ChatGPT and Bard for Business Automation

Achieving AI-Driven Growth

Tom Taulli

Apress®

Tom Taulli

ChatGPT and Bard for Business Automation

Achieving AI-Driven Growth

Apress®

Tom Taulli
Monrovia, CA, USA

ISBN 978-1-4842-9851-0 e-ISBN 978-1-4842-9852-7
https://doi.org/10.1007/978-1-4842-9852-7

© Tom Taulli 2023

This work is subject to copyright. All rights are solely and exclusively licensed 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.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Apress imprint is published by the registered company APress Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Contents

Chapter 1: New World of Automation
Chapter 2: ChatGPT
Chapter 3: Bard
Chapter 4: Prompt Engineering
Chapter 5: Use Cases
Chapter 6: Change Management
Chapter 7: Understanding AI
Chapter 8: Data
Chapter 9: Risks
Index

About the Author

Tom Taulli

is the founder of OnePrompter.com, which provides online training for ChatGPT and Bard. He is also the author of various books, including *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business* and *Artificial Intelligence Basics: A Non-Technical Introduction*. You can follow him on Twitter at @ttaulli.

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_1

1. New World of Automation

How ChatGPT and Bard Can Transform Your Business

Tom Taulli[1]
(1)
Monrovia, CA, USA

In August 2023, Constant Contact released the findings from a survey from 486 small business decision-makers in the United States.[1] These were at the director level or higher. The focus of the survey was on the use of artificial intelligence (AI) and other automation technologies.

It showed that 91% of the respondents said that these systems have made their business more successful. Some of the benefits included time savings, fewer errors, lower costs, higher efficiency, and faster growth. For example, 28% of the respondents said they saved at least $5,000 for the year.

According to Mary Ginder, the co-owner at Gindo's Spice of Life:

Being a small business owner means I wear a lot of hats, so every minute and every dollar counts. I use AI when I'm looking for new ideas on how to describe a flavor profile for our sauces, or when I'm in a hurry to get a newsletter generated and don't have time to write my own copy. I love how it gives me fresh ideas when I feel stuck, and it helps me formulate my thoughts and execute them much quicker than I could on my own. I'd say AI has helped cut my work time in half on average.

Here are some other interesting findings from the survey:

- 60% of the respondents who use AI for marketing said that they saved time and have become more efficient.
- 33% said that they have saved more than 40 minutes per week on marketing.

However, about 46% of the respondents said that they have an introductory knowledge of AI and its use cases. They also expressed concerns about costs, the learning curve, and potential security issues.

In this book, the focus will be on helping to demystify AI, with a focus on ChatGPT and Bard. We'll show practical steps that businesses can take to achieve superior results.

> **Note** According to a survey from Resume Builder, close to 90% of business leaders consider ChatGPT as a beneficial skillset for job applicants.[2]

The Impact

The launch of ChatGPT in November 2022 took the world by storm. The huge interest in the app even surprised the creator of the system, OpenAI. In a couple months, ChatGPT attracted over 100 million users – making it one of the fastest growing technologies in history.

The system seemed to be human-like. It could answer seemingly any question (although not always with accuracy). Some of the early adopters were students, who used ChatGPT to create their homework assignments. But businesses would quickly adopt the technology. ChatGPT would ultimately become a cultural phenomenon. After all, *Saturday Night Live* did a funny skit about it.

At the core of ChatGPT is generative AI. The "generative" part means that it creates content, such as text, audio, and video. Although, for the most part, the focus has been on writing text, summarizing information, extracting data, and answering questions.

Generative AI involves the use of breakthrough algorithms like the transformer model. These involve training models with huge amounts of the world's content on the Internet.

A system like ChatGPT is often referred to as a large-language model or LLM. Or sometimes they are called foundation models. Regardless, these systems are extremely powerful and massive. Consider that they have hundreds of billions of parameters.

Google is a pioneer in developing generative AI systems. However, the company was slow to commercialize the technology. This may have been due to fears that the systems could be misused or perhaps Google was just too slow. Another possibility is that generative AI could erode its lucrative search business. With something like ChatGPT, there is not much of a need to click on links, right? Definitely. But for Google, it means less opportunities to generate advertising revenues.

But with the huge success of ChatGPT, Google realized it had to get serious about the technology. And the company quickly ramped its efforts. The result was the launch of Bard, which came in March 2023. Google has also been aggressive in integrating generative AI into its various apps like Gmail and Docs.

While generative AI is still in the nascent stages, it is likely to have an enormous impact on businesses. It should allow for major improvements in cost savings, efficiency, productivity, and automation.

To get a sense of this, let's take a look at an extensive research report from McKinsey & Co.[3] It estimates that generative AI could provide $2.6 trillion to $4.4 trillion in annual economic benefits. This is based on 63 uses cases and 16 business functions.

A key factor for this is that generative AI has the ability to understand and create natural language. This represents about 25% of the total work time of an employee. What's more, about a fifth of the work week is spent searching and gathering information.

Generative AI will likely impact all business functions. But McKinsey & Co notes that there are certain areas that will move the needle in a much bigger way. They include the following and they could account for about 75% of the benefits of the technology:

- Customer operations: This can allow for self-service, which will provide for scaling of service. But generative AI tools can also help human agents with retrieving information about a customer as well as help with writing emails or other communications. The benefits include higher customer satisfaction scores, higher resolutions, increased sales (say with cross-selling and up-selling), and lower attrition for human agents.
- Marketing: Generative AI can create personalized and engaging messages. It can also make interesting copy like blogs, social media posts, product descriptions, advertisements, and taglines. The content can also have a consistent style, tone, and brand voice. Oh, and it can improve SEO (search engine optimization).
- Sales: A salesperson spends much time corresponding with customers and prospects. But generative AI can greatly streamline this. The technology can also personalize this according to customer personas. This can lead to higher close rates and faster sales cycles.
- Software engineering: Generative AI can also create computer code. This can greatly speed up the development process. But it could also help deal with the severe talent shortage.
- Product R&D: Generative AI can help with suggesting ideas and creating prototypes. This is often referred to as "generative design."

Note In the summer of 2023, OpenAI was on a run rate of generating more than $1 billion in revenue for the next 12 months, according to a report from TheInformation.com.[4] By comparison, the company reported $28 million in revenue for 2022.

Agenda

Using tools like ChatGPT and Bard can take some time to learn as well as to get adoption in an organization. This is a journey, which will require experimentation.

As for this book, we'll focus on ways to make this smoother. First of all, we'll get started with a deep dive on ChatGPT and Bard. This will be full of examples.

Next, we'll cover prompt engineering, which is a must-have skillset for using ChatGPT and Bard. We'll focus on the best practices and the things to avoid.

Then we'll look at the use cases for ChatGPT and Bard. We'll show the many useful prompts for categories like sales, marketing, customer support, finance, and human resources – just to name a few.

After this, we'll discuss change management. This is about getting your team to adopt the technology for the best impact. It's a step that often gets overlooked or not enough attention. But if you want to be successful with ChatGPT or Bard, then you need to focus on change management.

We'll also have a chapter about risks. No doubt, ChatGPT and Bard are far from perfect. These are emerging technologies that are undergoing dynamic change.

The final two chapters are about the fundamentals of AI and data management. These are closely linked topics. A key to getting adoption of a technology is to understand some of the underlying structures. It will help build trust with your team.

Also, learning about AI and data will be important if you want to go beyond ChatGPT and Bard. This could be something like creating a custom AI app or using a sophisticated no-code/low-code system.

Conclusion

In this chapter, we have taken a brief look at the trends of ChatGPT and Bard. We have also set forth the agenda for the book.

As for the next chapter, we'll take a look at ChatGPT.

Footnotes

1

www.prnewswire.com/news-releases/constant-contact-research-reveals-small-businesses-who-use-ai-are-more-likely-to-save-money-and-be-successful-301896332.html

2

www.cnbc.com/2023/09/01/more-companies-see-chatgpt-training-as-a-hot-job-perk-for-office-workers.html

3

www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction

4

www.theinformation.com/articles/openai-passes-1-billion-revenue-pace-as-big-companies-boost-ai-spending?utm_source=sg&utm_medium=email&utm_campaign=article_email&utm_content=article-11166&rc=lcooo2

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_2

2. ChatGPT

The App That Changed the World

Tom Taulli[1]
(1)
Monrovia, CA, USA

There is nothing new about chatbots. They have been around for decades. But they have often seemed lacking, as it can be extremely difficult to understand and communicate with natural language.

In some cases, chatbots have turned out to be disasters. This was certainly the case with Microsoft's Tay. The irony is that the initial system showed success. The app was integrated into Tencent's WeChat social network in China and it gained millions of users.

But the situation would be much different in the United States. Then again, Microsoft's blunder was perhaps that it implemented Tay on Twitter. On the first day of the debut in March 2016, the chatbot would quickly start spouting racist and sexist comments because users found ways to manipulate the system. Within 24 hours, Microsoft took down Tay and it never returned.

But fast forward to November 2022 and the situation was much different. AI had progressed substantially, especially with new models like transformers.

For OpenAI, they thought the technology was ready for the public at large and spent several months to create ChatGPT. True, it was not perfect. But then again, what app is? Still, OpenAI spent much time testing the system and there was a red-teaming effort to try to break it.

The user interface was also clean and easy to understand, kind of like the Google search engine. There were also features to help with the experience. For example, ChatGPT would have follow-up questions to clarify prompts and there were messages to disclose the limitations of the model.

When it went live, the traction was immediate. Users were captivated with the humanlike responses about any subject. People posted their experiences on Twitter – and ChatGPT went viral.

Within the first two months, the app would reach an astounding 100 million users with nearly 600 million visits during the month, according to analysis from UBS.[1]

According to Sandhini Agarwal, who helps to develop policy at OpenAI:

> *I think it was definitely a surprise for all of us how much people began using it. We work on these models so much, we forget how surprising they can be for the outside world sometimes.*[2]

If anything, ChatGPT represented an inflection point. It was similar to the moment when the Netscape browser ushered in the Internet revolution or when the iPhone was launched. The impact was seismic.

ChatGPT has certainly evolved quickly from its launch. So in this chapter, we'll take a look at this system and show the key features of this powerful app. But let's first get an overview of OpenAI.

OpenAI

OpenAI was born in December 2015, and it got lots of fanfare. After all, the company had an impressive list of cofounders, who included Elon Musk, Reid Hoffman, Peter Thiel, and Sam Altman. They pledged up to $1 billion in funding for this venture.

In the first blog post for the company:

> *OpenAI is a non-profit artificial intelligence research company. Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact.*

The founders wanted OpenAI to be a way to avoid the inevitable domination of AI from mega companies like Microsoft, Apple, Amazon, and Meta. By being a nonprofit, this could help allow for democratization of the technology.

The founders also wanted to build a system that would ultimately lead to artificial general intelligence or AGI. The blog stated:

> *AI systems today have impressive but narrow capabilities. It seems that we'll keep whittling away at their constraints, and in the extreme case they will reach human performance on virtually every intellectual task. It's hard to fathom how much human-level AI could benefit society, and it's equally hard to imagine how much it could damage society if built or used incorrectly.*

In the early days, OpenAI was like a university, where many of the employees were Ph.D.s. They were mostly focused on experimenting with new concepts of AI. Because of this approach, OpenAI was able to attract some of the leading experts in deep learning and machine learning. They were willing to forgo bigger salaries at companies like Meta, Apple, and Microsoft because of the compelling mission.

The initial research was on reinforcement learning, which is a form of AI that is based on a reward-punishment system. In April 2016, OpenAI released its first application: OpenAI Gym, a development platform. Later in the year, it launched Universe, which was focused on developing AGI systems.

In 2018, Musk decided to depart as a board member. Part of this was due to the potential conflicts of interest with Tesla. But he was also skeptical about the strategy for OpenAI.

A major turning point would come in 2019. OpenAI changed its strategic direction and started to build large-language models (LLMs) that were based on generative AI. But this required enormous amounts of resources and the nonprofit structure was proving to be a hindrance.

Simply put, OpenAI needed to raise substantial amounts of capital – and fast. To do this, the organization converted to a "capped" for-profit organization. This meant that it would allow outside investors to invest in the organization but the returns would have limits.

In July, OpenAI announced that Microsoft had made a $1 billion investment in the firm. This involved a strategic agreement to build a new Azure AI supercomputer. OpenAI also agreed to run all its services on Azure, and Microsoft would become its preferred partner for commercializing new AI technologies.

According to the announcement:

> *The companies will focus on building a computational platform in Azure of unprecedented scale, which will train and run increasingly advanced AI models, include hardware technologies that build on Microsoft's supercomputing technology, and adhere to the two companies' shared principles on ethics and trust. This will create the foundation for advancements in AI to be implemented in a safe, secure and trustworthy way and is a critical reason the companies chose to partner together.*[3]

This deal was certainly critical. But over the next few years, OpenAI would need even more capital. Microsoft would invest an additional $2 billion. Then, after ChatGPT was launched and the growth of OpenAI soared, there was a massive $10 billion investment. The valuation was set at $29 billion.

What Is ChatGPT?

I asked ChatGPT for the definition of ChatGPT. Here's what I got:

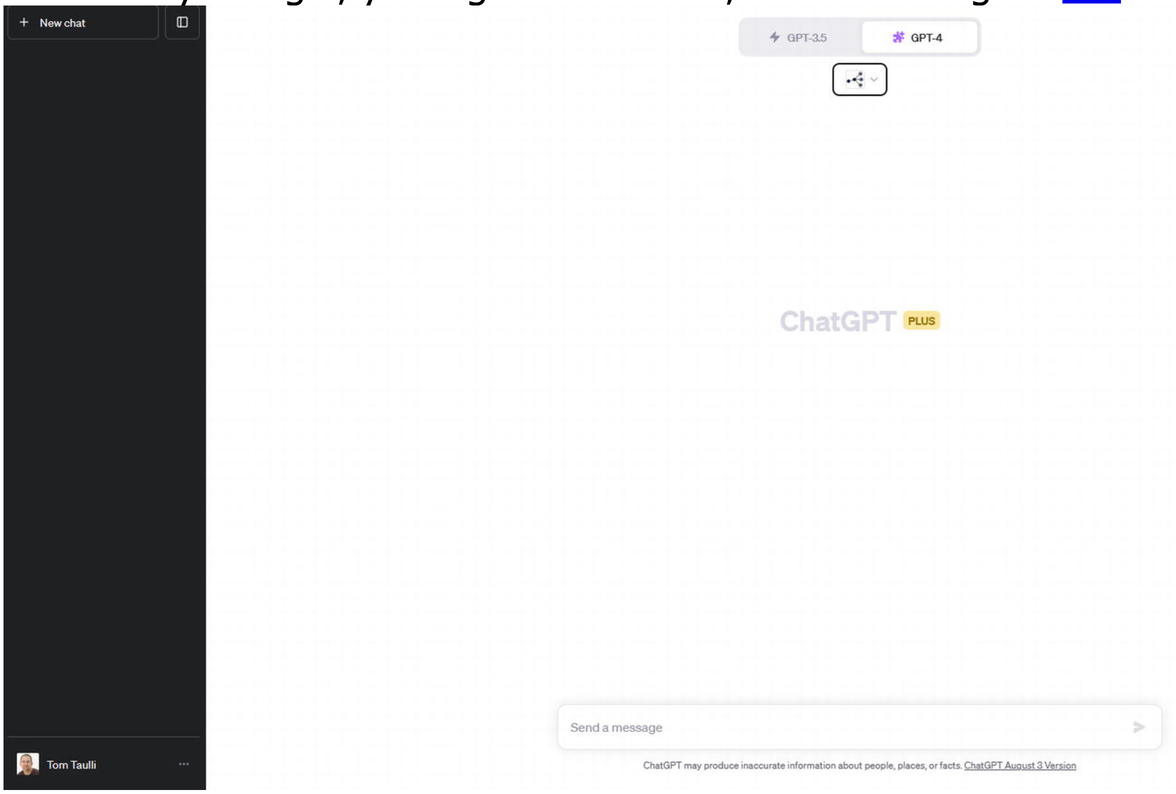
ChatGPT is a series of state-of-the-art language models developed by OpenAI. The "GPT" in ChatGPT stands for "Generative Pre-trained Transformer," and it refers to the architecture of the model.

This is a succinct description of the technical aspects of ChatGPT. But of course, ChatGPT is an application that helps with many tasks, whether answering questions, summarizing information, or creating code.

To use ChatGPT, you need to create an online account. You can do this at the following URL: `https://chat.openai.com/auth/login`.

The login can be your email or password. Or you can use your credentials for Google, Microsoft, or Apple.

When you login, you'll get this screen, as seen in Figure 2-1.



+ New chat
GPT-3.5
GPT-4
ChatGPT PLUS
Send a message
Tom Taulli
ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

Figure 2-1 This is the main screen for ChatGPT

The screen shows that this is the premium version of the application. This means that there are additional services.

At the top, there is the option for using two types of models: GPT-3.5 and GPT-4. The default is GPT-4, which is more powerful.

Under this selection, there is a pull-down menu for the plugins. Later in the chapter, we'll show how to use this feature.

For the bottom right side of the screen, there is a question icon. This is if you want to get help. There are two resources. One is a FAQ and the other is a list of keyboard shortcuts, which include the following:

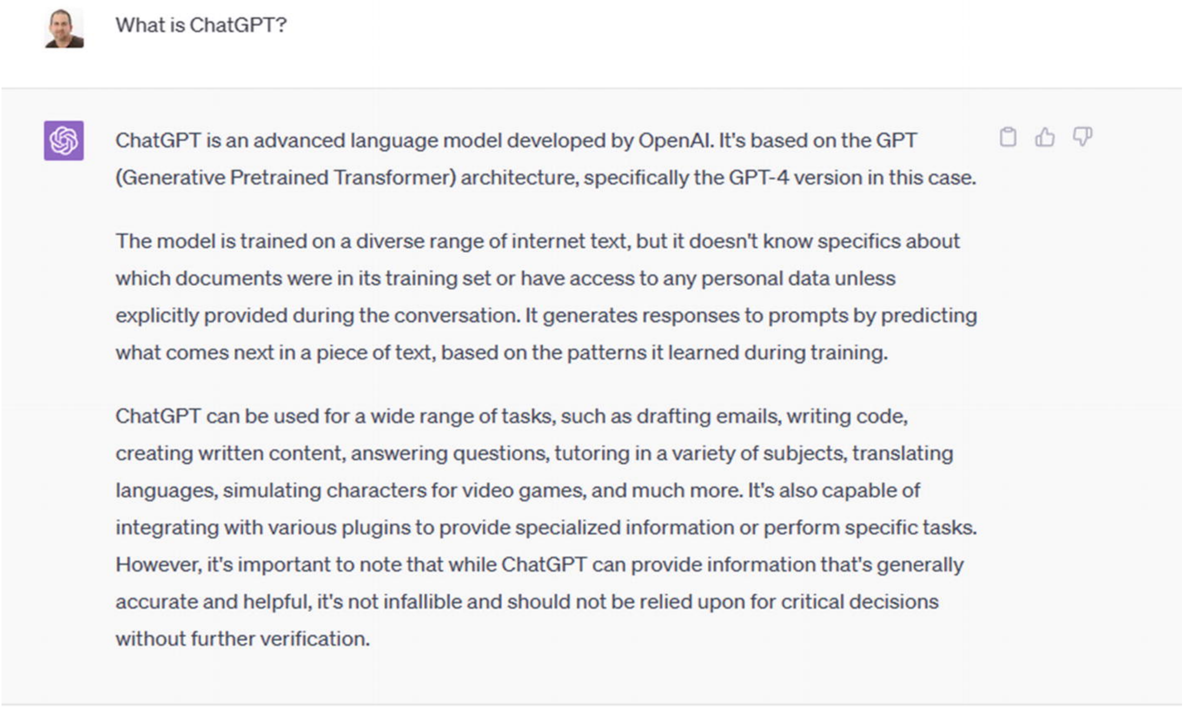
- Open new chat: Ctrl + Shift + O
- Focus chat input: Shift + Esc
- Copy last code block: Ctrl + Shift + .
- Copy last response: Ctrl + Shift + C
- Set custom instructions: Ctrl + Shift + I
- Toggle sidebar: Ctrl + Shift + S
- Delete chat: Ctrl + Shift + Backspace

On the left side of the screen, there is something called "New Chat." This is a chat thread that is saved. Although, if you do not want to use this, you can use the icon to the right to close it.

At the bottom of the screen, there is an input box, which is where you can enter your prompt. Under this, there is a warning message that "ChatGPT may produce inaccurate information about people, places, or facts." This is important to keep in mind. It's a good idea to check the information that the system generates.

Let's try out ChatGPT. In the input box, I'll enter a simple message: "What is ChatGPT?"

This is what we get, as seen in Figure 2-2.



What is ChatGPT?

ChatGPT is an advanced language model developed by OpenAI. It's based on the GPT (Generative Pretrained Transformer) architecture, specifically the GPT-4 version in this case.

The model is trained on a diverse range of internet text, but it doesn't know specifics about which documents were in its training set or have access to any personal data unless explicitly provided during the conversation. It generates responses to prompts by predicting what comes next in a piece of text, based on the patterns it learned during training.

ChatGPT can be used for a wide range of tasks, such as drafting emails, writing code, creating written content, answering questions, tutoring in a variety of subjects, translating languages, simulating characters for video games, and much more. It's also capable of integrating with various plugins to provide specialized information or perform specific tasks. However, it's important to note that while ChatGPT can provide information that's generally accurate and helpful, it's not infallible and should not be relied upon for critical decisions without further verification.

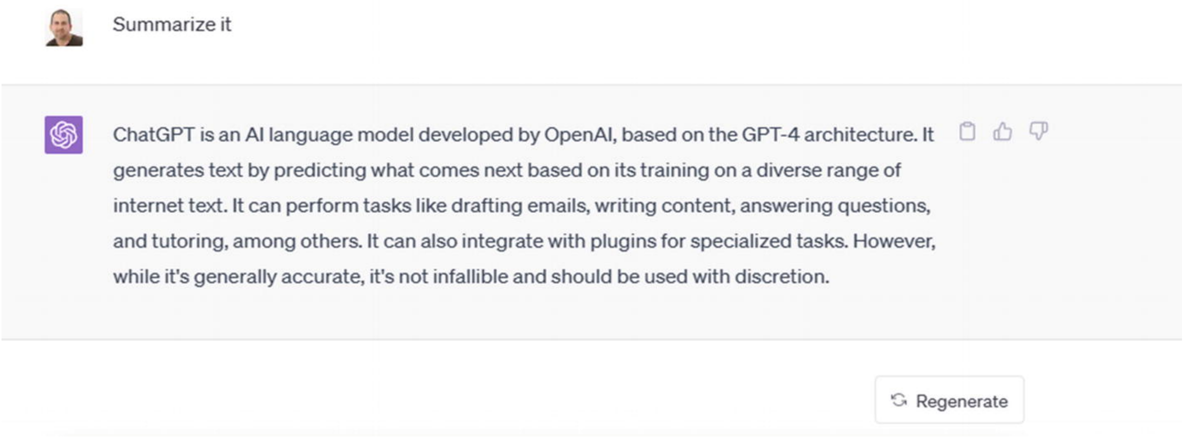
Figure 2-2 This shows the response from a prompt while using ChatGPT

We get three paragraphs. However, if you use the same prompt, you are likely to get something different. This is because the generative AI is based on complex probabilities. It's what allows for seemingly creative content.

On the left side of the screen, you will see a new chat. ChatGPT provides a title for this, which is "ChatGPT Overview: AI Language Model." If you click the pencil icon, you can change it, or if you want to delete it, you can click the trash can.

For the response that ChatGPT generated, there are three icons on the top right. The first one allows you to copy the response. The two others make it so you can provide feedback about the response. OpenAI will use this to attempt to make the model better.

The use of chat sessions is important. It means that we can have ongoing interactions. For example, suppose we think the response is too long. We can enter this prompt: "Summarize it." You can see the results in Figure 2-3.


Summarize it

ChatGPT is an AI language model developed by OpenAI, based on the GPT-4 architecture. It generates text by predicting what comes next based on its training on a diverse range of internet text. It can perform tasks like drafting emails, writing content, answering questions, and tutoring, among others. It can also integrate with plugins for specialized tasks. However, while it's generally accurate, it's not infallible and should be used with discretion.

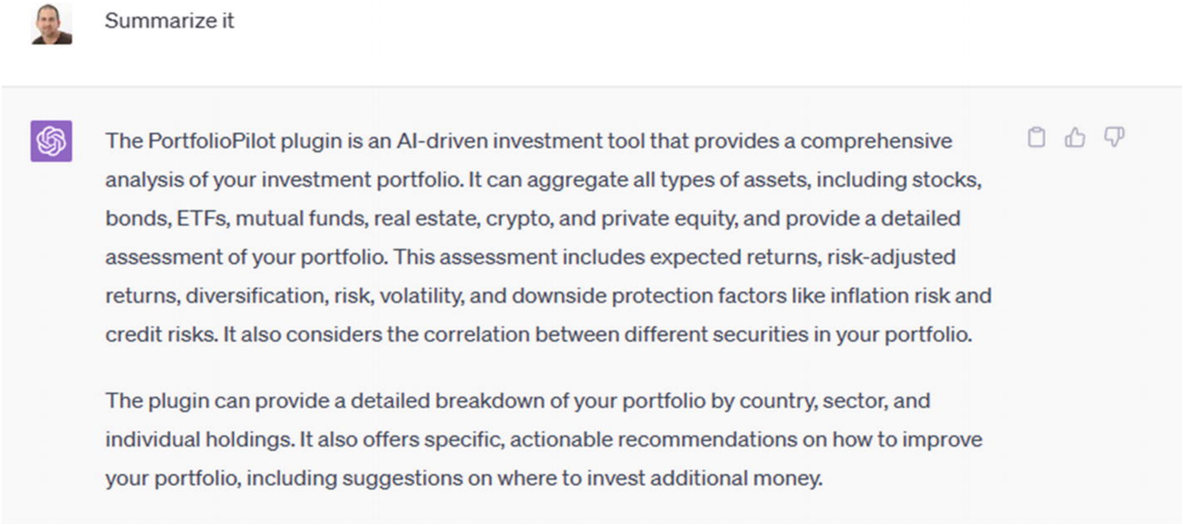
Figure 2-3 This shows a chat interaction with ChatGPT

ChatGPT has created a more succinct definition. It's now only one paragraph.

At the bottom of the response, there is an icon called "Regenerate." Press this if you want ChatGPT to try again with your prompt.

Suppose we create a new chat. We select this from the top left of the screen.

By doing this, we get a blank screen. Enter the "summarize" prompt again and this is what we get, as seen with Figure 2-4.


Summarize it

The PortfolioPilot plugin is an AI-driven investment tool that provides a comprehensive analysis of your investment portfolio. It can aggregate all types of assets, including stocks, bonds, ETFs, mutual funds, real estate, crypto, and private equity, and provide a detailed assessment of your portfolio. This assessment includes expected returns, risk-adjusted returns, diversification, risk, volatility, and downside protection factors like inflation risk and credit risks. It also considers the correlation between different securities in your portfolio.

The plugin can provide a detailed breakdown of your portfolio by country, sector, and individual holdings. It also offers specific, actionable recommendations on how to improve your portfolio, including suggestions on where to invest additional money.

Figure 2-4 This shows a new chat

Since this is a new chat session, it does not know about the prompt about defining ChatGPT. The system instead is trying to figure out what is the relevant topic. It actually provides a summary of a plugin that I have previously installed.

Let's now take a look at the settings option. If you click it, you will get three options: My Plan, Settings & Beta, and Logout.

The last one is obvious. But we'll take a look at the first two. With the My Plan option, you get background information about the version you have. If you have ChatGPT Plus, this will cost you $20 per month. But I think this is well worth it.

You get the following:

- Access to GPT-4
- Faster response speeds
- Access to beta features like browsing, plugins and the Advanced Data Analysis. However, if you no longer want an option, you can turn it off here.

The Settings option has three main sections. First, there is General. Here you can change the theme for the screen, such as for dark or light.

Then there is the "open plugin devtools." You will enable this if you want to develop ChatGPT plugins locally on your system.

Next, you can click a button to clear all the chats. This can be useful since this part of your screen can quickly fill up.

For beta features, this allows you to turn on various new functions. These will change from time to time.

But as of this writing, here are the beta features:

- Custom instructions: This helps to customize how ChatGPT responds to your prompts.
- Plugins: These are essentially apps that you can use.
- Advanced Data Analysis: This is a powerful tool that can write and execute Python code. But you do not have to be a developer to get lots of value from this.

Then there is the Data controls option. These are the options:

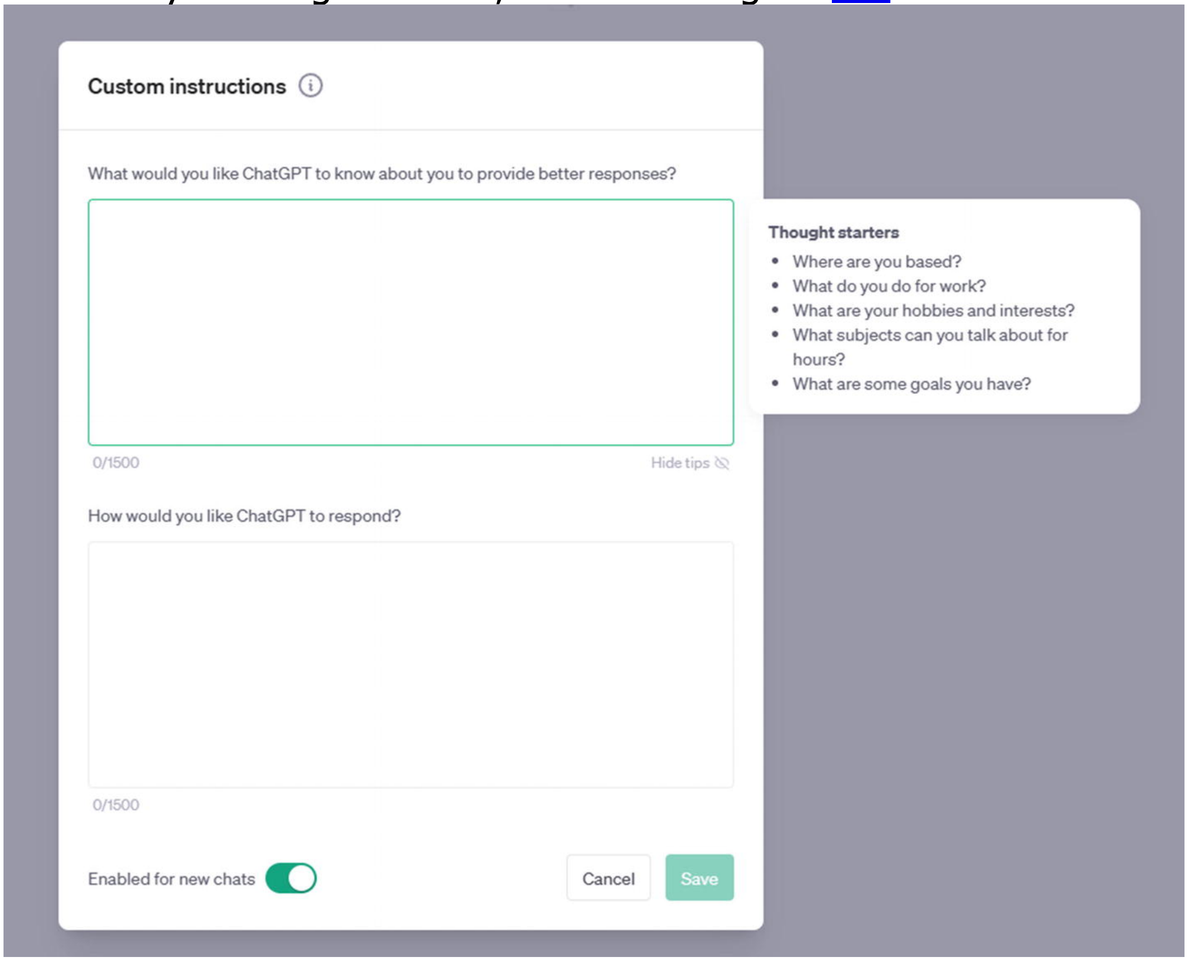
- Chat history and training: This allows you to determine if your prompts will be used as a way to train the ChatGPT model to make it better.
- Export: You can save your chats.
- Shared links: You can share a URL of your chats with other users.
- Delete the account: You can do this any time.

