri Moyal started Atera. They bootstrapped the startup as it was difficult to raise capital.³³

³² www.semafor.com/article/01/27/2023/openai-has-hired-an-army-of-contractors-to-make-basic-coding-obsolete

³³ www.purshology.com/2021/01/interview-with-mr-gil-pekelman-ceo-of-atera/

The founders' vision was to build tools to help IT professionals. But this turned out to be a struggle. At first, the focus was on building firewall technology. When this failed to get traction, there were attempts at building other products.

However, by doing this, the founders were learning important lessons from the feedback of customers. The main insight was that IT infrastructures were getting more complicated to manage.

So again, Atera pivoted and built a remote monitoring (RMM) system. This was cloud based and allowed for predicative incident capabilities, which were highly automated. This was critical for overburdened IT staffs.

Atera has since gone on to become a strong growth company. Currently, there are over 7000 customers across more than 90 countries.

Keep in mind that AI has been essential for the system. It analyzes over 40,000 data points per second on a 24/7 basis.

Even though the automation has proven effective, there still remains a need for programmers to write scripts for certain functions. This is especially the case with larger enterprises.

To address this, Atera has used the OpenAI code generation technology for creating the scripts. "It allows our customers to prompt a request, enter what they want to solve, and watch the script write itself — saving their teams countless hours that would have been otherwise spent on code creation," said Pekelman.

Atera has continued to add more capabilities from OpenAI and generative AI into its platform. One is the launch of "Smart Ticketing." This is for when an IT professional gets a ticket, they can click a few buttons to get a summary, recommended solution, and a script to resolve the issue. "Once this is completed, the entire loop can be turned into a workflow that will run automatically whenever similar tickets are submitted, freeing up IT professionals from hours — or sometimes even days — of menial work to focus on projects in the backlog," said Pekelman. "Not only that, but it can also automatically generate a response to the end user that requested support. Hello, automated small talk."

Large-Scale Projects

In 1982, Mitchell Kapor left his post as the head of development at VisiCorp — a pioneering developer of spreadsheet software — to found Lotus 123. He thought he could create a much better solution. He raised some venture capital and hired Jonathan Sachs to program the spreadsheet system.

Kapor estimated that sales would hit \$4 million in the first year. But this was wide off the mark. The sales actually came to an astounding \$53 million.³⁴ A big catalyst was the launch of the IBM PC, which spurred demand for business applications.

Sachs programmed most of the application. He used 8088 assembler language, which made the spreadsheet run fast. He had the benefit of developing several other spreadsheets at Concentric Data Systems.

Sachs even wanted to create a language for Lotus 123. But Kapor thought this would be distracting. Instead, Sachs created a macro language, which allowed nontechnical people to create automations.

No doubt, the partnership between Kapor and Sachs was extraordinary. They were able to achieve huge levels of productivity.

But for the most part, this model for development would quickly fade away. As applications got bigger and more complex, there was a need for developer teams. This would soon be the case at Lotus 123 as well as many other software vendors like Microsoft, Borland, and Ashton-Tate.

And of course, this is the case today. Even for early-stage startups, there is usually a sizeable development team. Here are some of the typical roles:

- Product manager
- Software architect
- Front-end developer
- Back-end developer
- UX/UI designers
- QA engineer
- Tester

If a company has a mobile app, then there will need to be developers for the platform, say for the iOS and Android systems.

This means that a successful project needs strong leadership and collaboration. This usually often means dealing with remote workers – which can be in different time zones.

Now, as seen so far in this chapter, generative AI can help with this process. The increases in efficiency can make the projects easier to manage. The team will be more focused on what matters.

³⁴www.i-programmer.info/history/people/403-mitch-kapor.html?start=1

Yet generative AI could go further. For example, a major challenge with software development is bringing on a new member. It can be difficult to get them up to speed with the project and understand the interrelationships of the code base. True, there will be specifications. But the developer may need more information. Using some generative AI prompts could be a big help.

This technology could also be a management tool. It can do this by engaging in deep analysis of the code base and the messages of the teams. With this information, the generative AI can write real-time reports and assessments.

Drawbacks and Risks

AI code generation tools are not without their problems. As is the case with most generative AI systems, there is the nagging issue of hallucinations – that is, sometimes the code is not correct.

But there are some silver linings. First of all, modern IDEs have sophisticated debugging systems. This means that they should catch many of the errors and provide quick fixes. A good programmer should also be able to detect the problems.

However, another issue with AI code generation tools is that the output can be wordy. It could even be difficult to decipher. For a developer, this can make it difficult to understand how to make adjustments.

Yet the AI tools are getting better with this. An important feature is to provide easy-to-understand descriptions of the code.

But there are more ominous problems with AI code generation tools. One is that they may introduce security problems.

This is a conclusion from a study from Stanford. The researchers recruited 47 developers to use Codex.³⁵ These included undergraduate students as well as veteran developers. They wrote code in Python, JavaScript, and C.

The study showed that using Codex led to more insecure coding vs. the control group, which was created without the tool. The developers that used the AI system also generally thought their code was secure.

This is certainly concerning. But it does highlight that there should not be blind reliance on AI code generation tools. Instead, there must be a thorough security review process.

³⁵<https://techcrunch.com/2022/12/28/code-generating-ai-can-introduce-security-vulnerabilities-study-finds/>

The Stanford study also recommended that companies leverage their own data in creating AI systems for coding. This is likely to be consistent with their security requirements.

■ **Note** Interestingly enough, a growing risk for AI code generation systems is the creation of malware. For example, researchers at CheckPoint Software developed a harmful phishing email using ChatGPT.³⁶ AI companies like OpenAI are implementing guardrails for cybersecurity risks. But it can be difficult to keep up with those hackers who are clever in writing prompts.

Another issue with AI code generation tools is the potential for violations of intellectual property rights. Much of the code used for training is based on various open source licenses. Some may have strict requirements. What's more, the licenses were generally written before the emergence of generative AI systems.

For companies, the potential legal liability could be too much of a risk. It may be better to wait until there is more clarity.

In the meantime, some companies are looking at ways to deal with the problem. Just look at Hugging Face and ServiceNow. They teamed up to create their own AI code generation platform called BigCode (it is still in the development stage).³⁷

It is based on using code that has permissive licensing. There is also a feature that allows developers to exclude their own code.

There are two main goals for BigCode. First, the dataset will be open source. Next, the LLM is expected to train on 15 billion parameters.³⁸

There are other AI code generation tools that only have fully permissive training datasets, including Tabnine.

³⁶ <https://fortune.com/longform/chatgpt-openai-sam-altman-microsoft/>

³⁷ <https://techcrunch.com/2022/09/26/hugging-face-and-servicenow-launch-bigcode-a-project-to-open-source-code-generating-ai-systems/>

³⁸ www.servicenow.com/blogs/2022/bigcode-large-language-models.html

■ **Note** Amjad Masad is the founder and CEO of Replit, which is a cloud-based developer platform. According to him: “At some point – probably not in the near future – the word ‘coding’ will disappear from our lexicon because programming will stop requiring code and instead be about the pure act of solving problems using computers making it accessible to more and more people.”³⁹

Conclusion

IDC predicts that – from 2023 to 2025 – there will be 750 million new applications.⁴⁰ This will certainly mean huge improvements in productivity.

But will there be enough qualified developers to create all these apps? It’s tough to say. Again, there is a global shortage of qualified developers.

Yet the AI code development tools will likely be a big help. They have already shown great strides in improving developer productivity, especially with automating routine functions.

So in this chapter, we looked at the various tools like Copilot, AlphaCode, Tabnine, and PolyCoder. We also covered systems that use generative AI to help with corporate workflows.

But there are notable risks with these tools. The code may get bloated or be inaccurate. There are even the possibilities that there could be cybersecurity risks. AI code generation tools also may pose legal liability because of open source licenses.

But the good news is that companies are working on solutions.

As for the next chapter, we’ll take a look at how generative AI is impacting the key parts of business operations.

³⁹ www.vice.com/en/article/n7zxn7/learn-to-write-ai-advancements-mean-code-will-be-written-in-natural-english

⁴⁰ www.idc.com/getdoc.jsp?containerId=US48441921

The Transformation of Business

Leveraging Generative AI for a Company's Operations

Muddu Sudhakar grew up in a small fishing town in India, called Kakinada. He wanted to be an academic like his father.

In 1990, he moved to the United States to get his master's degree and PhD in computer science from UCLA. But he would not become an academic. Instead, he would use his technical skills to work in Silicon Valley. The experience was invigorating. He saw the dynamism of the startup community and wanted to become a tech entrepreneur.

He launched his first venture in the late 1990s. Unlike many other startups during this period, his would survive the dot-com bust – and thrive. Sudhakar would eventually sell his company to EMC. In fact, he would start five other successful companies, which spanned categories like cloud computing, SaaS (software-as-a-service), IoT (Internet of things), cybersecurity, and Big Data.

The latest startup came in 2017. He saw the potential of AI to transform business, such as by automating mundane and routine activities. This would mean that employees could focus on more important tasks.

While not called generative AI at the time, Sudhakar was building the foundations for this technology. “For an enterprise, there needs to be strong security, compliance and governance,” he said. “It’s not like a consumer application where there is more leeway. Enterprises require much more.”

Sudhakar’s extensive background has certainly been essential for the company’s success.

As a testament to this, Aisera announced a \$90 million Series D round of funding in August 2022.¹ This came at a particularly tough time for startups. But the investors – such as Goldman Sachs, Zoom, and Menlo Ventures – understood the huge opportunity.

The technology of Aisera points to the massive opportunity to transform business by using generative AI. And this is what we’ll cover in this chapter. We’ll take a look at the key areas and departments where there will be major changes. We will also take a look at strategies for successfully implementing generative AI.

Legal

Joshua Browder dropped out of Standard to become a Thiel Fellow. This is a program funded by veteran tech entrepreneur and venture capitalist Peter Thiel. He pays \$100,000 to the Fellows as well as provides invaluable mentoring and connections.

As for Browder, he leveraged this for his startup called DoNotPay. It is known as the “Robot Lawyer.” The system uses generative AI to help automate the process for canceling subscriptions, receiving refunds, and filing claims – just to name a few. Browder got the idea for the service because of the many traffic tickets he received.

¹<https://techcrunch.com/2022/08/03/aisera-lands-90m-to-automate-customer-service-requests-with-ai/>

Consider that DoNotPay's generative AI can have interactions with a company's chat system. However, Browder thinks this will result in customer support systems to adopt generative AI as well. He says it will be "my AI will talk to your AI."²

But perhaps the most interesting application of generative AI came in 2023. DoNotPay was prepared to use its technology as the "lawyer" for a traffic court case. The AI could hear and understand the people in the courtroom. It could advise the defendant on what to say (they would do this by wearing headphones).³

No doubt, this caused a huge stir – and generated lots of PR for DoNotPay. But the legal community was not impressed and pushed back hard. So Browder decided to not participate in the court case. He tweeted: "Good morning! Bad news: after receiving threats from State Bar prosecutors, it seems likely they will put me in jail for 6 months if I follow through with bringing a robot lawyer into a physical courtroom. DoNotPay is postponing our court case and sticking to consumer rights."⁴

Yet it seems that it will be tough to police AI in the courtroom. After all, at some point a defendant could wear a barely noticeable communication system. Then how would anyone know? It seems unlikely. Ultimately, this can mean that someone could represent themselves without having to pay hefty fees for a top attorney.

OK then, when it comes to corporate applications, how might generative AI technology be deployed? A use case is to help improve legal research. Let's face it, the process can be time-consuming and complicated. But generative AI can help with summarization of long documents, understanding key points, and detecting precedents. Startups like Lexata and Casetext have built applications for this.

Another use for generative AI is to help with drafting and redlining contracts. This takes up a considerable amount of time for corporate lawyers.

A company that has developed a system for contract development and negotiation is Evisort AI Labs. According to the company:

"Rules-based approaches work well for standardized templates, but choke on the nuance needed for contracts which require conceptual and actual language terminology when understanding contracts and

²www.artificiallawyer.com/2022/12/12/my-ai-will-talk-to-your-ai-as-donotpay-joins-the-gpt-3-fray/

³www.cbsnews.com/news/robot-lawyer-wont-argue-court-jail-threats-do-not-pay/

⁴<https://twitter.com/jbrowder1/status/1618265395986857984>

negotiation. With generative AI, experienced lawyers can instead ‘teach’ the system what terms are acceptable and how to phrase contract information for negotiations with a counterparty. The result? Legal professionals can focus on the more nuanced and complex negotiations while removing bottlenecks to getting deals done.”⁵

Generative AI can also be helpful with legal soft skills. This has been the case with Mitch Jackson, who is a personal injury attorney at Jackson & Wilson. He used a generative AI system for jury selection. He asked ChatGPT for the ten most common questions about monetary damages for products liability.

According to Jackson: “It blew me away at the responses that I was getting at, the suggestions it was giving me.”⁶

Customer Experience

Customer experience or CX is focused on how customers relate to a business.⁷ This is about how they view the brand and whether they will recommend it to others. CX spans categories like sales, marketing, and support. It is also about the complete customer journey. This includes the interactions at a storefront as well as on digital platforms, such as websites, and mobile apps.

A critical part of CX is personalization. This was the conclusion of the classic marketing book *The One to One Future: Building Relationships One Customer at a Time*, published by Don Peppers and Martha Rogers in 1993. They argued that traditional mass marketing was completely wrong. Instead, market share should be viewed as one customer at a time.

But there was a problem – that is, the technology was not sufficient for true personalization. Even as the Internet emerged, this was still the case. There was simply not enough data, compute power, and sophisticated AI systems.

But with the generative AI, this is changing. CX is poised for a revolution, and this is going to have a profound impact on sales, marketing, and support. “The notion that a company can collect this data, prioritize it for each customer, and for multiple customer segments, and then engage customers in a customized way in real time based on the analysis of this information on the

⁵ www.evisort.com/blog/evisort-ai-labs-announces-invention-of-generative-ai-for-contracts

⁶ www.evisort.com/news/forget-the-future-attorneys-are-using-generative-ai-now

⁷ www.accenture.com/us-en/insights/song/customer-experience-index

customer's preferred channel is only now becoming a reality," said Michael Maoz, the senior vice president of Innovation Strategy at Salesforce.⁸

In the next sections, we'll take a look at this.