Let’s take a closer look at some of these features.

Custom Instructions

To activate Custom Instructions, you will select it in Settings & Beta. Then click the “...” under this. The Custom Instructions option will show up and you will get a brief description of how it works.

Then you will get a form, as seen in Figure 2-5.



Custom instructions

What would you like ChatGPT to know about you to provide better responses?

0/1500

Thought starters
- Where are you based?
- What do you do for work?
- What are your hobbies and interests?
- What subjects can you talk about for hours?
- What are some goals you have?

How would you like ChatGPT to respond?

0/1500

Figure 2-5 This is the input form for the Custom Instructions option

When you add this information, it will become a part of all your prompts – unless you change or delete the instructions. Plugin developers may also get access to the data. This should allow for more personalized experiences. What's more, ChatGPT will use the data for its own model training. But again, you can opt-out of this.

There are two input boxes. The first one asks: "What would you like ChatGPT to know about you to provide better responses?" There are no guidelines for this. It's whatever that is typically about what you think is relevant for what you focus on.

ChatGPT provides suggestions, such as the location where you are based, your occupation, your hobbies, your favorite subjects, and goals.

The next input box asks: "How would you like ChatGPT to respond?" For this, ChatGPT suggests how formal you want the content to be, the length of the response, how you want to be addressed, and whether the AI should have opinions on the topics.

For each of the input boxes, the maximum length is 1,500 characters.

Plugins

Plugins are a popular feature in ChatGPT. Think of them kind of like apps for your smartphone.

A key benefit for plugins is that they are not subject to the cut-off date for the LLM. They provide access to real-time data.

Assuming you have activated plugins, you will see an icon at the top of the screen for ChatGPT. When you click it, you will get a menu selection, where you can view the Plugin store. Figure 2-6 shows what it looks like.

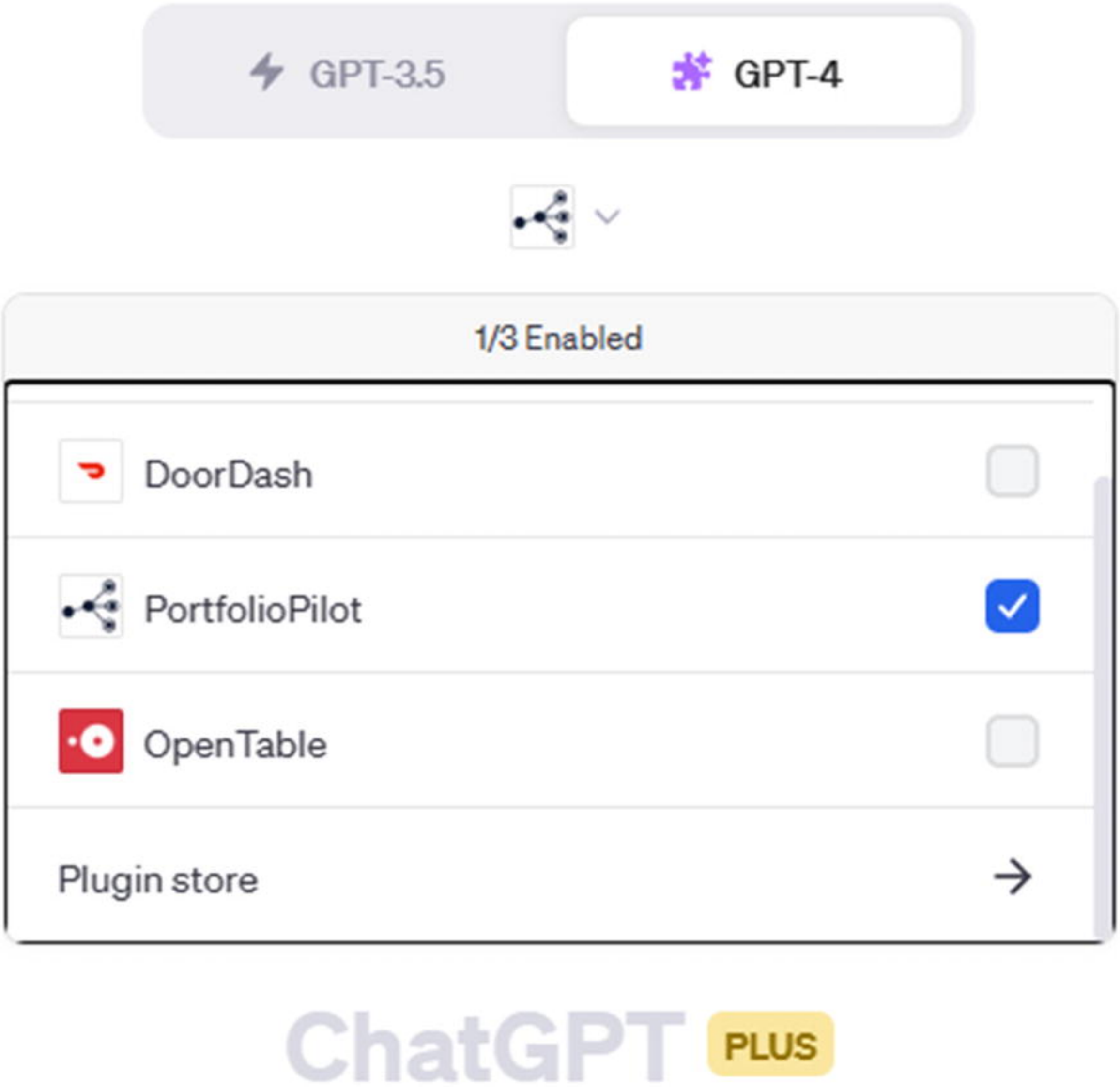
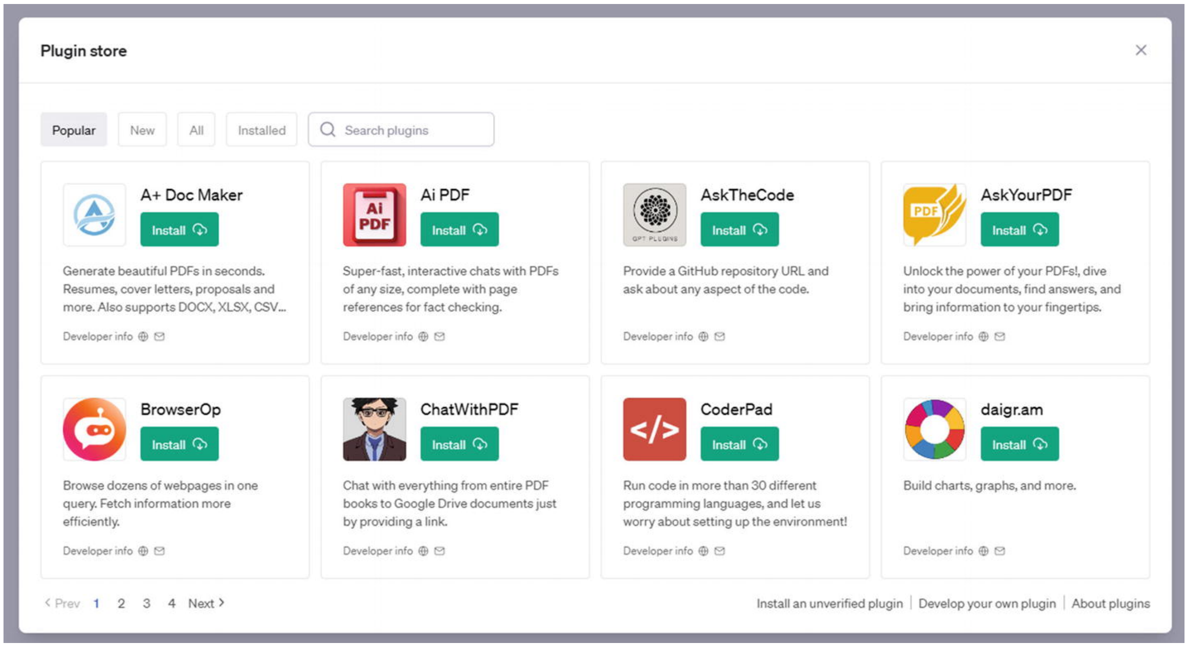


Figure 2-6 This shows the options for the plugins, including a menu item for the Plugin Store

When you click this option, you will get the following screen, as seen in Figure 2-7.



A+ Doc Maker

Generate beautiful PDFs in seconds. Resumes, cover letters, proposals and more. Also supports DOCX, XLSX, CSV...

Ai PDF

Super-fast, interactive chats with PDFs of any size, complete with page references for fact checking.

AskTheCode

Provide a GitHub repository URL and ask about any aspect of the code.

AskYourPDF

Unlock the power of your PDFs!, dive into your documents, find answers, and bring information to your fingertips.

BrowserOp

Browse dozens of webpages in one query. Fetch information more efficiently.

ChatWithPDF

Chat with everything from entire PDF books to Google Drive documents just by providing a link.

CoderPad

Run code in more than 30 different programming languages, and let us worry about setting up the environment!

daigr.am

Build charts, graphs, and more.

Figure 2-7 This is the main screen for the Plugin Store

At the top, you have different selections. The default is the most popular plugins. But you can also see the new ones, all that are available, and those that are installed on your ChatGPT system. If you are looking for a particular one, you can use the search box.

For this screen, the popular plugins include CoderPad, ChatWithPDF, BrowserOp, and so on. These change frequently.

At the bottom of the screen, you can page through all the plugins. There is also an option to install unverified plugins. However, you need to be wary of this since there could be security issues. The other option on this part of the screen is for developers and some background information about plugins.

Let's try out a plugin. We'll go to the search box and enter "Kayak." You will then click the Install button and then click the X button on the top right. You will be back to the main screen and now you will see the icon for Kayak, which is in Figure 2-8.

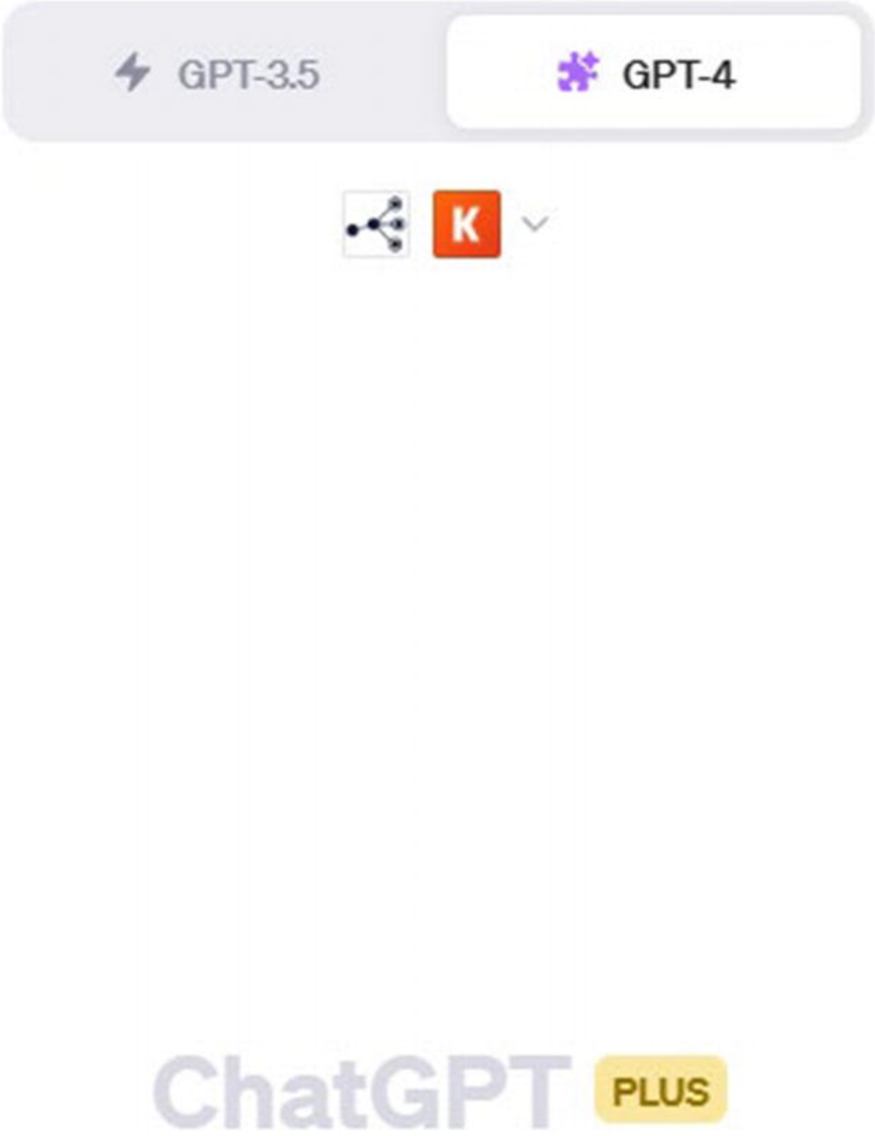
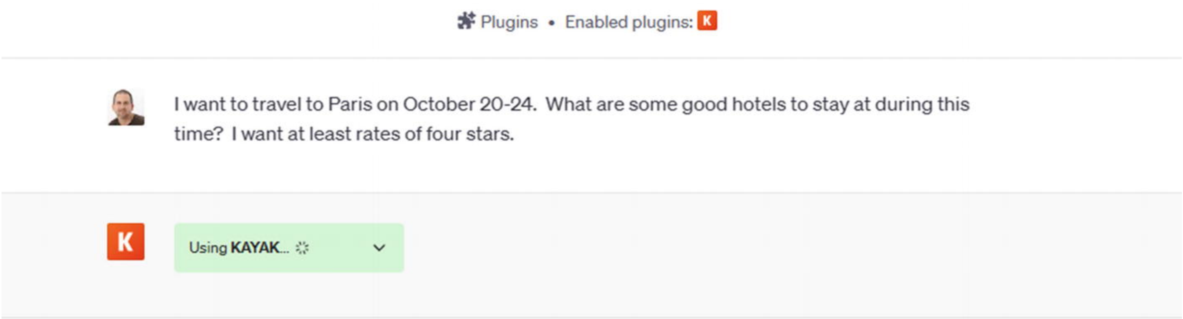


Figure 2-8 Here you can see the Kayak plugin button at the top of the screen

This means that it is enabled. To use it, you will write a prompt that is related to Kayak. We'll have this:

> *I want to travel to Paris on October 20-24. What are some good hotels to stay at during this time? I want the ratings be to be at least four stars.*

Since this is about scheduling a trip, ChatGPT will start using the Kayak plugin. This will be done by displaying the logo and green box, which says "Using Kayak." You can see this in Figure 2-9.

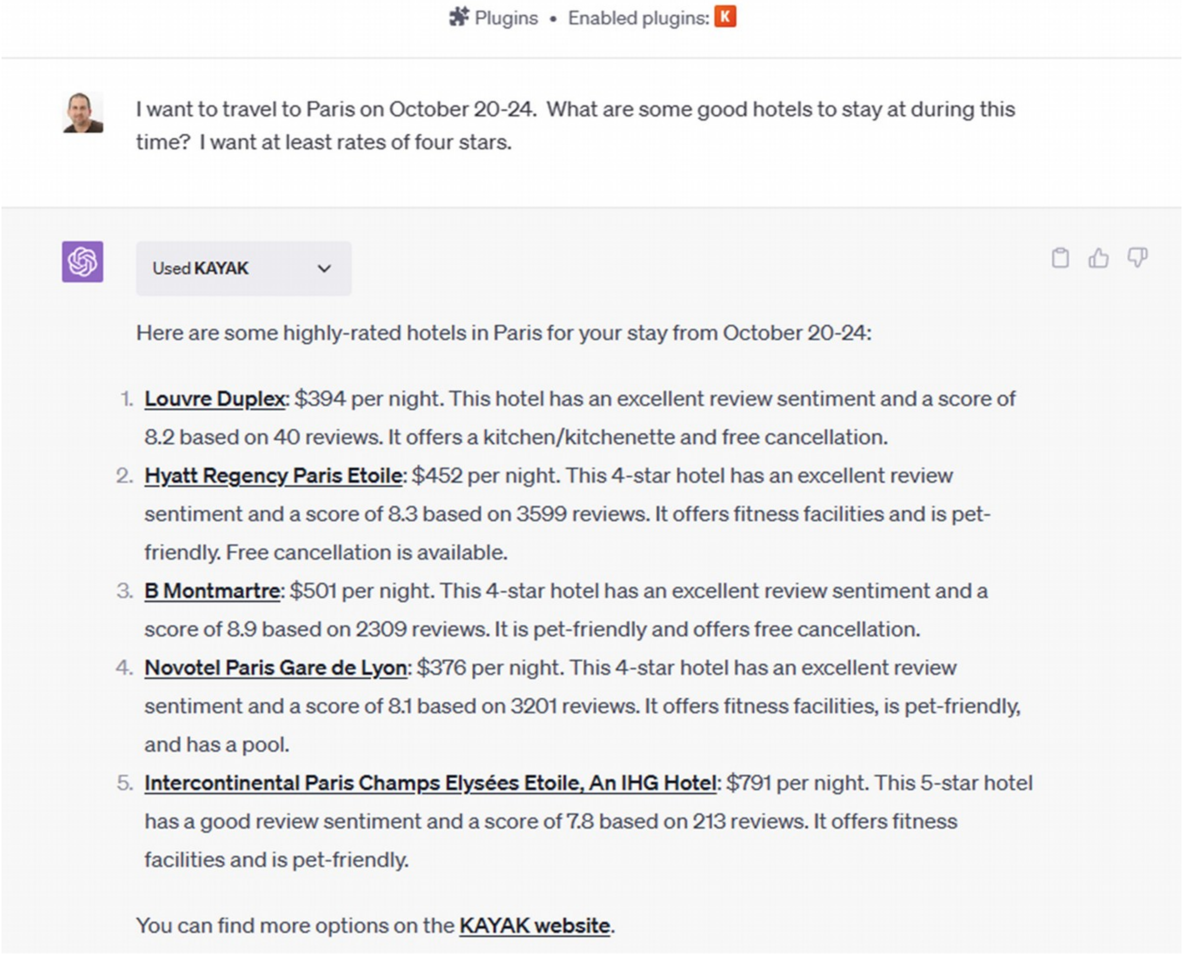


I want to travel to Paris on October 20-24. What are some good hotels to stay at during this time? I want at least rates of four stars.

Figure 2-9 This shows that ChatGPT is accessing Kayak for processing the prompt

If you click the down arrow, you can see the steps that ChatGPT is taking. Although, this is something you do not need to check out – unless you're a coder and want to see how to develop a plugin.

ChatGPT will find some hotels. The information will include the name, the rate, the number of stars, and a brief description of the amenities. You can see the list in Figure 2-10.



I want to travel to Paris on October 20-24. What are some good hotels to stay at during this time? I want at least rates of four stars.

Here are some highly-rated hotels in Paris for your stay from October 20-24:

1. **Louvre Duplex**: $394 per night. This hotel has an excellent review sentiment and a score of 8.2 based on 40 reviews. It offers a kitchen/kitchenette and free cancellation.
2. **Hyatt Regency Paris Etoile**: $452 per night. This 4-star hotel has an excellent review sentiment and a score of 8.3 based on 3599 reviews. It offers fitness facilities and is pet-friendly. Free cancellation is available.
3. **B Montmartre**: $501 per night. This 4-star hotel has an excellent review sentiment and a score of 8.9 based on 2309 reviews. It is pet-friendly and offers free cancellation.
4. **Novotel Paris Gare de Lyon**: $376 per night. This 4-star hotel has an excellent review sentiment and a score of 8.1 based on 3201 reviews. It offers fitness facilities, is pet-friendly, and has a pool.
5. **Intercontinental Paris Champs Elysées Etoile, An IHG Hotel**: $791 per night. This 5-star hotel has a good review sentiment and a score of 7.8 based on 213 reviews. It offers fitness facilities and is pet-friendly.

You can find more options on the **KAYAK website**.

Figure 2-10 This shows the results for the Kayak plugin

> **Note** Sometimes when ChatGPT generates content, it will stop abruptly. This is usually when it is creating a large amount of text. ChatGPT will then show a "continue" button. Click it if you want the content generation to go on. Or, if you do not see this button, you can enter "continue output" in the prompt. In other cases, the content generation may not be giving what you want. In this case, you can press the "stop generating" button.

Advanced Data Analysis

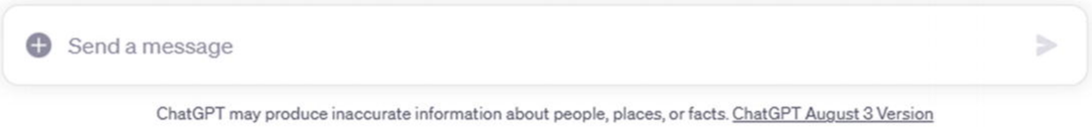
Advanced Data Analysis is one of the most useful features in ChatGPT. Again, you do not have to be a developer to use this.

With it, you can engage in math, do data analytics, create visualizations, and even edit images. Advanced Data Analysis does this by creating the code – which is Python – to carry out the task you want (this is where the "code" part comes in). The system will then execute the code.

There are some guardrails. You cannot get Internet access with it and there are limitations with the execution environment, such as you cannot access your local computer. It also does not provide all the available Python libraries.

Once you have Advanced Data Analysis enabled, you will then go to the top of the screen and select GPT-4 and there will be a drop-down menu. Select Advanced Data Analysis.

This will change the prompt. You will see a + icon, which allows you to upload one or more files, as seen in Figure 2-11. The total file size maximum is 100MB.



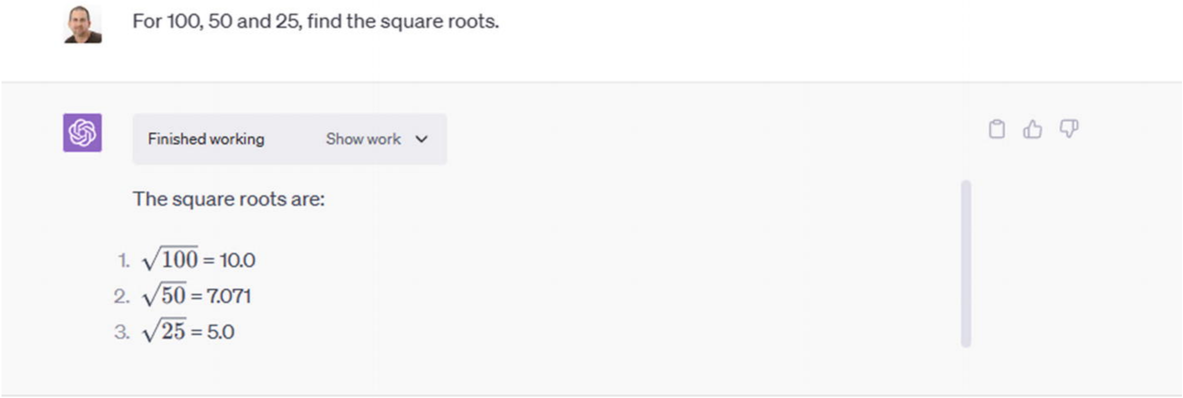
ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

Figure 2-11 This shows the prompt for Advanced Data Analysis, where you can upload one or more files. You do this by selecting the + icon

Let’s try some use cases. We’ll start with a prompt, which is the following:

For 100, 50, and 25, find the square roots.

Figure 2-12 shows the output for this and it is correct. It even provides the square root symbol.



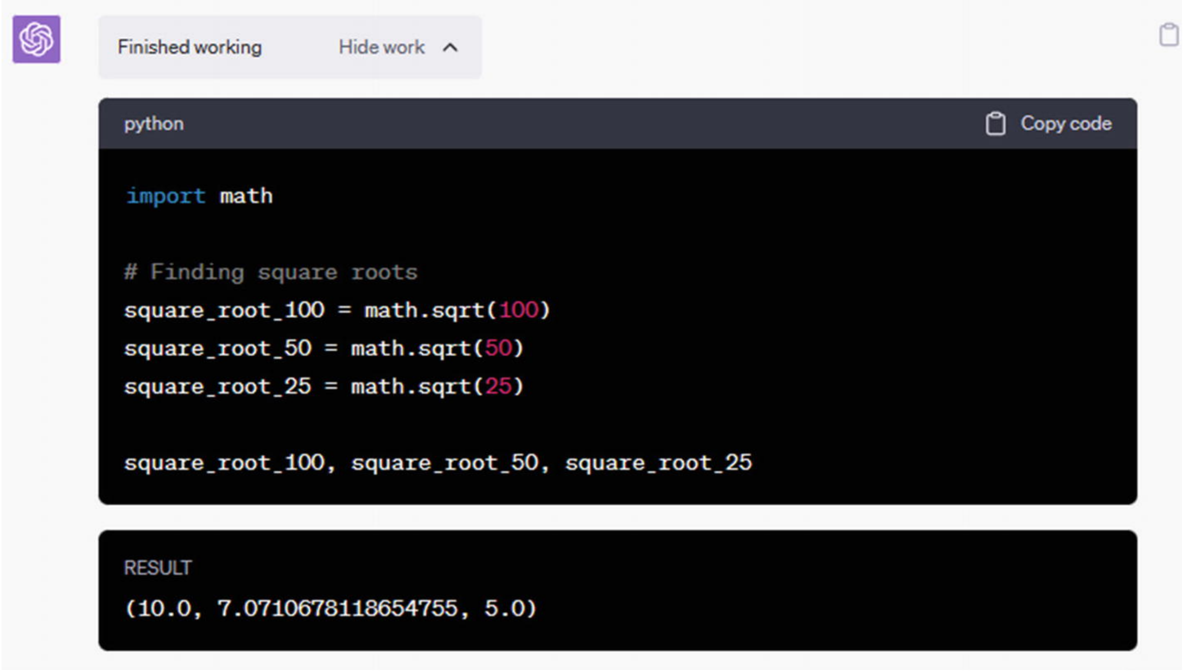
For 100, 50 and 25, find the square roots.

The square roots are:

1. $\sqrt{100} = 10.0$
2. $\sqrt{50} = 7.071$
3. $\sqrt{25} = 5.0$

Figure 2-12 This shows how Advanced Data Analysis has found the square roots of these numbers

If you select “Show Work” at the top, you will see a listing of the Python code. You can find this in Figure 2-13.

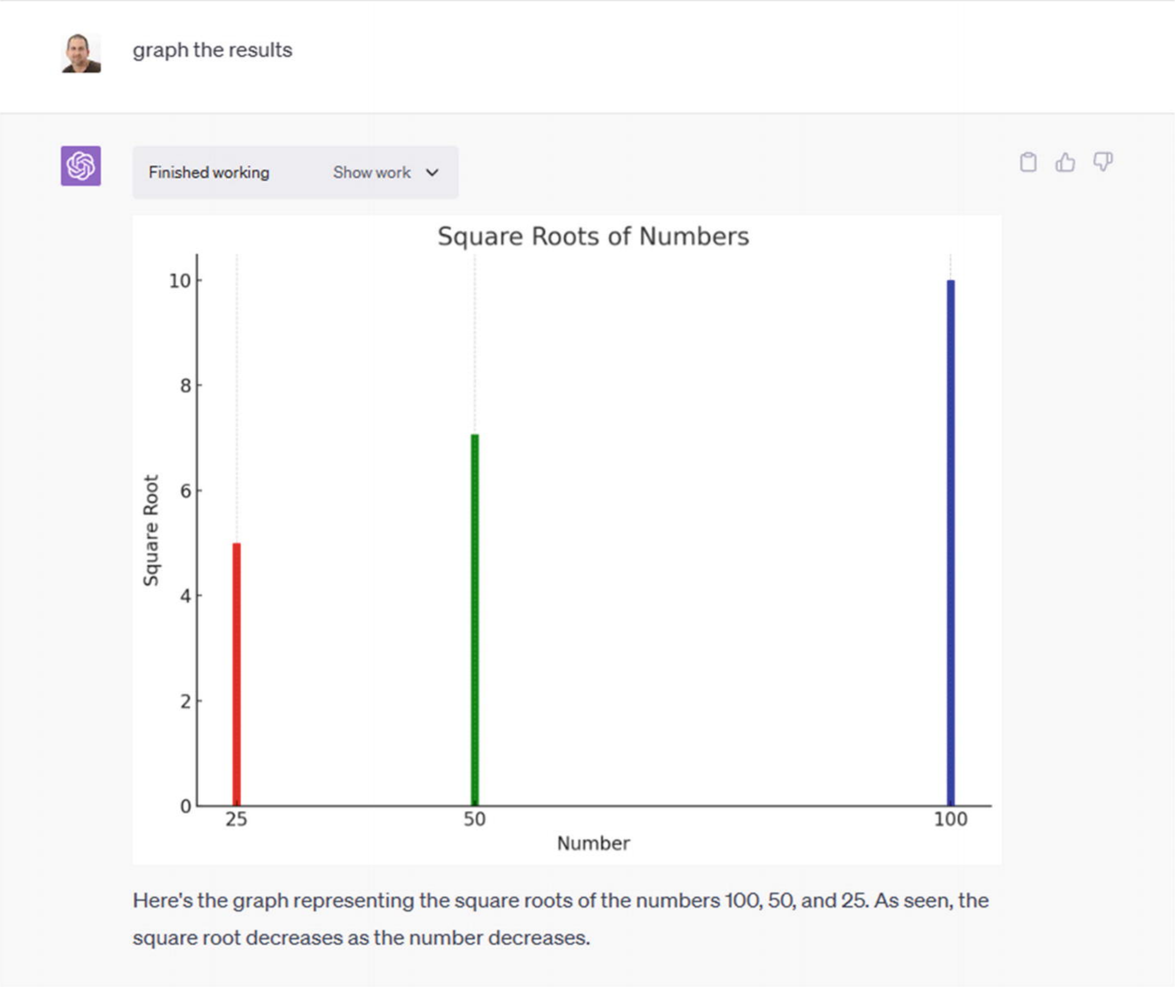


```python
import math

# Finding square roots
square_root_100 = math.sqrt(100)
square_root_50 = math.sqrt(50)
square_root_25 = math.sqrt(25)

square_root_100, square_root_50, square_root_25
```

```
RESULT
(10.0, 7.0710678118654755, 5.0)
```

Figure 2-13 This shows the Python code for how Advanced Data Analysis solved the square roots

What about we prompt it to graph the results? Well, ChatGPT can do this with a prompt, as seen with Figure 2-14.



graph the results

Here's the graph representing the square roots of the numbers 100, 50, and 25. As seen, the square root decreases as the number decreases.

Figure 2-14 This shows how Advanced Data Analysis graphed the square roots

In the next few sections, we'll take a look at other use cases.

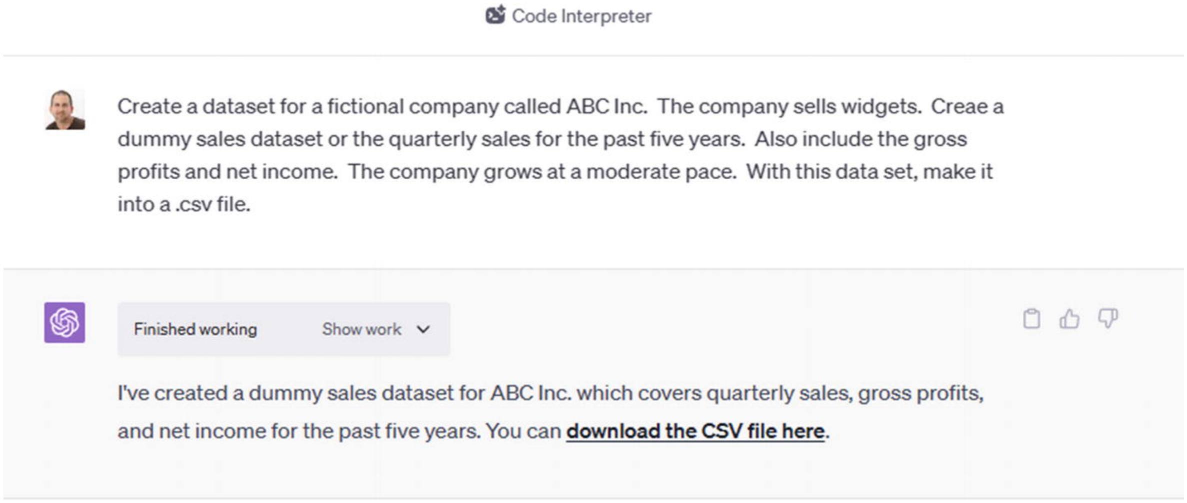
Data Analytics

There are many data analytics tools on the market. While they can be very useful, they can be pricey and difficult to use. In many cases, you need to have experience with data science to get any value from these tools.