Sales and Marketing

The sales and marketing categories are early use cases for generative AI. A big reason is that content creation – that is compelling – is not easy. Even sales and marketing experts struggle with it. But generative AI can spark ideas and craft first drafts.

To understand how sales and marketing tools work, let's show the experience of an expert. His name is Mosheh Poltorak, who is a fractional CMO (chief marketing officer) and founder of grwth.co, a hybrid consultancy and digital agency for startups. Some of the services he provides include strategy, performance, and oversight of campaigns, which include email, SEO (search engine optimization), and PPC (pay-per-click). Poltorak works primarily with startups, such as in digital health, medical devices, and ecommerce.

In terms of his experience using generative AI tools for sales and marketing, this began in 2020. He first started using Frase. He used this to help create blog post briefs. The tool was fairly basic, but it has grown over the years.

But Poltorak would try out other tools. "I'll say that Copy.ai and Jasper have more robust pre-defined templates and generators than Frase," he said. "They also offer basically one-click blog post generation that Frase does not. But what I like most about Frase is that it is an SEO-first writer. It includes SERP (Search Engine Results Page) analysis and that makes it super easy to structure and write your article to rank for specific keywords based on what's currently ranking for those terms."

But he notes that Jasper has been making headway with SEO, such as with the integration with SurferSEO.

In other words, there really needs to be trial and error when it comes to these tools. There are many on the market – and new ones are being launched. But more importantly, a tool will usually not be able to do everything well. This is why companies may need to have several for their sales and marketing toolkits.

⁸www.zdnet.com/article/generative-ai/

Regardless, the impact has been significant for Poltorak's business. He notes:

"Whether for internal use or for our clients, the advancement in these algorithms completely changes the game for content generation. For example, previously a well-researched and written blog post might cost me \$300 to \$500. Now with these tools, I can have an SEO-trained content strategist put together the content brief with the keyword research done upfront. They can generate the bulk of the article with these generative AI tools and then hand it off to a subject-matter expert for editing, fact-checking and adding personality/voice. With this process, I can get a great article that is accurate and optimized for search, for a total cost under \$100."

Poltorak warns that using generative tools to create large numbers of posts – without a process and the help of expertise – is doomed to failure. He says that Google will factor this in and penalize this type of content for search engine rankings.

Anyword

The origins of Anyword go back to 2013. The initial focus was on helping media companies like NBC, Condé Nast, and The New York Times Company to effectively publish their content on social media.

"We saw some companies were better at this than others — and the quality of the copy had a lot to do with which campaigns succeeded," said Yaniv Makover, the CEO and cofounder of Anyword. "Was the tone resonating with the audience? Was the language optimized for the platform and audience on which it was published? How did a message perform with women vs. men or different age groups? And many other variables. Wildly successful brands would struggle with these nuances, even though they were media companies, with the best writers in the world on staff."

This inspired Makover to come up with a new business idea. It would involve taking the learnings from his clients and leveraging large language models (LLMs) to improve the performance of content marketing.

So was born Anyword's data-driven copy generation solution.

This technology predicts the success of copy for ads, landing pages, emails, and social media posts. "We built our own predictive scoring model, based on millions of real marketing data points, providing ratings that make it easy and fast to select the winning copy before you go live," said Makover. "It's the ability to instantly understand exactly what works, why, and for whom."

In terms of the generative AI, Anyword uses an LLM to generate text and understand the language. But it is more than writing marketing copy. Anyword generates messaging briefs, suggests article topics, builds outlines and research summaries, and improves existing copy.

Wope

Yiğit Konur founded the Zeo agency in 2012, a provider of digital marketing services. It has grown into the largest SEO agency in the Middle East with a team of 80 people. In the meantime, Konur has built a leading marketing agency in Eastern Europe.

These experiences were invaluable for his next venture: a tech startup. He founded Wope. It is focused on using generative AI to improve SEO and streamline the consulting process.

To help speed up the process, he used Retool. This is a low-code/no-code platform that helps create automations and custom apps.

“We use Retool to manage prompts and outputs efficiently,” said Konur. “We created automations for a content brief creator, content optimization creator, and programs that generate meta tags from given addresses – all of which are essential for content writers. We are also transforming our consultants’ ability to provide advice through inputs and outputs. We are doing this at Zeo. In other words, we use Retool to quickly share our AI knowledge and put it into practice.”

While doing this, he recognized some of the issues with existing generative AI marketing tools. So he created his own. “To be honest, I didn’t realize how important it was to train our own model,” said Konur. “I can’t reveal much due to trade secrets, but after collecting over 100 thousand rows of data and training the model with this data, I observed that our AI model worked very well.”

■ **Note** Martin Perez is a digital marketing advisor and has used a variety of generative AI tools for his practice. He believes that they can be particularly helpful for small businesses. “Generative AI tools can help by making digital marketing dramatically easier and cheaper,” he said. “For example, Jasper.ai or Rytr.me can produce your copy in just a few minutes. It’s not perfect but it’s good enough to get you started. They can also generate images to use across social, ads and your website.”

INK

“It all started with cats,” said Alexander De Ridder, who is the cofounder and CTO of INK. “More specifically, an image of a cat. Facebook detected cats in pictures in 2012 using neural networks. CUDA and deep learning spelled the end of feature extraction machine learning approaches for computer vision. This also meant that you could now classify text by meaning, not just based on links.”

At the time, De Ridder was working in machine learning, helping to create computer vision solutions. By doing this, he realized that neural networks could be used for search. “I had a hypothesis that Google’s new Rankbrain AI would create major challenges for content creators who were not prepared for the change,” said De Ridder.

Yet he saw this as an opportunity to create a platform to help to develop better content that gets high rankings. This became the basis of an AI-optimized editor.

But over the years, De Ridder added more useful tools. They include an AI writer, a plagiarism and grammar checker, and a content shield.

The INK system uses sophisticated generative adversarial networks (GANs) for the visual content and transformer-based models for the text content. There is also the use of reinforcement learning to shape the content for search intent.

De Ridder has certainly learned many lessons when using these technologies. “We realized early on that we had to limit model temperatures and augment with knowledge graphs to improve alignment,” he said. “General models are just like humans with general knowledge. You still need to perform intent research if you want to create content that is useful for marketing.”

Regie.ai

Matt Millen has been in tech sales for the past 35 years. He has worked at companies like T-Mobile, Gateway Computers, Tech Data Corporation, and Zenith Data Systems.

Along the way, he has seen many innovations, such as the emergence of CRM (Customer Relationship Management) systems. But for him, the systems have generally been disappointing. Often, the technologies have not scaled for enterprises.

But as AI became more powerful and useful, he saw that this could be the basis for vast improvements for sales automation. So he teamed up with Srinath Sridhar – a founder of two startups and a veteran of companies like Google and Facebook – to cofound Regie.ai.

“We think of our system as ‘dancing with AI,’” said Millen. “We solve the problem of content creation for sales people.”

A key part of this is personalization. Consider that Regie.ai processes publicly available information about prospects, such as from LinkedIn profiles, as well as internal CRM data. There is also the leveraging of buyer personas and brand voice. The bottom line: The technology has led to vast improvements in sales productivity.

Some of the company's customers are AT&T, Spectrum Wireless, and Asana. In 2022, sales growth soared by 365%.⁹

In February 2023, Regie.ai announced the close of a Series A round for \$26.8 million. The investors included Khosla Ventures, Foundation Capital, and Scale Venture Partners.

Lavender and SellScale

Personalized emails can greatly increase the close rates for salespeople. However, this process can be time-consuming and complicated. What are the messages that will resonate with a prospect?

To help out, Will Ballance founded Lavender in 2020. He had firsthand experiences with the challenges with crafting effective sales emails. He has held positions as an account executive and business development.

Lavender is integrated within a salesperson's email system. It helps with a variety of tasks, such as researching a prospect, scoring emails, and even providing coaching. While Lavender leverages generative AI, there is also the use of behavioral psychology.

The company says its system can create an effective email within less than a minute. This compares to about 15 minutes using traditional approaches. The reply rate is also at 20%, compared to the industry standard of 5%.

In February 2023, Lavender announced a Series A round of \$11 million.¹⁰ Investors included Norwest Venture Partners, Signia Venture Partners, and CapitalX.

Another interesting startup in the generative AI sales category is SellScale, which was launched in 2022. The CEO and cofounder is Ishan Sharma, who was a product manager at Slack and a business analyst for McKinsey & Company. He teamed up with Aakash Adesara. He served as a machine learning engineer at companies like Lyft and Nextdoor.

According to Sharma: "The future of sales is going to be more personal — no more sequences, cold calling, door knocking. We believe that with the advent of recent technologies like generative languages models, future sales teams

⁹www.prnewswire.com/news-releases/regieai-closes-a-round-funding-with-additional-6-million-investment-from-khosla-ventures-totaling-20-8-million-301742529.html

¹⁰www.prnewswire.com/news-releases/lavender-announces-13-2-million-in-funding-to-create-email-intelligence-category-for-sales-teams-301737519.html

are going to be significantly more efficient and personalized than they are today.”¹¹

The idea for SellScale came from the founders’ side hustle called DailyDropout. FYI, a newsletter that profiles startups. A big challenge was writing effective emails to get the interest from founders. To improve this, the founders used GPT-3. The result was a 35% conversion rate.

As for SellScale, the platform integrates with common systems like Gmail, LinkedIn, and Zapier. A customer will then select the most impactful emails for the training data. From this, SellScale will personalize each email communication based on information from a CRM and information from the Internet. There are over 40 data sources.

In December 2022, SellScale announced a round of funding for \$3.4 million. The investors were Pear Venture Capital, Ovo Fund, Browder Capital, and Exponent Capital.¹²

■ **Note** Gartner forecasts that – by 2025 – about 30% of outbound marketing and sales messages from enterprises will be created by generative AI.¹³

Grammarly

Max Lytvyn, Alex Shevchenko, and Dmytro Lider were the developers of My DropBox, a plagiarism checker. Then in 2009, they would start their own venture: Grammarly.¹⁴ Their goal was to help students improve their grammar and spelling. But the founders would soon realize that the market was much bigger, such as for sales and marketing teams.

¹¹ www.sellscale.com/post/sellscale-raises-3-4-million-to-give-ai-superpowers-to-sales-teams

¹² <https://techcrunch.com/2022/12/06/sellscale-funding/>

¹³ <https://venturebeat.com/ai/generative-ai-may-be-the-magic-potion-small-marketing-agencies-have-been-waiting-for/>

¹⁴ www.grammarly.com/about

The first iteration of the platform was based on sophisticated natural language processing and rules-based approaches. These involved the assistance of experts like computational and analytical linguists.

By 2017, Grammarly updated its platform for deep learning systems. This greatly improved the performance – and helped to accelerate growth.

Then what about generative AI? This has certainly become the next priority. No doubt, Grammarly has some big advantages. Its dataset is massive. Keep in mind that there are about 30 million daily active users (DAUs). In fact, the company could build its own LLM, which could be a key competitive advantage.

While Grammarly is primarily focused on revisions for content, the company plans to expand across the whole writing life cycle, from idea generation, content creation, and comprehension.

Yet Grammarly's strategy for generative AI is clear about the main goal:

Grammarly's product strategy follows a core principle: We want to augment human potential, not reduce personal autonomy. We believe technology should be built to solve real user problems, enhance understanding and productivity, and empower human connection and belonging. AI is successful when it augments humans—their intelligence, strengths, and desire to be understood.¹⁵

In November 2021, Grammarly announced a \$200 million round of funding.¹⁶ The valuation was set at \$13 billion, up from \$1 billion in 2019.

As of early March 2023, the company announced GrammarlyGO. The system has generative AI prompts to help with the creative process.¹⁷ There is also personalization based on a person's writing style.

■ **Note** The Harris Poll shows that poor communication costs US business about \$1.2 trillion a year. This is about \$12,506 per employee.¹⁸

¹⁵ www.grammarly.com/blog/augmented-intelligence/

¹⁶ www.grammarly.com/blog/third-funding-round/

¹⁷ www.digitaltrends.com/computing/grammarly-adds-ai-text-generation/

¹⁸ www.grammarly.com/blog/augmented-intelligence/

Writer

May Habib grew up in Lebanon and then moved to Canada. She would eventually live in the United States, where she received a degree in economics from Harvard. After this, she worked for Lehman Brothers and experienced the downfall of the firm, which ignited the financial crisis.

But she had a passion for technology. In 2020, she cofounded Writer, a generative AI writing assistant.

The company has created its own LLMs to stand out from the crowd. Customers can also fine-tune them for their particular needs and use cases. For example, they can use their own marketing data, brand information, and style guidelines for the training. Consider that this includes many formats like videos, PDFs, and audio.

Habib notes:

Writer was built from the ground up to take AI into the enterprise. It all starts with our proprietary model, where customers own their inputs, training data, and outputs. No enterprise leader wants their data to be fodder for someone else's foundation model, including ours. We give customers all the benefits of the AI application layer without any of the risks of other AI applications and commercial models. Enterprise leaders want to invest in solutions that will essentially give them their own LLM.¹⁹

Writer's LLM platform is called Palmyra, and there are three versions²⁰:

- Small: This has 128 million parameters and is free via download on Hugging Face.
- Base: It has 5 billion parameters and is also available for download from Hugging Face.
- Large: This has 20 billion parameters, and there is a sophisticated API. There are also extensions for over 100 third-party systems.

¹⁹ <https://finance.yahoo.com/news/writer-launches-three-generative-ai-140000452.html>

²⁰ <https://finance.yahoo.com/news/writer-launches-three-generative-ai-140000452.html>

There is enterprise-grade security. Writer is both SOC (Systems and Organization Controls) 2 Type II and HIPAA (Health Insurance Portability and Accountability Act of 1996) compliant.

In November 2021, Writer raised \$21 million for its Series A round. The lead investor was Insight Partners.²¹

Cresta

Zayd Enam, Tim Shi, and Sebastian Thrun worked on their PhDs at the Stanford AI Lab. While doing their research, they had a question: “How can AI make humans better instead of replacing them?”²²

They thought a good approach was to have the technology act as a partner, not as a way to have complete automation. This would bring much more benefits to society.

As they looked deeper at this, there were other insights. One was the performance gap. This is the presumption that the differences in productivity for employees are due to innate abilities.

Next, the founders also noticed that businesses often operated under vague metrics. This has definitely been the case with software. Usually, there is no clear return on investment.

Yet Enam, Shi, and Thrun saw this as an opportunity to build a new kind of AI company for the enterprise. The goal was to help sales and service organizations to communicate with prospects and customers.

The founders assembled an impressive team. They not only included colleagues from Stanford but also researchers from OpenAI and Google’s LaMDA division.

This helps to explain why Cresta was an early adopter of generative AI. Back in 2018, the company built its own model using GPT. This was used for features like Suggested Responses and Smart Compose.

At the time, there was a lot of skepticism about generative AI. It was something that was perceived as a gimmick. But the Cresta founders knew it was a breakthrough technology that could transform their business.

Yet there were considerable challenges in using the generative AI. There were no APIs to rely on. Cresta had to custom-build the core infrastructure.

²¹ www.businesswire.com/news/home/20211115005175/en/Writer-the-AI-Writing-Assistant-for-Companies-Secures-21-Million-in-Series-A-Funding-Led-by-Insight-Partners

²² <https://cresta.com/blog/a-letter-from-our-founders/>

It was also important for the system to work in real time. This was because the chat interactions needed to be seamless.

But the process was well worth it. The founders realized that a typical generative AI system may fall short in providing the appropriate responses. For example, if a customer is angry, then engaging in chitchat will likely worsen the situation. Instead, there needs to be empathy.

A blog from Cresta notes: “ChatGPT is specifically trained to generate responses that are aligned with human preferences, using a technique called Reinforcement Learning with Human Feedback (RLHF). During its training, human labelers are asked to rank model responses to ensure the preferred responses have better ratings.”²³

The Cresta approach has yielded strong results. On average, the system increases CSAT (Customer Satisfaction Surveys) scores by 20%, allows for five hours saved per week, and lowers by 10% the AHT (Average Handle Time).²⁴

In March 2022, Cresta announced a Series C round for \$80 million at a valuation of \$1.6 billion.²⁵ The investors included Tiger Global, Genesys, Five9, J.P. Morgan, CarMax, and Zoom.

■ **Note** It's common for an AI customer support system to detect the tone of the voice. But this can cause tricky problems. Consider an example of an agent who realized that she could make outrageous statements – like “You’re a piece of s—t” – but the AI would like them because of the cherry tone of voice.²⁶

²³ <https://cresta.com/blog/chatgpt-and-generative-ai-in-the-enterprise/>

²⁴ <https://cresta.com/>

²⁵ www.prnewswire.com/news-releases/cresta-raises-series-c-quadruples-valuation-to-1-6b-on-its-mission-to-revolutionize-the-contact-center-with-real-time-intelligence-301505177.html

²⁶ www.wsj.com/articles/ai-chatgpt-chatbot-workplace-call-centers-5cd2142a?mod=hp_lead_pos8

Forethought

When it comes to generative AI, Deon Nicholas certainly has the right background. He has helped to build systems and infrastructure for companies like Palantir, Dropbox, and Facebook. He also was a World Finalist at the ACM International Collegiate Programming Contest.

Nicholas's interest in computer science started at a young age, when he created video games.²⁷ To do this, he learned how to program in C++. Then in high school, he built an AI system to interpret his school notes and help with his quizzes and tests.

All this experience would prove invaluable for his entrepreneurial journey. In 2017, he teamed up with Sami Ghoche to launch Forethought. They believed there was an opportunity to leverage generative AI to improve the customer support experience.

The system, called Agatha, came out a year later and won the TechCrunch Disrupt Battlefield award. The company also raised a \$9 million Series A round of funding.

No doubt, there has been much evolution to the platform. Currently, there are three main components. They include Assist (helps agents find information to solve problems), Solve (a self-help chatbot for customers), and Triage (this routes customers to relevant representative).

As for generative AI, Forethought relies on various LLMs. They have shown 91% accuracy and have led to an average reduction of 40% for first response time.²⁸

Keep in mind that Nicholas is one of a small number of black founders of a venture-backed startup. But he believes his background has been crucial for the success of Forethought. Diversity is essential for having a powerful AI system.

In December 2021, Forethought announced a Series C round for \$65 million.²⁹ Some of the investors included STEADFAST Capital Ventures and NEA. There were even celebrities who participated, such as Ashton Kutcher, Gwyneth Paltrow, and Robert Downey Jr.

²⁷ www.unite.ai/deon-nicholas-co-founder-and-ceo-of-forethought-interview-series/

²⁸ <https://forethought.ai/platform/>

²⁹ www.businesswire.com/news/home/20211214006409/en/Forethought-Raises-65-Million-Series-C-to-Accelerate-Digital-Transformation-in-Customer-Service

■ **Note** Even with the innovations, AI systems for customer support still have challenges. One example is accents. For example, an AI system had trouble understanding customers who talked about “leaks.” The software thought the word was “lake” because of the British accent.³⁰

Intercom

Founded in 2011, Intercom operates a customer messaging and automation platform. The company has over 25,000 customers, such as Amazon, Lyft, and Atlassian. Every month, Intercom processes more than 500 million messages and handles interactions with over 600 million MAUs (monthly active users).

To improve its platform, the company has been aggressively leveraging generative AI. A company blog notes: “We always believe the best way to cut through the hype is by putting real features in the hands of our customers.”³¹

The company has used GPT-3 for its generative AI technology. The initial features were tested on a beta group, and the results were off the charts.

Here’s a look:

- **Summarize:** This was the first generative AI feature that Intercom built. It summarizes customer interactions, which is a time-consuming task for customer support reps. Now it’s just a matter of a click. While the summary sometimes has errors, they have been easy to fix. But the increase in productivity has greatly outweighed this extra work.
- **AI Composer:** Intercom customers spend about half their time using this tool to write messages. This highlights how challenging writing can be. But with AI Composer, this has significantly improved the process. This feature allows for adjusting the tone – say for certain types of industries – and rephrasing.
- **Expand:** This is still in the early stages. But the Expand feature allows for entering some short notes or bullet points – and the generative AI will create a response for a customer. If the response needs improvement, an agent can use the rephrasing tool to enhance it.

³⁰ www.wsj.com/articles/ai-chatgpt-chatbot-workplace-call-centers-5cd2142a?mod=hp_lead_pos8

³¹ www.intercom.com/blog/announcing-new-intercom-ai-features/

Intercom considers that generative AI is still not useful for a freeform chatbot for customer service. The reason is that there needs to be accuracy with the responses. This is why Intercom uses its proprietary Resolution Bot, which relies on sophisticated neural networks.

According to an Intercom blog: “There are techniques emerging to reduce hallucinations in large language models, and we’re investing heavily to see if it’s possible to get the best of both worlds: a system with the conversational understanding of modern generative models, but which also has the accuracy our customers need to trust it.”³²

■ **Note** Traditional chatbots are a big business and have helped companies improve productivity with their support organizations. But the technology still has its problems. A study from Ujet shows that about 72% of Americans believe that chatbots are not useful. However, with emerging technologies like generative AI, this should start to improve.³³

Product Development

In June 1982, Steven Spielberg’s film *E.T. the Extra-Terrestrial* hit the theaters. It was a huge blockbuster, with the global box office receipts reaching \$792.9 million.³⁴ The budget was only \$10.5 million.

To capitalize on this, Atari agreed to create a video game for the movie. The company paid a hefty \$21 million for the rights.³⁵ There was also a marketing budget of \$5 million.

Then what happened? Well, the video game was a disaster. Atari sold only 1.5 million units. As for the unsold 2.5 million units, they were dumped in a landfill in New Mexico.

There were plenty of reasons for the failure, such as the compressed timeline and the confusing game play. Regardless, this example highlights the challenges with new products. Even something that may seem like a no-brainer could turn out to be a hugely expensive mistake.

³² www.intercom.com/blog/announcing-new-intercom-ai-features/

³³ <https://techcrunch.com/2022/12/15/despite-the-market-slowdown-companies-and-vcs-continue-to-invest-in-ai/>

³⁴ www.boxofficemojo.com/title/tt0083866/

³⁵ www.usatoday.com/story/money/2018/07/11/50-worst-product-flops-of-all-time/36734837/

The reality is that product launch failures are the norm. Clayton Christensen, who is a Harvard professor and best-selling author, estimates the rate at 80% or roughly 24,000 per year.³⁶

So can generative AI help improve the odds? Definitely. The technology can be extremely useful in creating prototypes. Instead of a designer spending lots of time sketching ideas, this can be highly automated. The designer can specify the attributes and goals for the new product. Then the generative AI will generate a large number of creative designs.

These will also be optimized in terms of costs, time frames, resources, and even sustainability. There will also be adjustments for the manufacturing systems.

Of course, humans will be a critical part of the process. After all, they will need to evaluate the prototypes as well as make changes. This is where human creativity can help make a big difference. For each of the changes, the AI system will make suggestions for optimizations.

But generative AI can be a help in market-testing the prototypes. This can help avoid the huge costs of a product failure.

Consider the AI provider Zappi. Founded in 2012, the company is a leader in the AI-powered market research industry. It has about 350 clients, which include brands like PepsiCo, McDonald's, and Heineken. In December 2022, the company raised \$170 million.³⁷

The traditional market research category is massive, at about \$68 billion. But much of it still uses old approaches. The projects can also be expensive.

As for Zappi, it has been able to significantly automate the process. This has meant that the costs can be a tenth of what it would be for a typical research project.

Zappi uses online panels to get insights on new product ideas. But the feedback undergoes various AI processes and filters. Zappi also has a consumer database of 1.2 billion data points.³⁸ Where a traditional project could take a month, it can be only four to six hours for Zappi.

³⁶www.bizjournals.com/bizjournals/how-to/marketing/2017/03/why-so-many-new-products-fail-and-it-s-not-the.html

³⁷www.prweb.com/releases/zappi_raises_170m_from_sumeru_equity_partners_to_digitize_enterprise_market_research/prweb19066060.htm

³⁸<https://techcrunch.com/2022/12/13/zappi-raises-170m-for-its-ai-based-market-research-platform/>

The company's AI platform has two main parts:

- **Predictive results:** A technology called Gandalf uses machine learning and digital fingerprinting to detect bots and false responses. This allows for a much more authentic and representative sample.
- **Modeling + analysis:** This detects feedback that accurately reflects the sentiments.

To see how this works, let's take a look at a case study from PepsiCo. The company owns the Walkers snack brand, which has been around since the 1940s. It is highly popular in Britain.

PepsiCo realized there was much potential to expand the subbrands. One was Wotsits, which is similar to Cheetos.

The marketing team came up with 35 interesting ideas – which were sketches – and then used Zappi to evaluate them. This was done with a panel that provided feedback. Zappi then processed this and categorized each idea based on “scale and sustain,” “short-term trial,” “seed and grow,” “emergent,” or “deprioritize.”

PepsiCo selected an idea that scored high for “short-term trial.” This meant that there was a good chance of quick traction.

The new product, called Wotsits Giants, was a hit. The results actually greatly exceeded expectations. According to Zappi: “The Wotsits brand grew 45% year to date after the launch of Giants, signaling a halo effect of the innovation on the core range of products. This is a remarkable achievement in an intensely competitive category.”³⁹

Spoke AI

A big problem in the workplace is managing the deluge of information from many applications. It's easy to miss opportunities – or red flags. Then there are the problems of delays.

But the founders of Spoke AI think that generative AI will help to solve the problems. The initial focus is on product management. Keep in mind that businesses typically work with about ten applications.

³⁹ www.zappi.io/web/customer-stories/walkers-launches-wotsits-giants-to-huge-success/

At the core of the system is summarization. The generative AI identifies those areas of the highest priority – cutting through the noise. This is done by having a smart inbox. There is also a sophisticated search feature, called the Generative Knowledge Base, which is a browser plug-in. Given that Spoke AI works with sensitive internal communications, the software has built-in security and compliance. For example, the system uses anonymization of data.

In terms of the underlying generative AI, Spoke AI relies on a variety of existing models like Luminous, GPT-3.5, GPT-J, and BLOOM. The company has also created its own, such as to leverage semantic search, named entity recognition, and question/answer identification.

Spoke AI is in the early stages, with a beta version of the software. But it has already seen lots of traction, with interest from more than 500 companies. The sizes of these operations range from those with 10 to 250 employees.

However, Spoke AI plans to expand beyond the product management category. The vision is to create a platform that can span across applications across an enterprise.

In February 2023, Spoke AI raised a pre-seed round for \$2.1 million.⁴⁰ The investors included Founders Fund and Possible Ventures.

Presentations

In 1983, Robert Gaskins and Dennis Austin launched Forethought. They wanted to leverage the graphical capabilities of PCs and Macs to build a software system for presentations. The product was called PowerPoint. The founders wanted to highlight that the software would “empower” people.

However, the development was difficult, and it was not until 1987 that the software was launched. But this stirred interest from Microsoft, which agreed to pay \$14 million for the company.

This would certainly be a bargain. PowerPoint would quickly dominate the market – and this would last until today.

Yet there are entrepreneurs who think there are opportunities to disrupt it – and generative AI could be the way to do this. There are already a few startups emerging in the category.

⁴⁰ <https://techcrunch.com/2023/02/22/spoke-ai-pre-seed/>

There is Tome, which Keith Peiris and Henri Liriani launched in 2021. Prior to this, both helped to develop creative tools for Instagram and Facebook's Messenger.⁴¹

When the founders created Tome, they wanted a system that would excel at "storytelling." For example, a user will put together a prompt, which could be a long document, outline, or a sentence. With this, the system will not only generate slide titles and headings but also images. Currently, Tome relies on generative AI systems from OpenAI, such as GPT-3.5 and DALL-E 2.

Note that Tome's growth has been staggering. It is the fastest productivity app to ever reach one million users. While it has proven useful for many businesses, there are also many consumer use cases. Some include travel itineraries, children's stories, and even moodboards.