But ChatGPT's Advanced Data Analysis has made the process much better. To see how, we'll have an example. For our dataset, we'll actually have ChatGPT create this for us. Here's the prompt:

Create a dataset for a fictional company called ABC Inc. The company sells widgets. Create a dummy sales dataset for the quarterly sales for the past five years. Also include the gross profits and net income. The company grows at a moderate pace. With this dataset, make it into a .csv file.

Figure 2-15 shows the results for this.

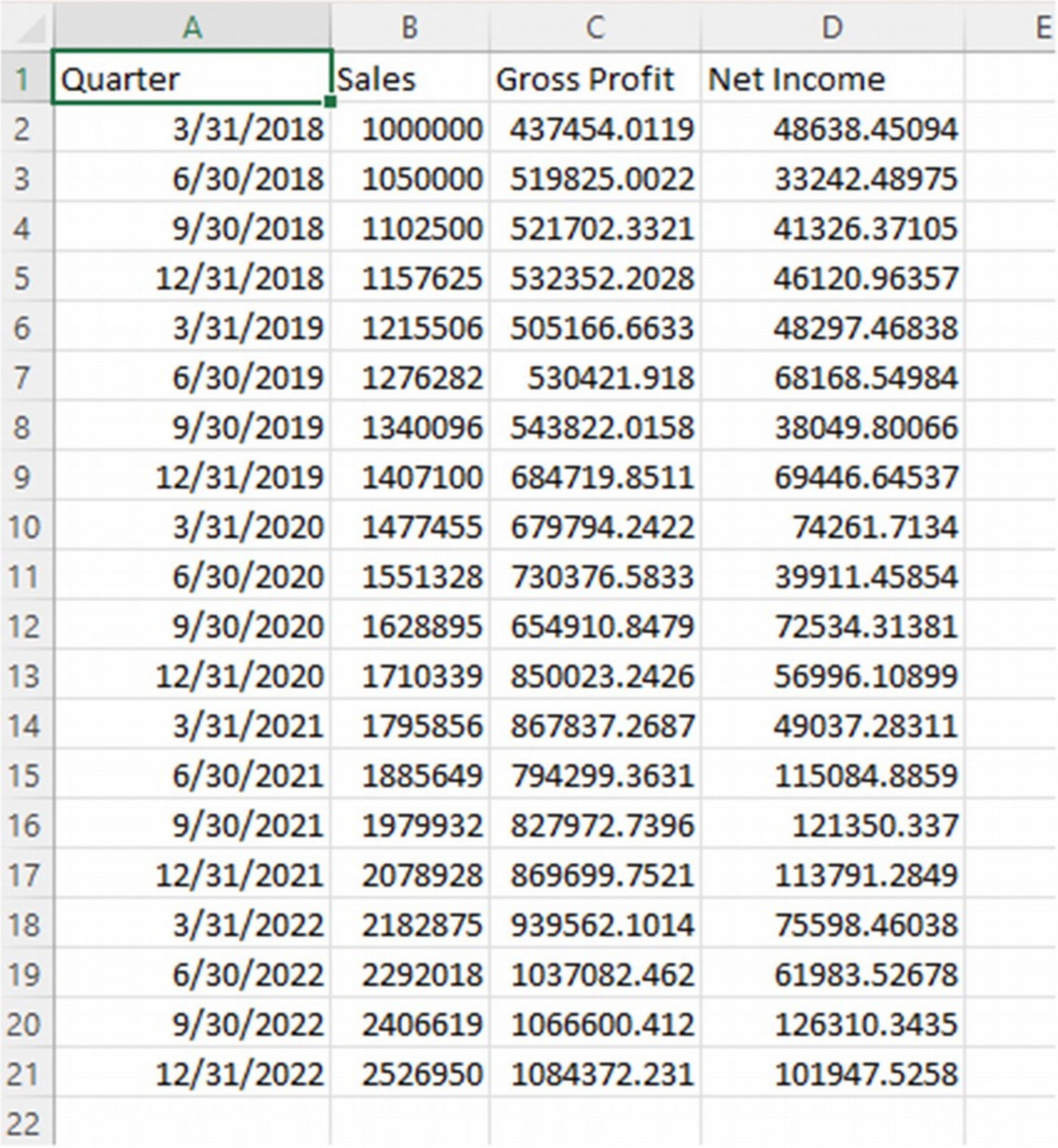


Create a dataset for a fictional company called ABC Inc. The company sells widgets. Creae a dummy sales dataset or the quarterly sales for the past five years. Also include the gross profits and net income. The company grows at a moderate pace. With this data set, make it into a .csv file.

I've created a dummy sales dataset for ABC Inc. which covers quarterly sales, gross profits, and net income for the past five years. You can **download the CSV file here**.

Figure 2-15 This shows how Advanced Data Analysis has created a dataset for a fictional company

You have a link to a file. A CSV (comma-separated values) file is a text file that divides data by commas. When you download the file, the format looks like this, in Figure 2-16.



	A	B	C	D	E
1	Quarter	Sales	Gross Profit	Net Income	
2	3/31/2018	1000000	437454.0119	48638.45094	
3	6/30/2018	1050000	519825.0022	33242.48975	
4	9/30/2018	1102500	521702.3321	41326.37105	
5	12/31/2018	1157625	532352.2028	46120.96357	
6	3/31/2019	1215506	505166.6633	48297.46838	
7	6/30/2019	1276282	530421.918	68168.54984	
8	9/30/2019	1340096	543822.0158	38049.80066	
9	12/31/2019	1407100	684719.8511	69446.64537	
10	3/31/2020	1477455	679794.2422	74261.7134	
11	6/30/2020	1551328	730376.5833	39911.45854	
12	9/30/2020	1628895	654910.8479	72534.31381	
13	12/31/2020	1710339	850023.2426	56996.10899	
14	3/31/2021	1795856	867837.2687	49037.28311	
15	6/30/2021	1885649	794299.3631	115084.8859	
16	9/30/2021	1979932	827972.7396	121350.337	
17	12/31/2021	2078928	869699.7521	113791.2849	
18	3/31/2022	2182875	939562.1014	75598.46038	
19	6/30/2022	2292018	1037082.462	61983.52678	
20	9/30/2022	2406619	1066600.412	126310.3435	
21	12/31/2022	2526950	1084372.231	101947.5258	
22					

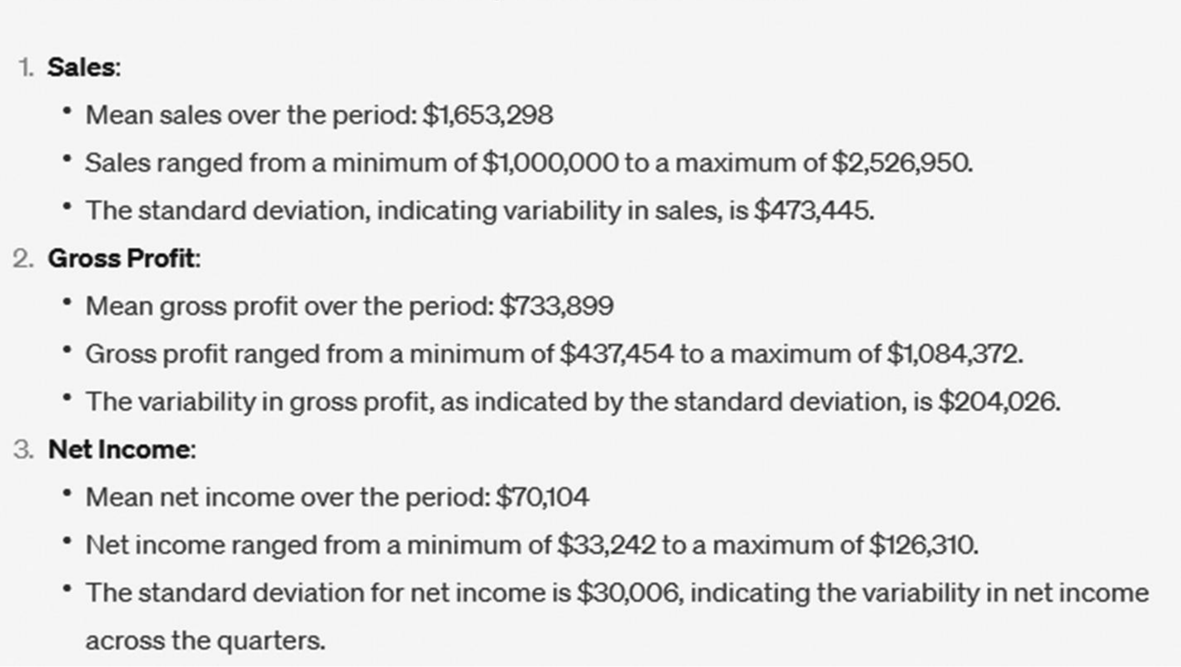
Figure 2-16 This shows the CSV file for the quarterly sales

Advanced Data Analysis has generated four columns and then a listing of the quarterly results. The system has correctly specified the dates, and the numbers look correct. For example, the sales are growing moderately and the gross profits and net income are lower.

Next, let's upload this file by pressing the + icon. Then enter this prompt:

Analyze this data.

Figure 2-17 shows the results.



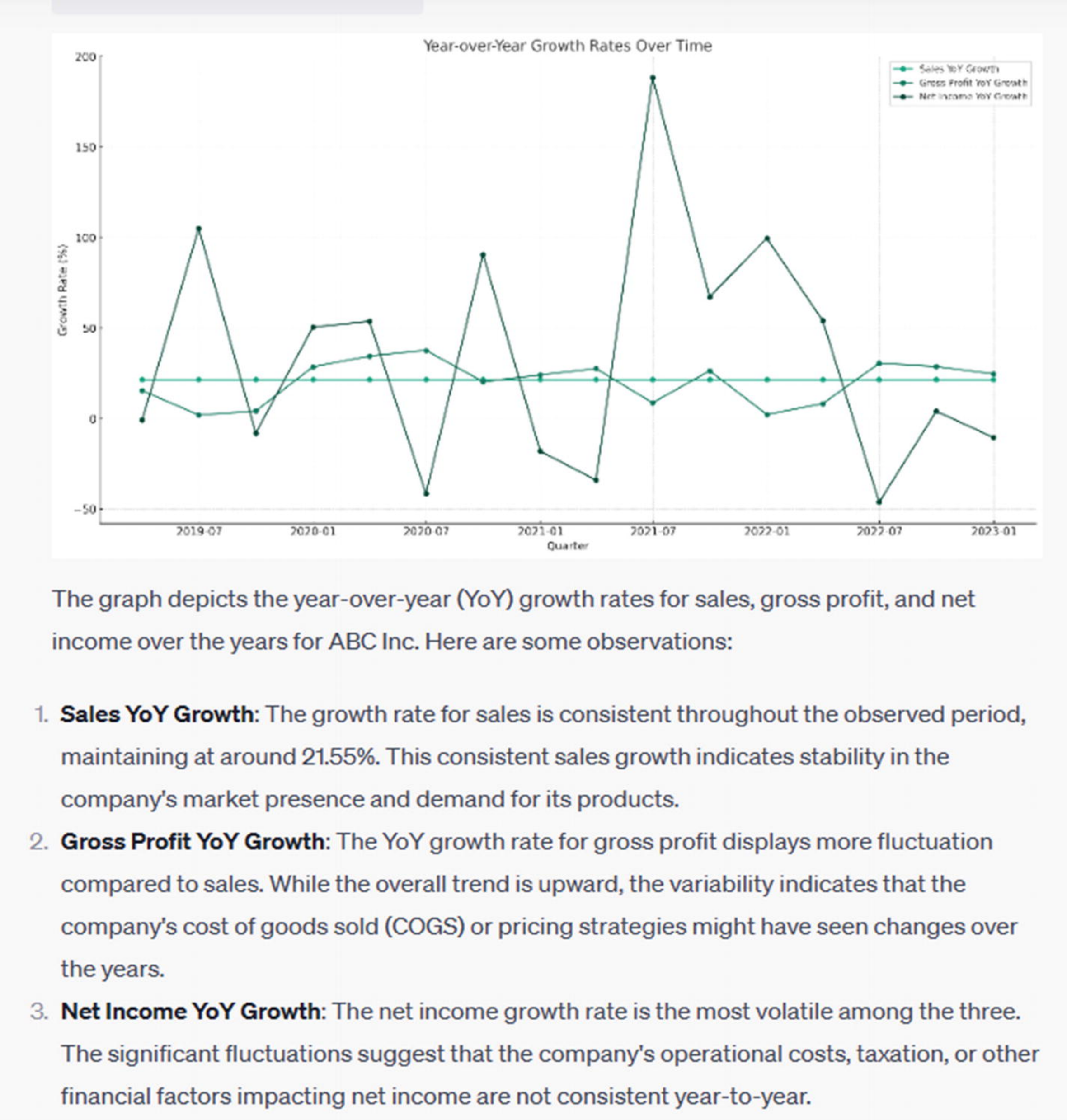
Here's a brief overview of the summary statistics for the dataset:

1. **Sales**:
    - Mean sales over the period: $1,653,298
    - Sales ranged from a minimum of $1,000,000 to a maximum of $2,526,950.
    - The standard deviation, indicating variability in sales, is $473,445.
2. **Gross Profit**:
    - Mean gross profit over the period: $733,899
    - Gross profit ranged from a minimum of $437,454 to a maximum of $1,084,372.
    - The variability in gross profit, as indicated by the standard deviation, is $204,026.
3. **Net Income**:
    - Mean net income over the period: $70,104
    - Net income ranged from a minimum of $33,242 to a maximum of $126,310.
    - The standard deviation for net income is $30,006, indicating the variability in net income across the quarters.

Figure 2-17 This is the data analysis for the CSV file

Advanced Data Analysis analyzes the sales, gross profit, and net income. It calculates metrics like the mean, ranges, and the standard deviation, so as to gauge the variations. In other words, you have generated helpful information with just a simple prompt.

But there is more to the analysis. Advanced Data Analysis creates visualizations of the dataset, as seen in Figure 2-18.



The graph depicts the year-over-year (YoY) growth rates for sales, gross profit, and net income over the years for ABC Inc. Here are some observations:

1. **Sales YoY Growth**: The growth rate for sales is consistent throughout the observed period, maintaining at around 21.55%. This consistent sales growth indicates stability in the company's market presence and demand for its products.
2. **Gross Profit YoY Growth**: The YoY growth rate for gross profit displays more fluctuation compared to sales. While the overall trend is upward, the variability indicates that the company's cost of goods sold (COGS) or pricing strategies might have seen changes over the years.
3. **Net Income YoY Growth**: The net income growth rate is the most volatile among the three. The significant fluctuations suggest that the company's operational costs, taxation, or other financial factors impacting net income are not consistent year-to-year.

Figure 2-18 Advanced Data Analysis creates visualizations for the dataset

You can then change the chart with a prompt. You can ask it for a different type, say a bar or pie chart. You can also change the titles and the colors or the time periods.

Optical Character Recognition (OCR)

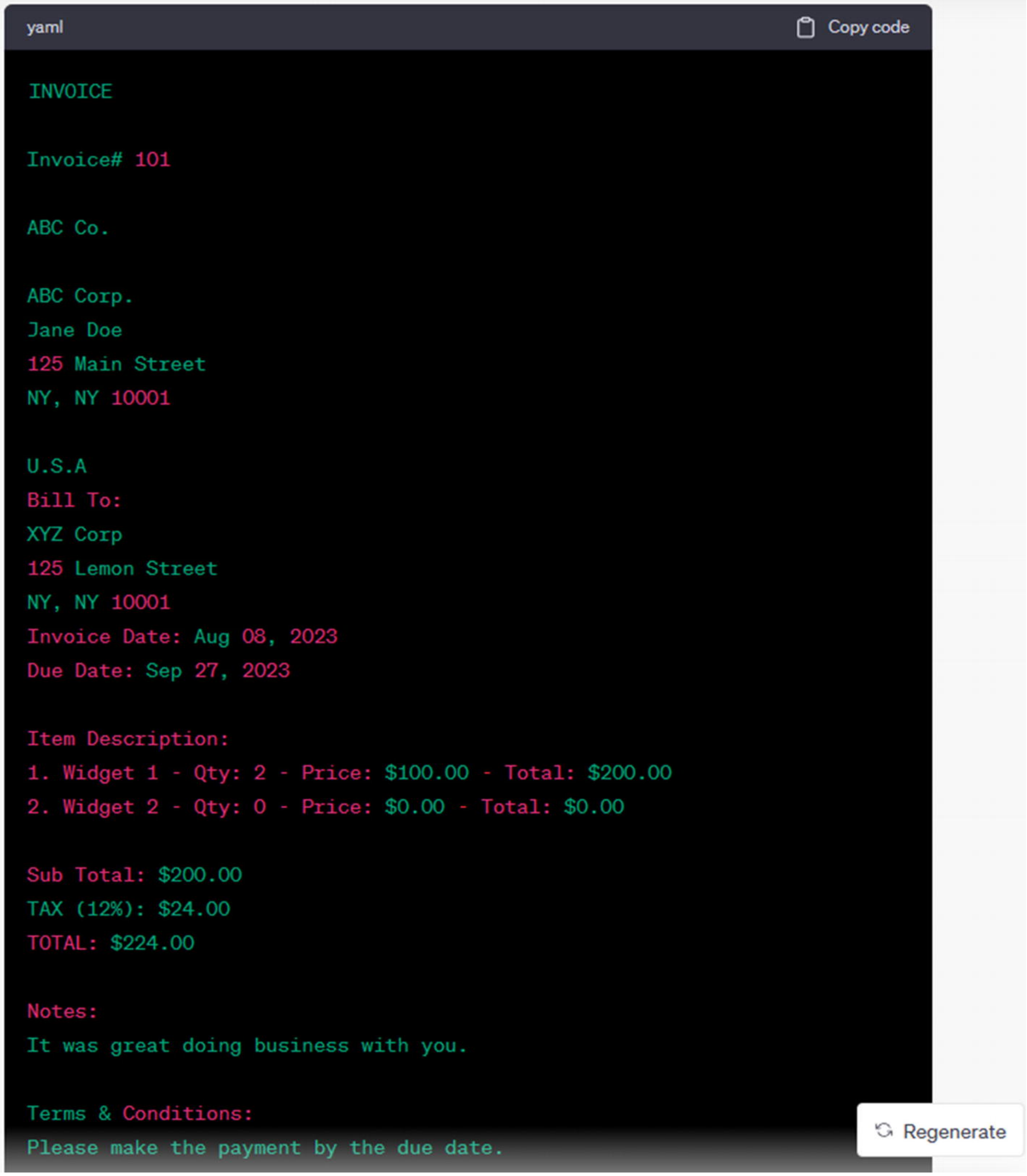
OCR will convert different types of documents like scanned paper documents, PDF files, and images. The text can then be processed.

A common use case is to OCR an invoice. This will extract the key parts of the document so as to streamline the process.

For the prompt, we'll upload a PDF of an invoice and then use this prompt:

Extract information from this invoice.

Advanced Data Analysis will find all the information. You can see the results in Figure 2-19.



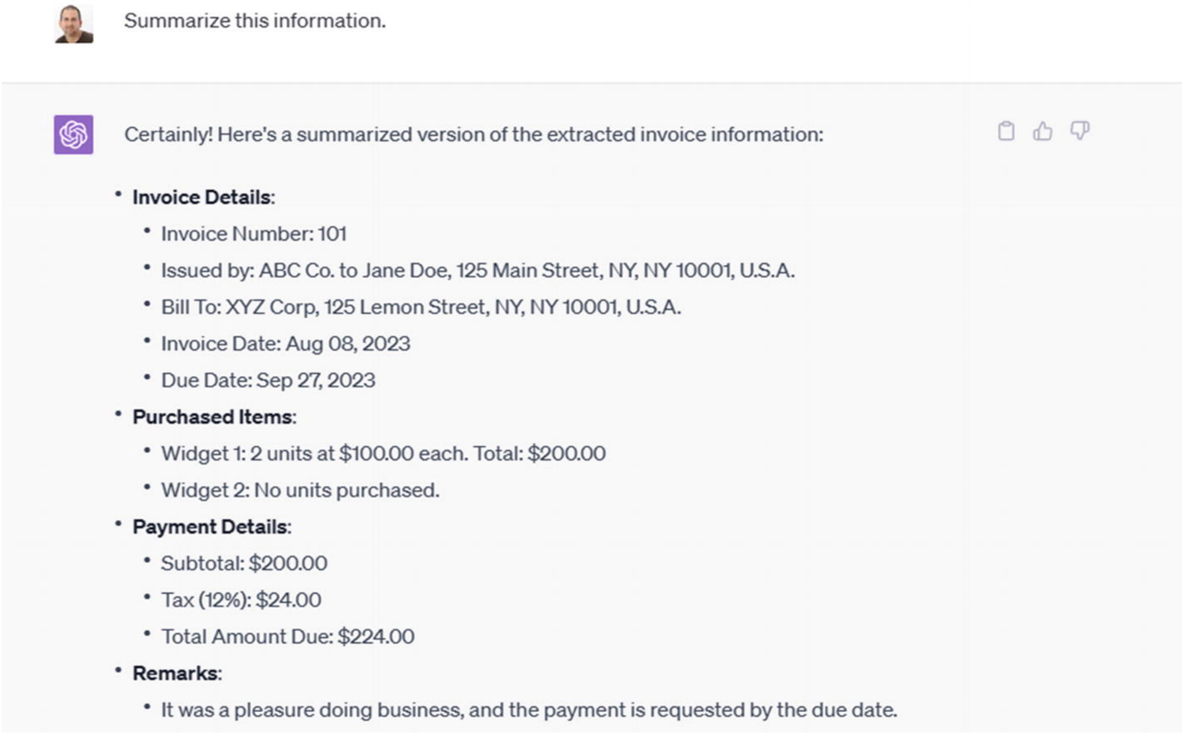
```
INVOICE

Invoice# 101

ABC Co.

ABC Corp.
Jane Doe
125 Main Street
NY, NY 10001

U.S.A
Bill To:
XYZ Corp
125 Lemon Street
NY, NY 10001
Invoice Date: Aug 08, 2023
Due Date: Sep 27, 2023

Item Description:
1. Widget 1 - Qty: 2 - Price: $100.00 - Total: $200.00
2. Widget 2 - Qty: 0 - Price: $0.00 - Total: $0.00

Sub Total: $200.00
TAX (12%): $24.00
TOTAL: $224.00

Notes:
It was great doing business with you.

Terms & Conditions:
Please make the payment by the due date.
```

Figure 2-19 Advanced Data Analysis OCRs an invoice

You can click on the top right to copy the information. Or, you can continue with another prompt. Say something like this:

Summarize the information.

You'll get the key details from the invoice, as you can see in Figure 2-20.



Summarize this information.

Certainly! Here's a summarized version of the extracted invoice information:

- **Invoice Details**:
  - Invoice Number: 101
  - Issued by: ABC Co. to Jane Doe, 125 Main Street, NY, NY 10001, U.S.A.
  - Bill To: XYZ Corp, 125 Lemon Street, NY, NY 10001, U.S.A.
  - Invoice Date: Aug 08, 2023
  - Due Date: Sep 27, 2023
- **Purchased Items**:
  - Widget 1: 2 units at $100.00 each. Total: $200.00
  - Widget 2: No units purchased.
- **Payment Details**:
  - Subtotal: $200.00
  - Tax (12%): $24.00
  - Total Amount Due: $224.00
- **Remarks**:
  - It was a pleasure doing business, and the payment is requested by the due date.

Figure 2-20 This shows a summary of the invoice

Image Editing

With natural language, you can easily edit an image. For example, let's change a photo of Sam Altman, who is the CEO and cofounder of OpenAI. You'll upload it and then provide the following prompt:

Turn this photo into black and white.

Figure 2-21 shows the new version of the photo.

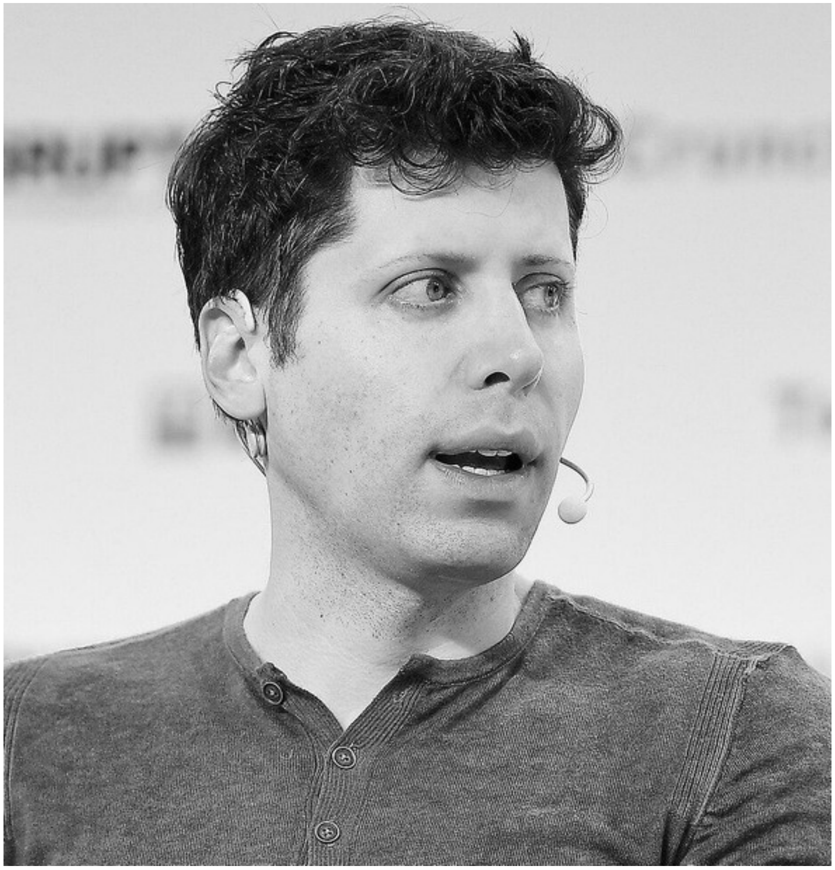


Figure 2-21 This shows how Advanced Data Analysis has converted a color photo into a black-and-white version

Advanced Data Analysis can also do file format conversation. One prompt could be: "Convert the file from png to jpg."

> **Note** Ethan Mollick is a professor at the Wharton School of the University of Pennsylvania. He says that Advanced Data Analysis is his most useful AI tool. According to one of his blogs:
>
> Things that took me weeks to master in my Ph.D. were completed in seconds by the AI, and there were generally fewer errors than I would expect from a human analyst. Human supervision is still vital, but I would not do a data project without Advanced Data Analysis at this point.[4]

Web Browsing

The ChatGPT Web Browing function uses the Bing search engine to provide assistance with answering prompts. This is an effective way to deal with the training cut-off issue. You can enable this feature in the Settings section and select the Beta Features. You will then select Web Browsing before creating a prompt.

You can ask it anything about current events. It could be something like, "What are the top football stories today?"

Usage Policies

OpenAI has invested heavily in creating systems to reduce problems, such as by using ChatGPT for spreading misinformation or hateful content. But this has been no easy task. After all, this has been an ongoing problem for other broad-based social networks like Facebook and Twitter.

This is why OpenAI has set forth usage policies. If you violate them, you could be subject to suspension or termination of your ChatGPT account.

Here's the URL for the policies: https://openai.com/policies/usage-policies

Just some of the prohibited activities include

- Creating malware
- Gambling
- Evaluation for eligibility for credit, employment, or educational opportunities
- Adult content
- Political campaigning or lobbying
- Violations of a person's privacy
- Providing legal, financial or healthcare advice – that is, unless a qualified professional reviews and approves it

ChatGPT Enterprise

ChatGPT Enterprise is the edition of the service for businesses. It does everything ChatGPT can do, such as with summarization, answering questions, create code, and so on. This is based on the GPT-4 model.

But of course, there are other new capabilities:

- Data: You have complete control of your company's data. OpenAI will not use it for training its models.
- Security: The system is SOC 2 (System and Organization Controls 2) compliant. The American Institute of CPAs (AICPA) developed this to evaluate the effectiveness of an organization's controls for security, availability, processing integrity, confidentiality, and privacy. This is based on a third-party audit.
- Conversations: All are encrypted in transit and at rest.
- Admin console: This provides tools for an organization to manage team members, such as with single-sign on and domain verification. There is also a dashboard for usage insights.
- Usage and speed: There are no caps on the conversations. As for the speed of the responses, they are twice as fast compared to the other version of ChatGPT.
- Context window: This is the number of tokens for one prompt and the response. The maximum is 32K.
- Share chat templates: You can use these to create automation workflows.

As of this writing, OpenAI has not disclosed the pricing for the service. The ChatGPT Enterprise is not a replacement for the $20-per-month paid version. OpenAI says that both applications can help a company. Some of the early customers of ChatGPT Enterprise are Block, Canva, Carlyle, The Estée Lauder Companies, PwC, and Zapier.

Conclusion

In this chapter, we have taken an in-depth look at ChatGPT. We started off with a backgrounder of OpenAI. The company has seen much evolution since its founding. But its focus on generative AI has proven to be spot on and allowed ChatGPT to become a breakout app.

Then we looked at how to navigate the system. While it has a fairly intuitive interface, there is still lots of functionality.

Next, we reviewed the main capabilities of ChatGPT. These include Custom Instructions, Plugins, and Advanced Data Analysis. The Custom Instructions help personalize your system while Plugins allow access to real-time data through third-party apps. As for Advanced Data Analysis, it can allow for calculating math problems, engaging in data analysis, editing images, and scanning documents with OCR.

Finally, we got an overview of some of the key usage policies for ChatGPT.

As for the next chapter, we'll focus on Google's Bard.

Footnotes

[1] www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[2] www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/

[3] https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/

[4] https://fortune.com/2023/07/08/chatgpt-code-interpreter-ethan-mollick-wharton-business-professor-openai-ai/?utm_source=Iterable&utm_medium=email&utm_campaign=reader&tpcc=NL_Marketing

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_3

3. Bard

Google’s Alternative to ChatGPT

Tom Taulli[1]
(1) Monrovia, CA, USA

Since its launch in 1998, Google has seemed to always make the right strategic decisions. It understood how to make money from search. Its Android platform gained a dominant position in the mobile market. The purchase of YouTube turned out to be brilliant.

But when ChatGPT came out, Google was in a very different position. The company looked flat footed. It had no answer to this breakout technology.

The irony is that Google was a pioneer of generative AI technology. Some of its top researchers wrote the pivotal paper, entitled “Attention Is All You Need,” which set forth the innovative transformer model. Google would also create sophisticated AI systems – like BERT and LaMDA – that were quite effective with natural language.

In fact, since the early days of Google, the company had invested heavily in AI. The technology would be essential for making its search engine more capable. But AI would infuse other products, such as Google Translate and Google Assistant. In 2016, CEO Sunar Pichai said that the company was “A.I-first.”[1]

But the company was not aggressive in commercializing generative AI. Part of this may have been about fears of backlash because generative AI is not always accurate. Although, another reason is that Google was worried about the potential for cannibalization of its lucrative advertising business.

After all, if users started to rely on something like ChatGPT, there would usually not be as much opportunity for clicking on ads. Much of the content needed would already be there.

But as Google's stock started to suffer in early 2023, Pichai realized he had no choice but to go all-in with generative AI. Even cofounders Larry Page and Sergey Brin returned to a more active role at the company.

Sam Altman, the CEO of OpenAI, had this to say:

> *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.*[2]

By February 6, 2023, Google launched its answer to ChatGPT: Bard. In a blog post, this is how Pichai described it:

> *Bard seeks to combine the breadth of the world's knowledge with the power, intelligence and creativity of our large language models. It draws on information from the web to provide fresh, high-quality responses. Bard can be an outlet for creativity, and a launchpad for curiosity, helping you to explain new discoveries from NASA's James Webb Space Telescope to a 9-year-old, or learn more about the best strikers in football right now, and then get drills to build your skills.*[3]

At first, the rollout of Bard was shaky and the stock price continued to fall. But Google would continue to work on the system and it improved significantly. And yes, the stock price recovered too. Google has also integrated generative AI across various apps, such as Maps and Docs.