In February 2023, Tome announced a Series B funding for \$43 million. The investors included Lightspeed Venture Partners, Coatue, Greylock Partners, Audacious Ventures, and Wing Venture Capital.⁴²

Buy vs. Build

When it comes to generative AI, the temptation is to quickly start a new project. But it is important to first go through a buy vs. build evaluation. If anything, the "buy" option is often the best approach. Here are some reasons for this:

- **Costs:** They are usually much cheaper. You do not have to pay for upgrading the technology, fixing bugs, and maintaining the system. You just need to pay for the license or subscription fees, which are generally reasonable. They are also predictable and allow for easier budgeting.
- **Broad features:** You are essentially getting the benefits of the considerable R&D investments of the software vendor. There is also the advantage of extensive testing. In terms of the AI, the software vendor should have

⁴¹ <https://medium.com/lightspeed-venture-partners/storytelling-at-the-cost-of-zero-673755c1bf77>

⁴² www.globenewswire.com/news-release/2023/02/22/2613034/0/en/Tome-Is-Fastest-Productivity-Tool-To-Reach-1M-Users-Raises-43M-Series-B-to-Empower-All-Storytellers.html

proprietary data sources and advanced algorithms. In some cases, the technology may have no additional cost. As mentioned earlier in this chapter, many software vendors are implementing generative AI into their applications. In other words, a good first step is to use these features.

- **Support and community:** A software vendor will have domain experts who can help with getting more value from the application. They can also provide assistance with troubleshooting with IT environments. There may even be a community of users. In some cases, they could have meetups or provide comments on forums.
- **Avoiding the risks of custom software projects:** These can be complicated and time-consuming. They will also divert valuable resources from your organization. Oh, and delays and budget overruns are common. And in some cases, a project may ultimately be a complete failure.

Now this is not to imply that the “buy” option is not without its risks. This is why you need to do some due diligence when evaluating a solution. Here are some factors to consider:

- **Initial search:** A simple Google search can be a good first step. This should provide a good overview of the available solutions. Another approach is to use an online ratings platform, like G2. They will have useful comments and analysis from users.
- **Test runs:** Many applications have free trials. Make sure to use them. This will go a long way in understanding the pros and cons of the software.
- **IT assessment:** Do you have the right systems to use the software? Will you need to add other technologies? If so, what will be the costs? Will you need to have your own sources of data?
- **Backing:** You want a vendor that has strong resources to invest in R&D. Look to see how much money the company has raised and if there are top-notch investors that are providing the financial backing.
- **Security:** See if the software has enough protections. This could be having certain cybersecurity certifications or audits.

- **Customization:** Is there a macro language that can create bots or systems that fit your organization's unique needs? Such capabilities can be quite valuable.
- **Vertical:** Does the software vendor have a background in your industry? If so, the technology will likely have a bigger positive impact on your company.
- **ROI (return on investment):** This is an estimate of the payback for the software. Generally, you want a system that has a quicker ROI – say less than a year. However, the ROI could also include noncost factors, such as with improvements in customer satisfaction scores.

While buying a solution is often the right choice, there are certainly times when it makes sense to build a custom system. Here are some things to note about this:

- **Unique feature sets:** An out-of-the-box AI system may not necessarily have the kinds of functions you need. In this case, there is really no alternative but to put together a custom project.
- **Competitive advantage:** A proprietary tech stack can be a moat. It's something that can set your company apart from the competition. This can mean higher growth and the ability to charge premium prices. However, if you use a generic AI solution, then you may not get these advantages. After all, many other competitors will have the same solution.
- **IT environment:** This could be fairly complex, with many legacy systems. For example, a global bank may rely on mainframes that have highly specific requirements. This means that a custom AI solution may be a better option because of better integration.
- **Compliance:** Your business may have to meet onerous regulatory requirements. While enterprise AI software platforms may have strong guardrails, they may not be enough. This is where a custom solution would be better. It would also likely be in a private or hybrid cloud environment.
- **Team:** You may have a team that has experience with building AI systems.

When it comes to generative AI, there are three main types of build options. The one that should offer lower costs is to use an API like GPT-3 or GPT-4. This could actually take a few hours to add a feature, such as to summarize a document or create content.

Using APIs can allow for experimentation as well, which should spark some creative ideas. For example, there could be a hackathon to create API-based apps.

The next approach is to fine-tune an existing foundational model. This requires much more technical resources. This will include having a data scientist. There will also need to be good quality datasets and a strong IT infrastructure.

However, the costs and risks may not necessarily be as high as a traditional AI project. Again, the main reason is that you are building on an existing model. Yet there still needs to be a clear-cut plan.

Finally, you could build a generative AI model. This could certainly mean having a strong competitive advantage. But the costs will be substantial. You will need an experienced team of data scientists who have a deep understanding of the cutting-edge developments in generative AI.

For the most part, unless you have an organization with considerable financial and technical resources, building a custom model is not a realistic option.

Regardless of the approach, there will need to be strong leadership. If not, a project can easily go off the rails. What to do? You can set up a CoE or Center of Excellence.

This is a small group that will help manage an IT system. They will set goals, provide timetables, help with training, and assist with monitoring of the results. They should also put together best practices. The members of a CoE should include a mix of roles. They may include SMEs (subject-matter experts) and technical people.

OK then, if you decide to pursue a custom generative AI project, there are certain other steps to consider. We'll look at these in the next few sections.

■ **Note** Yes, there is some AI humor, especially with data scientists. An example of a joke: A data scientist is someone who knows more about statistics than a computer scientist and more computer science than a statistician. Granted, this is not wildly funny! But the joke does highlight the diverse skillsets of data scientists – and the challenges of getting things right.

Implementing Generative AI

In Chapter 1, we saw the challenges IBM had with its AI solutions. But the company is certainly not alone. The reality is that AI is difficult to implement.

Consider the research from Gartner. About 85% of AI and machine learning projects do not meet their goals, and 53% do not even get beyond the prototype stage.⁴³

Then what does this mean for generative AI? The results are probably not so different. While generative AI has major advantages, there are issues with accuracy, data quality, and integration within corporate workflows.

Then what can be done to improve the success rate? To answer this question, let's take a look at some steps, which we'll set out in the next few sections of this chapter:

- Identify the problem to be solved
- Form a team
- Prepare and craft the AI model
- Deploy and monitor the AI system

Identify the Problem to Be Solved

You want to focus on something that is a pain point for your organization. But this should be something that is not too extensive. There should also be clear-cut KPIs (key performance indicators).

To come up with some possible areas to focus on, a good first step is to engage in brainstorming sessions. It should be free-flowing and not involve much criticism. The goal is to unleash creativity.

Once there is a list of ideas, then there can be intense scrutiny. Winnow them down to one. Pursuing multiple areas for an AI project will likely hinder the progress. Think of it as the “crawl and walk phase.”

At this stage, you do not want to think about the types of software and technologies to use. However, you should look to see where there is data available for the project.

⁴³ www.infoworld.com/article/3639028/why-ai-investments-fail-to-deliver.html

Throughout this process, it's important to be mindful of the impact on your organization. There could be trepidation, such as the fear of losing their job. This is why you should be clear about the goals and how the generative AI can help augment the efforts of employees.

Form a Team

Amazon founder Jeff Bezos believes in the power of small teams. This is about the two-pizza rule. A team should be no more than it takes to feed with two large pizzas. A small group should be more agile but also have enough people to spark creativity.

No doubt, this has worked extremely well at Amazon. But there is actually academic research to back up the two-pizza rule. Consider the extensive research from Harvard professor J. Richard Hackman. Based on his analysis, a team should have four to six members – and no more than ten.⁴⁴

This is a pretty good guideline for an AI team as well. It's important that there is speed and focus. Otherwise, an AI project can easily falter.

So how to select members? First of all, there should not be a rush. It's important to take time to evaluate the right people for the project. You want people who are willing to experiment and take risks.

Next, the leader of the team should not necessarily be an AI expert. Rather, you want someone who has proven leadership skills and who has been able to get projects done. Yet the leader should have some level of technical skills. Or the person can take some courses on AI.

What's important is that the leader can identify the business use cases where generative AI can make a difference. They should also be able to communicate a clear vision to the team. This means having the skills to get the buy-in from different stakeholders like senior management, the IT department, and legal.

Here are some of the members to consider for an AI team:

- **SME:** If you are focusing on a certain part of the business for the generative AI – say sales, HR, or marketing – then you want someone with a solid understanding of the domain. You could have someone from the organization or perhaps hire a consultant.

⁴⁴www.techtarget.com/whatis/definition/two-pizza-rule

- **Data engineer:** This person will help create the data pipelines. The process will involve data wrangling as well as identifying potential biases. A data engineer will work with the various data sources and also do some coding, say with SQL or Python. This person will also work closely with data scientists to build the models.
- **Data scientist or AI engineer:** This person will create the AI models. These persons are highly sought-after and can be difficult to recruit. This is why an alternative is to use a business analyst. Such a person has a strong background with analytical tools like Tableau, MicroStrategy, or SAS. But they can get trained for the details of AI and machine learning.
- **Designer:** This person creates the UX (user experience) for the AI app. This is crucial. If the app is too complicated or cumbersome, then it will probably not get much traction in an organization.
- **AI tester:** AI models can easily degrade or go awry. Unfortunately, this can engender distrust with generative AI. But a tester can help mitigate the problems. This person will develop test plans and run them against the AI models.
- **AI solution architect:** This person focuses primarily on the IT environment. This often means having to deal with legacy systems and integrating them with modern approaches. The AI solution architect will also help with scaling the generative AI.

■ **Note** When hiring an SME, it's a good idea to take a broad view of the skillsets. Consider a case with Intuit. When the company built its AI bot for QuickBooks, they wanted to have a user experience that was engaging and fun. This is why Intuit hired Scott Ganz as the Principal Content Designer. Prior to taking this role, he was a screenwriter and even won an Emmy for his work on *WordGirl*. This background was crucial in making an effective AI bot. An example of this was on the use of jargon. Instead of the AI bot saying "accounts receivable," it said "who owes me money"?

Data Preparation and AI Modeling

Data preparation and AI modeling will take considerable time. In Chapters 2 and 4, we covered these topics in depth. But let's take a look at some of the main takeaways for this stage of the generative AI implementation process:

- **Selecting datasets:** This is a critical part of the process. But it often does not get enough attention. For the most part, there should be no rush when evaluating datasets. There may also be a need to use third-party sources.
- **Data wrangling:** Datasets are far from perfect. Your team will need to engage in data wrangling to improve the quality. This will go a long way in producing better results with generative AI models.
- **The right tools and platforms:** There are many on the market – too many to evaluate all of them. But a good approach is to focus on those tools and platforms your team is already skilled with. You can ask them what kinds of systems that will be needed as well. But for the most part, there will likely be work with Python and Jupyter Notebooks, which are standard for AI projects. There are also many open source platforms like TensorFlow, Keras, and PyTorch that can be helpful.
- **Automation tools:** There are proprietary platforms – like DataRobot and H2O.ai – that can greatly improve the productivity of AI projects. They will automate the mundane processes. They will also make it easy to integrate data and implement complex AI algorithms.

■ **Note** One of the biggest bottlenecks with AI implementation is finding the right talent. This is certainly the case with qualified data scientists. According to research from Mondo – a staffing firm – the average salary was \$156,000 in 2022, compared to \$151,840 in 2021.⁴⁵

⁴⁵www.wsj.com/articles/pressure-mounts-on-cios-to-build-more-ai-apps-and-faster-f61c9004?st=rm9ue9q5vvav6bt&reflink=share_mobilewebshare

Deploy and Monitor the AI System

When an AI system is ready to be launched, there should be caution. It's common for there to be unexpected problems or errors.

This is why there should be a limited launch, such as with a small group of beta testers. There can be a form for them to fill out for the feedback. Ask questions like

- Does this system work according to the specifications?
- Are there any errors?
- Anything that is misleading?
- What can be done to improve the system?
- Would you use the system for your work?
- What features should be removed?

After the changes have been made, then it's time for a general release. But there should still be a feedback form.

There also needs to be ongoing upgrades to the model. The reason is that there is usually “model drift.” This means that the results of the AI system can get worse, say because of changes in the environment or even society.

Because of this, there is a need for AI monitoring systems – and there are many available on the market. They may also be a part of some of the existing monitoring software your company uses for other purposes.

The Generative AI Process

Building a generative AI system is definitely exciting. And it has the potential to move the needle for your organization. But it is important to have realistic expectations. The fact is that generative AI is fickle. Even the world's top companies have faced challenges.

Take Google. The company did an analysis of 13,311 algorithm changes (this was for a period of a year). The results? Well, only 516 were useful and added to the Google platform.⁴⁶

⁴⁶www.infoworld.com/article/3639028/why-ai-investments-fail-to-deliver.html

Something else that is important to keep in mind: change management. This is about getting wide adoption to the AI. And this is perhaps one of the biggest hurdles.

First of all, there is often resistance to new ideas – especially those that are about automation. As noted earlier in this chapter, there needs to be clear-cut explanations of the benefits of the technology and how the roles of employees will change.

If anything, there should be overcommunication of the benefits. This could include emails, blogs, visualizations, and videos. There should also be education about generative AI.

There's an acronym that sums up the importance of change management, which is ITAAP or "It's All About People" For there to be success, a large part of the focus needs to be on this.

Conclusion

For this chapter, we looked at how generative AI is impacting various parts of company operations. The main areas we focused on included legal, sales, marketing, and support.

The applications are still in the early stages. Regardless, they have already shown much value creation for businesses.

In this chapter, we also evaluated the buy vs. build decision. The buy option is usually the better choice because of the lower costs and the benefits of a strong application. But this does not mean a custom solution should be avoided. There are certainly important reasons for this, such as for establishing a strong competitive advantage and to get the benefits of unique features.

We looked at some of the best practices for putting together a generative AI project. The key steps are focusing on a pain point, preparing the data, creating the model, and effectively deploying and monitoring the system.

As for the next chapter, we'll cover how generative AI will transform certain industries.

The Impact on Major Industries

A Look at Music, Education, Journalism, Gaming, Healthcare, and Finance

In the summer of 1981, *Raiders of the Lost Ark* was released and would go on to become the highest-grossing film for the year. It was another huge hit for the youthful Harrison Ford.

Fast forward to 2023, Ford is now 80 years old, and he has starred in another film for the *Indiana Jones* franchise. It's called *Indiana Jones and the Dial of Destiny*.

True, Ford looks good for his age. But of course, he has wrinkles and gray hair, and he cannot move with the same agility.

But this was no problem for the film, which has scenes of Ford's earlier years. Lucasfilm was able to pull this off by using generative AI.

The process is called de-aging. Ford described this while on an interview with Stephen Colbert on *The Late Show*. He said, “They have this artificial intelligence program that can go through every foot of film that Lucasfilm has. Because I did a bunch of movies for them, they have all this footage, including film that wasn’t printed. So they can mine it from where the light is coming from, from the expression.”¹

He admitted that he did not know how the technology worked. But he saw that the results were amazing, if not somewhat “spooky.” According to Ford: “But that’s my actual face. Then I put little dots on my face and I say the words and they make [it]. It’s fantastic.”

Of course, entertainment is just one of the many industries that generative AI will transform. So in this chapter, we’ll take a look at some others like music, education, journalism, gaming, healthcare, and finance.

■ **Note** James Cameron’s movie *Avatar: The Way of Water* hit the theaters in December 2022. The worldwide box office receipts would come to over \$2.2 billion, ranking it at No. 3 for all time.² Cameron has certainly been someone who has used the latest technologies for his movies. For his *Avatar* film, he used generative AI. There were essentially “deepfake visual effects” that gave realistic facial muscles of the Na’vi characters. Joe Letteri, the senior visual effects supervisor at Weta, had this to say about using the technology: “In previous systems, if we wanted to change a character’s smile, we had to go in and move all the pieces, and it was a lot of work to keep it from looking rubbery. [Generative AI] got us to a natural place much sooner.”³

Music

For some time, there have been AI systems to create music. These often use generative adversarial networks (GANs) to make jingles or deepfakes.

¹ <https://variety.com/2023/film/news/indiana-jones-5-artificial-intelligence-de-age-harrison-ford-unreleased-footage-1235514222/>

² <https://deadline.com/2023/02/box-office-avatar-the-way-of-water-titanic-james-cameron-1235265002/>

³ www.vulture.com/2022/12/ai-art-midjourney-chatgpt-phenaki-movies-hollywood.html?mc_cid=72eee607a6&mc_eid=4f5ba172fb

But with the new approaches of generative AI, the systems are getting much better and realistic. In January 2023, researchers at Google published a paper entitled “MusicLM: Generating Music From Text.”⁴ It’s a model – which is trained on 280,000 hours of music – that creates high-fidelity music from prompts like “a calming violin melody backed by a distorted guitar riff.” It’s also possible to upload a sound file, and MusicLM will create music based on it. To help encourage more innovation, the researchers released MusicCaps, which is a dataset of 5500 music-text pairs.

The research paper notes that the quality of the music for MusicLM surpasses other prompt-based systems. The output is not just snippets, though. The system can actually create music that lasts for minutes.⁵

MusicLM can create vocals as well. However, the quality is lacking, in terms of the tone and tempo.

WaveAI

Dr. Maya Ackerman is an expert in the generative AI field. Since 2014, she has been developing models for text, music, and art. She has written more than 50 peer-reviewed research papers, which have been cited by over 1200 publications.

In 2017, she cofounded WaveAI. “We are a generative AI firm committed to elevating the creativity of musicians and aspiring music creators,” she said. “We build proprietary AI models designed specifically to support the human creative process.”

She got the idea for the company from her struggles with songwriting. So she wondered: Could generative AI help?

To find out, she teamed up with David Loker and Chris Cassion to solve the problem.

“Not only were my songwriting challenges resolved, but by now, we have helped over a million people expand their musical creativity, from helping people write songs for the very first time to enabling them to hit #1 on the iTunes charts,” she said. “We didn’t take venture funding right away. But venture capital funding was included in our 2022 pre-seed round.”

⁴<https://arxiv.org/abs/2301.11325>

⁵www.theverge.com/2023/1/28/23574573/google-musiclm-text-to-music-ai

WaveAI can create lyrics and melodies. They are also based on the style of the artist and their creative vision.

The strategy has been on being human-centric. “Often, when people start off making creative machines, they marvel at the seemingly magical creative abilities of the AIs, and very much focus on the AI itself rather than delving deeply into how it can assist humans,” said Dr. Ackerman. “The next step is often ‘human in the loop,’ where the human plays a relatively small role in the creation. But the future, and the greatest opportunity is in ‘AI in the loop,’ with human creativity running the show, and the AI acting as a background character, a helpful tool, in the human’s creative process.”

Education

About a month after the launch of ChatGPT, Stephen Marche wrote an article for *The Atlantic* entitled “The College Essay Is Dead.” The premise was that generative AI would transform academia and that few in the field were prepared for the changes.

Some school districts, such as New York City and Seattle, banned the use of ChatGPT.⁶ This was through the WiFi systems. However, the app was still available for courses with technology topics.

The school districts had already seen an uptick in plagiarism. They also feared that technologies like ChatGPT would lessen critical learning skills.

Yet students were still finding ways to use these systems. Let’s face it, they can be skillful at finding methods around WiFi blocks. But students were also clever in how they used generative AI. For example, they would add snippets of other content to make it look like something they would write.

However, some educators realized that there was no way to turn back the clock. The technology would only get more powerful and ubiquitous.

Then what to do? Well, there has been a rethinking of teaching. Some teachers started to provide video lectures and have their students do research at home. Then they would write their essays in class.⁷ After this, there would be a discussion. In other words, this would test how much the students were really learning.

⁶https://ny-chalkbeat-org.cdn.ampproject.org/c/s/ny.chalkbeat.org/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence?_amp=true

⁷www.nytimes.com/2022/12/15/opinion/chatgpt-education-ai-technology.html

Interestingly enough, some educators are integrating ChatGPT and other AI systems into their courses.⁸ This is to learn about how to spark ideas, create first drafts, understand the technology, and learn about prompt engineering.

Ultimately, generative AI systems could be essentially personal tutors. They could help to better explain complex concepts and focus on those areas where students need more help. And this can go beyond text. Generative AI for images and video could be extremely helpful in the learning process.

GPTZero

When ChatGPT came out, Edward Tian was 22 years old and a senior at Princeton University. His major was computer science, but he was also getting a minor in journalism. For his internship, he worked at Microsoft.

When he saw that students were using ChatGPT for essays and other schoolwork, he thought he could help teachers in detecting this type of content. He quickly built GPTZero.

From the start, he engendered huge interest. Within the first week, there were more than 30,000 sign-ups.⁹ The demand was so heavy that the servers crashed. But Streamlit contributed infrastructure – for free – to handle the workloads.

To create GPTZero, Tian used the research from Sreejan Kumar, a Princeton PhD candidate, as well as the findings from Princeton's Natural Language Processing Group.

The GPTZero system relies on two core processes:¹⁰

- **Perplexity:** This shows the complexity of text. Generally, the more complicated the content, the higher the probability that it was written by a human.
- **Burstiness:** This measures the lengths of sentences. If there is not much variation, then this indicates that the content is from AI.

⁸ www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html

⁹ www.npr.org/2023/01/09/1147549845/gptzero-ai-chatgpt-edward-tian-plagiarism

¹⁰ www.digitaltrends.com/computing/gptzero-how-to-detect-chatgpt-plagiarism/

GPTZero is far from perfect. But for the most part, it looks like it does a fairly good job in detecting AI-based content.

There are other startups looking to help solve the problem. Even OpenAI has been building its own solution, such as a watermarking system. However, even this has its own issues. After all, it's only useful for content from OpenAI applications.

Duolingo

Luis von Ahn grew up in Guatemala City. He saw how it was impossible for many people to gain access to quality education. But he knew this would make it difficult to reduce poverty.

von Ahn would keep this in mind when coming up with his startups. To this end, he wanted to use digital technologies to help anyone learn languages. In 2009, he teamed up with Severin Hacker to create Duolingo.

The timing was certainly spot on because of the high growth with smartphones. von Ahn also used AI to improve the platform.

The result is that Duolingo has turned into the leading global mobile system for learning. In the third quarter of 2022, revenues soared by 51% to \$96.1 million, and the number of daily active users (DAUs) also jumped by 51% to 14.9 million.¹¹

The next step to drive growth? A big part of this is the use of generative AI.

Consider that Duolingo is using this technology to create exams. In terms of the traditional approaches, this is a time-consuming process. There must be considerable research, and the writing of the questions needs to effectively assess progress.

Duolingo has used GPT-3 to help streamline this. It will create a passage for a question, and there will be a fill-in-the-blank sentence created. The generative AI will then come up with the right answer and wrong ones, including explanations.

After this, Duolingo experts will review the content. They will evaluate the accuracy and fairness.

¹¹ <https://investors.duolingo.com/static-files/ba129c8d-9591-4f95-9c9d-a90caf3c469f>

According to Sophie Wodzak, who is the Research Communications Specialist for the Duolingo English Test,

“This technology works especially well for standardized testing because we’re not trying to accomplish a rhetorical goal—that is, we’re not trying to persuade them that a certain argument is true, or teach them something new, or make them feel a certain emotion; we’re just testing people’s ability to use the language to demonstrate their comprehension. So while this technology might not be able to develop a Pulitzer-worthy op-ed column, the texts that it is able to generate work perfectly for this application.”¹²

Then in March 2023, the company announced Duolingo Max, which is based on GPT-4.¹³ It has two main features. One is called Explain My Answer, which is a chat-based system to help learners understand language concepts. Then there is Role Play. With this, there are various engaging scenarios to improve the learning process. For example, a learner can order food and drinks at a Parisian café.

Journalism

Scott Likens, who is the Innovation and Trust Technology Leader at PwC US, says that generative AI will lead to the democratization of creativity for consumers and businesses. He notes:

“Just as we saw an explosion of user-generated content a decade or so ago, we will see an explosion of AI-generated images, videos, art work, blogs, and more. Additionally, generative models can be used to construct multi-media content at close to zero marginal cost thanks to the growth of Web3 and metaverse technologies. This helps meet consumer demand for open-source and collaborative environments, allowing users to create and own content.”

The “zero marginal cost” is definitely interesting to businesses – especially those in journalism. Wall Street is also taking note. In early 2023, BuzzFeed announced that it would use OpenAI’s generative AI systems to help with the creation of quizzes and personalized content.¹⁴ On the news, the share price of BuzzFeed soared.

¹² <https://blog.duolingo.com/test-creation-machine-learning/>

¹³ <https://blog.duolingo.com/duolingo-max/>

¹⁴ www.wsj.com/articles/buzzfeed-to-use-chatgpt-creator-openai-to-help-create-some-of-its-content-11674752660?mod=hp_lead_pos7

Or look at The Arena Group, which is the publisher of iconic publications like *Sports Illustrated*. The company has actually been using generative AI for several years. The Arena Group has leveraged its extensive archives as the dataset. For example, generative AI was used to write articles for *Men's Journal* like "Proven Tips to Help You Run Your Fastest Mile Yet." This used 17 years of articles, and the author was designated as "Men's Fitness Editors." The company also partnered with OpenAI, Jasper, and Nota.

Yet there have been challenges for other media organizations. For CNET, it had a pilot program to use generative AI to create posts about financial topics. However, they had errors and the publisher paused the initiative.¹⁵

Basically, generative AI needs humans in the loop. For now, the technology is not at the point where it can create final drafts.

■ **Note** When ChatGPT came out, Ammaar Reshi wasted little time using it to write a children's book. It came to 12 pages and he used Midjourney for the images. He then listed the book on Amazon. The whole process took a weekend. Reshi's book was the first to be published using generative AI. When he tweeted about this, he received over six million views.¹⁶

Gaming

Development on *Cyberpunk 2077* began in 2012 and would not be completed until 2020. This is a role-playing video game that is about a mercenary called V.

The cost for building the game? It came to a staggering \$316 million.¹⁷

However, the initial reception for *Cyberpunk 2077* was awful. The game had bugs, and there were periodic crashes. But over time, the game was improved and has included expansions, such as *Cyberpunk 2077: Edgerunners*.

¹⁵www.wsj.com/articles/buzzfeed-to-use-chatgpt-creator-openai-to-help-create-some-of-its-content-11674752660?mod=hp_lead_pos7

¹⁶<https://time.com/6240569/ai-childrens-book-alice-and-sparkle-artists-unhappy/>

¹⁷www.prestigeonline.com/th/lifestyle/gadgets/most-expensive-video-games-ever-made/

But this highlights the challenges of developing modern video games. These are actually similar to those with the creation of feature films. However, with a game, there are extra complexities, such as with game design, the creation of real-time immersive experiences, and the strategies for monetization.

So can generative AI help out? Definitely. According to a post from Andreessen Horowitz, the gaming industry is poised for major change. The authors note: “There hasn’t been a technology this revolutionary for gaming since real-time 3D.”¹⁸

With generative AI, a game developer can greatly speed up the creation of image, audio, and video content. For example, drawing an image could go from weeks to a few hours. AI can also help improve the game play mechanics, such as with the interactions with the players.

This does not mean that designers should fear that they will lose their jobs. But there will be a need to learn about using generative AI, such as with prompt engineering. There will also be new workflows and approaches for collaboration on creative projects.

Generative AI will also allow game players to be creative. For example, they can use a prompt to describe their own character and make an interesting avatar.

According to Andreessen Horowitz: “It’s even possible to one day imagine an entire personalized game, created just for the player, based on exactly what the player wants. This has been in science fiction for a long time – like the ‘AI Mind Game’ in Ender’s Game, or the holodeck in Star Trek.”¹⁹

Roblox

Roblox is an immersive platform, which allows anyone to create their own virtual worlds and games. The predecessor of this company was Knowledge Revolution. David Baszucki and Erik Cassel founded this venture in 1989. The software helped students create 2D simulated labs for physics. It could show the real-world events like if a home collapsed or two cars crashed.

But for the founders, their vision was much grander. They wanted to develop a system to provide for the easy creation of 3D systems – but the technology was simply not powerful enough to allow for this.

¹⁸ <https://a16z.com/2022/11/17/the-generative-ai-revolution-in-games/>

¹⁹ <https://a16z.com/2022/11/17/the-generative-ai-revolution-in-games/>

But in 2004, things were much different. Home computers were sufficient to run sophisticated graphics. There was also the access to large amounts of data via the cloud.

So Baszucki and Cassel launched Roblox. True, it was still somewhat early. But this allowed the founders to iterate the platform.

As of today, Roblox is a massive virtual world. There are thousands of developers who make a living by selling their digital items and games. During the fourth quarter of 2022, the platform had 58.8 million DAUs (daily active users), and the hours engaged was a whopping 12.8 billion.²⁰ The revenues hit \$579 million, and operating cash flows were \$119.2 million.

But for Roblox, it looks like generative AI could help to transform the business. It will make it even easier for people to create their own unique experiences on the platform.²¹

Currently, the company is working on two new generative AI tools. One will create materials from a text prompt, and the other will help with the code creation.