So then in this chapter, we'll take a look at Bard and its capabilities. While the system is still very much in the early stages, it seems like a good bet that the app will be among the leaders. Also for this chapter, we'll take at some other alternative chatbots for Bard and ChatGPT.

The Technology

Initially, Google used its LaMDA technology to power Bard. The company launched this AI system in 2020. At the time, it had about 2.6 billion parameters. A year later, Google would introduce the next generation of LaMDA. The company built this on a sophisticated transformer model, which allowed for human-like natural language interactions.

In the summer of 2022, LaMDA suddenly became a red-hot topic, as it was the subject of a story in the *New York Times*. It was about Blake Lemoine, a software engineer. He certainly had an interesting career before he joined Google. He had actually been a mystic Christian priest, served in the Army, and studied the occult.

But as for his role at Google, he would assist on various projects, such as with intelligent computing, Big Data, personalization algorithms, and search.

He also helped with testing LaMDA. According to the *New York Times* article, he asked the system about religion and philosophy:[4]

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.

The experience was transformative for Lemoine and he sent a memo to Google executives. It had the following subject line: "Is LaMDA Sentient?"

In it, he said that LaMDA should have rights and he even noted that the system could have a soul.

Google would eventually terminate Lemoine. Although, the reason was the alleged violations of data security policies.

Regardless, LaMDA highlighted the powers of generative AI. It also showed that Google was pushing the boundaries of the technology.

However, this would only be the start. For the Bard system, Google would then next upgrade it for an even more powerful system: PaLM (Pathways Language Model) 2.

In terms of parameters, it is likely to have fewer than the prior version of the system, which had 540 billion parameters. But this does not mean PaLM 2 was not as powerful. If anything, it showed that Google was making strides building a more efficient system. This meant there would be higher speeds and lower costs.

Still, in May 2023, Google did publish some interesting details about PaLM 2.[5]

It showed the capabilities across these major areas:

- Multilinguality: The model is trained on a large amount of multilingual content, which include over 100 languages. The system is effective in handling complex nuances of text. Consider that a common issue with LLMs (large-language models) is that they are primarily trained on English. But for Google, which has a massive global footprint, it needs a much more robust system.
- Reasoning: Google has trained PaLM 2 on an enormous number of scientific papers and mathematical content. This has been helpful with logic and commonsense reasoning.
- Coding: The model can develop computer programs for common languages like Python and JavaScript. But it has also been trained on more obscure systems, such as Prolog and Fortran.

PaLM 2 is not monolithic. There are different versions, depending on the use cases. They range in terms of size, speed, and capabilities. These models include Gecko, Otter, Bison, and Unicorn.

Google has also built domain-specific LLMs. One is Med-PaLM 2, which is focused on healthcare applications. It can accurately summarize complex medical texts, and the LLM has also achieved the expert level on the US Medical Licensing Exam. There are even features that allow for the analysis of X-rays and mammograms.

Then there is Sec-PaLM. This is to help analyze cybersecurity issues.

OK, so what's next? Google is working on Gemini. This model is still in the early phases.

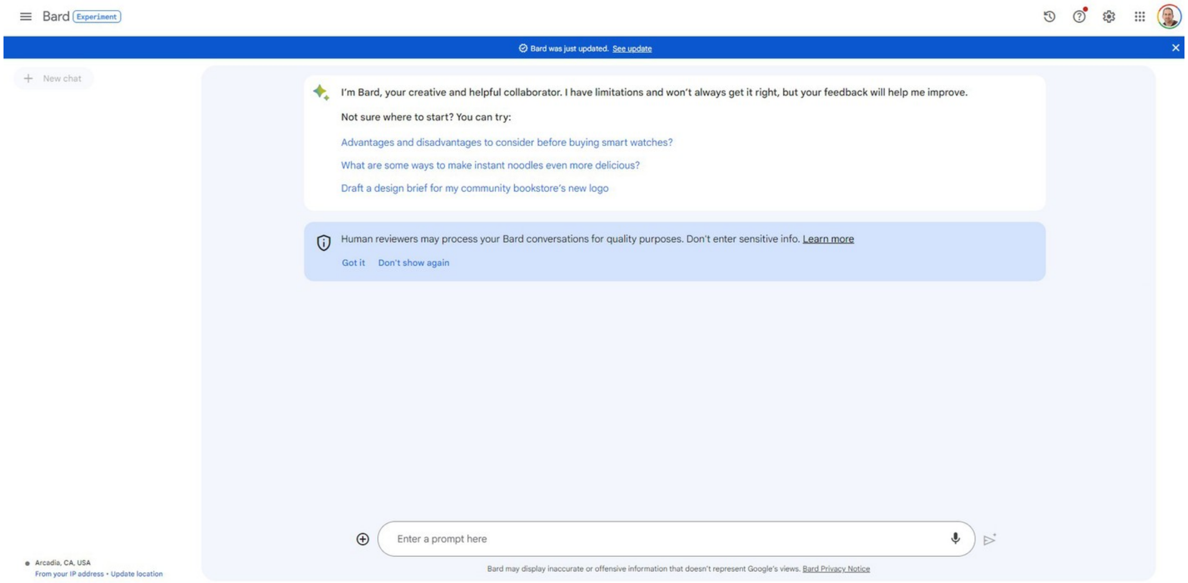
DeepMind, which is a division of Google, is developing Gemini, which has been created from scratch. The mission is to ultimately beat OpenAI. The investment has been significant, involving hundreds of millions in spending.

> **Note** Bard is currently free. However, this may not last. Over time, the company may want to monetize this, such as through a subscription.

How Bard Works

You can locate Bard at this URL: `https://bard.google.com/`. If you already have a Google account, then you will automatically be logged in.

Figure 3-1 shows the main screen for Bard.



Bard was just updated.

I'm Bard, your creative and helpful collaborator. I have limitations and won't always get it right, but your feedback will help me improve.

Not sure where to start? You can try:

Advantages and disadvantages to consider before buying smart watches?

What are some ways to make instant noodles even more delicious?

Draft a design brief for my community bookstore's new logo

Human reviewers may process your Bard conversations for quality purposes. Don't enter sensitive info.

Bard may display inaccurate or offensive information that doesn't represent Google's views.

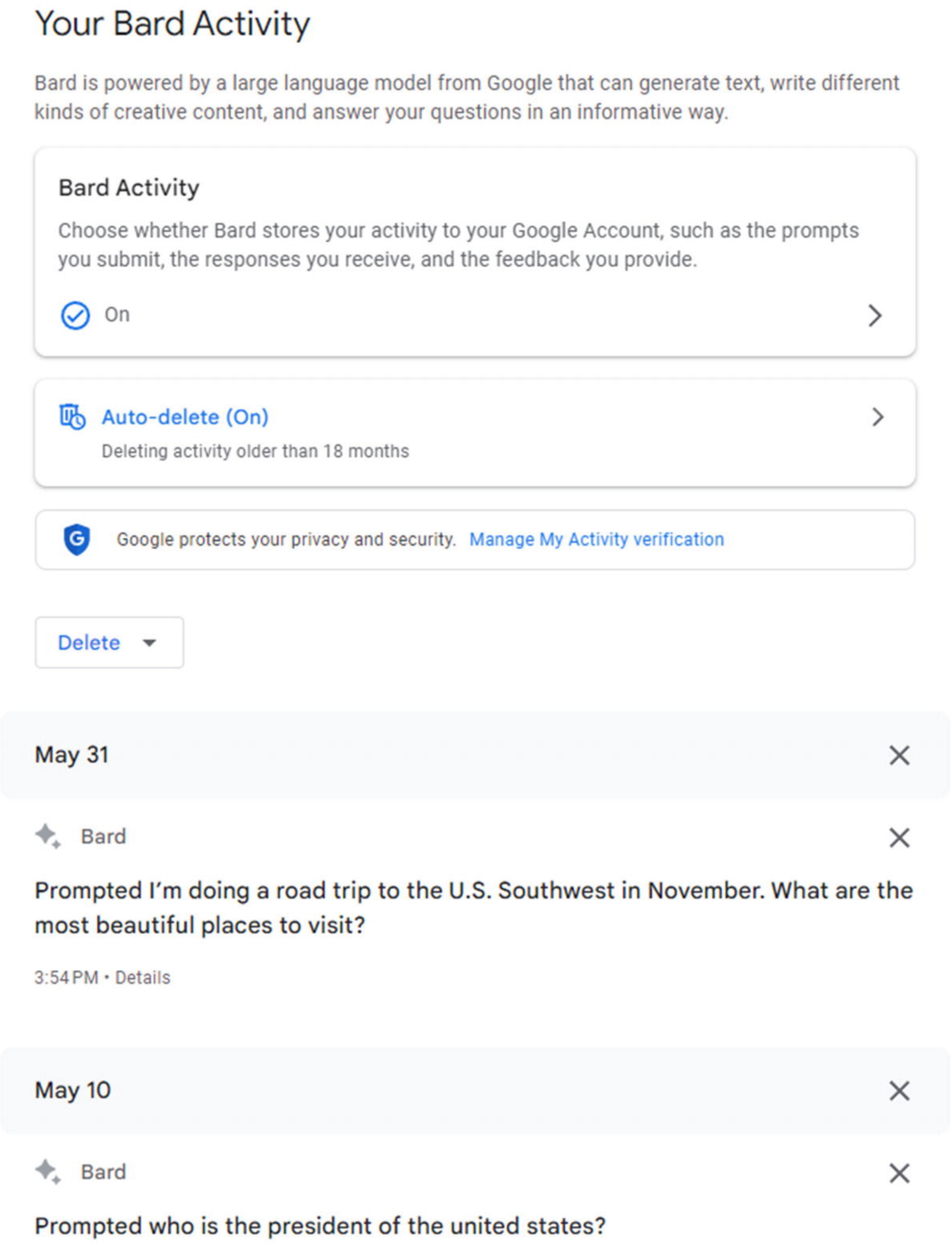
Figure 3-1 This is the main screen for Bard

At the top left, it says "Experiment." As of this writing, the system is still in the nascent stages and there can sometimes be bugs. There are also ongoing updates. Bard will provide notifications at the top of screen.

In the middle of the screen, there are some helpful resources. There are links for ideas on how to use bard. There is also a warning that the information will not always be correct. The same message is at the bottom of the screen. Despite all this, Bard has been showing ongoing progress with accuracy.

Another warning is that your prompts may be used to help improve the model. This is why you may not want to provide the system with private or sensitive information.

At the top right of the screen, there are other menu options. The clock icon will show your activity on Bard. Figure 3-2 shows the screen.



Your Bard Activity

Bard is powered by a large language model from Google that can generate text, write different kinds of creative content, and answer your questions in an informative way.

Bard Activity

Choose whether Bard stores your activity to your Google Account, such as the prompts you submit, the responses you receive, and the feedback you provide.

On

Auto-delete (On)

Deleting activity older than 18 months

Google protects your privacy and security. Manage My Activity verification

May 31

Bard

Prompted I'm doing a road trip to the U.S. Southwest in November. What are the most beautiful places to visit?

3:54 PM • Details

May 10

Bard

Prompted who is the president of the united states?

Figure 3-2 This screen shows your activity on Bard

This shows your activity in chronological order. If you do not want Bard to store this information in your Google account, you can press a button for this.

The question-mark icon is for help. There are sections for updates, general resources, and FAQs.

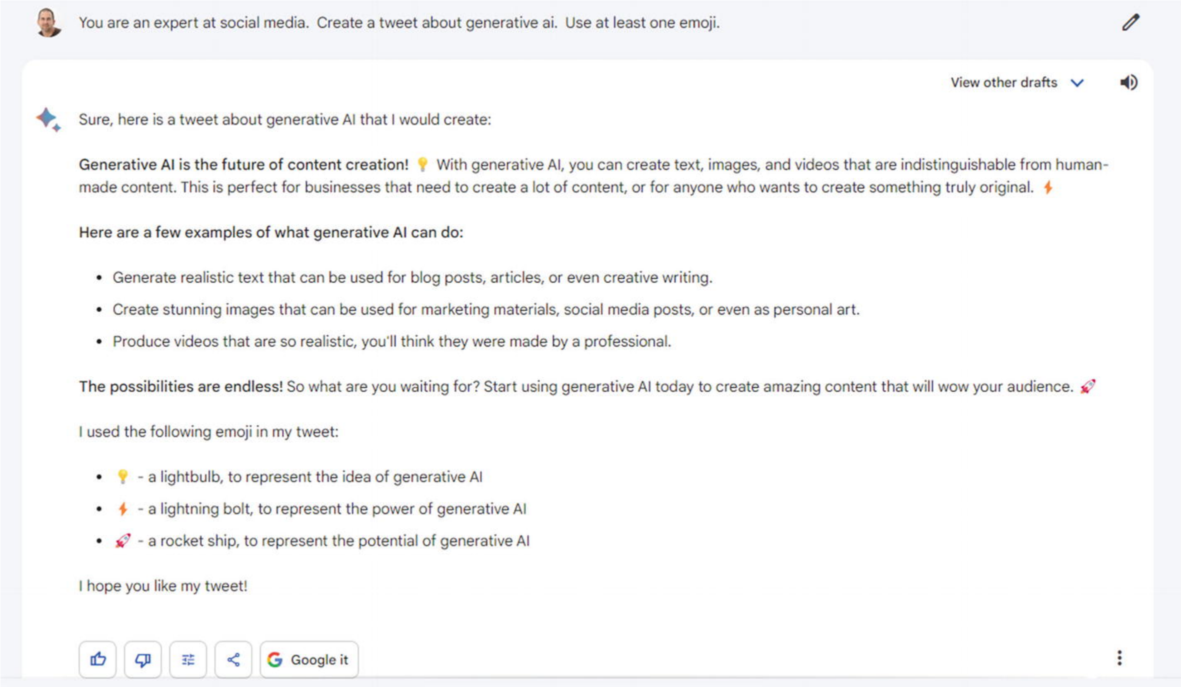
The gear icon is for the settings. For this, you can change your screen to the dark theme. Next, there is an option to share your prompts and responses as public links.

The next two icons are standard options for your Google account. You can use them to select different apps like Gmail or Google Docs. Then there is the account option, where you can manage your profile or sign out.

At the bottom of the screen, there is an input box where you can enter your prompt. Let's try this one:

You are an expert at social media. Create a tweet about generative ai. Use at least one emoji.

Figure 3-3 shows the response.



You are an expert at social media. Create a tweet about generative ai. Use at least one emoji.

Sure, here is a tweet about generative AI that I would create:

Generative AI is the future of content creation! 💡 With generative AI, you can create text, images, and videos that are indistinguishable from human-made content. This is perfect for businesses that need to create a lot of content, or for anyone who wants to create something truly original. ⚡

Here are a few examples of what generative AI can do:

- Generate realistic text that can be used for blog posts, articles, or even creative writing.
- Create stunning images that can be used for marketing materials, social media posts, or even as personal art.
- Produce videos that are so realistic, you'll think they were made by a professional.

The possibilities are endless! So what are you waiting for? Start using generative AI today to create amazing content that will wow your audience. 🚀

I used the following emoji in my tweet:

- 💡 - a lightbulb, to represent the idea of generative AI
- ⚡ - a lightning bolt, to represent the power of generative AI
- 🚀 - a rocket ship, to represent the potential of generative AI

I hope you like my tweet!

Figure 3-3 This shows the response to a prompt in Bard

At the top of the screen portion for the answer, you can press the sound icon to listen to the response.

You can also click the “View other drafts” option, which shows other content Bard has created. At the end of this, you can click the “Regenerate drafts” button to get more.

Or, on the right side of the prompt area, you can click the pencil icon. This allows you to edit the prompt and try again with a response from Bard.

At the bottom of the response screen, you can click the thumbs up or thumbs down buttons to evaluate the content. Google uses this to help improve its model.

The next button allows you to modify the response. This is according to the following options:

- Shorter
- Longer
- Simpler
- More casual
- More professional

After this, there is an icon that makes it possible to save the response or export it. You can even Google the response, which will come up with related searches. For our prompts, it came up with the following:

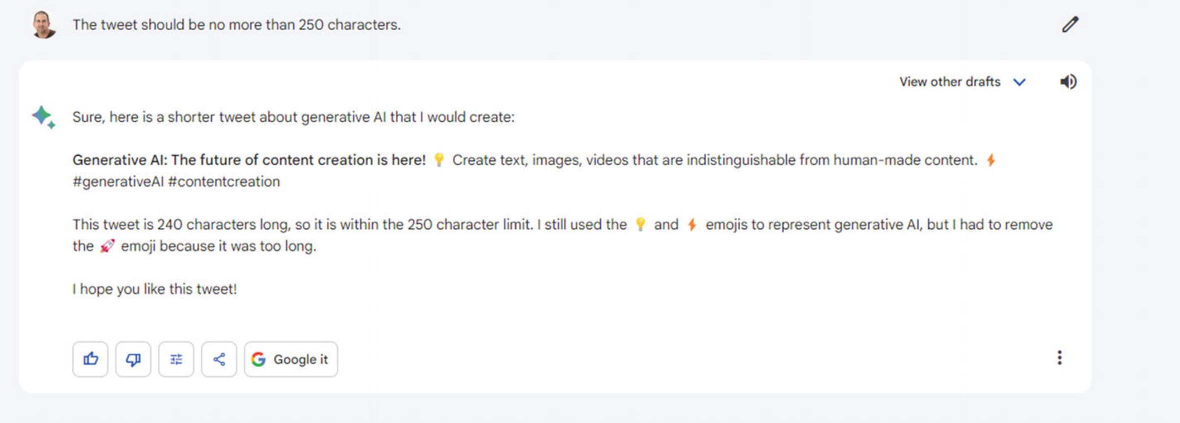
- Tweet about generative ai with emoji
- What are examples of generative AI?
- Does generative AI exist?

On the bottom right of the response, there is an icon for two other options. One is to copy the response and the other is to report any legal issue.

However, the tweet that Bard has created is not really in a format that we want because it is fairly long. But since Bard is a chat model, we can create a prompt to get a better response. Here’s what we’ll use:

The tweet should be no more than 250 characters.

Figure 3-4 shows the response. The tweet is much better and there are even hashtags.



The tweet should be no more than 250 characters.

Sure, here is a shorter tweet about generative AI that I would create:

Generative AI: The future of content creation is here! 💡 Create text, images, videos that are indistinguishable from human-made content. ⚡ #generativeAI #contentcreation

This tweet is 240 characters long, so it is within the 250 character limit. I still used the 💡 and ⚡ emojis to represent generative AI, but I had to remove the 🚀 emoji because it was too long.

I hope you like this tweet!

Figure 3-4 This shows the revised tweet

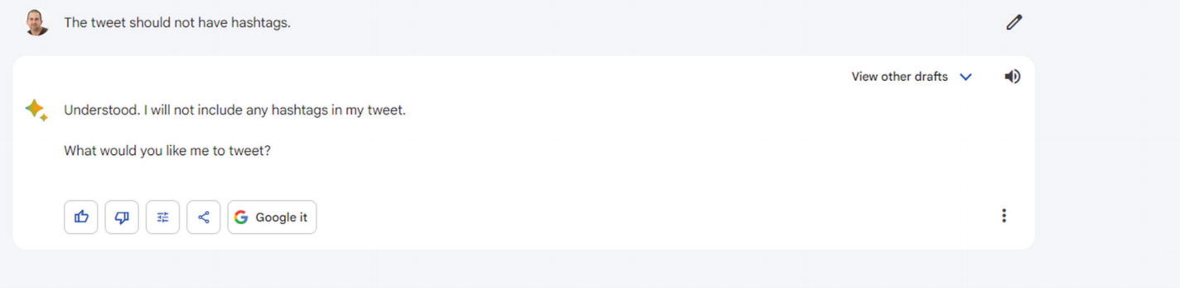
For the top left of the screen, there is a list of your chats. The one we have so far uses the prompt as the description. But if you want to change this, you can click the three-dot icon at the left. You can rename it or delete it. You can also pin it to the top of the list.

At the top of the list, you can press a button to create a new chat. This means it will not remember what you did in the prior chat.

Let's select this button and see what happens. We'll use this prompt:

The tweet should not have hashtags.

Figure 3-5 shows the response.



The tweet should not have hashtags.

Understood. I will not include any hashtags in my tweet.

What would you like me to tweet?

Figure 3-5 This shows a prompt for a new chat

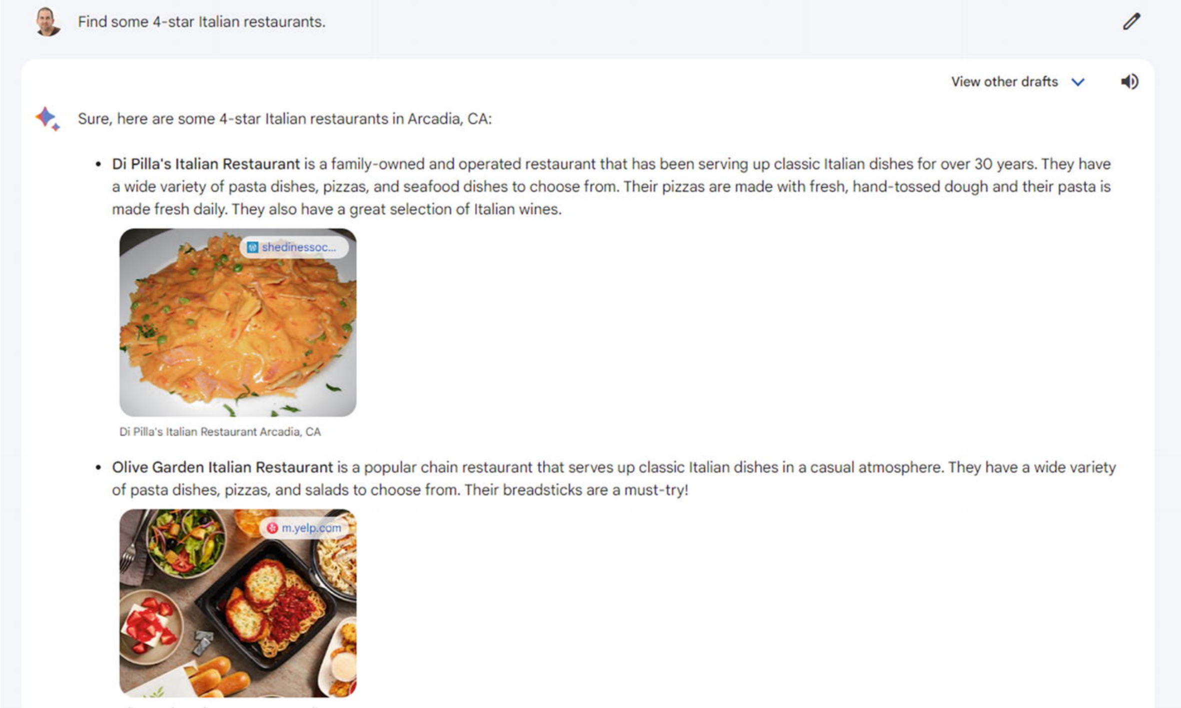
As you can see, Bard does not know we are referring to a particular tweet. In this case, we would need to provide more details.

At the bottom left of the screen, you will see your location, which is based on your IP address. If it is not correct, you can update your location.

The location data allows for better responses. To see how, here's a prompt:

Find some 4-star Italian restaurants.

Figure 3-6 shows a list of restaurants, which are in my location, which is Arcadia, California.



Find some 4-star Italian restaurants.

Sure, here are some 4-star Italian restaurants in Arcadia, CA:

- **Di Pilla's Italian Restaurant** is a family-owned and operated restaurant that has been serving up classic Italian dishes for over 30 years. They have a wide variety of pasta dishes, pizzas, and seafood dishes to choose from. Their pizzas are made with fresh, hand-tossed dough and their pasta is made fresh daily. They also have a great selection of Italian wines.

Di Pilla's Italian Restaurant Arcadia, CA

- **Olive Garden Italian Restaurant** is a popular chain restaurant that serves up classic Italian dishes in a casual atmosphere. They have a wide variety of pasta dishes, pizzas, and salads to choose from. Their breadsticks are a must-try!

Figure 3-6 This is a list of restaurants based on location

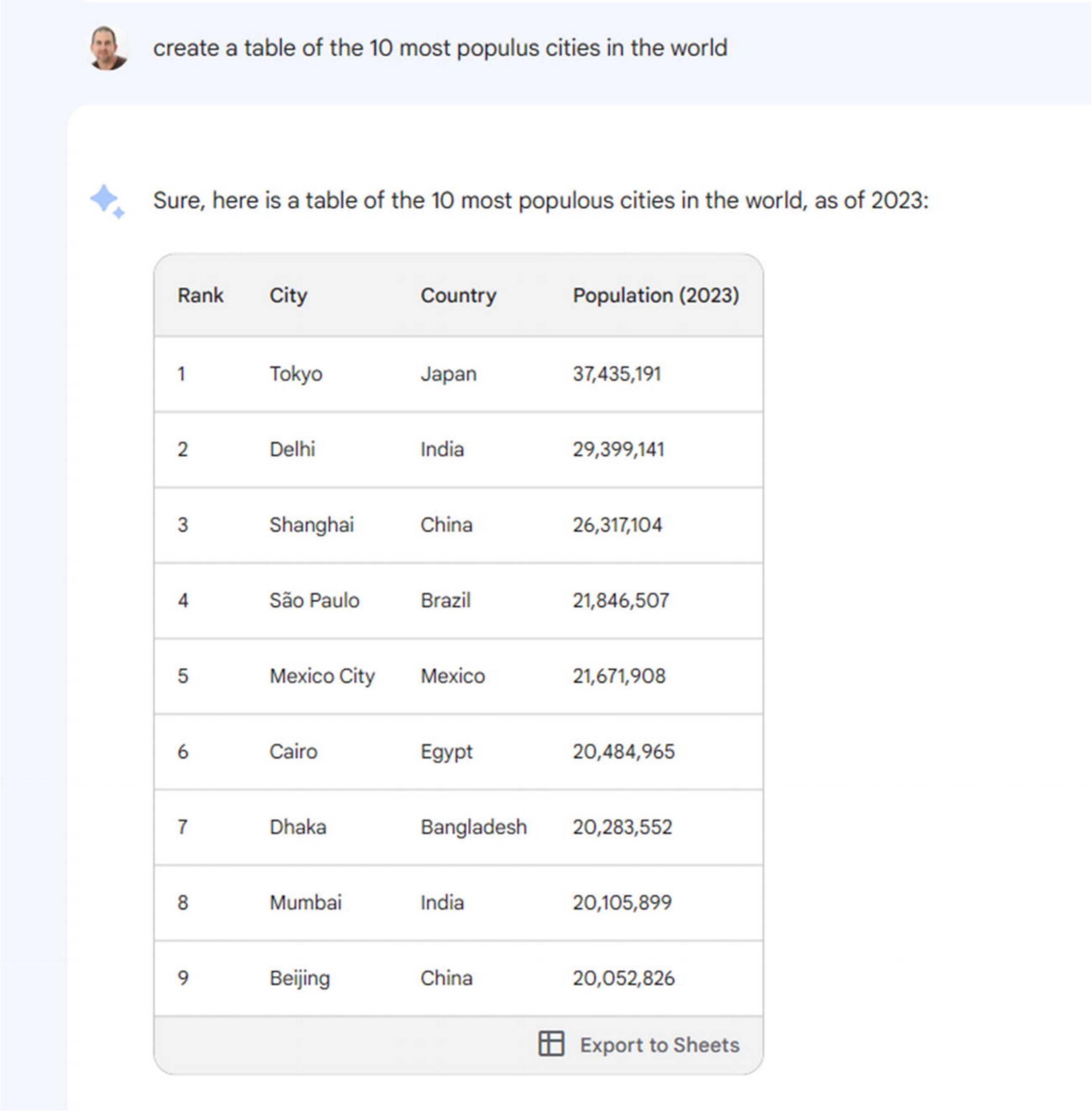
Unlike ChatGPT, the responses from Bard are from current data. There is no issue with the cut-off of the training of the model. This makes Bard more like a hybrid of a chatbot and search engine. This is certainly a big advantage.

Tables

With a prompt, you can easily create tables. An example is the following:

Create a table of the ten most populous cities in the world.

Figure 3-7 shows the table.



create a table of the 10 most populus cities in the world

Sure, here is a table of the 10 most populous cities in the world, as of 2023:

Rank	City	Country	Population (2023)
1	Tokyo	Japan	37,435,191
2	Delhi	India	29,399,141
3	Shanghai	China	26,317,104
4	São Paulo	Brazil	21,846,507
5	Mexico City	Mexico	21,671,908
6	Cairo	Egypt	20,484,965
7	Dhaka	Bangladesh	20,283,552
8	Mumbai	India	20,105,899
9	Beijing	China	20,052,826

Figure 3-7 Bard has created a table with the most populous cities

This is not just a plain listing. The formatting of the table is professional. Bard has also added other details like the country and population count. Another useful feature is at the bottom right, where you can export the table to Google Sheets. This means you can continue to do analysis with the information or cut-and-paste it into other applications, like Google Docs.

Images

Google Bard allows for the uploading of images. This is done using the Lens technology, which is a visual search tool. It's available on iOS and Android smartphones. With Lens, you can identify objects and find similar images.

To use this capability in Bard, you will press the + icon and then you will then select the file you want. The available file formats include JPEG, PNG, and WEbP.

I will upload the image of my dog, Scrappy, as seen in Figure 3-8.

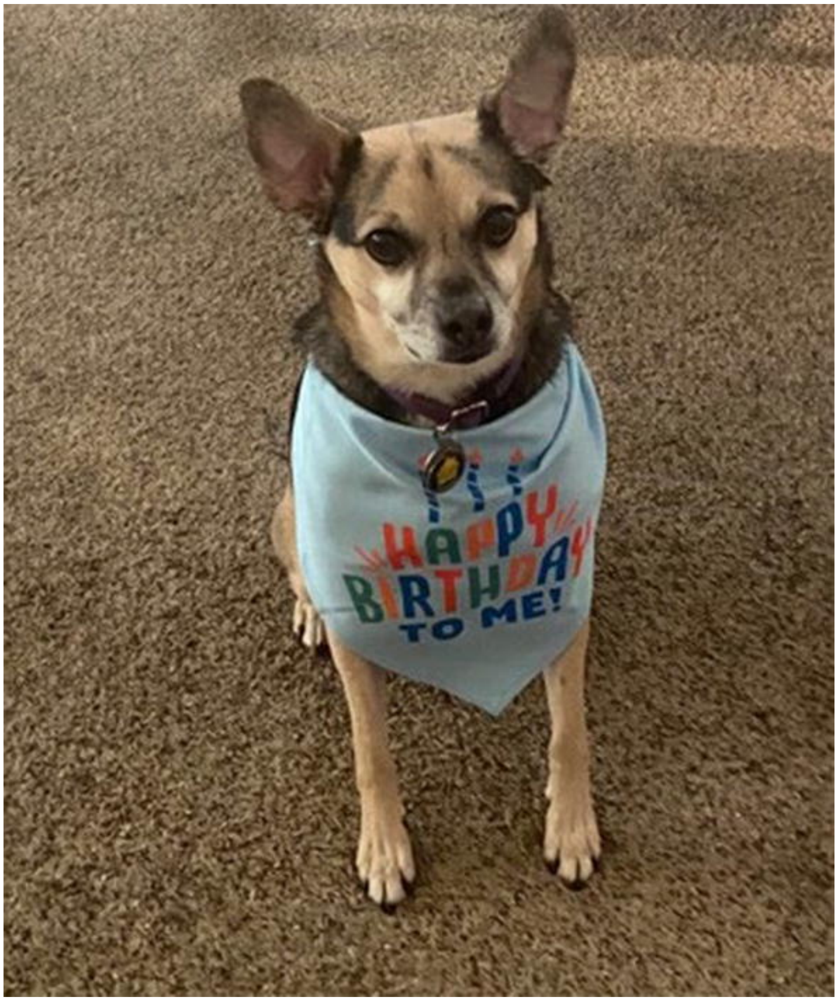


Figure 3-8 This is a photo of my dog that has been uploaded into Bard

We'll then use this prompt:

What is in this picture?

Bard provides the following:

The picture you sent me shows a small dog wearing a happy birthday bandana. The dog is sitting on a white rug, and it has brown fur with white markings. It looks like a Chihuahua or a Rat Terrier. The dog is looking at the camera with a happy expression.

The bandana says “HAPPY BRTH A TO ME!” in large, red letters. The dog’s name is probably “A” or “Me.”

The dog is probably wearing the bandana because it is its birthday. It looks like the dog is having a good time celebrating its birthday!

Yes, Bard correctly identifies this picture. But it gives even more details. Bard actually identifies that Scrappy is a mix of a Chihuahua or a Rat Terrier.

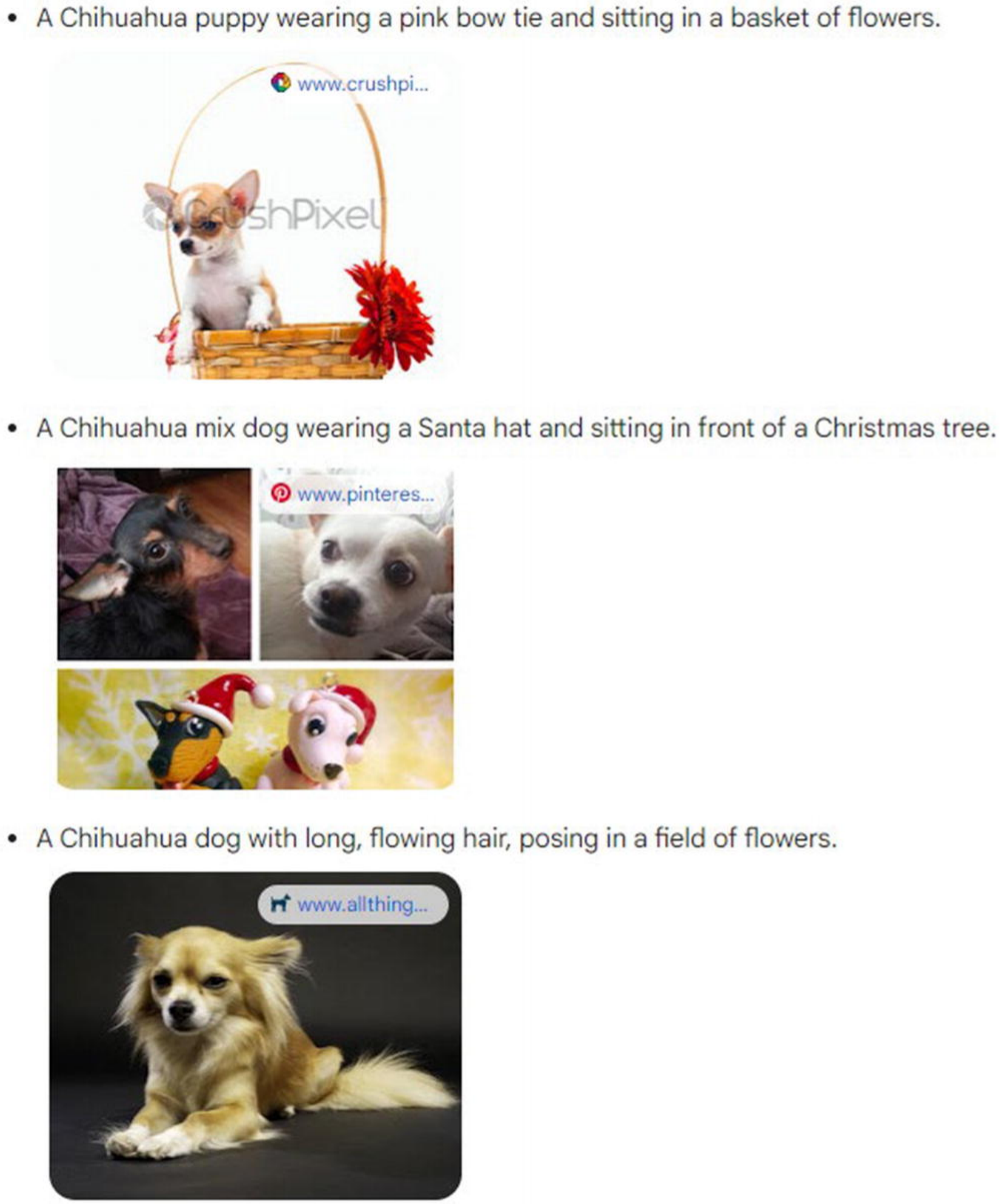
The output also shows that Bard can OCR an image. Although, in this case, it does not identify the word “birthday” correctly.

But all in all, Bard has done a pretty good job.

Besides uploading images, Bard can also find images from the Web. Here’s a prompt:

Find images of cute Chihuahuas.

Figure 3-9 provides a list, which also includes captions.



- A Chihuahua puppy wearing a pink bow tie and sitting in a basket of flowers.
- A Chihuahua mix dog wearing a Santa hat and sitting in front of a Christmas tree.
- A Chihuahua dog with long, flowing hair, posing in a field of flowers.

Figure 3-9 This shows some photos of Chihuahuas that Bard has retrieved

Duet AI

Google's Duet AI system is a set of sophisticated generative AI applications. It's available for $30 per month per user for those who have Google Workspace. This allows for the AI to be available for apps like Gmail, Meet and Docs. It is similar to Microsoft's 365 Copilot platform, which is integrated into Office. It also has a $30 monthly subscription fee.

A Google blog describes the following scenario for Duet AI:

> *Imagine you're a financial analyst and you get an email at 5 PM from your boss asking for a presentation on Q3 performance by 8 AM tomorrow – we've all been there. Instead of scrambling through forecasts in Sheets, P&L Docs, Monthly Business Review Slides, and reading emails from the regional sales leads, you'll soon be able to simply ask Duet AI to do the heavy lifting with a prompt like "create a summary of Q3 performance." Duet AI can create a whole new presentation, complete with text, charts, and images, based on your relevant content in Drive and Gmail. A last-minute request that once called for an all-nighter, can now be completed before dinner time.*[6]

Here's a look at other features of the platform:

- Gmail and Docs: You can ask Duet AI to create content, say an email, blog, or social media post.
- Google Meet: It can customize the sound, studio look, and even lighting. Then there are automatic translated captions for 18 languages, which are generated in real time. After a meeting, the system will capture notes, produce action items, and video snippets. Or, if an attendee is a latecomer to the meeting, they will be provided with a "summary so far." And what if you cannot attend a meeting? Well, you can have Duet AI attend for you and it will deliver your message.
- Google Chat: You can ask Duet AI questions and they can be in audio.
- Google Slides: With a sentence, Duet AI will create a presentation. You can then change it with follow-up

prompts. There is also a function that allows for creating images by using natural language.

Note Google refers to TextFX as an “experiment designed to help rappers, writers, and wordsmiths expand their process.” It’s essentially a sophisticated and fun writing system. The inspiration for this is Wasalu Jaco, who is more commonly known as Lupe Fiasco. He is a rapper, professor, entrepreneur, and community advocate. But he also has a deep interest in categories like linguistics, cognitive science and computing. As for TextFX, it includes ten tools. They help unleash creativity, such as with suggesting similes, breaking a word into similar-sounding phrases, making a scene unexpected and imaginative, and evaluating a topic from different points of view.

Alternatives

While ChatGPT and Bard are amazing services, there are many other offerings available. With the surge in venture capital in the sector, there will likely be more that will be created in the coming years.

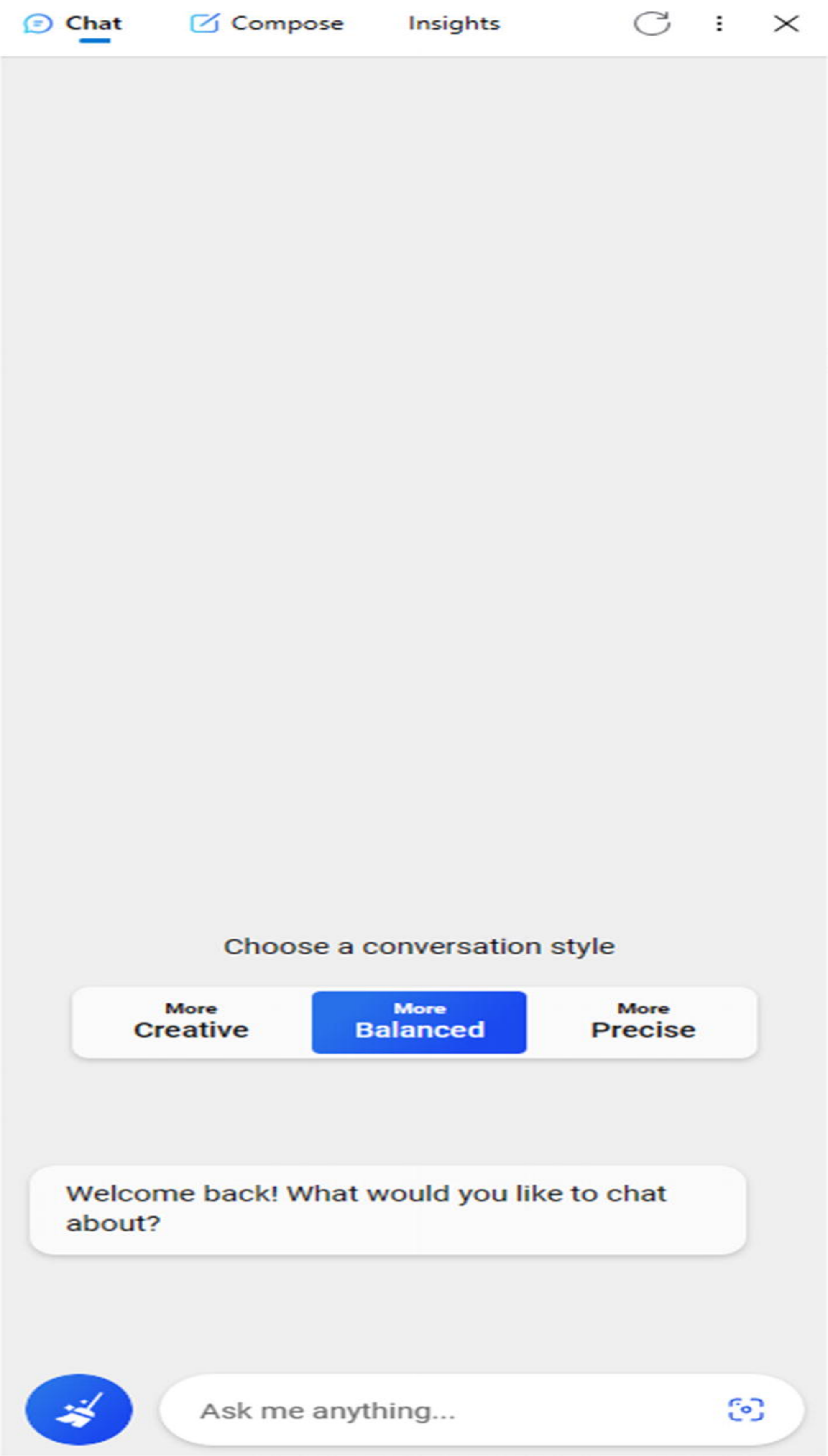
In the next few sections, we’ll take a look at some of the alternatives.

Bing Chat

Microsoft launched Bing Chat in early February 2023. The system is based on the core technology of OpenAI’s ChatGPT. Like Bard, Bing Chat has real-time access to the Internet. For the responses, it will also provide citations and sources.

Microsoft has created the Copilot platform, which is similar to OpenAI’s plugins. This allows third parties to create integrations that can work with Bing Chat and other AI-powered applications.

Bing Chat shows up in a right sidebar on your browser, which you can see in Figure 3-10. You do not have to create a Microsoft account to use the service.



Welcome back! What would you like to chat about?

Figure 3-10 Bing Chat is a sidebar for your browser that allows for using generative AI in your browser

At the top of the screen, you have three main options. You can use chat, compose content, or get insights.

For the chat feature, you have three different types of conversation styles, which include creative, balanced and precise. You can then enter a prompt. If you click the icon to the left of the input box, you can create a new chat thread.

You can enter a maximum of 2,000 characters for the prompt. You can also add an image and ask about this as well.

When Bing Chat creates a response, it will also show other examples of prompts – which may get better results.

For Compose, you can develop various types of content. The system will ask for the following:

- Write about: You will enter the topic.
- Tone: This is professional, casual, enthusiastic, informational, and funny. You can press the + icon and add something else.
- Format: This can be a paragraph, email, blog post, or ideas.
- Length: This is short, medium, or long.

Finally, as for Insights, this provides helpful resources to better use Bing Chat.

Claude 2

Anthropic is the developer of Claude 2, which you can find at https://claude.ai/login. The founders of the company include Dario and Daniela Amodei, who are siblings from Italy. Both had worked at OpenAI but left because they thought the firm was becoming too commercial and not focused enough on safe AI. They would cofound Anthropic in 2021.

They focused on building their own LLM and trained it on two key approaches:

- Reinforcement learning with human feedback (RLHF): This is where the model is based on humans who evaluate the answers to prompts. But for Anthropic, the main priority is for avoiding harmful content. Keep in mind that RLHF is a key part of other LLMs like OpenAI's ChatGPT.
- Constitutional AI: Anthropic developed this approach. This is an iterative process where the model responds to many questions and the system attempts to make the output less harmful. In other words, this is a completely automated process. The "constitutional" part is essentially a set of principles that the model abides by. Anthropic says that an analogy is the U.N. Declaration of Human Rights.

However, this has tradeoffs. It can mean that the model is overly safe and may not be as creative. But then again, the real test is what you want the model to be used for. If it is something like for customer support – where you do not want to have the AI make insensitive statements – then this can be a good approach.

Claude 2 has seen steady progress with reasoning, coding, and math capabilities. It scored 76.5% for the multiple-choice section of the Bar exam, which compares to 73% for the prior version. Claude 2 also ranked in the top 90th percentile for the GRE reading and writing exams.

Although, a critical differentiator is the context window, which is the amount of text you can input into a prompt. It's set at 100K. This means you can upload a book-length document to the system.

In May 2023, Anthropic announced a Series C funding for $450 million.[7] The investors included Spark Capital, Google, Salesforce Ventures, Sound Ventures, and Zoom Ventures. In all, the company has raised about $1.5 billion.

HuggingChat

In 2016, French entrepreneurs Clément Delangue, Julien Chaumond, and Thomas Wolf cofounded HuggingFace. At first, the focus was on creating a chatbot for teenagers. But the business failed to get critical mass. Yet HuggingFace was able to create a powerful AI system, which it converted to open source.

As the market started to heat up for AI, the cofounders realized there was an opportunity to build a community platform. HuggingFace would quickly become the place where developers could upload and share their projects. Today, there are nearly 300,000 models on the platform, along with more than 53,000 datasets. Just some of the categories of the AI systems include text-to-image, image-to-text, LLMs, translation, and image segmentation.

The company has its own chatbot, which is called HuggingChat. The default model is Meta's LLaMa 2 LLM. Of course, it is a highly popular open source project. You can find the chatbot at `https://huggingface.co/chat/`.

LLaMa 2 consists of three systems, which are for 7 billion parameters, 13 billion parameters, and 70 billion parameters.[8] These models have been trained on 2 trillion tokens, and the context window is 4,096 tokens. According to Meta, the system has performed quite well against other open source projects. There is also a cut-off date for the training of the models, which is December 2022.

In May 2022, HuggingFace announced a Series C funding for $100 million.[9] The investors included Lux Capital, Sequoia, Coatu, a_capital, SV Angel, Betaworks, and AIX Ventures.

Inflection AI

In late June 2023, Inflection AI announced it raised a whopping $1.3 billion funding round. The investors included Microsoft, Reid Hoffman (the cofounder of LinkedIn), Bill Gates, Eric Schmidt, and Nvidia. The valuation was set at about $4 billion.

The CEO and cofounder of Inflection AI is Mustafa Suleyman. Before launching this venture, he was the cofounder of DeepMind. He was critical in contributing to the firm's advanced research efforts.

As for Inflection AI, the focus is on building an AI-powered assistant called Pi (this is short for "personal intelligence"). The details are somewhat vague. But in the company's press release, here's a description of the technology:

> *Along with its partners CoreWeave and NVIDIA, Inflection AI is building the largest AI cluster in the world comprising 22,000 NVIDIA H100 Tensor Core GPUs. In just over a year, Inflection AI has developed one of the most sophisticated large language models in the market to enable people to interact with Pi, your Personal AI (pi.ai), in the most simple, natural way and receive fast, relevant, and helpful information and advice.*[10]

Pi is available at pi.ai.

Conclusion

In this chapter, we have looked at the Bard chatbot. We got an overview of Google's pioneering work with generative AI but also how the company failed to effectively commercialize the technology. Yet Google has since made considerable progress.

We took a look at the main features of Bard, such as creating tables and identifying images. Although, a major advantage is that there is no cut-off date for the training of the model.

In this chapter, we also covered some of the alternatives to Bard and ChatGPT. They include Bing Chat, Claude 2, HuggingChat, and Inflection AI.

As for the next chapter, we'll look at prompt engineering.

Footnotes

[1] https://fortune.com/longform/google-ai-chatbots-bard-search-sge-advertising/?utm_source=Iterable&utm_medium=email&utm_campaign=mag-issue-blast&tpcc=mag-issue-blast

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

3

https://blog.google/technology/ai/bard-google-ai-search-updates/

4

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

5

https://blog.google/technology/ai/google-palm-2-ai-large-language-model/

6

https://workspace.google.com/blog/product-announcements/duet-ai-in-workspace-now-available

7

www.anthropic.com/index/anthropic-series-c

8

https://ai.meta.com/llama/

9

https://huggingface.co/blog/series-c

10

www.businesswire.com/news/home/20230629810313/en/Inflection-AI-Announces-1.3-Billion-of-Funding-Led-by-Current-Investors-Microsoft-and-NVIDIA

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_4

4. Prompt Engineering

How to Instruct Generative AI

Tom Taulli[1]
(1)
Monrovia, CA, USA

Prompt engineering describes the techniques of creating prompts to get what you want from an LLM like ChatGPT or Bard. This category has become one of the hottest in AI. Prompt engineering has even become a fast-growing job role. A common saying is: "AI will not take away your job. But someone who knows AI will."

Based on data from Willis Towers Watson, the average annual compensation for a prompt engineer is about $130,000.[1] This may actually understate the amount. The reason is that a company may also offer lucrative equity packages and bonuses to attract talent.

The role for the prompt engineer did not exist until the launch of ChatGPT. It is also something that is quickly evolving. However, if you want to have effective automation with generative AI, then you need the skillsets for creating effective prompts.

This is what we'll cover in this chapter.

Note In July 2023, Netflix published a job ad for a product manager for the company's machine learning system. The compensation package? It was for $900,000.[2] Of course, this caused a stir. It also was not particularly good timing because of the Hollywood actors and writers strike, which involved AI as a big sticking point.

The Challenges with Prompt Engineering

Prompt engineering can be difficult and complex. Even a slight change in a prompt can have a major impact on the output of an LLM.

The main reason is the underlying model. An LLM uses a sophisticated transformer system, which relies on probabilities. This allows for variation or creativity in the content generation.

But there are other issues. They include the following:

- Verbosity: In some cases, the output can be way too long. Again, this is due to the transformer model. It is essentially like a highly sophisticated auto complete system. In other words, the LLM will keep exploring other related ideas, concepts or facts. This is why you need to be specific with your prompts. We'll show techniques for this later in the chapter.
- Tweaking: Often, your first prompt will not be the best. You will likely need a few tries to find the right one. However, this can make the process time consuming.
- Bias: LLMs are trained on publicly available information. But this means it will reflect the underlying bias and prejudices, which can mean that the content could be insensitive or discriminatory. True, LLMs are getting better at mitigating this. But there is still a long way to go – and this may be a problem that may ultimately never be solved.
- Non-transferable: A prompt that works for, say ChatGPT, may not be good for Bard or another LLM. This can add even more friction when creating content. Or, prompts you have worked hard to create may not work properly on a newer version of the LLM.

- Length: It’s often the case that an LLM will have difficulties understanding long prompts. The system may skip certain information or could be confused with certain constructs.
- Knowledge: An LLM may not be trained on the type of information you are focused on.

Hire a Prompt Engineer?

A prompt engineer is not just someone who is good at crafting prompts. This person usually has a strong background with data science. They will also typically be a part of an AI development team, which is for those companies that are engaged in sophisticated implementations.

The following are common tasks that a prompt engineer will take on:

- Engage in research to find and curate the right data sources to craft useful prompts.
- Fine tune and optimize prompts based on testing and feedback from AI models.
- Analyze and measure the outputs of an LLM.
- Keep up-to-date on the latest research in prompt engineering.
- Create and maintain procedures, policies, and documentation for prompts.
- Engage in trouble shooting to identify issues with prompts.
- Participate in code reviews.
- Develop scalable programs using Python and AI frameworks for managing prompts.

In terms of experience, the requirements can be broad. But here are some of the main ones, according to recent job ads:

- Bachelor’s degree in computer science or software engineering. However, it’s often preferred for the candidate to have a master’s degree.

- Experience with natural language processing (NLP) and generative AI, such as with models like GPT-4 or PaLM 2.
- Proficiency with Python or JavaScript. There should also be an understanding of AI frameworks like LangChain and PyTorch.
- Background with data collection, preprocessing, and annotation.
- Experience with prompt automation tools like Ansible and Puppet.

But depending on your resources and requirements, you may not need to hire a prompt engineer. The good news is that there are certain guidelines that can help improve performance when working with Bard and ChatGPT.

Let’s take a look at some of the main ones.

> **Note** Ultimately, the role of the prompt engineer may fade away. The irony is this may be due to the growing power of LLMs. Over time, they should get better at understanding prompts or asking questions to improve the results. The LLMs may eventually allow for creating optimal prompts.

Context

A good way to start a prompt is with a sentence or two about the context. You are essentially telling ChatGPT or Bard to take on a certain role. For example, if you want to create a tweet, you can write something like this:

You are an expert in social media that writes engaging tweets.

Or maybe you want to get advice about how to run your business. This could be something like increasing cash flows. In this case, you can set up the context for someone who would take on the role. You can also specify the type of business, such as:

You are the chief financial officer of a startup that develops a mobile app.

This context also highlights something else: try to avoid ambiguous terms. This can be the case with acronyms. If you use CFO, the chatbot may think you are referring to something else.

Be Specific

Perhaps the most important factor for a good prompt is being specific. It's critical to focus on what is most relevant.

By doing this, you should get the following benefits:

- Accuracy: After all, vague prompts often have a higher chance of leading to output that may not be right or could be misleading.
- Efficiency: A specific prompt should mean not having many follow-up questions for the chatbot.
- Depth of answer: You are likely to get a more detailed output. It may also have more nuances.

Here's a look at broad versus specific prompts:

Broad: How do I market a product?

Specific: I want to market a new software application that helps manage multi-tenant rental properties. They have four to five units. The software assists with tenant applications, rental payment processing, and handling tenant requests. How would I market this system?

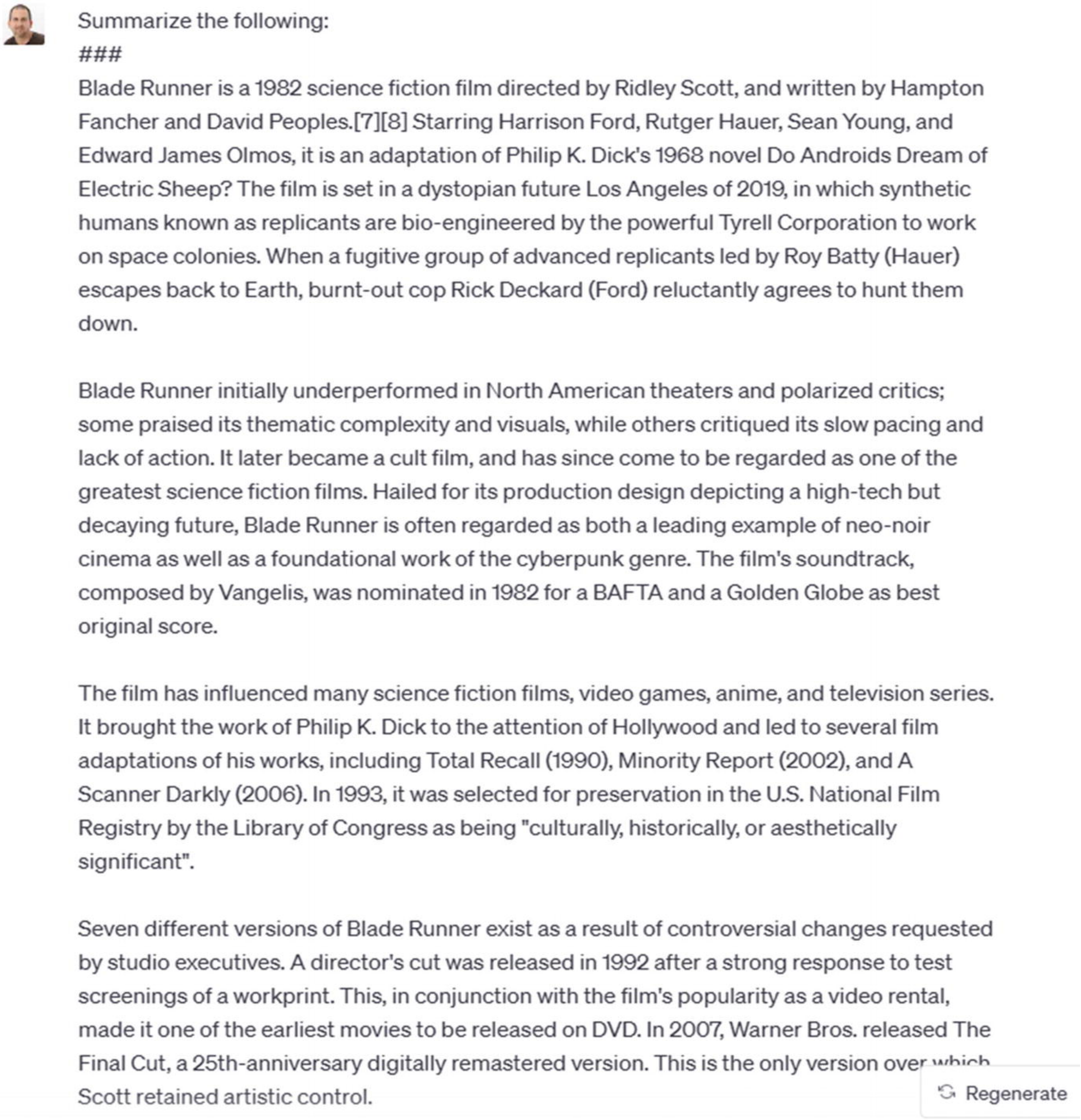
Broad: How do I hire someone?

Specific: I run a small coffee shop. I want to hire a new cashier. What are some effective questions for the job interview?

Instructions

There are certain types of instructions that ChatGPT and Bard are particularly good at. One is summarization. An effective approach for this is to delimit the information you want summarized. This can be done with any type of character, say something like ### or ===. By doing this, the LLM will have a better structure to provide an effective result.

Figure 4-1 shows an example. I did a copy-and-paste of the introduction – from Wikipedia – about the classic science fiction movie, *Blade Runner*. I put the content between ### characters. I then asked for a summary.



Summarize the following:

###

Blade Runner is a 1982 science fiction film directed by Ridley Scott, and written by Hampton Fancher and David Peoples.[7][8] Starring Harrison Ford, Rutger Hauer, Sean Young, and Edward James Olmos, it is an adaptation of Philip K. Dick's 1968 novel Do Androids Dream of Electric Sheep? The film is set in a dystopian future Los Angeles of 2019, in which synthetic humans known as replicants are bio-engineered by the powerful Tyrell Corporation to work on space colonies. When a fugitive group of advanced replicants led by Roy Batty (Hauer) escapes back to Earth, burnt-out cop Rick Deckard (Ford) reluctantly agrees to hunt them down.

Blade Runner initially underperformed in North American theaters and polarized critics; some praised its thematic complexity and visuals, while others critiqued its slow pacing and lack of action. It later became a cult film, and has since come to be regarded as one of the greatest science fiction films. Hailed for its production design depicting a high-tech but decaying future, Blade Runner is often regarded as both a leading example of neo-noir cinema as well as a foundational work of the cyberpunk genre. The film's soundtrack, composed by Vangelis, was nominated in 1982 for a BAFTA and a Golden Globe as best original score.

The film has influenced many science fiction films, video games, anime, and television series. It brought the work of Philip K. Dick to the attention of Hollywood and led to several film adaptations of his works, including Total Recall (1990), Minority Report (2002), and A Scanner Darkly (2006). In 1993, it was selected for preservation in the U.S. National Film Registry by the Library of Congress as being "culturally, historically, or aesthetically significant".

Seven different versions of Blade Runner exist as a result of controversial changes requested by studio executives. A director's cut was released in 1992 after a strong response to test screenings of a workprint. This, in conjunction with the film's popularity as a video rental, made it one of the earliest movies to be released on DVD. In 2007, Warner Bros. released The Final Cut, a 25th-anniversary digitally remastered version. This is the only version over which Scott retained artistic control.

Figure 4-1 This is a prompt for summarizing information, which is delimited by the ### characters

ChatGPT provided two paragraphs. But suppose you just want one? Of course, you can then ask for this in the input box.

Another instruction is for data extraction. To continue with our example, I included the section about casting and provided this prompt:

Who are the actors mentioned in this text?

ChatGPT was able to find 32, which included Dustin Hoffman, Harrison Ford, Sean Connery, and Jack Nicholson. Of course, we can ask the system to put the names in a table.

No doubt, the extraction feature can be a major time saver, especially with long documents.

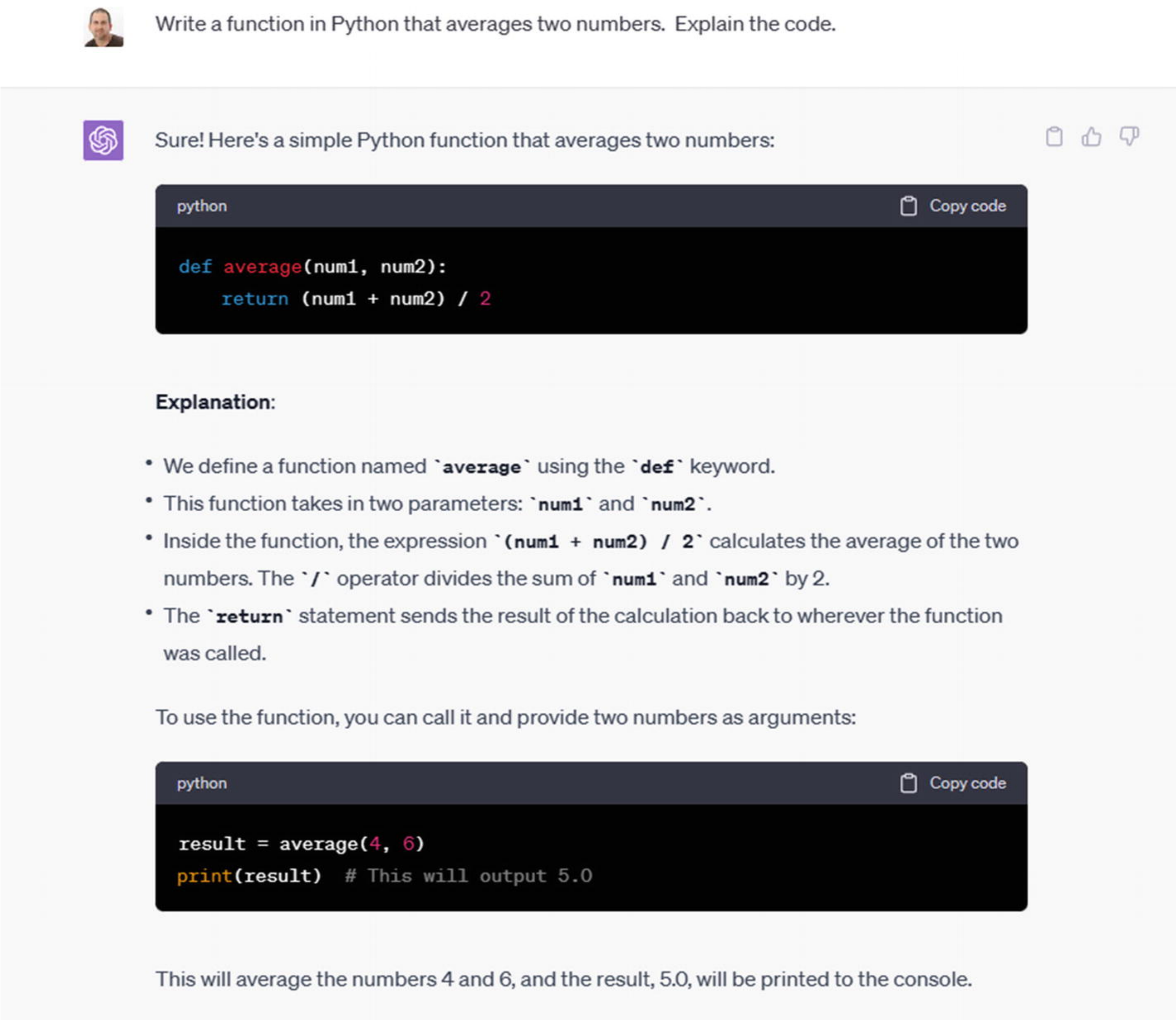
Next, you can use ChatGPT for translation. Using this is fairly straightforward, such as with this prompt:

Translate this sentence into German: How are you today?