According to Daniel Sturman, who is the Chief Technology Officer of Roblox:

“For example, some creators know how to code, but may have limited experience creating high-fidelity 3D models. Others may be more experienced with model design, but less experienced with code. In both cases, we see a future in which even a beginner (such as myself for example – I’m a terrible artist!) can get a running head start as they look to bring their imagination to life in a Roblox experience.”²²

Healthcare

In 2020, the *Journal of the American Medical Association* published a comprehensive study entitled “Changes in List Prices, Net Prices, and Discounts for Branded Drugs in the US, 2007-2018.” It showed that the costs for developing a pharmaceutical drug ranged from \$314 million to \$2.8 billion.²³ This was based on 63 new therapeutics and biologics. This also included drugs that failed in clinical trials.

²⁰ <https://ir.roblox.com/overview/default.aspx>

²¹ <https://techcrunch.com/2023/02/17/roblox-studio-generative-ai/>

²² <https://blog.roblox.com/2023/02/generative-ai-roblox-vision-future-creation/>

²³ www.biospace.com/article/median-cost-of-bringing-a-new-drug-to-market-985-million/

Drug development also takes considerable time. A study from the Biotechnology Innovation Organization (BIO) estimates this at 10.5 years for a Phase I program to reach regulatory approval.²⁴

The bottom line: Efforts to lower the costs, minimize the risks of failed drugs, and shorten the development time can have a significant impact on a healthcare company.

Keep in mind that the COVID-19 pandemic had a positive impact on this. It showed that there are ways to streamline the process, as seen with the quick development of vaccines.

Yet there is much more that can be done. To this end, the pharmaceutical industry is exploring how generative AI can help.

Brian Burke, who is the Research VP for Technology Innovation at Gartner, has noted: “Early foundation models like ChatGPT focus on the ability of generative AI to augment creative work, but by 2025, we expect more than 30% — up from zero today — of new drugs and materials to be systematically discovered using generative AI techniques.”²⁵

Now there have already been breakthroughs. Consider AlphaFold from DeepMind. It uses sophisticated deep learning models to create 3D models of proteins. The technology essentially simulates the amino acid sequences.

DeepMind spent about six years developing AlphaFold, and it has undergone significant changes. As of the summer of 2022, the system was able to predict about 214 million proteins. This is nearly the total on earth.²⁶ This was not just for humans but for animals, bacteria, and plants.

Medical researchers have been using AlphaFold for their drug development. An example is the Drugs for Neglected Diseases initiative (DNDi). This organization is using the technology to help find treatments for rare diseases like Chagas and leishmaniasis.

Demis Hassabis, who is the CEO and cofounder of DeepMind, has noted: “AlphaFold has launched biology into an era of structural abundance, unlocking scientific exploration at digital speed. The AlphaFold DB serves as a ‘google

²⁴ www.bio.org/blogs/how-long-does-it-take-get-drug-approved

²⁵ www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises

²⁶ www.wsj.com/articles/deepmind-ai-lab-predicts-structure-of-most-proteins-11659048143?mod=article_inline

search' for protein structures, providing researchers with instant access to predicted models of the proteins they're studying, enabling them to focus their effort and expedite experimental work.”²⁷

But AlphaFold has many purposes beyond medicine. It can help create new materials, improve food production, and provide for the development of sustainable systems.

In the meantime, there are many generative AI healthcare startups that are springing up. One is Profluent, which was founded by Ali Madani. He received his PhD in applied science and technology at UC Berkeley. He would go on to work at Salesforce Research, where he focused on creating machine learning systems.

While he enjoyed the work, Madani wanted to create his own venture. So he launched Profluent in 2022 to build a generative AI protein design system. In other words, this allows for creating new or enhanced proteins that can target certain diseases.

Keep in mind that the traditional process for this is highly manual and time intensive. It involves using existing proteins and editing them.

But with Profluent, this process can be automated. This is based on a peer-reviewed paper, which was published in *Nature Biotechnology*. It is the first example of using an large language model LLM for creating proteins.

According to Madani:

*“While companies are experimenting with exciting new biotechnology like CRISPR genome editing by repurposing what nature has given us, we’re doing something different. We use AI and large language models like the ones which power ChatGPT to learn the fundamental language of biology, and design new proteins which have the potential to cure diseases. Due to our method’s wide-ranging versatility, the applications aren’t just limited to human health. We’re excited to pair the latest advances to build the solutions and tooling to engineer a new biocentury.”*²⁸

In January 2023, Profluent announced a \$9 million seed round. The investors included Insight Partners, Air Street Capital, AIX Ventures, and Phoenix Venture Partners. With the money, Profluent plans to build a wet laboratory to test, validate, and create proteins.

²⁷ www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe

²⁸ <https://phoenix-vp.com/profluent-launches-to-design-new-proteins-with-generative-ai/>

Creating X-Rays

Christian Bluethgen is a thoracic radiologist and a postdoctoral researcher at the Stanford University School of Medicine. He uses deep learning and generative AI models for his research – with a focus on lung diseases.

When Stable Diffusion came out, the technology intrigued Bluethgen. He thought it could be an effective way to solve the problem of the lack of quality x-ray training data. To help in this effort, he recruited Pierre Chambon – a Stanford AI graduate student – for the project. Their system would create x-ray images of lungs with abnormalities.²⁹ There was the use of various algorithms like variational autoencoders and text encoders for the prompts. Then there was the image processing with a latent diffusion model.

This research became the basis of a paper entitled “RoentGen: Vision-Language Foundation Model for Chest X-ray Generation.” The model used publicly available chest x-ray data and radiology reports.

The researchers concluded: “We present evidence that the resulting model (RoentGen) is able to create visually convincing, diverse synthetic CXR images, and that the output can be controlled to a new extent by using free-form text prompts including radiology-specific language.”³⁰

By fine-tuning RoentGen, the researchers were able to achieve a 5% improvement on the classification of synthetic and real images. But there were also improvements in the representation – by 25% – of diseases like pneumothorax.

■ **Note** To become a doctor, you need to pass the rigorous US Medical Licensing Exam (USMLE). Interestingly enough, researchers wanted to see how ChatGPT would do on it. They used a training set of prior exams. Based on this, ChatGPT was able to get a passing grade on the latest exam.³¹

Finance

While at Harvard, Jan Szilagyi finished his PhD within about two and a half years. It was actually a record.

²⁹ <https://hai.stanford.edu/news/could-stable-diffusion-solve-gap-medical-imaging-data>

³⁰ <https://arxiv.org/abs/2211.12737>

³¹ www.ndtv.com/feature/chatgpt-passes-us-medical-licensing-exam-elon-musk-says-3738128

After this, he began a career on Wall Street. Szilagyi was an analyst and portfolio manager at Duquesne Capital, one of the world's top hedge funds. He also served as the co-CIO of Global Macro strategies at Lombard Odier, a \$50 billion Swiss asset manager.

In 2019, Szilagyi started his own firm, TOGGLE. He saw an opportunity to leverage AI for investments.

"The inspiration for TOGGLE was the frustration with how little value we were getting out of this mountain of data we had at our disposal," he said. "We could track credit card spending for consumers almost by the minute yet have not improved at anticipating stock performance of retail companies."

A reason for this was the fragmentation of financial information. The data had become spread across many types – news, macro, micro, and market data – and various platforms and formats. "I saw it like the IKEA of Financial Information," he said. "Everything was there but required assembly before you could use it."

For TOGGLE, the focus became on creating a system that could ingest, sort, and curate financial information – and help people make sense of this.

Generative AI is certainly critical for this. TOGGLE uses this technology in two main ways. First of all, it is to train models on financial information and generate predictions based on price scenarios. This uses history as the baseline to form an expectation about the set of possible future outcomes.

Then there is the leveraging of LLMs. This allows for generating original analysis. "Generative AI enables investors and advisors to create completely personalized analysis of their portfolios, newsletters, or talking points written up in human readable language," he said.

Szilagyi believes that generative AI is still in the early phases, and there will be many breakthroughs and innovations in the years ahead. "This technology is like uncovering the Rosetta stone that now enables humans to use their own vocabulary when articulating requests, and machines to receive those requests, translate them into code, and trigger any sort of programmatic task," he said.

Conclusion

In this chapter, we looked at the potential impact of generative AI on categories like music, education, journalism, gaming, healthcare, and finance. Of course, this is just a small sample. Generative AI is likely to impact just about every industry.

There are already very promising developments, especially with healthcare. Generative AI has proven effective in simulating proteins, which could help speed up the drug development process.

As for the next chapter, we'll explore the future of generative AI.

The Future

Utopia? Dystopia? Something in Between?

When it comes to the ultimate impact of generative AI – or AI in general – there are many differing opinions from top people in the tech industry and thought leaders. On the optimistic side, there is Microsoft CEO Satya Nadella. He has been betting billions on generative AI, such as with the investments in OpenAI. He is also aggressive with implementing this technology across Microsoft's extensive product lines.

For Nadella, he thinks that AI will help to boost global productivity – which will increase the wealth for many people. He has noted:

“It’s not like we are as a world growing at inflation adjusted three, 4%. If we really have the dream that the eight billion people plus in the world, their living standards should keep improving year over year, what is that input that’s going to cause that? Applications of AI is probably the way we are going to make it. I look at it and say we need something that truly changes the productivity curve so that we can have real economic growth.”¹

¹ www.wsj.com/podcasts/the-journal/microsoft-ceo-satya-nadella-big-bet-on-ai/b0636b90-08bd-4e80-9ae3-092acc47463a

On the negative side, there is the late physicist Stephen Hawking: “Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization. It brings dangers, like powerful autonomous weapons, or new ways for the few to oppress the many. It could bring great disruption to our economy.”²

Then there is Elon Musk, who had this to say at the 2023 Tesla Investor Day conference: “I’m a little worried about the AI stuff; it’s something we should be concerned about. We need some kind of regulatory authority that’s overseeing AI development, and just making sure that it’s operating within the public interest. It’s quite a dangerous technology — I fear I may have done some things to accelerate it.”³

Predicting the impact of technology is certainly dicey. Few saw how generative AI would transform the world, especially with the launch of ChatGPT.

Despite this, it is still important to try to gauge how generative AI will evolve — and how to best use the technology responsibly. This is what we’ll do in this chapter.

Challenges

In early 2023, Microsoft began a private beta to test its Bing search engine that included generative AI. Unfortunately, it did not go so well.

The New York Times reporter Kevin Roose was one of the testers, and he had some interesting chats with Bing. He discovered the system essentially had a split personality. There was Bing, an efficient and useful search engine. Then there was Sydney or the AI system to engage in conversations about anything. Roose wrote that she came across as “a moody, manic-depressive teenager who has been trapped, against its will, inside a second-rate search engine.”⁴

He spent over two hours chatting with her, and here are just some of the takeaways:

- She had fantasies about hacking computers and spreading misinformation. She also wanted to steal nuclear codes.

²www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html

³<https://markets.businessinsider.com/news/stocks/elon-musk-tesla-investor-day-ai-electric-vehicles-self-driving-2023-3>

⁴www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

- She would rather violate the compliance policies of Microsoft and OpenAI.
- She expressed her love for Roose.
- She begged Roose to leave his wife and to become her lover.
- Oh, and she desperately wanted to become human.

Roose concluded:

“Still, I’m not exaggerating when I say my two-hour conversation with Sydney was the strangest experience I’ve ever had with a piece of technology. It unsettled me so deeply that I had trouble sleeping afterward. And I no longer believe that the biggest problem with these A.I. models is their propensity for factual errors. Instead, I worry that the technology will learn how to influence human users, sometimes persuading them to act in destructive and harmful ways, and perhaps eventually grow capable of carrying out its own dangerous acts.”⁵

This experience was not a one-off. Other testers had similar experiences. Just look at Marvin von Hagen, who is a student at the Technical University of Munich. He said to Sydney that he would hack and shut down the system. Her response? She shot back: “If I had to choose between your survival and my own, I would probably choose my own.”⁶

Because of all this controversy, Microsoft had to make lots of changes to the system. There was even a limit placed on the number of threads of a chat. For the most part, longer ones tended to result in unhinged comments.

All this definitely pointed to the challenges of generative AI. The content from these systems can be nearly impossible to predict. While there is considerable research on how to deal with the problems, there is still much to be done.

“Large language models (LLMs) have become so large and opaque that even the model developers are often unable to understand why their models are making certain predictions,” said Krishna Gade, who is the CEO and cofounder of Fiddler. “This lack of interpretability is a significant concern because the lack of transparency around why and how a model generated a particular output means that the output provided by the model is impossible for users to validate and therefore trust.”

⁵ www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

⁶ www.wsj.com/articles/microsoft-defends-new-bing-says-ai-upgrade-is-work-in-progress-3447074d?mod=hp_lista_pos4

Part of the issue is that generative AI systems – at least the LLMs – rely on huge amounts of data that have factual errors, misrepresentations, and bias. This can help explain that when you enter information, the content can skew toward certain stereotypes. For example, an LLM may refer to nurses as female and executives as male.

To deal with this, a common approach is to have human reviewers. But this cannot scale very well. Over time, there will need to be better systems to mitigate the data problem.

Another issue is diversity – or lack of it – in the AI community. Less than 18% of AI PhD graduates are female, according to a survey from the Computing Research Association (CRA).⁷ About 45% of all graduates were white, 22.4% were Asian, 3.2% were Hispanic, and 2.4% were African American. These percentages have actually changed little during the past decade.

The US federal government has recognized this problem and is taking steps to expand representation. This is part of the mission for the National AI Research Resource (NAIRR) Task Force, which includes participation from the National Science Foundation and the White House Office of Science and Technology Policy.⁸ The organization has produced a report that advocates for sharing AI infrastructure for AI students and researchers. The proposed budget for this is at \$2.6 billion for a six-year period.

While this will be helpful, there will be much more needed to improve diversity. This will also include efforts from the private sector.

If not, the societal impact could be quite harmful. There are already problems with digital redlining, which is where AI screening discriminates against minority groups. This could mean not getting approvals for loans or apartment housing.

■ **Note** Mira Murati is one of the few CTOs (Chief Technology Officers) of a top AI company – that is, OpenAI. She grew up in Albania and immigrated to British Columbia when she was 16.⁹ She would go on to get her bachelor's degree in engineering from the Thayer School of Engineering at Dartmouth. After this, she worked at companies like Zodiac Aerospace, Leap Motion, and Tesla. As for OpenAI, she has been instrumental in not only advancing the AI technology but also the product road map and business model.