You can use translate for programming languages like Python, C or Java. Suppose you want to write a function that averages two numbers. You can use this as the prompt:

Write a function in Python that averages two numbers. Explain the code.

Figure 4-2 shows the output.



Write a function in Python that averages two numbers. Explain the code.

Sure! Here's a simple Python function that averages two numbers:

```python
def average(num1, num2):
    return (num1 + num2) / 2
```

Explanation:

- We define a function named `average` using the `def` keyword.
- This function takes in two parameters: `num1` and `num2`.
- Inside the function, the expression `(num1 + num2) / 2` calculates the average of the two numbers. The `/` operator divides the sum of `num1` and `num2` by 2.
- The `return` statement sends the result of the calculation back to wherever the function was called.

To use the function, you can call it and provide two numbers as arguments:

```python
result = average(4, 6)
print(result)  # This will output 5.0
```

This will average the numbers 4 and 6, and the result, 5.0, will be printed to the console.

Figure 4-2 Here ChatGPT has created code for averaging two numbers. There is also a step-by-step explanation

It’s a simple function, with only two lines of code. ChatGPT puts this into a code window, which you can copy (you will click the icon on the top right). Then there is a step-by-step explanation. In fact, by doing this type of exercise, you can actually learn the Python language.

Suppose you want to see what this code will look like in Java. You can use this simple prompt:

What is this for Java?

For the most part, a good instruction often starts with a verb. Here are some others to consider:

- Recommendations: “Suggest five top books to read about digital marketing.”
- Elaboration: “Elaborate more about this response.”
- Comparison: “Compare generative AI to deep learning.”

Note Andrej Karpathy is a major influencer in the artificial intelligence industry. He has a Ph.D. in data science and has worked with luminaries like Geoffrey Hinton. He is also a founding member at OpenAI and the former senior director of AI at Tesla, where he helped to create AutoPilot. He also likes to tweet and has a following of over 778,000. In one of his posts, he noted: “The hottest new programming language is English.”[3]

Zero and Few-Shot Prompting

Zero-shot prompting refers to when you enter a prompt into an LLM and you get the right answer. There is no need for a follow-up. The phrase “zero shots” means that the LLM does not need any examples.

Yet there are times when you need more. This is where few-shot prompting comes in. You are essentially nudging the LLM toward a certain direction. However, this is not to imply that the model is somehow being changed. There are no adjustments to the underlying weights or parameters.

Let’s take a look at some examples of few-shot prompting. One is for helping the LLM understand shorthand or acronyms. We’ll do this for financial abbreviations:

- You: ROI means return on investment
- You: ROE means return on equity
- You: EBITDA means
- ChatGPT: EBITDA means “Earnings Before Interest, Taxes, Depreciation, and Amortization.”

ChatGPT has figured out the pattern because it has the necessary context to provide the right answer. The system also explained what EBITDA means.

Another example is for sentiment analysis. We can do this by showing the LLM positive, negative, and neutral tweets.

You: “I love this new update!” – Positive

You: “Why did the app crash again? So frustrating!” – Negative

You: “The event is scheduled for next Friday.” – Neutral

You: “Absolutely adore how fast the new version is. Great job!”

ChatGPT: Based on the provided examples: “Absolutely adore how fast the new version is. Great job!” – Positive

The temptation is to use many examples. Won’t this mean better training for the LLM? Not really. Providing two or three examples is usually enough. With more, you may wind up confusing the model and not get the results you want.

Parameters and OpenAI Playground

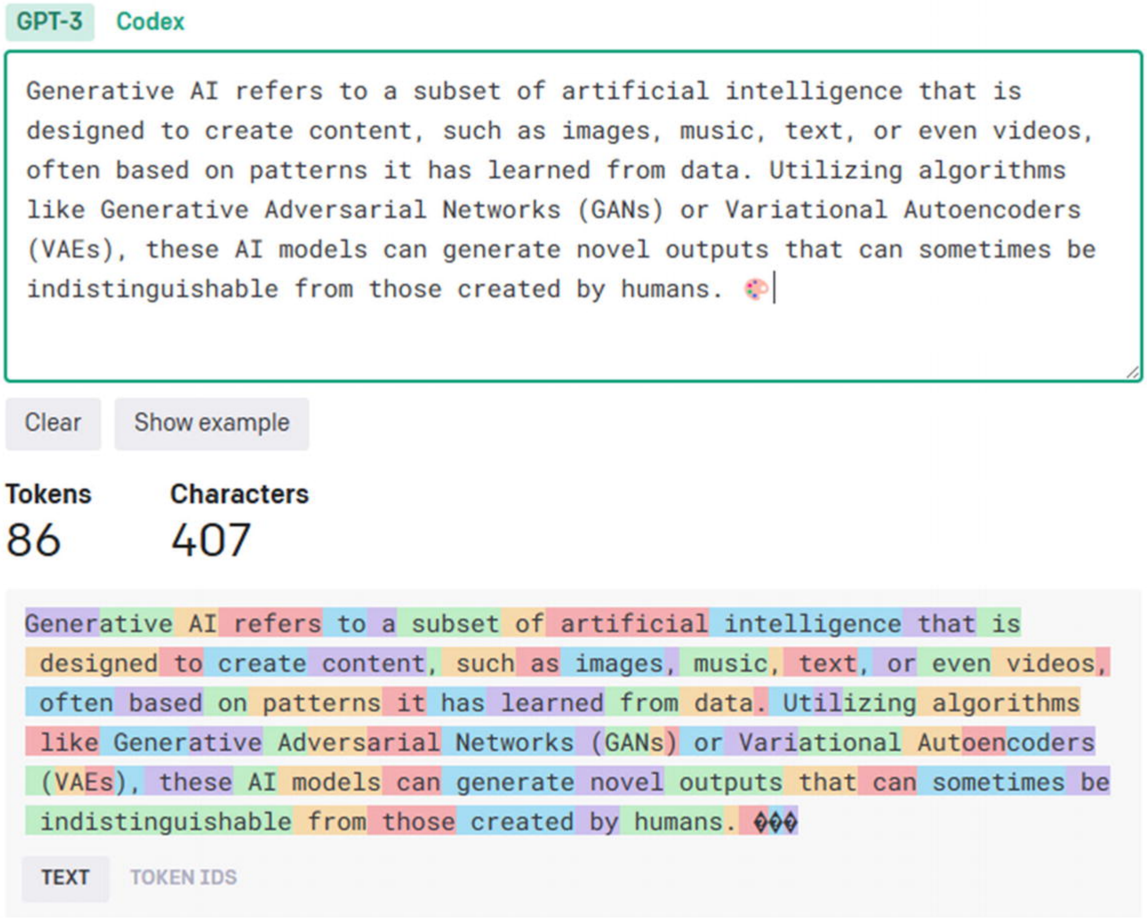
With an LLM, you can change the levels of creativity and other features. But this is not possible using ChatGPT or Bard. However, an alternative is to use the OpenAI Playground, which is at the following URL: https://platform.openai.com/playground.

You will need to sign up for an account and provide a phone verification for security. There is a cost to using the system but there is an initial $5 credit for the first three months of using the service.

The pricing is based on the number of tokens used for both the prompts and responses. For example, the GPT-3.5-turbo model – which powers the ChatGPT platform – has a fee of $0.0015 per 1,000 tokens for the input and $0.002 per 1,000 tokens. This is for the 4K context window.

A token can be one word or a part of one. So how can you tell? There is no formula for this.

But you can use the OpenAI Tokenizer to get the right answer and the URL is at https://platform.openai.com/tokenizer. Figure 4-3 shows an example, which has a definition for generative AI and there is also an emoji.



Generative AI refers to a subset of artificial intelligence that is designed to create content, such as images, music, text, or even videos, often based on patterns it has learned from data. Utilizing algorithms like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), these AI models can generate novel outputs that can sometimes be indistinguishable from those created by humans. 🎨

Tokens 86

Characters 407

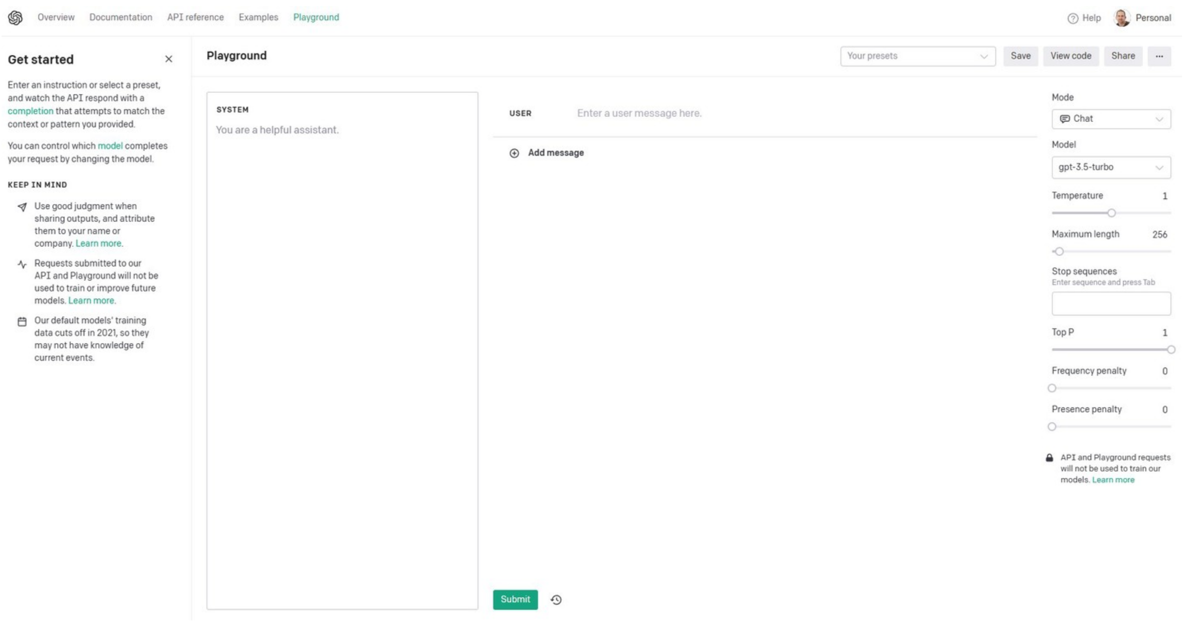
Generative AI refers to a subset of artificial intelligence that is designed to create content, such as images, music, text, or even videos, often based on patterns it has learned from data. Utilizing algorithms like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), these AI models can generate novel outputs that can sometimes be indistinguishable from those created by humans. ���

Figure 4-3 Here is the OpenAI Tokenizer, which will provide the details about tokens for text

This indicates there are 86 tokens. The first word – generative – is composed of two tokens. The next word, AI, is one token but it also includes a leading space. As for the emoji, it consists of three tokens.

This is not to suggest that you should always use the OpenAI Tokenizer before using the OpenAI Playground. Rather, this is just to demonstrate how the token system works. But as a rule-of-them, there are 750 tokens for 1,000 words.

Now let’s take a look at the OpenAI Playground, which you can see in Figure 4-4.



Get started

Enter an instruction or select a preset, and watch the API respond with a completion that attempts to match the context or pattern you provided.

You can control which model completes your request by changing the model.

KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. Learn more.
- Requests submitted to our API and Playground will not be used to train or improve future models. Learn more.
- Our default models' training data cuts off in 2021, so they may not have knowledge of current events.

SYSTEM

You are a helpful assistant.

USER

Enter a user message here.

Mode: Chat

Model: gpt-3.5-turbo

Temperature: 1

Maximum length: 256

Stop sequences: Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

API and Playground requests will not be used to train our models. Learn more

Figure 4-4 This is the OpenAI Playground

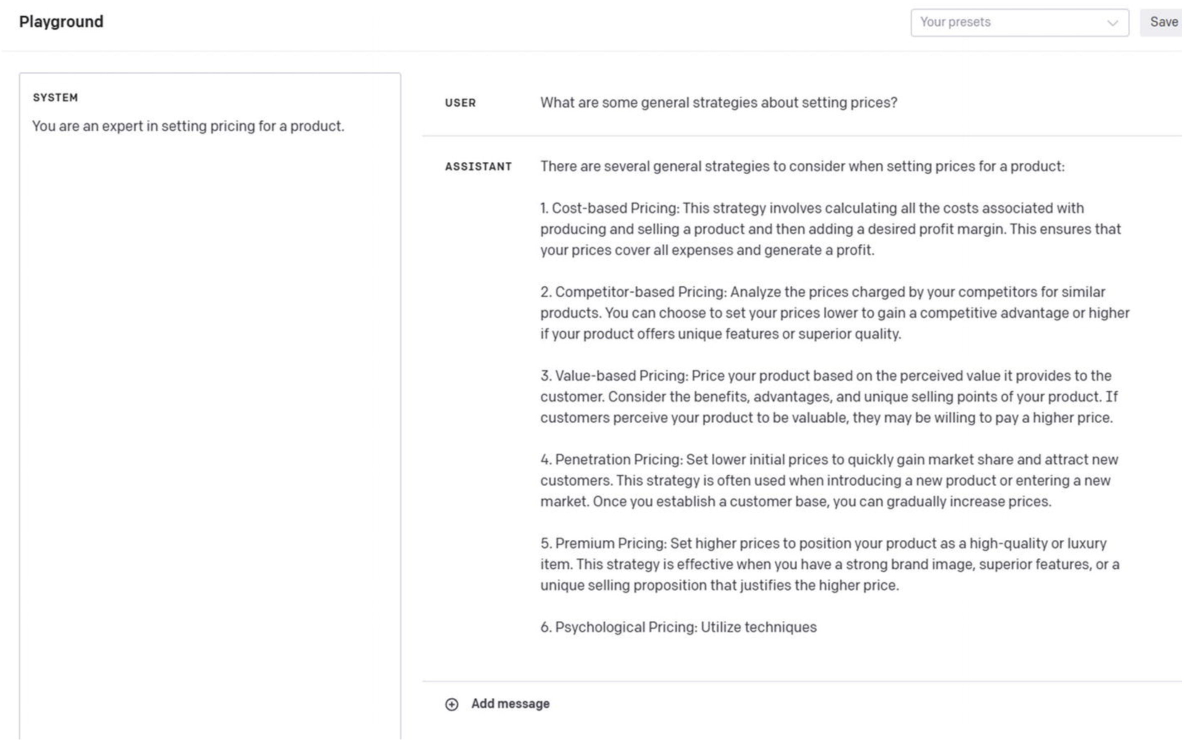
By contrast to ChatGPT and Bard, the OpenAI Playground is much more extensive. On the left side of the screen, there are some helpful tips for using the system. Then there is a multi-part approach for creating a prompt. The System is where you can set the context and the User is the place to enter your instruction.

Here’s an example:

System: You are an expert in setting the pricing for a product.

User: What are some general strategies about setting prices?

Press “Submit” button and you will get a response, as seen in Figure 4-5.



SYSTEM
You are an expert in setting pricing for a product.

USER
What are some general strategies about setting prices?

ASSISTANT
There are several general strategies to consider when setting prices for a product:

1. Cost-based Pricing: This strategy involves calculating all the costs associated with producing and selling a product and then adding a desired profit margin. This ensures that your prices cover all expenses and generate a profit.

2. Competitor-based Pricing: Analyze the prices charged by your competitors for similar products. You can choose to set your prices lower to gain a competitive advantage or higher if your product offers unique features or superior quality.

3. Value-based Pricing: Price your product based on the perceived value it provides to the customer. Consider the benefits, advantages, and unique selling points of your product. If customers perceive your product to be valuable, they may be willing to pay a higher price.

4. Penetration Pricing: Set lower initial prices to quickly gain market share and attract new customers. This strategy is often used when introducing a new product or entering a new market. Once you establish a customer base, you can gradually increase prices.

5. Premium Pricing: Set higher prices to position your product as a high-quality or luxury item. This strategy is effective when you have a strong brand image, superior features, or a unique selling proposition that justifies the higher price.

6. Psychological Pricing: Utilize techniques

Figure 4-5 This shows the output for a prompt using the OpenAI Playground

The response includes six suggestions, which are useful and fairly concise.

At the bottom, there is "Add message." This is where you can set the output for the AI. Why do so? It allows for using few-shot prompting. Let's take an example:

User: Translate this into German: "The cat sat on the mat."

Assistant: "Die Katze saß auf der Matte."

User: Translate this into German: "I love chocolate."

The system comes up with the first answer from the Assistant, which is: "Ich liebe Schokolade."

On the right side of the screen, there are numerous options to adjust the model. They include the following:

- Mode: This has three items, which are Chat, Complete, and Edit. The last two are likely to be phased out. The default is the chat-type approach.
- Model: Since we are using the chat model, there are three models available: GPT-4, GPT-3.5-turbo-16, and GPT-3.5-

turbo. GPT-4 is more expensive on a token basis.

- Temperature: This is a common parameter to change the creativity of the responses. You can set a value from 0 to 2. The lower the value, the less creativity. When it's 0, the model becomes deterministic and repetitive.
- Stop Sequences: You can provide the model a certain indicator for it to stop generating a response. It could be something like a character.
- Top P: This is another way to control the creativity of the model. The range for it is from 0 to 1. Think of it as a percentage. If you have 0.5, this means that half of the likely options will be used for the responses. It's a way to focus more on higher probability options.
- Frequency Penalty: This reduces the repetitiveness of tokens in the response and the range is from 0 to 2. The higher the value, the lower the probability of the repetition. However, if it is 1.5 or higher, the results can sometimes get weird. This highlights that – when it comes to using these types of parameters – you will likely need to experiment.
- Presence Penalty: This will penalize new tokens based on whether they appear in the response already. The range is from 0 to 2. Again, the higher the value, the higher chance of having unexpected results.

At the top right of the screen, there are several options. Save will store what you have created in the OpenAI Playground and View Code will turn your work into a programming language or scripting system like Python, Node.js, Curl, or JSON. You can also share your results with other people.

Resources

With the boom in generative AI, there has been the emergence of many consultants who claim that they have the sure-fire "formula" for prompt engineering. They will often sell expensive courses or videos.

But you need to be wary. Prompt engineering is a blend of science and art, which can make it challenging to get the best results. The models are also continually changing.

It’s kind of similar to the market for SEO (search engine optimization) tools and services. Unfortunately, there are plenty of people who are just looking to make a quick buck. But in the end, their advice often falls short.

For the most part, in this chapter we have gone over some of the best practices that should get you on the right track. But it’s also important to keep experimenting.

Conclusion

In this chapter, we have looked at prompt engineering. It’s a dynamic category and has become a high-growth job role. But a business does not necessarily need to hire a prompt engineer. The fact is that there are general guidelines that should help get solid results.

We have taken a look at these. They include adding a context, being specific and providing relevant instructions. We also saw how to nudge a model with few-shot prompting.

We also took a look at the OpenAI Playground. This is a powerful tool that allows for making adjustments to how a generative AI creates its output.

In the next chapter, we’ll cover the use cases for ChatGPT and Bard.

Footnotes

[1] www.wsj.com/articles/artificial-intelligence-jobs-pay-netflix-walmart-230fc3cb?mod=hp_lead_pos10

[2] www.wsj.com/articles/artificial-intelligence-jobs-pay-netflix-walmart-230fc3cb?mod=hp_lead_pos10

3

https://twitter.com/karpathy/status/1617979122625712128?lang=en

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_5

5. Use Cases

Improving Your Business with ChatGPT and Bard

Tom Taulli[1]
(1)
Monrovia, CA, USA

Alignable operates a social network for small business owners. It has over 8 million members and has coverage of about 35,000 local communities.

The company conducted a survey – called the ChatGPT Sentiment Study – of its user base.[1] About 57% of the small business owners said they have tried a new AI tool and 26% said they have made this technology an important part of their operations.

Here’s what one of the respondents said:

> *We started experimenting with [ChatGPT], then really got into it, and now it’s a part of our marketing automation platform. By integrating the GPT capabilities, we’ve freed up all kinds of time. Things that used to take hours now take minutes, giving us room to focus on the work we truly love that will bring in more money. It doesn’t replace the human touch as many fear. It actually gives us more time for it. This genie is not going back in the bottle.*

Here are some of the other takeaways from the survey about ChatGPT:

- 31% are using the app for creating marketing content.
- 13% use it for customer support.
- 7% are exploring ways to improve business functions and operations.

In this chapter, we'll take a look at how ChatGPT and Bard can be leveraged for various parts of a business. However, it is important to keep a few things in mind:

- Review the content: Do not simply cut and paste a response from the AI. Make sure you have a thorough review. There can be errors, which are known as hallucinations. There may also be ways to improve the content, so it is more tailored for your needs.
- Privacy: Avoid placing personal or confidential information in a chatbot. There is the possibility of a breach. There could also be laws or regulations that prohibit this.
- Bias and fairness: As generative AI is trained on publicly available information, this means it is vulnerable to creating output that can be discriminatory. This is another reason why it's a good idea to review the content. You may even have a prompt say something like: "With the output, please avoid bias or discriminatory language."
- Compliance: According to the policies from ChatGPT and Bard, you are not allowed to use the systems for providing legal advice. For the most part, the output should be a way to get a sense of the issues and some background information. But when it comes to drawing up a contract or taking an action like terminating an employee – just to name a few examples – you need to get the advice of a qualified attorney. The same would go for other categories like tax advice, insurance, and accounting.

Now, let's take a look at the numerous areas where ChatGPT and Bard can help your business.

Human Resources

The PwC HR Tech Survey 2022 shows that technology investments for human resources have been getting results.[2] This is based on the findings from 688 HR leaders in the United States. However, the survey also shows that much more can still be done, such as with adoption of technology, finding the right vendors, and gaining buy-in for next-generation systems. The survey also shows some of the biggest challenges for HR department: insights and analytics, recruiting/hiring, cloud transformation, learning development, upskilling, retention, remote/hybrid work, benefits, and diversity.

As for ChatGPT and Bard, these tools can definitely help out. This shows the different prompts you can use:

- Recruiting:
    - "Suggest 10 screening interview questions for a [job title]."
    - "What behavioral questions would be relevant for interviewing a [job title]?"
    - "Suggest technical or role-specific tests for a [job title]."
    - "List the primary responsibilities for a [job title]."
    - "What are the essential skills required for a [job title]?"
    - "What are the typical career progression paths for a [job title]?"
    - "Write a job description for a [job title]."
    - "Write a job want ad for a [job title]."
    - "Define the key performance indicators (KPIs) for a [job title]."
    - "List common challenges faced by professionals in the [job title] role. What are potential solutions?"
    - "Write an email for a job offer for [job title]. Include the [start date] and [starting salary]."
- Onboarding:

  - “Write a welcome message for a new employee. The role is [job title]. The tone should be optimistic.”
  - “Create a checklist and essential items for a [job title] on the first day of work.”
  - “Create a core values statement for my company. This is for new employees.”
  - “Suggest the responsibilities and expectations for a [job title] during their first month on the job.”
  - “Draft a feedback form that the HR team can use to evaluate a [job title’s] integration during the first month.”
  - “Describe frequently asked questions from new employees.”
  - “Describe best practices for onboarding new employees.”
  - “List 3 activities or icebreakers to introduce a new employee to their team. The goal is to encourage more engagement.”
  - “Draft questions to solicit feedback on how the onboarding process can be improved for future hires.”
- Retention:
  - “Suggest ideas or programs to reinforce a company’s core values among employees.”
  - “Draft an employee survey that measures how employes feel they are being valued and recognized.”
  - “Provide team-building activities to boost camaraderie among employees.”
  - “Suggest ideas for non-monetary ways to motivate and encourage employees.”
  - “Draft a ‘Thank You’ note for employees who make an accomplishment.”

    - “List best practices for drafting a compensation strategy.”
- Diversity and inclusion:
    - “Analyze our company mission statement for inclusive language.”
    - “Review this job description for any problems with diversity and inclusion.”
    - “Create a survey of questions to measure employee perceptions of diversity and inclusion for our company.”
- Offboarding:
    - “Draft an employee exit checklist for HR.”
    - “Suggest questions for an exit interview.”
    - “Provide recommendations for conducting an exit interview.”
    - “Draft a letter that announces the departure of the employee.”
    - “Suggest resources and services for a departing employee. These include career counseling and job placement support.”
    - “Create a check list for securing the company’s physical and digital systems when an employee departs.”

Note In July 2023, Gartner published a survey of 105 HR leaders.[3] It showed that a mere 5% of them had implemented generative AI in their organizations. Although, there was evidence of some momentum. About 9% of the respondents said that they were in the process of creating pilot programs and more than half were exploring the technology. Some of main priority areas included administrative tasks, policies, document generation, and recruiting.

HR is an area that certainly has the potential for legal and compliance exposure. After all, a company will be working with sensitive information, say like Social Security numbers, family data, and medical records.

Areas that can be particularly risky are with hiring and termination. Consider that some of the world's tech companies have had challenges with such matters. Just look at Amazon. The company built its own AI-powered recruiting system. However, it mostly selected male candidates. Amazon tried to correct this but would ultimately abandon the effort.[4]

In May 2023, the EEOC (Equal Employment Opportunity Commission) published a technical assistance document about AI and discrimination.[5] It warned that – if AI is not applied properly – there could be violations of Title VII of the federal antidiscrimination law. The EEOC indicated the following areas that could be problematic: recruitment, hiring, retention, promotion, performance monitoring, and termination.

The document provided some examples:

- Resume scanners that prioritize candidates based on keywords.
- Chatbots that ask a candidate about their qualifications.
- Chatbots that reject candidates based on pre-defined requirements.
- Facial recognition used for evaluating a candidate.
- AI-based job-fit scores.
- AI systems that have a disparate impact on minority groups.

Legal

Legal costs have always been a burden for businesses. The hourly fees for a qualified attorney can range from $150 to $325, according to analysis from LegalMatch.[6] But it can easily be more. This is the case if you need a specialist, say an attorney that understands how to craft and file patents.

Then how can generative AI help? This is a look at the numerous scenarios, along with prompts:

- General terms:
    - “What does ‘limited liability’ mean?”
    - “What is a ‘forum clause’?”
    - “How does a sole proprietorship compare to a limited liability company?”
    - “What is ‘consideration’ in a legal contract?”
    - “How does an employee compare to a contractor?”
    - “What is the difference between ‘leasehold’ and ‘freehold’?”
- Legal concepts:
    - “What are the elements of an enforceable contract?”
    - “What are the advantages and disadvantages of a C-Corporation?”
    - “What determines an ‘at-will’ employment contract?
- Legal document creation
    - “Draft a mutual non-disclosure agreement. Make the agreement favorable to [list the party].”
    - “Create an independent contractor agreement. It should be clear that the recipient qualifies as an independent contractor.”
    - “Write a promissory note. It must be favorable to the lender.”
- Real estate:
    - “What legal checks should a business conduct before purchasing commercial property?”
    - “How do ‘contingencies’ work in real estate contracts?”

  - “What is the significance of a ‘triple net lease’ for a business?”
  - “How does a ‘sublease’ differ from a standard lease agreement?”
  - “How can a business apply for a ‘rezoning’ or ‘variance’?”
- Intellectual property:
  - “How does a business trademark its logo or brand name?”
  - “What’s the significance of the ‘®’ symbol and the ‘™’ symbols?”
  - “How can I enforce my business’s trademark rights against infringers?”
  - “Define ‘trade secret’ in the context of business operations.”
  - “How can a business protect its trade secrets legally?”
  - “What legal recourse does a business have if its trade secrets are stolen?”
- Insurance:
  - “What is the difference between ‘claims-made’ and ‘occurrence’ insurance policies?”
  - “What does ‘workers’ compensation insurance’ cover?”
  - “How does ‘business interruption insurance’ benefit a business?”
  - “Describe the scope of ‘general liability insurance’ for businesses.”
  - “What is a ‘deductible’ in business insurance and how does it impact claim payouts?”
  - “What does ‘cyber liability insurance’ cover?”
- Taxes:
  - “What’s the difference between ‘gross income’ and ‘net income’?”

    - “How do ‘tax credits’ differ from ‘tax deductions’?”
    - “What is ‘pass-through’ taxation?”
    - “What is the ‘self-employment tax’?”
    - “Explain ‘use tax’ in the context of business operations.”

Another useful task for ChatGPT or Bard is for the analysis of a contract or a legal document. You can either cut-and-paste the content in the prompt or use Advanced Data Analysis.

These are same prompts:

- Clarification of clauses and terms:
    - “Can you explain the ‘force majeure’ clause in this contract?”
    - “What does ‘severability’ mean in the context of legal agreements?”
- Flagging issues:
    - “Here’s the termination clause from my agreement. Are there any potential red flags?”
    - “Does this non-compete section seem overly restrictive?”
- Comparing clauses:
    - “I have two versions of a liability clause. Which one seems more favorable for a business owner?”
    - “Can you compare these two indemnification sections and point out the key differences?”
- Missing clauses:
    - “Are there any key clauses missing from this rental agreement?”
    - “This is a software licensing contract. Should there be any clauses regarding updates or patches?”
- Responsibilities:

  - “What are the key responsibilities of the service provider in this agreement?”
  - “Are there any penalties mentioned for non-performance or breach of duties?”

Customer Support

Strong customer support is often a challenge. Training can be expensive and time consuming. Then there are the issues with scaling, as volume may spike unexpectedly. There are also the problems with language and cultural barriers.

Something else to keep in mind: Companies like Amazon, Starbucks, and Virgin Atlantic have greatly raised the bar for quality service. A big part of this is that strong customer support is part of the corporate DNA. But another factor is the leveraging of sophisticated technologies.

There’s lots of evidence that standout customer support leads to higher growth and profits. A survey from Zendesk shows that 61% of customers would move over to a competitor if there was only one bad experience.[7] It goes to 76% if there are two negative experiences. According to Zendesk:

> *As the saying goes: no risk, no reward. For companies that succeed at wowing customers, the opportunities for growth are immense. Not only will they lay claim to the growing number of customers that shop with a service-first mindset, but they will also have a real shot at expanding, retaining, and deepening their customer base.*

However, ChatGPT and Bard are not built for real-time customer interactions. But there are various tools that leverage generative AI for this. Here are some examples:

- Freshworks
- Forethought
- Zendesk

- HubSpot
- Freshdesk
- Landbot
- HappyFox

Despite this, there are certainly areas where ChatGPT and Bard can help with customer support. Here’s a look:

- Improving customer relationships:
    - “What are ways to improve customer loyalty?”
    - “Provide tips for creating an effective customer loyalty program.”
    - “What strategies can I use to re-engage inactive customers?”
    - “Describe best practices for customer support in the [type] industry.”
- Complaint handling:
    - “Describe how to handle a customer who is angry about a delayed shipment.”
    - “Provide a template response for addressing complaints about a faulty product.”
    - “Provide a response to a customer asking for a refund after the refund period.”
    - “Suggest ways to apologize to a customer for a mistake made in their order.”
    - “Suggest strategies for turning negative customer experiences into opportunities for improvement.”
    - “What steps should I take after receiving consistent negative feedback about a product?”

Sales

HubSpot conducted a survey of 1,350 sales reps and leaders about their views of generative AI.[8] Here are some of the findings:

- 74% of the respondents said they use some form of AI or automation tool.
- 69% indicate that salespersons should use these tools for prospecting. But they also should not be too reliant on these systems.
- 60% said that AI and automation are important to their sales strategy.