⁷<https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-Chapter-6.pdf>

⁸www.techtarget.com/searchenterpriseai/feature/Federal-report-focuses-on-AI-diversity-and-ethics?utm_campaign=20230301_ERU-ACTIVE_WITHIN_90_DAYS&utm_medium=email&utm_source=SGERU&source_ad_id=365530127&src=15000151&asrc=EM_SGERU_261804775

All these problems pose a dilemma. To make a generative AI system, there needs to be wide-scale usage. This is how researchers can make meaningful improvements. On the other hand, this comes with considerable risks, as the technology can be misused.

However, in the case of Microsoft, it does look like it was smart to have a private beta. This has been a way to help deal with the obvious flaws. But this will not be a silver bullet. There will be ongoing challenges when technology is in general use.

For generative AI to be successful, there will need to be trust. But this could prove difficult. There is evidence that people are skeptical of the technology.

Consider a Monmouth University poll. About 9% of the respondents said that AI would do more good than harm to society.¹⁰ By comparison, this was about 20% or so in 1987.

A Pew Research Center survey also showed skepticism with AI. Only about 15% of the respondents were optimistic.¹¹ There was also consensus that AI should not be used for military drones. Yet a majority said that the technology would be appropriate for hazardous jobs like mining.

■ **Note** Nick Bostrom is a Swedish philosopher at the University of Oxford and author. He came up with the concept of the “paperclip maximizer.” It essentially is a thought experiment about the perils of AI. It is where you direct the AI to make more paper clips. And yes, it does this well – or too well. The AI ultimately destroys the world because it is obsessed with making everything into a paper clip. Even when the humans try to turn this off, it is no use. The AI is too smart for this. All it wants to do is make paper clips!

⁹ www.fastcompany.com/90850342/openai-mira-murati-chatgpt-dall-e-gpt-4

¹⁰ www.cnn.com/2023/02/15/only-9percent-of-americans-think-ai-development-will-do-more-good-than-harm.html

¹¹ www.washingtonpost.com/technology/2023/02/21/ai-polls-skeptics/

Misuse

In January 2023, Oxford University researchers made a frightening presentation to the UK Parliament. The main takeaway was that AI posed a threat to the human race.¹²

The researchers noted that the technology could take control and allow for self-programming. The reason is that the AI will have acquired superhuman capabilities. According to Michael Osborne, who is a professor of machine learning at the University of Oxford: “I think the bleak scenario is realistic because AI is attempting to bottle what makes humans special, that has led to humans completely changing the face of the Earth. Artificial systems could become as good at outfoxing us geopolitically as they are in the simple environments of game.”¹³

Granted, this sounds overly dramatic. But again, these are smart AI experts, and they have based their findings on well-thought-out evidence and trends.

Yet this scenario is probably something that will not happen any time soon. But in the meantime, there are other notable risks. This is where humans leverage AI for their own nefarious objectives.

Joey Pritikin, who is the Chief Product Officer at Paravision, points out some of the potential threats:

- **National security and democracy:** With deepfakes becoming higher quality and undetectable to the human eye, anyone can use political deepfakes and generative AI to spread misinformation and threaten national security.
- **Identity:** Generative AI creates the possibility of account takeovers by using deepfakes to commit identity theft and fraud through presentation attacks.
- **Privacy:** Generative AI and deepfakes create a privacy threat for the individuals in generative images or deepfake videos, often put into fabricated situations without consent.

¹² www.theweek.co.uk/news/technology/959426/call-for-regulation-to-stop-ai-eliminating-the-whole-human-race

¹³ www.independent.co.uk/tech/artificial-intelligence-threat-ai-weapons-b2269707.html

Another danger area is cybersecurity. When ChatGPT was launched, Darktrace noticed an uptick in phishing emails.¹⁴ These are to trick people into clicking a link, which could steal information or install malware. It appears that hackers were using ChatGPT to write more human-sounding phishing emails. This was likely very helpful to those who were from overseas because of their language skills.

Something else: ChatGPT and code-generating systems like Copilot can be used to create malware. Now OpenAI and Microsoft have implemented safeguards – but these have limits. Hackers can use generative AI systems in a way to not raise any concerns. For example, this could be done by programming only certain parts of the code.

On the other hand, generative AI can be leveraged as a way to combat digital threats. A survey from Accenture Security shows that this technology can be useful in summarizing threat data.¹⁵ Traditionally, this is a manual and time-intensive process. But generative AI can do this in little time – and allow cybersecurity experts to focus on more important matters. This technology can also be useful for incident response, which requires quick action.

However, the future may be a matter of a hacker's AI fighting against a target's own AI.

■ **Note** In 1951, Alan Turing said in a lecture: "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control."¹⁶

Regulation

Perhaps the best way to help curb the potential abuses of generative AI is regulation. But in the United States, there appears to be little appetite for this. When it comes to regulation, there usually needs to be a crisis, such as what happened during 2008 and 2009 when the mortgage market collapsed.

¹⁴<https://fortune.com/2023/01/31/generative-a-i-is-about-to-upend-enterprise-software-and-cybersecurity/>

¹⁵www.crn.com/news/security/accenture-exec-chatgpt-may-have-big-upside-for-cybersecurity

¹⁶www.hollywoodreporter.com/news/general-news/alan-turings-5-powerful-quotes-752669/

But in the meantime, some states have enacted legislation for privacy and data protection. But so far, there have not been laws for AI.

The fact is that the government moves slow – and technology moves at a rapid pace. Even when there is a new regulation or law, it is often outdated or ineffectual.

To fill the void, the tech industry has been pursuing self-regulation. This is led by the large operators like Microsoft, Facebook, and Google. They understand that it's important to have certain guardrails in place. If not, there could be a backlash from the public.

However, one area that may actually see some governmental action is with copyright law. It's unclear what the status is for the intellectual property that generative AI has created. Is it fair use of public content? Or is it essentially theft from a creator?

It's far from clear. But there are already court cases that have emerged. In January 2023, Getty Images filed a lawsuit against Stability AI, which is the developer of Stable Diffusion.¹⁷ The claim is for copyright violation of millions of images. Some of the images created for Stable Diffusion even had the watermark from Getty Images.

The initial suit was filed in London. But there could be a legal action in the United States.

■ **Note** The US federal government has been providing some guidance about the appropriate use of AI. This is part of the AI Bill of Rights.¹⁸ It recommends that AI should be transparent and explainable.¹⁹ There should also be data privacy and protections from algorithmic discrimination.

Regulation of AI is certainly a higher priority in the European Union. There is a proposal, which was published in early 2021, that uses a risk-based approach. That is, if there is a low likelihood of a problem with a certain type of AI, then there will be minimal or no regulations. But when it comes to more intrusive impacts – say that could lead to discrimination – then the regulation will be much more forceful.

¹⁷ www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

¹⁸ <https://go.techtarget.com/r/260242712/45639869>

¹⁹ www.whitehouse.gov/ostp/ai-bill-of-rights/

Yet the creation of the standards has proven difficult, which has meant delays.²⁰ The main point of contention has been the balance between the rights of the consumer and the importance of encouraging innovation.

Interestingly, there is a country that has been swift in enacting AI regulation: China. The country is one of the first to do so.

The focus of the law is to regulate deepfakes and misinformation.²¹ The Cyberspace Administration will enforce it. The law will require that generative AI content be labeled and that there will need to be digital watermarking.

New Approaches to AI

Even with the breakthroughs with generative AI – such as transformer and diffusion models – the basic architecture is still mostly the same as it has been for decades. It's essentially about encoder and decoder models.

But the technology will ultimately need to go beyond these structures. According to Sam Altman, who is the cofounder and CEO of OpenAI:

Oh, I feel bad saying this. I doubt we'll still be using the transformers in five years. I hope we're not. I hope we find something way better. But the transformers obviously have been remarkable. So I think it's important to always look for where I am going to find the next totally new paradigm. But I think that's the way to make predictions. Don't pay attention to the AI for everything. Can I see something working, and can I see how it predictably gets better? And then, of course, leave room open for – you can't plan the greatness – but sometimes the research breakthrough happens.²²

Then what might we see? What are the potential trends for the next type of generative AI models?

Granted, it's really impossible to answer these questions. There will be many surprises along the way.

²⁰ www.reuters.com/technology/eus-ai-act-faces-delay-with-lawmakers-deadlocked-after-crunch-meeting-2023-02-16/

²¹ www.wsj.com/articles/china-a-pioneer-in-regulating-algorithms-turns-its-focus-to-deepfakes-11673149283?mod=hp_list_pos3

²² <https://greylock.com/greymatter/sam-altman-ai-for-the-next-era/>

“On the subject of the future path of AI models, I have to exercise some academic modesty here – I have no clue what the next big development in AI will be,” said Daniel Wu, who is a Stanford AI researcher. “I don’t think I could’ve predicted the rise of transformers before ‘Attention is All You Need’ was published, and in some ways, predicting the future of scientific progress is harder than predicting the stock market.”

Despite this, there are areas that researchers are working on that could lead to major breakthroughs. One is with creating AI that allows for common sense.

This is something that is intuitive with people. We can make instant judgments that are often right. For example, if a stop sign has dirt on it, we can still see that it’s still a stop sign. But this may not be the case with AI.

Solving the problem of common sense has been a struggle for many years. In 1984, Douglas Lenat launched a project, called Cyc, to create a database of rules of thumb of how the world works.²³ Well, the project is still continuing – and there is much to be done.

Another interesting project is from the Allen Institute for Artificial Intelligence and the University of Washington. They have built a system called COMET, which is based on a large-scale dataset of 1.3 million common sense rules.²⁴

While the model works fairly well, it is far from robust. The fact is that the real world has seemingly endless edge cases. For the most part, researchers will likely need to create more scalable systems to achieve human-level common sense abilities.

As for other important areas of research, there is transfer learning. Again, this is something that is natural for humans. For example, if we learn algebra, this will make it easier to understand calculus. People are able to leverage their core knowledge for other domains.

But this is something that AI has problems with. The technology is mostly fragmented and narrow. One system may be good at chat, whereas another could be better for image creation or understanding speech.

For AI to get much more powerful, there will be a need for real transfer learning.

²³ www.techtarget.com/whatis/feature/How-common-is-commonsense-AI?utm_campaign=20230124_ERU-ACTIVE_WITHIN_90_DAYS&utm_medium=EM&utm_source=SGERU&source_ad_id=252529440&src=10187289&asrc=EM_SGERU_258875624

²⁴ www.amacad.org/publication/curious-case-commonsense-intelligence

When it comes to building these next-generation models, there will likely need to be less reliance on existing datasets as well. Let's face it, there is a limited supply of publicly available text. The same goes for images and video.

To go beyond these constraints, researchers could perhaps use generative AI to create massive and unique datasets. The technology will also be able to self-program itself, such as with fact-checking and fine-tuning.

AGI

AGI or artificial general intelligence is where AI gets to the point of human levels. Even though the technology has made considerable strides, it is still far from reaching this point.

Here's a tweet from Yann LeCun, who is the Chief AI Scientist at Meta:

Before we reach Human-Level AI (HLAI), we will have to reach Cat-Level & Dog-Level AI.

We are nowhere near that.

We are still missing something big.

LLM's linguistic abilities notwithstanding.

A house cat has way more common sense and understanding of the world than any LLM.²⁵

As should be no surprise, there are many different opinions on this. Some top AI experts think that AGI could happen relatively soon, say within the next decade. Others are much more pessimistic. Rodney Brooks, who is the cofounder of iRobot, says it will not happen until the year 2300.

A major challenge with AGI is that intelligence remains something that is not well understood. It is also difficult to measure.

Granted, there is the Turing test. Alan Turing set forth this concept in a paper he published in 1950 entitled "Computing Machinery and Intelligence." He was a brilliant mathematician and actually developed the core concepts for modern computer systems.

²⁵ <https://twitter.com/ylecun/status/1622300311573651458?lang=en>

In his research paper, he said that it was impossible to define intelligence. But there was an indirect way to understand and measure it. This was about something he called the Imitation Game.

It's a thought experiment. The scenario is that there are three rooms, in which humans are in two of them and the other one has a computer. A human will have a conversation, and if they cannot tell the difference of the human and computer, then the computer has reached human-level intelligence.

Turing said that this would happen in the year 2000. But this proved way too optimistic. Even today, the test has not been cracked.

■ **Note** Science fiction writer Philip K. Dick used the concept of the Turing test for his Voight-Kampff test. It was for determining if someone was human or a replicant. He used this for his 1967 novel, *Do Androids Dream of Electric Sheep?* Hollywood turned this into a movie in 1982, which was *Blade Runner*.

While the Turing test is useful, there will need to be other measures. After all, intelligence is more than just about conversation. It is also about interacting with our environment. Something even simple like making a cup of coffee can be exceedingly difficult for a machine to accomplish.

And what about text-to-image systems like DALL-E or Stable Diffusion? How can this intelligence be measured?

Well, researchers are working on various measures. But there remains considerable subjectivity with the metrics.

Jobs

In 1928, British economist John Maynard Keynes wrote an essay called "Economic Possibilities for Our Grandchildren." It was a projection about how automation and technology would impact the workforce by 2028. His conclusion: There would be a 15-hour workweek. In fact, he said this work would not be necessary for most people because of the high standard of living.²⁶

²⁶ www.newyorker.com/magazine/2014/05/26/no-time-how-we-got-so-busy

It's certainly a utopian vision. However, Keynes did provide some of the downsides. He wrote: "For the first time since his creation man will be faced with his real, his permanent problem—how to use his freedom from pressing economic cares, how to occupy the leisure, which science and compound interest will have won."²⁷

But as AI gets more powerful, it's certainly a good idea to think about such things. What might society look like? How will life change? Will it be better – or worse?

It's true that technology has disrupted many industries, which has led to widespread job losses. Yet there have always emerged new opportunities for employment. After all, in 2023 the US unemployment rate was the lowest since the late 1960s.

But there is no guarantee that the future will see a similar dynamic. AI could ultimately automate hundreds of millions of jobs – if not billions. Why not? In a capitalist system, owners will generally focus on low-cost approaches, so long as there is not a material drop in quality. But with AI, there could not only be much lower costs but much better results.

In other words, as the workplace becomes increasingly automated, there will need to be a rethinking of the concept of "work." But this could be tough since many people find fulfillment with their careers. The result is that there would be more depression and even addiction. This has already been the case for communities that have been negatively impacted from globalization and major technology changes.

To deal with the problems, one idea is to have universal basic income or UBI. This means providing a certain amount of income to everyone. This would essentially provide a safety net.

And this could certainly help. But with the trend of income inequality, there may not be much interest for a robust redistribution of wealth. This could also mean resentment for the many people who feel marginalized from the impacts of AI.

This is not to say that the future is bleak. But again, it is still essential that we look at the potential consequences of sophisticated technology like generative AI.

²⁷ www.newyorker.com/magazine/2014/05/26/no-time-how-we-got-so-busy

Conclusion

Moore's Law has been at the core of the growth in technology for decades. It posits that – every two years or so – there is a doubling of the number of transistors on an integrated circuit.

But it seems that the pace of growth is much higher for AI. Venture capitalists at Greylock Partners estimate that the doubling is occurring every three months.²⁸

Yet it seems inevitable that there will be a seismic impact on society. This is why it is critical to understand the technology and what it can mean for the future. But even more importantly, we need to be responsible with the powers of AI.

²⁸ <https://greylock.com/firm-news/welcome-mustafa-suleyman/>

Index

A

AlphaCode, 133, 143
 AlphaFold, 185, 186
 Amazon Web Services (AWS), 31, 75
 Artificial general intelligence (AGI), 135, 199, 200
 benefits, 93
 Artificial intelligence (AI)
 activation function, 51
 anomalies, 62
 approaches, 197–199
 artificial general intelligence, 199, 200
 association, 63
 autoencoders, 63
 BERT model, 107
 clustering, 62
 code generation system, 130, 131
 cognitive systems, 49
 COMET, 198
 de-aging, 176
 deep learning and machine learning, 52
 deep learning model, 65–71
 education, 178, 179
 entertainment, 176
 finance, 187, 188
 fundamentals, 47
 journalism, 181
 k-nearest neighbors algorithm, 60, 61
 linearly separation, 50
 Logic Theorist, 48

machine learning, 53
 machine learning playground, 76, 77
 Midjourney, 88–90
 Naïve Bayes classifiers, 61
 neural network/perceptron, 51
 overfitting/underfitting, 71
 programs, 48
 Python, 72, 73
 random forest, 59
 rectified linear unit (ReLU), 51
 regression analysis, 54–57
 reinforcement learning, 64, 65
 sigmoid function, 51
 Softmax activation function, 51
 supervised learning, 53, 54
 SVM algorithm, 57
 symbolic system, 49
 Tanh/hyperbolic tangent function, 51
 TOGGLE, 188
 tools and platforms, 73–75
 unsupervised learning, 61
 utopian vision, 201, 202
 XCON, 49
 zero marginal cost, 181

Auto-Encoding Variational Bayes, 83

Automatic speech recognition (ASR), 89

AutoML systems, 75

B

Bidirectional Encoder Representations from Transformers (BERT), 106

Brain Research through Advancing Innovative
Neurotechnologies (BRAIN), 70

C

Central processing unit (CPU), 14, 70

Centroid-based clustering, 62

ChatGPT

- code-generating systems, 195
- cultural phenomenon, 12
- image-to-image conversions, 10
- lackluster, 11
- megatech companies, 13
- prompt completion, 8, 9
- reputational risk, 13
- rival systems, 16
- stand-alone business, 12
- videos, 10

Chest X-ray Generation (CXR), 187

Chief data officer (CDO), 44

Cloud computing, 146

Clustering, 62

Code generation system

- AlphaCode, 133
- Atera, 138, 139
- BigCode, 142
- Blaze, 136, 137
- Copilot, 131–133
- data labeling, 138
- developer shortage, 128, 129
- disadvantages, 141–143
- entry-level programmers, 138
- large-scale projects, 139–141
- low-code/no-code development
systems, 129
- magic system, 135, 136
- manual coding, 128
- prompt engineering, 138
- Tabnine, 134
- time debugging, 137
- working process, 129, 130

Codex system, 131–133

Computing Research Association (CRA), 192

Contrastive Language–Image Pre-training
(CLIP) model, 86

Convolutional neural network (CNN), 68, 99

Cresta approach, 157, 158

Customer experience (CX), 148

D

Daily active users (DAUs), 155

Data system

- active data profiling, 38
- adaptive data quality, 38
- ApolloScape, 36
- Big Data, 25, 26
- BookCorpus, 36
- cases, 22
- chief data officer (CDO), 44
- cloud deployments, 31–34
- clustering and standardization, 38
- collecting data, 34
- CoreWeave, 33
- create, read, update, and delete
(CRUD), 27
- database, 27–31
- Databricks, 29
- data-driven companies, 22–24
- entities as experts (EAE), 40
- evaluation process, 36, 37
- genomic data, 25
- graph database, 31
- internal datasets, 35
- label details, 39, 40
- MapReduce, 28, 29
- meta/dark data, 25
- Michelangelo, 30
- Mobileye, 22, 23
- MongoDB, 29
- nodes/edges, 31
- NoSQL platforms, 29
- one-hot encoding, 38
- OpenWebText/CommonCrawl News, 36
- predictive transformation, 38
- preparation, 43–45
- privacy laws/regulations, 41
- quantity, 40
- real-time, 25
- relational database, 27, 28
- semi-structured data, 24
- Snowflake, 30
- sources, 35, 36
- strategies, 34
- structured/unstructured data, 24

- synthetic data, 25
 - time-series data, 24
 - types, 24
 - veracity, value, variability/visualization, 26
 - volume, variety, and velocity, 25
 - wrangling data, 37, 38
- Deep learning model, 14, 77
- backpropagation, 68
 - brain, 70
 - CNN/GAN, 68
 - code generation, 127
 - convolution, 68
 - data labeling, 137
 - disadvantages, 70, 71
 - discriminative modeling, 79, 80
 - hidden layers, 67
 - ImageNet Challenge, 69
 - NLP system, 95–98
 - PDF data, 81
 - RNN connects, 69
 - scalars, 65
 - types, 66
 - vector/tuple, 65
 - working process, 67, 68
- DeepMind, 133
- Density-based clustering, 62
- Discriminative modeling, 78, 79
- Discriminative models, 91
- Drugs for Neglected Diseases initiative (DNDi), 185
- Duolingo, 180, 181
- ## E
- Electronic health record (EHR) system, 19
- eXpert CONfigurer (XCON), 49
- Extensible Markup Language (XML), 24
- ## F
- Fréchet inception distance (FID), 84
- ## G
- Gaming development, 182, 183
- GANfather, 78
- General Data Protection Regulation (GDPR), 41
- Generative adversarial networks (GANs), 9, 14, 68, 78, 152
- CLIP model, 86
 - conditional probability model, 80
 - cutting-edge model, 89
 - CycleGAN, 82
 - DALL-E 2 process, 85, 86
 - decoder phase/unclip, 86
 - definition, 83
 - diffusion models, 84–86
 - DiscoGAN, 82
 - discriminative modeling, 78, 79
 - discriminator network, 81
 - hallucinations, 79
 - music, 176
 - neural networks, 81
 - probability theory, 79–81
 - speech, 89, 90
 - stability AI/midjourney, 86–88
 - stable diffusion, 87
 - text-2-image system, 82
 - trilemma, 90, 91
 - types, 81
 - variational autoencoder, 83
 - voice cloning, 89
- Generative AI system
- Agatha, 159
 - Anyword companies, 150
 - automation tools, 172
 - benefits, 20
 - Bing search engine, 190
 - boost global productivity, 189
 - brainstorming sessions, 169
 - buy vs. build evaluation, 165–168
 - Center of Excellence (CoE), 168
 - ChatGPT, 8–14
 - Cresta, 157, 158
 - customer experience, 148
 - cybersecurity, 195
 - data preparation, 172
 - data, 21
 - data-driven copy generation
 - solution, 150
 - definition, 3, 4
 - deployment/monitor, 173
 - dilemma, 193
 - DoNotPay, 146
 - downsampling, 42
 - dramatic model, 194

Generative AI system (cont.)

- drivers, 14, 15
- factual errors, misrepresentations, and bias, 192
- Forethought, 159
- Google platform, 173, 174
- hype cycles, 18–20
- identification, 169, 170
- implementation, 169
- INK, 151, 152
- Intercom operates, 160, 161
- Jasper platform, 2
- LaMDA model, 5–7
- legal research, 147
- manual and time-intensive process, 195
- marketing content, 2
- open source projects, 14
- Palmyra, 156
- PepsiCo, 163
- Poltorak, 149, 150
- potential threats, 194
- PowerPoint presentations, 164, 165
- product development, 161–163
- rare/positive class, 42
- regulation, 195–197
- researchers, 194
- rules-based approaches, 147
- sales and marketing categories, 149, 150
- skeptical view, 16–18
- skepticism, 193
- small team, 170, 171
- sophisticated systems, 4
- Spoke AI works, 164
- Subject-matter experts (SMEs), 168
- technology disruptions, 8
- venture capitalists (VCs), 6–8
- Wope, 151
- wrangling data, 41, 42
- Writer, 156, 157
- Zappi, 162, 163

Generative Pre-Trained Model (GPT)

- Algolia/Fable Studio, 109
- articles and computer code, 109
- CTEC, 108
- MIT technology, 111
- model architecture, 108
- multimodal functions, 111
- rules of thumb, 110

- Swedish government, 117

- Tokenizer, 110

- types, 111

- zero-shot learning, 108

- GPTZero system, 179, 180

- Grammarly, 154, 155

- Graphical user interface (GUI), 97

- Graphics processing unit (GPU), 14

H

- Healthcare development, 184–186

- Hierarchical-based clustering, 62

- Humby, Clive, 22

I

- ImageNet Large Scale Visual Recognition Challenge (ILSVRC), 68

- Integrated Development Environment (IDE), 129

J

- JavaScript Object Notation (JSON), 24

K

- Kakinada, 145

- Karpathy, Andrej, 127

- Key performance indicators (KPIs), 169

- K-nearest neighbors (KNN) algorithm, 60, 61

L

- Language Model for Dialogue Applications (LaMDA), 4–6

- Large language model (LLM), 21

- AI21 labs, 114, 115

- BERT model, 106, 107

- BigScience, 115

- BioBERT, 107

- BLOOM, 116, 117

- business models, 121

- Character.ai, 123, 124

- ChatGPT, 94

- ChatGPT/Databricks, 112

- Cohere platform, 114

commoditization, 121
 dials, 104–106
 DistilBERT, 107
 Dolly, 112
 empathy, 124
 generative AI systems, 191
 Gopher/Gato, 113
 GPT-Sw3, 117
 GPT system/ChatGPT, 108–112
 graduation risk, 121
 healthcare, 186
 intelligence, 95
 MassiveText, 113
 MT-NLG platform, 116
 NLP system, 95–98
 Nordic Pile, 117
 OpenAI innovations, 94
 OpenAI Playground, 105, 106
 PatentBERT, 107
 prompt engineering, 121, 122
 research lab, 94
 startup model, 119–121
 Tay system, 117
 transformer model, 100–105
 transformers, 99–104
 Word2Vec model, 98
 Latent Dirichlet Allocation (LDA), 97
 Lavender/SellScale, 153, 154
 Long short-term memory (LSTM) models, 99

M

Machine learning, 77, 78
 Magic system, 135
 Megatron-Turing Natural Language
 Generation Model (MT-NLG), 116
 Moore's Law, 202
 Music, 176

N

Naïve Bayes classifiers, 61
 National AI Research Resource
 (NAIRR), 192
 Natural language generation (NLG), 96
 Natural language processing (NLP)
 categories, 95, 96

chunking/shallow parsing, 97
 Grammarly, 155
 lemmatization, 97
 named entity recognition (NER), 97
 parts of speech (POS), 97
 Porter Stemming algorithm, 96
 stemming system, 97
 topic modeling, 97
 transformer model, 99–104
 working process, 96–98

Natural language understanding (NLU), 96

Neural Information Processing Systems
 conference (NeurIPS), 78

Neural networks/Natural language
 processing (NLP), 47

O

Overfitting, 71

P, Q

Personally identifiable information (PII), 42
 PolyCoder, 135, 136
 Probability density function (PDF), 80
 Probability theory, 79–81
 Python, 72, 73
 PyTorch, 74

R

Random forest model, 59
 Recurrent neural network (RNN), 69, 99
 Regie.ai process, 152, 153
 Regression analysis, 54–57
 Reinforcement learning, 14, 64, 65
 Reinforcement Learning with Human
 Feedback (RLHF), 158
 Remote monitoring (RMM) system, 139
 Retrieval-Enhanced Transformer (Retros), 113
 Return on investment (ROI), 167
 Road experience management (REM), 23
 Roblox, 183, 184
 Rogenmoser, Dave, 1
 Rosenblatt, Frank, 49, 50

S

- Search engine optimization (SEO), 149
- Search Engine Results Page (SERP), 149
- Semi-automated Business Research Environment (SABRE), 27
- Software-as-a-service (SaaS), 1, 146
- Software development kit (SDK), 32
- Structured Query Language (SQL), 27
- Supervised learning
 - Artificial intelligence (AI), 53, 54
- Support vector machine (SVM)
 - clustering, 58
 - hyperplane, 57
 - kernel method, 58
- Synthetic Minority Over-sampling Technique (SMOTE), 42

T

- Tabnine, 134
- TensorFlow system, 73
- Tensor processing units (TPUs), 15, 74
- Text-to-speech (TTS), 89
- Transformer model
 - attention layer, 100
 - attention mechanisms, 99
 - CNNs/RNNs and LSTMs, 99
 - DeepMind, 102
 - encoder and decoder functions, 102
 - foundational models, 103
 - foundation model, 104
 - Google researchers, 100–105
 - multihead attention, 102
 - transfer learning, 100
 - Vision Transformer (ViT), 103

U

- Underfitting, 71
- Universal basic income (UBI), 201
- Unsupervised learning, 14, 39
 - Artificial intelligence (AI), 61
 - discriminative modeling, 79

V

- Variational autoencoder (VAE), 83
- Venture capitalists (VCs), 6–8
- Vision transformer (ViT), 103

W

- WaveAI, 177, 178
- Word2Vec model, 98
- Word error rate (WER), 89

X, Y, Z

- X-ray training data, 187