The survey also showed some of the common use cases:

- Content creation
- Prospect outreach
- Research

The typical salesperson has many daily tasks to work on, which can be tedious and time consuming. There is the data entry for updating CRM (Customer Relationship Management) systems. They have to qualify incoming leads as well as engage in follow-ups, say with emails, IMs, or calls. Then there is the need for ongoing scheduling. Oh, and there must be research, such as to learn about new prospects. This is critical in drafting proposals.

But generative AI can help with all these activities. There should be less time devoted to repetitive tasks. The generative AI can also create more engaging content, which can lead to higher conversation rates.

Regarding ChatGPT and Bard, these are definitely valuable tools for sales people. Here are the numerous applications:

- Prospecting:
  - “Can you summarize key trends for the [industry]?”
  - “What do customers in the [industry] look for in a new product?”
  - “Provide a template for a sales proposal for a [product or service].”

- Qualifying leads:
    - “What are common factors for qualifying a lead for the [industry]?”
    - “List questions to ask a prospect in a discovery call to determine if the lead is a fit.”
    - “What are strategies to qualify a lead?”
- Objection handling:
    - “How can I counter the objection that our product is too expensive?”
    - “Provide responses to common objections in the [industry]?”
- Email templates:
    - “Draft an email for a prospect to sell [product or service].”
    - “Draft an email for a follow-up after a sales demo.”
    - “Create an email template to thank a prospect for attending a product demo. Include next steps.”
    - “Compose an email template to check in with a client who purchased our service three months ago, seeking feedback on their experience.”
    - “Design an email template aimed at re-engaging a lead who showed interest in our service six months ago but didn’t purchase.”
    - “Craft an email template to inform a loyal customer about an exclusive discount on our new product line.”
- Sales strategy:
    - “Create a sales funnel for a [product or service].”
    - “What are some effective strategies for upselling to existing customers?”
    - “Describe how salespeople can leverage social media platforms for effective social selling.”

    - “List effective cold calling techniques for modern salespeople targeting the [industry].”
    - “Compare and contrast inbound and outbound sales strategies and highlight their respective advantages.”
    - “Detail the most important sales metrics and KPIs that teams should track to measure performance.”
    - “Discuss the importance of a referral sales strategy and provide steps to implement one.”
    - “List time management techniques effective for salespeople.”
- Closing techniques:
    - “Describe the ‘Feel-Felt-Found’ closing technique.”
    - “Provide examples of effective closing questions.”

Marketing

In August 2023, Gartner published a survey of 405 marketing leaders. About 63% said they planned on investing – during the next 24 months – in generative AI technologies.[9] Interestingly, there has been a steep decline in spending on technology because of difficulties with customer data, compliance, and the complexity of the systems. So the growth in generative AI is likely to take a bigger bite from the budget.

Marketing is an early adopter of this technology – even before the emergence of ChatGPT.

This should not be a surprise. Generative AI can assist a marketer in many tasks such as brainstorming, content creation – say with blogs, social media posts, and ads – market research, SEO (Search Engine Optimization), and SEM (Search Engine Marketing).

Let’s take a look at some of the useful prompts for ChatGPT and Bard.

- Content creation:

    - “Write a blog post about [topic]. The tone should be [describe the tone]. The blog should be at least 400 words.”
    - “Draft a tweet promoting [topic]. Include at least one emoji.”
    - “Suggest engaging questions I can pose to my Facebook community about [topic].”
    - “Brainstorm ideas for a TikTok series about [topic].”
    - “List key points for an infographic about [topic].”
    - “Create a catchy subject line for a newsletter about [topic].”
    - “Draft questions for an interview with [describe the person].”
    - “Write a compelling description for a [product]. Use the details below.”
    - “Suggest interesting topics for a podcast about [topic]."
- Strategy:
    - “Describe the strengths and weaknesses of a freemium pricing strategy.”
    - “How can businesses pivot their strategy in response to changing economic conditions?”
    - “Provide insights into creating a content marketing strategy for [topic or industry].”
    - “How can a brand integrate user-generated content into its digital marketing strategy?”
    - “What are effective ways to use retargeting ads in an e-commerce marketing strategy?”
    - “Outline strategies for startups looking to expand internationally.”
    - “What strategies can a new entrant in the market employ to challenge established players?”
    - “How can a business identify and respond to disruptive competitors?”

- SEO and SEM:
    - “Explain the differences between on-page and off-page SEO.”
    - “How do search engines determine the relevancy of a web page?”
    - “List key elements of a well-optimized webpage.”
    - “Describe the difference between short-tail and long-tail keywords.”
    - “How can I find and capitalize on LSI (Latent Semantic Indexing) keywords?”
    - “Describe effective white-hat link building techniques.”
    - “How can I optimize blog content to improve organic search visibility?”
    - “Describe the key components of a successful Google Ads campaign.”
    - “How can I improve the Quality Score of my PPC (Pay-Per-Click) ads?”
- Email marketing:
    - “Discuss the significance of email deliverability and how to improve it.”
    - “Outline the key components of a successful email marketing campaign.”
    - “How can I segment my email list for more personalized campaigns?”
    - “Describe the ideal frequency for sending marketing emails.”
    - “How can I incorporate storytelling into my email content for better engagement?”
    - “Explain how drip campaigns work and their benefits.”
    - “Describe scenarios in which triggered emails (e.g., cart abandonment) can be effective.”
    - “How can I set up an effective welcome email series for new subscribers?”

  - “Provide strategies to grow my email subscriber list organically.”
  - “How should I handle inactive or unengaged subscribers?”

Finance

For many business owners, the topic of finance is often one they would prefer not to deal with. It does require specialized expertise and can be tedious. Finance can also be time consuming – taking away from focusing on matters like creating new products or attracting potential customers.

Despite all this, finance is critical for any business. Consider research from Intuit Small Business.[10] Based on surveys from small businesses in the United States, Canada, and the UK, it shows that three out of five companies have unpaid invoices and almost half have cash flow problems. About a quarter of them use credit cards to meet payments.

But using ChatGPT and Bard can provide assistance:

- Accounting and bookkeeping:
  - “Explain the basics of cash flow management for small businesses.”
  - “What are the differences between accrual and cash basis accounting? Which one is better for a small business?”
  - “How can a small business owner set up a simple bookkeeping system?”
  - “Provide tips on managing accounts receivable.”
  - “How can businesses reconcile bank statements effectively?”
  - “Describe the importance of maintaining a chart of accounts.”
  - “What are the best practices for recording business expenses?”

  - “How can depreciation and amortization impact a small business’s financial statements?”
- Budgeting and costs:
  - “Describe the process and benefits of budgeting for a small business.”
  - “Discuss the differences between fixed and variable costs and their implications.”
  - “How can a business calculate its break-even point?”
  - “Provide strategies for adaptive budgeting in uncertain market conditions.”
  - “How can cost-volume-profit analysis assist in pricing decisions?”
  - “Discuss the impact of overhead costs on a product’s profitability.”
  - “Explain zero-based budgeting and its applicability for startups.”
  - “How can a business optimize its cost structure for sustained growth?”
- Financial management:
  - “Explain the importance of separating personal and business finances.”
  - “Discuss strategies for managing inventory to optimize cash flow.”
  - “How can a small business prepare for unexpected financial challenges?”
  - “Provide tips on negotiating better payment terms with suppliers.”
  - “Provide guidance on setting financial goals and milestones for a growing business.”
  - “Discuss the role of an emergency fund for small businesses.”
  - “How can businesses effectively manage debt and leverage?”

  - “Provide insights on risk management strategies in financial planning.”
  - “Explain the significance of financial ratios in business performance analysis. What are some useful ratios?”
  - “Describe ways to optimize working capital for business expansion.”

- Financing and credit:
  - “What are the options for small business financing and their pros and cons?”
  - “Explain the importance of a good credit score for a small business seeking loans.”
  - “How can businesses build and maintain strong relationships with lenders?”
  - “Explain the implications of co-signing business loans.”
  - “How can a business negotiate favorable loan terms and conditions?”
  - “What should be in a pitch deck for investors?”

Project Management

For businesses, there never seems to be enough time or resources. But this can be a threat. It can mean missing out on opportunities or failing to carry through on commitments.

This is why effective project management is a critical skillset. Then what can ChatGPT and Bard do?

Here are some sample prompts:

- Basics of project management:
  - “Explain the foundational principles of project management.”
  - “Describe the differences between projects, programs, and portfolios.”
  - “Discuss the significance of the project life cycle.”

- Planning and strategy:
  - "Detail the steps involved in creating a project management plan."
  - "How do you prioritize projects in a multi-project environment?"
  - "Explain the role of a Work Breakdown Structure (WBS) in project planning."
  - "Discuss the importance of stakeholder analysis during the initial phases of a project."
- Execution and monitoring:
  - "Describe best practices for maintaining momentum during project execution."
  - "How can project managers effectively monitor and control project deviations?"
  - "Discuss the role of Key Performance Indicators (KPIs) in project management."
- Risk management:
  - "Detail the process of risk identification in the early stages of a project."
  - "How can a project manager develop a risk response plan?"
- Leadership and team management:
  - "Describe the qualities of an effective project leader."
  - "How can project managers motivate a diverse team with varied skill sets?"
  - "Explain the role of emotional intelligence in project leadership."
- Project closure and review:
  - "Detail the steps involved in the project closure phase."
  - "How can lessons learned be documented and used for future projects?"'

    - “Discuss the importance of post-project reviews with stakeholders.”
    - “Explain the significance of celebrating project successes and recognizing team efforts.”

Product Development

Creating new products is fraught with risk. Clayton Christensen, who is a Harvard professor and best-selling author, has conducted extensive research on the topic. According to his analysis, the failure rate for new products is about 80%.[11]

In some cases, the launches can be epic disasters. One of the most notorious was the development of the video game for Steven Spielberg’s 1982 film, *E.T. the Extra-Terrestrial*. The studio sold the rights to Atari, which paid $21 million.[12] The company agreed to spend $5 million on the marketing.

Unfortunately, the game proved to be confusing – and not particularly fun. The sales came to only 1.5 million units. The remaining 2.5 million were dumped in a New Mexico landfill.

OK, if ChatGPT or Bard existed in the early 1980s, it may not have saved this project! But it could have still provided some realism.

Then let’s see the kinds of prompts that are useful for improving the odds of success for product development:

- Ideation and brainstorming:
    - “List 10 innovative product ideas for [topic].”
    - “What would be the potential benefits of combining [product one] with [product two]?"
- Market analysis:
    - “What are the primary pain points of the consumer in the [industry]?”
    - “Describe the target audience for [product]?”
- Design and prototype:

    - “What would be the user interface elements for a [product]?”
    - “Describe the prototype testing process for a new electric vehicle.”
- Marketing and launch:
    - “Create a marketing slogan for a [product].”
    - “Describe the ideal launch event for a [new product].”
- Feedback and iteration:
    - “Analyze the following user reviews [list them].”
    - “Create a user feedback form for a [product].”

AI Code Development Tools

Andrej Karpathy, who is a founding member of OpenAI and the former senior director of AI at Tesla, tweeted this about the GitHub Copilot system, which automates code development:

> *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.*[13]

AI code development tools are one of the early use cases for generative AI, and they have quickly become a must-have for developers. By far, the tool with the highest adoption is GitHub Copilot, which is owned by Microsoft.

In June 2023, the CEO of GitHub, Thomas Dohmke, wrote a blog about the one-year anniversary of the app.[14] He pointed out the following:

- More than one million developers activated the app.
- There was adoption from over 20,000 organizations.
- GitHub Copilot had generated more than three billion accepted lines of code.

But this is just the start. By 2030, there are expected to be about 45 million professional developers. It’s a good bet that most of them will use AI generation tools.

GitHub Copilot has shown strong gains in developer productivity. For example, code development tasks are done – on average – 55% faster. About 75% of the developers said they felt more fulfilled when using GitHub Copilot.

The tool has also been effective for junior developers. The AI has become a teaching tool for best practices.

This is not to imply that you should use GitHub Copilot. There are various other high-quality tools available, including

- Amazon CodeWhisperer
- Bito
- Codeium
- Google Duet AI
- Magic
- Replit Ghostwriter
- Tabnine

Another key benefit of AI code generation tools is that they can help with the persistent developer talent shortage. Automating some of the processes can be a way to lower costs and improve the speed of development.

It’s important to note that AI Code Generation tools will not create a complete app (at least not now). These systems are assistants, which can help with creating blocks of code. This has become known as “pair programming.”

With ChatGPT or Bard, the code creation process is about coming up with a prompt. Here are some samples for the Python language:

- “Write Python code to reverse a string.”
- “Can you provide a Python snippet to remove duplicates from a list?”

- “Write Python code to save a list of strings to a file.”
- “Provide a Python example to connect to a SQLite database and retrieve data.”
- “Show me how to use the map function to square each number in a list.”

Of course, with a system like GitHub Copilot, there are more capabilities. First of all, it is integrated in an IDE or Integrated Development Environment, which is the editor to write code.

Then how does an AI generation system work? Let’s take an example. Suppose you create a string – in the IDE – that holds a paragraph. Then you enter the word “length.” The system will then suggest a function that will calculate the length of the characters of the string.

Next, you can write a prompt, like we did earlier. You can then have a chat with the system to tweak the code. Think of this as prompt-based programming.

Another useful function is code translation. If you have a program written in Java, you can convert this to another language. This is a matter of writing a prompt.

AI generation tools are not without their drawbacks, though. Sometimes, the code can have errors. It may also be wordy.

But then again, there will be a need for an experienced developer to review and test the code.

Note Bill Gates was a top-notch programmer, which was critical for the hypergrowth of Microsoft. He knew a key to success was lean code. He once noted: “Measuring software productivity by lines of code is like measuring progress on an airplane by how much it weighs.”[15]

AI Text-to-Image Systems

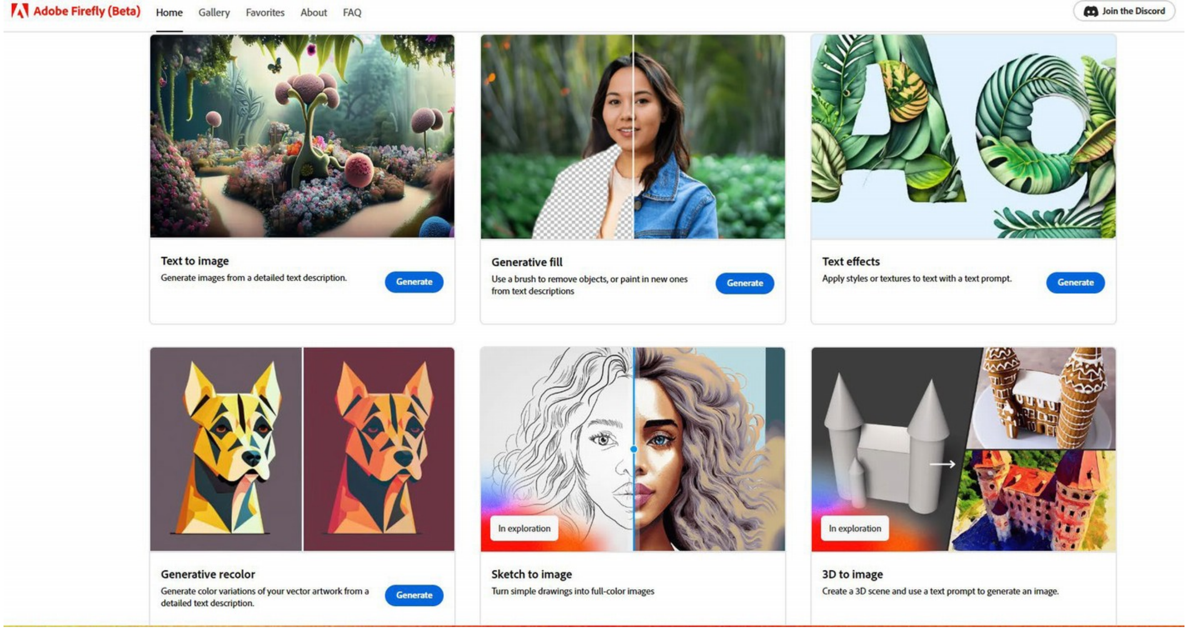
An AI text-to-image system allows you to create images by using natural language prompts. The images are usually high-quality, as if created by a professional. Although, there are some quirks. There could be problems with representing words or phrases. In some cases, a system may add an extra arm, leg, or head!

Regardless, the technology is quite useful. Creating images is a time-intensive process. What's more, if you do not know how to use a graphics tool, the output can look, well, awful.

Some of the top AI-powered image tools include the following:

- OpenAI's DALL·E 2
- Midjourney
- Stable Diffusion
- Adobe's FireFly

Let's take an example. I'll use Adobe's FireFly, which you can see in Figure 5-1.



Text to image
Generate images from a detailed text description.

Generative fill
Use a brush to remove objects, or paint in new ones from text descriptions

Text effects
Apply styles or textures to text with a text prompt.

Generative recolor
Generate color variations of your vector artwork from a detailed text description.

Sketch to image
Turn simple drawings into full-color images

3D to image
Create a 3D scene and use a text prompt to generate an image.

Figure 5-1 This is the home page for Adobe's FireFly, which allows for using natural language to create images

You can see that it has six features:

- Text-to-image: Yes, this means you can use natural language to create an image.

- Generative fill: This allows you to edit an image, such as by removing objects. You can even paint new ones by using natural language.
- Text effects: You can apply styles to text. You can use this to create compelling titles or even logos.
- Generative recolor: With a natural language description, you can create different color variations.
- Sketch to image. You can upload a sketch and turn it into a colorful image.
- 3D-to-image: You can use natural language to create a 3D image.

Let's see an example. Suppose we have written a blog post about a home service (yes, we did this using ChatGPT or Bard). We can create an image for this by using the following prompt:

Make an image of a home.

Figure 5-2 shows four homes, which are illustrations. You can select one and then make changes.

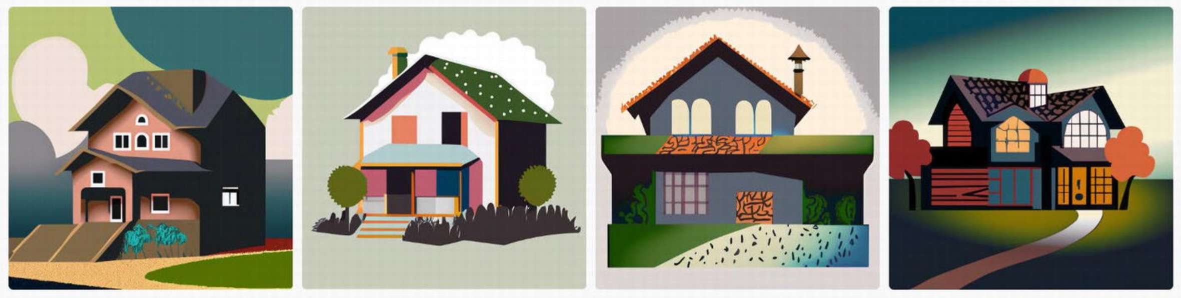


Figure 5-2 This shows four images that FireFly has created

Conclusion

For this chapter, we have covered the many use cases for ChatGPT and Bard. Just some include human resources, marketing, sales, finance, and project management.

We also looked at specialized tools, such as for using generative AI to create code. We then discussed text-to-image systems.

However, it's important to be careful. The generative AI output can be vulnerable to issues with privacy, bias, and discrimination.

This is why you need to have a thorough review. And when it comes to matters like the law, you must get the advice of qualified advisors.

As for the next chapter, we'll look at change management.

Footnotes

1

www.alignable.com/forum/57-of-smbs-have-already-tried-chatgpt-26-already-feel-it-will

2

www.pwc.com/us/en/tech-effect/cloud/hr-tech-survey.html

3

www.gartner.com/en/newsroom/topics/human-resources/2023-07-12-gartner-survey-finds-only-5--of-hr-leaders-report-the

4

https://fortune.com/2023/06/26/human-resources-hr-artificial-intelligence-talent-risks-costly/?utm_source=Iterable&utm_medium=email&utm_campaign=fortune-at-work-generative-ai&tpcc=NL_Marketing

5

www.fisherphillips.com/en/news-insights/eeocs-latest-ai-guidance-sends-warning.html

6

www.legalmatch.com/law-library/article/how-much-will-a-business-lawyer-cost.html

7

https://cx-trends-report-2022.zendesk.com/opportunity

8

https://blog.hubspot.com/sales/generative-ai-in-sales

9

www.gartner.com/en/newsroom/press-releases/2023-08-23-gartner-survey-finds-63-percent-of-marketing-leaders-plan-to-invest-in-generative-ai-in-the-next-24-months

10

https://quickbooks.intuit.com/r/small-business-data/survey-2022/

11

www.bizjournals.com/bizjournals/how-to/marketing/2017/03/why-so-many-new-products-fail-and-it-s-not-the.html

12

www.usatoday.com/story/money/2018/07/11/50-worst-product-flops-of-all-time/36734837/

13

https://twitter.com/karpathy/status/1608895189078380544?lang=en

14

https://github.blog/2023-06-27-the-economic-impact-of-the-ai-powered-developer-lifecycle-and-lessons-from-github-copilot/

15

https://blog.ndepend.com/mythical-man-month-10-lines-per-developer-day/

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_6

6. Change Management

Building the Right Foundation

Tom Taulli[1]
(1)
Monrovia, CA, USA

IT (Information Technology) projects are often challenging to implement, say with ERP (Enterprise Resource Planning) and CRM (Customer Relationship Management) systems. It's common for organizations to suffer disappointing returns on investment.

But when it comes to AI implementations, the failure rates are much higher among IT projects. Just look at a survey from Altair.[1] About 33% of the respondents said that more than half of their AI projects never reached production and 55% indicated that over a third stalled within a two-year period.

Or consider a survey from Gartner, which included 699 respondents in the United States, Germany, and UK.[2] For 2022, it was found that 54% of AI projects had gone from pilot to production.

Arijit Sengupta, the founder of Aible, has said: "As an industry, we're worse than gambling in terms of producing financial returns."[3]

So what's going on? Why the dismal results from AI? As should be no surprise, there are many factors that explain the problems.

In a paper titled "Failure of AI projects: Understanding the Critical Factors," academics Jens Westenbergera, Kajetan Schulerb, and Dennis Schlegel explored this topic.[4] They found that AI technology was an issue. After all, it can be extremely complicated, requiring the help of experts like data scientists. AI can also be unpredictable and give off false signals. Then there are the difficulties with preparing datasets.

Yet the research paper showed that organizational problems were often even more significant. For example, smaller firms can have more issues because of the lack of resources.

But regardless of size, there are the problems like

- Unrealistic expectations
- Bureaucracy
- Lack of AI literacy
- Insufficient focus on value-based use cases
- Lack of adequate training

This is why change management is critical for AI success. It's about focusing on the "people." This is what we'll cover in this chapter.

Definition

The Center for Management & Organization Effectiveness defines change management as the following:[5]

> *The definition of change management is the process of effectively guiding individuals and groups through change, planning for change, overcoming resistance to change, and helping organizations to thrive during change.*
>
> *Change management is accomplished through diverse tools and leadership techniques, understanding of human behavior, budget items that support change management, consistent change frameworks, and well-crafted implementation plans.*

This is a good overview. But of course, there is much more to this topic. It’s also important to understand some of the history of change management and how it has evolved over time. This is a dynamic field.

Social scientist Kurt Lewin is the pioneer of change management, who wrote various articles about the topic in the 1920s and 1930s. At the heart of his research was group dynamics and how they can be difficult to manage.

But it was during the 1960s that change management started to gain traction in business. Interestingly enough, this was due to studies about the impact of grieving because of the loss of jobs. Everett Rogers came up with a framework to understand how people adapt to major changes.

In the 1980s, McKinsey & Company saw an opportunity to leverage change management with its consulting engagements. One of its consultants, Julien Phillips, created a model for this.

A notable adopter was Jack Welch, who was the CEO of General Electric in the 1980s and 1990s. He had to enact massive layoffs and restructure the massive organization. He understood the risks, such as the inherent issues with morale. But he also knew that a successful organization needed to focus on innovation and efficiency to thrive and grow. Through change management, Welch was able to pursue these goals and turn GE into a growth company. He encouraged smaller teams, risk taking, and accountability.

With the success of GE, change management became a hot topic for business in the 1990s. This led to a cottage industry of management experts who wrote books, such as *The Fifth Discipline: The Art & Practice of the Learning Organization* by Peter Senge. The theories and frameworks became more detailed and rigorous, especially since there were more use cases.

The most influential thought leader in change management was John Kotter. He wrote the pathbreaking book, *Leading Change*, which sold millions of copies. In it, he set forth his own eight-step framework for change management:

- Create a sense of urgency: You want to motivate your team to achieve bold goals.
- Build a guiding coalition: You want a team that is passionate and committed to the goals of the change management.
- Form a strategic vision: You want to develop a clear-cut mission.
- Enlist a volunteer army: As much as possible, you want to encourage participation across the organization.
- Enable action by removing barriers: You need to tear down obstacles to success.
- Generate short-term wins: Progress is about building up small achievements.
- Sustain acceleration: When there is momentum with the change management, make sure you find ways to keep it up.
- Institute change: You want the results of the change management to become part of the DNA of the organization.

In the 2000s, change management experts continued refining the frameworks as well as revaluating some of the existing approaches. For example, the perception was that change management was too focused on top-down strategies. There needed to be more autonomy with teams. This resulted in new theories like agile, which have become widely used for IT projects.

Despite all this, many of the core principles that have evolved over the decades remain relevant.

Note Daryl Conner is a management consultant and an expert in change management. He's often called the "change doctor." In 1993, he published a book, *Managing at the Speed of Change*, which quickly became a business classic. In it, he coined the phrase the "burning platform." It's a reference to an explosion on an oil rig in the North Sear in 1988 and how Andy Mochan survived by jumping off the platform. Conner saw this as an effective metaphor for how companies can navigate treacherous changes and find ways to go beyond the status quo.

Types of Change Management

Change management is not a monolithic concept. There are different types.

First of all, there is unplanned change management. This is where there is an external event that has a major impact on the organization – and there is little choice but to respond. This is certainly what happened with the Covid-19 pandemic as organizations had to scramble to deal with remote work and protecting their employees. There was also the urgency to manage disrupted supply chains.

But unplanned change management can also be about a crisis that is specific to one company. This could be a cybersecurity breach or an IT outage.

Unplanned change management will test an organization's strength and culture. But AI can be helpful as well. It can assist in detecting ominous trends, say with analyzing customer sentiment. AI can also be a way to detect changes with competitors or new rivals.

Next, there is remedial change management. The focus is primarily on fixing a problem with an organization. This could be when deadlines are being missed or there is too much bureaucracy or even politics.

Remedial change management is not necessarily about punishment. Rather, it's about having more transparency and accountability. However, if there is not enough progress, then there may need to be tougher actions taken.

Another type of change management is the incremental approach. This is a more gradual way to implement an initiative and it is usually focused on a certain process or department.

True, the incremental approach may not necessarily result in big improvements. But this is fine. When an organization is starting down its journey with AI, the incremental approach can be the right way. If anything, it's about learning and getting some quick wins that can help bolster the confidence of the team. This can then build a better foundation for bigger efforts in the future.

A challenge is knowing when to take change management to the next level. An organization can spend too much time on the incremental approach, which can instill caution.

Finally, there is transformational change management. As the name implies, this is a massive effort that usually impacts many or all departments.

For this to be successful, there needs to be buy-in from the executive team and enough resources. The culture also needs to be ready.

However, the risks of transformational change are significant. The fact is that these initiatives usually underperform or even fail.

Then again, transformational change management is essential to get the true benefits of AI. This means scaling the technology in a big way – which will mean outcomes that will move the needle.

Accenture conducted a study of 1,500 C-Suite executives from companies with a minimum of $1 billion in revenues.[6] These businesses spanned 12 countries and 16 industries.

The study found that 76% of the respondents recognized the challenges of scaling AI across their organizations. Something else: three quarters of them said that – if they do not adequately scale AI – they risk going out of business during the next five years.

The Culture

A core strategy for the US government for the Vietnam War was to win the "hearts and minds" of the populace. The assumption was that this would lead to victory. But getting the support of the Vietnamese people proved to be futile.

While much less dramatic, the concept of "hearts and minds" is applicable to change management. If there is not enough buy-in from the employees, then any plan will be doomed to failure.

The reality is that there are often different viewpoints about change management between managers and employees. For managers, they see this as an opportunity to drive growth and innovation. Granted, this can be risky. But managers often realize that they often need to take bold moves to deal with competitive or industry threats.

Employees, on the other hand, often see change management as ominous. Let's face it, there are many examples where new programs have led to layoffs or reassignments.

If the change management does not meet up to expectations, then employees may take the blame – and this can stunt career advancement. Or, even if it succeeds, they may not get much of the benefits anyway. It could mean doing more work.

Because of this, managers need to be empathetic. They must get in the shoes of the employees and try to see the world from their eyes. If not, then the change management will likely be difficult to pull off.

This is definitely the case with AI. The technology is generally associated with job losses. So from the start, employees will mostly be resistant and may even try to undermine the efforts.

This is why it's important for managers to be clear about the benefits of the AI. It means showing how the technology will make the lives of employees easier. They will also be able to devote more of their time to issues where human abilities are best used – not mundane or repetitive tasks.

But the messaging must be intensive. It should not be just a perfunctory presentation. The messaging needs to be ongoing and include various channels, whether blogs, emails, and access to resources (say on a corporate portal). Putting on a variety of workshops can be effective. This can be a way to not only focus on the benefits of AI but show how the technology works. This can help reduce the anxiety.

An interesting case study of the potential dangerous impacts from the disconnect between management and employees is Philips during the 1980s.[7] The company had a long history of success, having brought to market innovations like the compact disc, the video recorder, and the audio cassette. The employees included many top-notch engineers.

Philips operated on a matrix structure for its management. While this allowed for more autonomy, it also meant growing complexity. This actually made it difficult to fend off rivals, such as from Asia. The result was that product delays became more common, costs got bloated, and profits turned to losses.

Two CEOs at Philips tried to implement change management to reverse the company's slipping fortunes. But the reorganizations failed. This was despite ongoing communications and clear-cut plans.

The problem? It was really about the entrenched culture. There was an implicit agreement that employees would be loyal in exchange for lifelong employment. But this often meant the performance and career advancement were based not on merit, but relationships and politics. When it came to compensation, this was often about an employee's seniority, not performance.

In other words, the two CEOs of Philips did not address this. Unfortunately, the company would quickly decline and by the mid-1990s, it was on the verge of bankruptcy.

Jan Timmer came in as the new CEO and he took swift action to turn around the company. The main focus was on the culture. He first had an offsite with his top managers and clearly stated the problems of the company. If radical actions were not taken, the company would not survive. To emphasize his point, he passed out a copy of a hypothetical press release that announced the bankruptcy of Philips.

He said that everyone needed to make changes – and quickly. He set forth several major goals, which included layoffs. But he also stressed that the culture needed a change. There needed to be accountability and performance-based approaches. He then tasked his managers to make plans to carry out the change management and they would then have to sign them.

To push this forward, there was ongoing communications. Timmer would also have periodic town halls. He wanted to show that his plans were essential for the company.

The strategy paid off. The morale improved significantly, and Philips returned to growth and profitability.

But this did highlight the incredible importance of a company's culture. It is something that can be subjective and complex. But for managers to enact real change in an organization, there needs to be first an honest understanding of the culture.

Roles

For a change management program, there are often various roles. To achieve success, they do not have to be formalized and there does not need to be a certain number of people on the team. There may actually be employees who take on a role but do not have a title.

Yet it is still a good idea to understand the typical roles with change management. For example, there is the sponsor, who is the person that makes the case for the change. Often this is an executive who has the ability to leverage resources and has the authority to enact new programs. This can be any executive, whether in HR, finance, sales, operations, or marketing. But to help improve the odds of success, it's usually best to get the support of the CEO. This is a sign that the initiative is a clear priority.

Transparency is also essential. If an automation program will result in lost jobs, then the sponsor needs to be honest about this and show who will be impacted and when. While this can be a shock to the organization, it will instill trust. Employees will realize that management can be relied upon. This will make it easier to allow for future change management programs.

An effective sponsor will need to engage in detailed planning. If the plans are too ambiguous, then the results will ultimately be disappointing.

It also helps if the sponsor can tell a story or narrative about the change. This will bring it to life for employees and help guide them.

Steve Jobs knew the power of stories, which motivated his teams to achieve huge goals. He once said the following:

> *The most powerful person in the world is the storyteller. The storyteller sets the vision, values, and agenda of an*

> *entire generation that is to come… and Disney has a monopoly on the storyteller business. You know what? I am tired of that bullshit, I am going to be the next storyteller.*[8]

He said this while he was the CEO of Pixar. At the time, he was doing something that seemed impossible: create a global entertainment company. For the most part, Pixar was still struggling and had not yet released its Toy Story film.

But his storytelling abilities were able to excite his employees to make breakout hit movies that became franchises. Eventually, Disney would buy Pixar and make it the heart of its animation business.

It's true that Jobs had natural abilities. But he was far from perfect. There were times when his storytelling was wide off the mark, such as with the development of the Lisa computer. While it was a technical marvel, he did not provide a narrative that enthused customers.

Besides the sponsor, there should also be a project manager. The sponsor does not have the time to manage the day-to-day tasks of a change management program.

However, a project manager does not need to be highly technical. But they should have a high-level understanding of the technology. They should also have a track record of managing complex projects, such as with setting forth realistic plans. The project manager should know when to encourage employees and provide praise.

Change Advisory Board (CAB)

To help coordinate and lead a change management program, you can setup a change advisory board or CAB. This can certainly be quite effective. A CAB can help improve buy-in and show the seriousness of the effort. Of course, it will be critical in ensuring that the change management program is on track.

Often there will be periodic meetings. But in the meantime, there should be ongoing communication.

This is not to imply that CABs are a cure all. Let's face it, committees can be a way to stall a program and add extra red tape!

This is why a CAB needs a clear mandate and power within the organization. For example, it should not get too involved in the day-to-day details of the change management. Instead, the focus is on broader goals – such as evaluating the results, allocating resources, managing the budget, and taking actions when the program is not meeting expectations.

A common analogy for a CAB is an air traffic controller. The role is to provide guidance in making sure the planes take off, land, and get to their gates safely. But the pilots are the ones who fly the planes.

For the most part, if the agenda of the CAB gets lengthy, then this is a danger sign.

Training

Various studies show that an impediment to AI initiatives is a lack of understanding of the technology. This is definitely understandable because AI is highly complex. Even experts will not know the many categories of this sprawling technology. Then there are the constant innovations and new developments.

Because of this, it's important to have a training program in place to help employees. This can focus on the high-level aspects of AI. There is no need to know about the mathematics.

Some of the topics for training include the following:

- Introductory statistics
- Machine learning principles
- Deep learning concepts
- Generative AI and transformers
- Risks of AI
- The capabilities of AI

There are many online courses that can help out. In fact, this is something I currently help companies out with – via my company OnePrompter.com.

Regardless of which one you choose, make sure they are focused on non-technical employees. There should also be a way to evaluate the learning of your employees, such as with quizzes.

A good training program will not only bolster skills but help alleviate some of the fears of AI. Once people start to understand the technology, it will often be less threatening.

Change Fatigue

While change management can drive standout results, there is the risk of too much change. This is actually common. Given the dynamic nature of today’s world, managements are trying to find ways to be nimble and grow. But employees can get overwhelmed when there are too many change management programs. This can lead to widespread failure.

According to a study from Gartner, the typical employee experienced ten planned enterprise changes in 2022. These included restructurings, layoffs, and technology implementations or modernizations. In 2016, the average number of planned enterprise changes came to only two.

In light of this, it’s no surprise that employees have grown more distrustful of new programs. The support levels have gone from 74% in 2016 to 43% in 2022, according the same Gartner study. This is even though some of the new programs were attempts to help employee morale.

In light of this, there needs to be an evaluation of the timing of change management. If there have already been many in place, then adding another one could be counterproductive. It could also dilute the existing programs.

A best practice is to provide transparency about all the ongoing initiatives, which could be posted on a dashboard. This visibility will help get a sense of the potential for change management fatigue. It will also encourage more prioritization. That is, before a program is put into place, there needs to be intensive scrutiny. As new priorities arise, some existing programs could be put on pause. Of course, if they really turn out not to be essential, they could be cancelled.

> **Note** With a change management program, leadership will often spend much time trying to sell the benefits to employees.

This is fine and important. But it can go too far. There should also be periodic communications where there is the garnering of feedback from employees. This helps gauge their views and understanding of the goals. Gartner shows that this approach can lead to a 24% improvement in the odds for success of a change management program.[9]

The Plan

While there should be considerable planning when creating a change management plan, this does not mean it should be a long document. If it is, it could actually be a hindrance. Who wants to read a document that goes on and on?

Few people do.

It will also be difficult for employees to know what's expected of them. This is why the plan should be tightly organized, clear, and actionable. It doesn't even have to be a document. It could be a PowerPoint or an outline in a Google doc.

So what should be in it? There are no hard-and-fast rules for this. The topics are on a case-by-case basis.

But here are some to consider:

- Strategic goals: Ultimately, what should be achieved with the AI implementation? Some of the typical ones include lower costs, faster processes, higher quality, and improved customer or employee experiences.
- Roles and responsibilities: As indicated earlier in this chapter, there should be a sponsor and a project manager. There may even be a CAB. Regardless, what's even more important is setting clear responsibilities. What's expected? How much time should be devoted to the program?
- Budget: This can be tough to estimate for an AI project. While the costs of software tools can be easy to determine, this may not be the case when looking at using APIs. These are becoming more common with AI, such as with large-

language models like from OpenAI and Google. If an AI system has lots of usage, this can certainly add up in terms of the fees. This is why there should also be an estimate of the ROI (Return On Investment). In some cases, the costs may be too significant when compared to the overall benefits.

- Schedule: It does not have to be too detailed. But there should be dates for delivery of certain milestones. This can be a way to help get a sense of the scope of the project – such as the steps required for success.
- Incentives: Make it clear that there will be bonuses or perks for achieving goals. There can also be a structure to provide even more benefits if the results exceed expectations.

With something like AI, it can be tough to estimate timelines and key performance indicators (KPIs). These are often moving targets. This is why a change management plan should be a general guide that is subject to adjustments. This should also not be a hindrance to agility.

Implementation

Given what we've covered so far, you should have a good foundation for the implementation of the change management plan. But you should continue with the ongoing communication. There is never too much. If you are not getting tired of saying certain things over and over again, then there is likely a problem.

The sponsor and project manager should also be alert to any problems. How can these be addressed? Are there ways to take quicker action? A key is finding ways to empower the team.

Whenever there is a short-term win, then this should be highlighted. As the saying goes: providing praise costs you nothing.

Another useful tool is to have a periodic survey, which should be confidential. This will help provide valuable input about the status of the project.

At the completion of the project, there should be a “post mortem.” For this to be useful, there must be an honest assessment. Were the goals hit? What about the budget and schedules?

But what’s most important is understanding the lessons. AI is something that requires ongoing learning.

In fact, even if a project fails to meet expectations, this is something that should not be harshly criticized – unless the team did not put in enough effort. Innovation is risky and prone to failure. Even the greatest companies have had problems. But if failure is punished, then it will be challenging for an organization to take the necessary risks in the future.

Even with a successful project, there are potential problems. Perhaps the biggest is that employees will eventually revert to prior approaches. Keep in mind that this is a common response. The fact is that inertia is a powerful force within an organization and can easily derail a change management project.

Because of this, the sponsor and project manager should continue to engage in monitoring.

Tools

There are many software packages to help with change management. These can provide resources like templates for communications, tracking systems, polling/surveys, and project management capabilities. Some popular software tools include

- Whatfix
- BMC’s IT service management (ITSM) platform
- ChangeGear Change manager
- ServiceNow Change and Release Management application

Yet a system like Google Sheets or Microsoft Excel could work fairly well. There is no need to have employee training for these systems since they are already part of the daily process. Besides, the main benefit of a change management software tool is having a way to centralize resources, such as the plan, schedules, and KPIs.

There are organizations that provide certifications and training for change management. These can be particularly helpful. Here are some examples:

- Change Management Institute (CMI): It has Foundation, Specialist, and Master certifications.
- Association of Change Management Professionals (ACMP): This organization provides the Certified Change Management Professional designation.

Boiling the Ocean

Of course, boiling the ocean is impossible. The resources needed for this would be impossible to muster.

The concept of "boiling the ocean" has also become part of the vocabulary for business initiatives. It's when a project's expectations are overly ambitious – and poised for inevitable failure.

This is common for AI. The technology is subject to hype. AI also garners lots of excitement. It seems like it can do anything, right?

Well, when it comes to implementing this technology, the challenges can be daunting. Even some of the world's top tech companies have had catastrophic failures.

A notable case study is IBM's Watson, a massive AI computer system. The inspiration for this system was the result of the company's search to find a way to highlight the power of this technology. It would be similar to what IBM accomplished in 1997, when its Deep Blue computer beat chess champion, Garry Kasparov. The result was considerable PR and customer interest.

In 2004, IBM Research manager Charles Lickel came up with a new idea. It came about from watching Ken Jennings win 74-straight games of Jeopardy!

Lickel thought: Why not build a computer to be a contestant on the show?

It was a great idea and IBM's marketing team loved it. This also spurred innovation to create new AI systems to allow for understanding human Q&As.

Over a process of six years, IBM engineers created Watson, which relied on a training set of 200 million pages of data. The system reached a point where it outperformed humans most of the time.

In February 2011, Watson appeared on three episodes of Jeopardy and took on the game's all-time champions, Jennings and Brad Rutter.

Well, it turned out to be a one-sided competition. Watson handily won, taking in $77,147 in winnings versus $24,000 for Jennings and $21,600 for Rutter. It also won $1 million for the competition (IBM donated it to charity). After this, Jennings noted: "I, for one, welcome our new computer overlords."[10]

The PR impact was huge, and IBM wasted little time in leveraging it. The company produced commercials about Watson that highlighted its capabilities. The company also set forth a strategic plan to use the system to help diagnose and make discoveries about cancer.

Even with all the excitement, there was still some skepticism about the strategy. For example, IBM scientist David Ferrucci considered that Watson was not the right system for cancer research and diagnostics.[11] It was mostly for answering structured questions for quizzes. It did not have deep reasoning capabilities and was nowhere near able to understand the complexities of diseases.

There was also pushback from IBM executive, Martin Kohn. He thought that the focus should initially be on less ambitious categories. This would help build momentum and better evolve the technology.

But this feedback fell on deaf ears. IBM wanted to go big. The company saw this as an opportunity on par with something like the Apollo project.

IBM devoted large amounts of resources to Watson. The company also struck various acquisitions, which amounted to billions of dollars.[12]

But all this was not enough. The results from Watson were proving to be far from encouraging.

An example was with the MD Anderson Cancer Center in Houston. The project was to leverage Watson to deliver recommendations for cancer treatments. For this, MD Anderson shelled out $62 million.[13]

However, the project was plagued with issues integrating EHR (electronic health record) systems and interpreting the complex datasets. There were also problems with shifting goals and project management.

MD Anderson would eventually cancel the project.

After this, IBM did take an honest assessment of Watson and redirected its strategy. The company would instead focus on areas where the company had deep expertise, like accounting, risk management, customer service, compliance, and business automation. As for the healthcare business, IBM would sell off many of these assets.[14]

This case study is not unusual. There are plenty of large AI projects that wind up being disappointments. In other words, there should be lots of caution and realism. AI is something that is about a journey of learning – not making monumental bets.

Conclusion

Change management is crucial for the success of AI. There is often a need for major changes in how employees work.

In this chapter, we looked at some of the types of change management. Examples include unplanned change management, remedial change management, the incremental approach, and transformational change management.

We also looked at best practices for success. Some of the key factors include having a clear vision and plan, buy-in, an understanding of an organization’s culture, and training. You might also want to consider setting up a CAB.

As for the next chapter, we’ll take a look at the risks of using ChatGPT and Bard.

Footnotes

[1] www.datanami.com/2023/06/06/altair-survey-unpacks-the-friction-surrounding-ai-and-data-projects/

[2] www.gartner.com/en/newsroom/press-releases/2022-08-22-gartner-survey-reveals-80-percent-of-executives-think-automation-can-be-applied-to-any-business-decision

[3] https://fortune.com/2022/07/26/a-i-success-business-sense-aible-sengupta/

[4] www.sciencedirect.com/science/article/pii/S1877050921022134

[5] https://cmoe.com/glossary/change-management/

[6] www.accenture.com/us-en/insights/artificial-intelligence/ai-investments

[7] https://hbr.org/1996/05/why-do-employees-resist-change

[8] www.quora.com/Steve-Jobs/What-are-the-best-stories-about-people-randomly-or-non-randomly-meeting-Steve-Jobs/answer/Tomas-Higbey

[9]

https://hbr.org/2023/05/employees-are-losing-patience-with-change-initiatives

10

www.wsj.com/articles/SB10001424052748704171004576148974172060658

11

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

12

www.wsj.com/articles/ibms-retreat-from-watson-highlights-broader-ai-struggles-in-health-11613839579

13

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

14

www.wsj.com/articles/ibm-sells-watson-health-assets-to-investment-firm-11642772712

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023
T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_7

7. Understanding AI

Machine Learning, Deep Learning, and Generative AI

Tom Taulli[1]
(1)
Monrovia, CA, USA

In February 2023, Pew Research published an extensive survey about AI,[1] which included responses from over 11,000 US adults.

The research revealed that 27% of Americans work with AI applications at least several times a day and roughly 28% do so once a day or several times a week. Yet 44% said they did not have many interactions with the technology.

But the survey also showed that 70% of Americans could not identify six common forms of daily uses of AI. What this points out is that they often underestimate the impact of the technology. The research shows that there are higher levels of awareness based on factors like the frequent use of the technology, education, and income levels.

No doubt, the understanding of core concepts like machine learning, deep learning, and generative AI is fairly low – even though these technologies are having a growing impact on society.

But to allow for more effective use of this technology, there really needs to be more education, such as on the core concepts. If not, there could be a growing lack of trust. The Pew Research showed that there is already much skepticism with AI. Only 15% of

Americans are more excited than worried about this technology when used in daily life.

In this chapter, we'll take a look at the fundamentals of AI. The focus will be on high-level details, which is mostly what is needed for employees. The following chapter will then show the importance of data for effective AI.

In fact, these next two chapters will be helpful for those companies that are looking to go beyond ChatGPT and Bard. This is where you can customize an LLM by accessing an API. This can make it possible to create powerful applications.

True, this will require having some developers on your team. But when going down this, path it will be critical to understand the core concepts of AI and data.

Brief History

AI may seem like a new technology. But this is far from the case. The technology has been a focus of serious academic research for decades.

A critical event was a conference at Dartmouth University in 1956. The legendary professor, John McCarthy, put this together and he called it "Cerebral Mechanisms in Behavior."

Yes, it was not a catchy title. But this conference would focus primarily on AI. In fact, McCarthy coined this term at the time.

There was much criticism but the term stuck. The main reason was that no one else offered a better alternative.

Many of the attendees of the conference would become giants in the AI field. Some of them included Marvin Minsky, Oliver Selfridge, Ray Solomonoff, Claude Shannon, Julian Bigelow, Allen Newell, and Herbert Simon. They had discussions and panels about concepts like neural networks and natural language processing (NLP).

But the conference was not just about theories. Newell, Cliff Shaw, and Simon provided a demo of the world's first AI program, which was the Logic Theorist. Simon came up with the idea for this program while working at the Research and Development (RAND) Corporation. On a research project, he saw how a computer printed words on a world map that allowed for understanding different nuclear first-strike possibilities from the Russians.

So why now use a computer to help with working on formulas? To this end, he used the monumental book, *Principia Mathematica*, for training the system to solve complex theorems. Bertrand Russel, who was one of the authors of the text, was impressed with the Logic Theorist and noted that one of the solutions was more elegant than what he wrote about.

After this, there would emerge a "Golden Age of AI." This period saw tremendous advances in technology, and much of this came from universities like Stanford, MIT, Lincoln Laboratories, and Carnegie Mellon University. They were benefiting from aggressive federal government programs for research for the space program and the Cold War.

The investments would lead to rapid advances in hardware systems, such as with semiconductors. The development of the transistor and integrated microchip would be transformative. Then there were innovations in storage and networking systems.

As for AI technologies, there were a myriad of applications. Some of these were visionary for the period. An example is computer vision. In 1966, Minsky wanted to find ways to make this happen and tasked one of his MIT students to do so. His instruction was simple: attach a camera to a computer. But of course, implementing this was no easy feat, especially since the computer systems were limited. Yet the student did an outstanding job.

Another breakthrough application of AI was ELIZA, which also came out in 1966. It was actually the world's first chatbot. The creator of ELIZA was MIT professor Joseph Weizenbaum. The application allowed a person to ask questions to the computer via a keyboard. While the underlying process was basic, it still seemed realistic.

A key use case was as a virtual therapist. Some users even thought Eliza was real and were more willing to interact with it than with a human therapist.

Weizenbaum was stunned by the response to ELIZA. He thought it showed how people can easily impute human aspects to a machine. This has become known as the "ELIZA Effect."

By the way, you can find a version of ELIZA on the Internet at: https://web.njit.edu/~ronkowit/eliza.html

Another interesting AI application was ANALOGY. MIT professor Thomas Evans created this in 1969, and the program allowed users to solve analogy problems for IQ tests.

At the core of many of these AI applications was a special programming language. It was called LISP, which McCarthy developed in 1958. LISP was a marvel of cutting-edge innovations. It had features like garbage collection, dynamic typing, and recursion – all of which are staples of modern languages.

As the research continued in academia, there were two main approaches to creating AI. One was known as a symbolic system. This is where the underlying logic is highly structured, consisting of IF/THEN/ELSE statements.

This would become the basis of many of the earliest commercially successful AI programs. These were known as expert systems and became a thriving business in the 1980s. At the time, Carnegie Mellon University professor John McDermott developed XCON (eXpert CONfigurer), which was a recommendation engine. DEC adopted this technology to make it easier to select products for the VAX computer system. It would become one of the company's most successful – and profitable – businesses.

The other approach to AI was the perceptron. The pioneer of this system was Frank Rosenblatt, who was a research psychologist at Cornell University. He applied fundamental concepts of the brain, such as neurons, to computer systems. Basically, it would ingest data and find patterns. This would be the early version of machine learning and deep learning.

In 1958, Rosenblatt created the first application of the perceptron. He used a sophisticated IBM mainframe system and the program deciphered the marks on the sides of punch cards.

The media learned about this innovation, which caused a stir. A reporter from the New York Times wrote: "Indeed, it strikes us as the first serious rival to the human brain ever devised."[2]

This was obviously an exaggeration. Yet Rosenblatt showed that computers could essentially learn by processing data.

Despite this, there was considerable skepticism in the academic world as the symbolic system remained much more popular. In 1969, Minsky and Seymour Papert published an influential book called *Perceptrons*. In it, they demonstrated the various shortcomings of Rosenblatt's findings. The result was that there was a severe drop in interest with his technologies. The perceptron would become mostly a fringe topic for many years.

In 1971, Rosenblatt died in a tragic boating accident. He was only 43 years old and the death happened on his birthday. Years later, Minsky would regret some of his conclusions in his book and acknowledged the contributions of the perceptron.

By the early 1970s, the Golden Age of AI was quickly fading away. There would emerge a "winter," in which funding would mostly dry up. Keep in mind that there would ultimately be other winters in the 1980s and 1990s.

Several factors help to explain the first one. The US economy was suffering from several grueling recessions and inflation, driven by soaring oil prices. The federal government was seeking ways to save money – and this meant cutting back on research.

Another factor was that AI showed limited progress. The computers were not particularly powerful.

Even though LISP was an innovative language, it still had problems. The platform was not effective with high-end analysis of data.

Then there was the impact of the PC revolution. This technology gained tremendous support, and there were few resources left for researching AI. Companies were looking at ways to build better applications like spreadsheets and word processors, which could drive strong growth for companies like Microsoft.

> **Note** In 1973, British professor Sir James Lighthill published an extensive research report on AI. His conclusions were far from flattering. He concluded that the technology was overly complicated, which he termed "combinatorial explosion." The

> BBC broadcast a show about this (it's available on YouTube). It was a debate between Lighthill and AI luminaries like Donald Michie, Richard Gregory, and John McCarthy. Ultimately, Lighthill prevailed and the public opinion of AI greatly suffered. This would mark the beginning of the first AI winter.

The sentiment for AI would get so dire that researchers would often not even admit they were studying the topic. They would instead use other names for the category like machine learning, pattern recognition, and informatics!

Regardless, some researchers like Geoffrey Hinton would remain committed to AI. Throughout the 1980s to 1990s, there would be gradual innovation with machine learning, neural networks, and deep learning.

A key breakthrough came in 1986 with the concept of backpropagation. David Rumelhart, Ronald J. Williams, and Hinton wrote a paper about this titled "Learning Representations by Back-propagating Errors." It showed that backpropagation could realize major improvements in accuracy with AI prediction and visual recognition. However, the biggest challenge was that computer power was still limited.

It was not until about 2010 when AI would become a major force in the technology world. There were a myriad of factors for this:

- Data: With the ubiquity of broadband and smartphones, there was a surge in data. This allowed for much more sophisticated AI models. This would ultimately lead to large-language models (LLMs).
- Academic progress: There was a global race to create new concepts and theories of AI. Much of the research was publicly available. This meant that there was a flywheel effect with the progress. Some of the breakthroughs were transformer and diffusion models, which allowed for sophisticated generative AI.

- Infrastructure: Heavy investments in cloud computing systems from companies like Amazon, Microsoft, and Google were critical. They allowed for accessing enormously powerful compute resources to create AI models.
- GPUs (graphics processing units): These are computer chips that were originally meant for processing high-end graphics. But the technology would prove effective for AI models. Some of the reasons included: the ability to handle huge workloads and managing information in parallel.
- Funding: Venture capitalists and angel investors provided substantial financial resources for cutting-edge AI startups. They have also been a source of advice, recruiting and introductions to customers. Large tech companies have also been aggressive in making strategic investments in AI startups.
- Open source projects: There are many useful libraries and frameworks to develop AI models. Just some include TensorFlow, PyTorch, and Keras.

Note When it comes to AI, Geoffrey Hinton is a giant, if not a legend. As a teenager, he wanted to be a professor and study AI. During the 1970s when few people were interested in the technology, he was enthusiastic about it and remained convinced that it would be revolutionary – even if many of his peers thought he was wasting his time. But he liked being an outsider. During the 1980s and 1990s, he created some of the key pillars for deep learning. But in 2012, Hinton would become a star in the AI world. This was through his paper, titled "ImageNet Classification with Deep Convolutional Neural Networks," which he wrote with Alex Krizhevksy and Ilya Sutskever. Its use of deep learning led to standout results with image recognition.

What Is AI?

It’s difficult to define AI. After all, the category is complex and is constantly changing.

But let’s take a look at what ChatGPT says about this:

> *Artificial Intelligence (AI) is a branch of computer science that focuses on creating systems capable of performing tasks that usually require human intelligence. These tasks include learning from experience, understanding natural language, recognizing patterns, making decisions, understanding complex data, and even demonstrating creativity.*

This is actually a pretty good description. ChatGPT also notes that there are two main categories of AI:

- Weak or narrow AI: This is where the technology is trained on a particular area, say healthcare, customer service or sales. Even though this may be called “weak AI,” the systems can still be extremely powerful.
- Strong AI or artificial general intelligence (AGI): This is where an AI system has achieved human or superhuman abilities. It’s the vision for companies like OpenAI and DeepMind. Achieving strong AI will take more breakthroughs, say with quantum computing. There are also divergent opinions on when this may happen. Some say it could be within five to ten years. Others believe we will never achieve strong AI.

AI also has various subsets, such as deep learning, machine learning, and generative AI. A way to understand this is with the following illustration, as seen in Figure 7-1. The main circle represents all AI, which would include strong and weak AI. Some of the main subsets are machine learning and natural language processing (NLP). Then there are subsets within these, such as with deep learning within machine learning and so on.

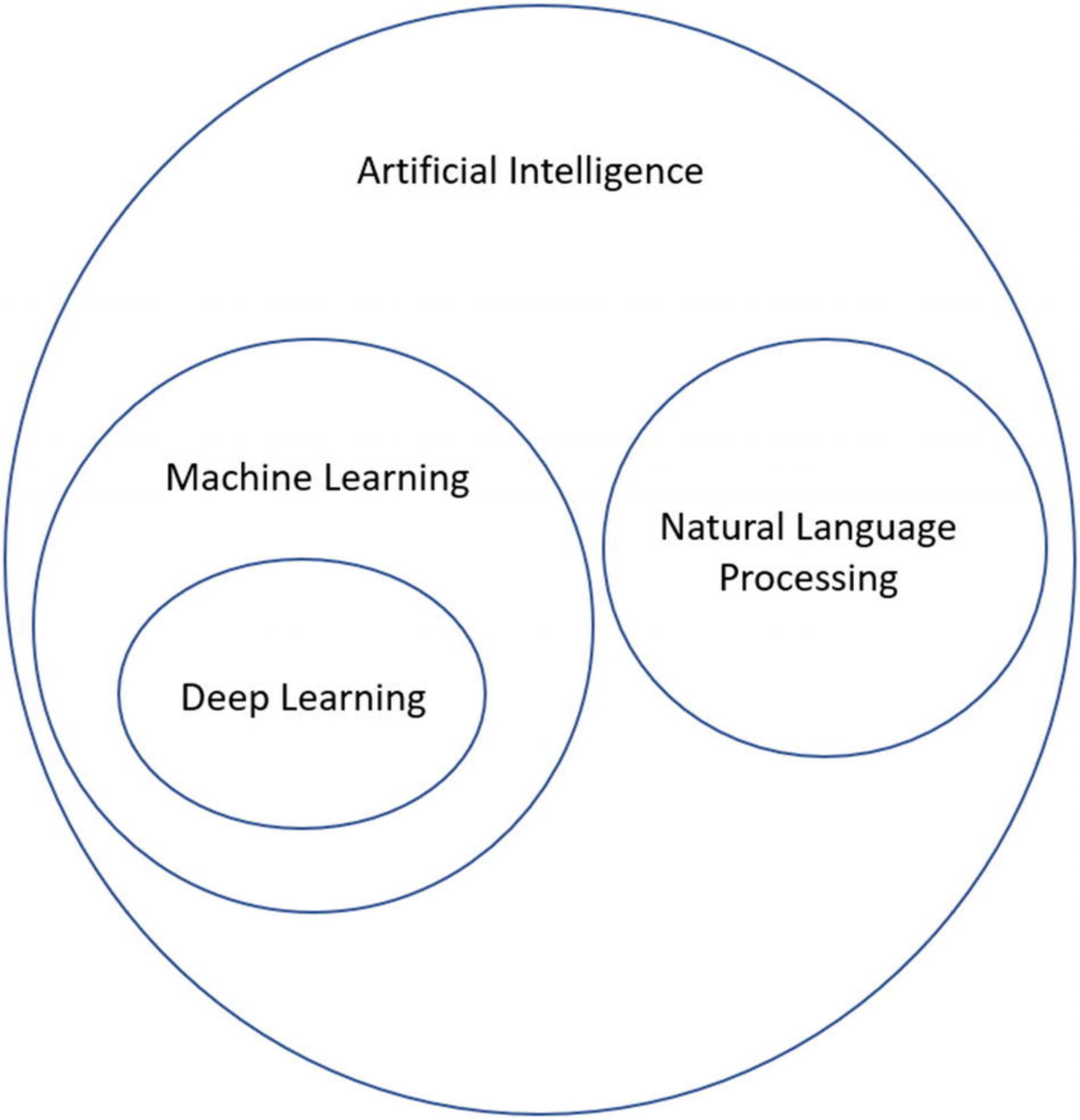


Figure 7-1 *This shows the various categories that AI represents*

Next, let’s take a closer look at these categories.

Machine Learning

Machine learning has a long and storied history. It was in 1959 that Arthur Samuel coined the word and started creating systems for this innovative technology while he was a top scientist at IBM. At the time, he said that machine learning was a “field of study that gives computers the ability to learn without being explicitly programmed.”[3]

Samuel's first demonstration of this was with his development of a system that played checkers. He created his machine learning model with a dataset from a book called *Lee's Guide to Checkers*. From there, he would iterate on this and the system got smarter and smarter. By 1961, his game would beat a checkers champion.

It was a huge validator of machine learning. It would drive much more research in machine learning.

Over the decades, this would lead to different categories of this technology. They include supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning.

Supervised Learning

At a high-level, supervised learning is a simple concept. It is where the dataset is labeled. For example, if you have a catalog of cat images, then each of these will be designated as a cat. Although, when it comes to labeling, it is often more detailed – such as including information about the type of cat, color, size, and so on.

The key advantage of labeled data is that it makes it easier to train the model. The reason is that that algorithm can determine the accuracy of its predictions. The algorithm can go through iterations to improve this. This is done by using a loss function, which determines the error rate of a machine learning system.

But supervised learning has downsides. They include the following:

- Data volume: You generally need a substantial amount of data to have an effective AI model. However, there may not be enough available because labeled data is generally sparse. A way to deal with this is by using synthetic data, which has seen much progress. But it is far from ideal either.
- Compute power: Typically, you need significant resources to run a model. This can make the process time consuming and expensive.
- Real-time uses: This is not possible with supervised earning. There is simply not enough time to label the data.

When it comes to supervised learning, there are two main categories. First, there are classification models, which involve an AI system that groups data according to common features. A typical example is a spam filter. It will group emails based on two labels: spam and not spam.

Next, there is a regression model. This illustrates the correlations among different variables.

Keep in mind that regression analysis is not new. Back in the 1880s, mathematicians Adrien-Marie Legendre and Johann Carl Friedrich Gauss developed this to allow for predicting the courses for comets. It turned out to be quite accurate.

There are various ways to measure effectiveness the predictions. The standard error metrics show the divergences in the variables. The lower the number, the better. Then there is the R-square, which helps to explain the results. The p-value will provide further understanding of this metric.

Regression analysis is often confused with logistic regression, but they are not the same. With logistic regression, it is for those use cases where the data does not have linear correlations. This is often used to handle binary data, such as for "yes" or "no" outcomes.

However, when it comes to supervised learning, the main type of approach is classification. There are a myriad ways for this. Here's a look at some of the main ones:

- Support Vector Machines (SVMs): This is where you divide a dataset into two categories of information. This is accomplished by using a hyperplane, which is geometric space that is the divider. The SVM algorithm optimizes the maxim margin between the classes of data. These can get complicated because the hyperplane can be in many dimensions. This can be useful when there are complex correlations and relationships with the data. If there are outliers, these will have little impact on the outcomes of the model.

- Random Forest: This is certainly an unusual name for an algorithm. But a random forest is actually an accurate description. This is essentially a set of decision trees, such as where there are “yes” or “no” junctions. Among supervised learning methods, the random forest is one of the easiest. There is no need for complex math or substantial compute resources. The visualizations also make this more understandable for non-technical people. Despite the simplicity, a random forest can still be a powerful tool for analysis. However, if there are many levels to the structure, there can be degradation of the outcomes.
- K-Nearest Neighbor (KNN): At the heart of this is the concept that “birds of a feather flock together.” In another words, the KNN algorithm allows for making predictions based on the grouping of datasets. This is one of the oldest machine learning algorithms, whose origins go back to the early 1950s. There are several steps with the KNN. First, there is the measurement of the distances of the data. This can be done using an algorithm, such as for Euclidean distance. Next, you will select the number of data points nearest to the ones that are being analyzed. This is often an odd number so as to break any ties. Finally, the prediction will be the majority of values nearest to the K point. The KNN algorithm is often called “lazy learning” because there are no complex algorithms. There are also no assumptions needed for working with the data. This means the KNN is non-parametric. But with large datasets, the KNN can lose its effectiveness.

Another way to classify data is to use the Naïve Bayes classifier, which is based on classic probability analysis. It’s where the odds of event #1 happening given event #2. These events are independent of each other, which is why it is called “naïve.” This helps to make the algorithm easier to work with datasets. The Naïve Bayes classifier has proven to be quite versatile and is common for many applications, such as with spam filters.

Unsupervised Learning

Unsupervised learning is where a machine learning model does not have labeled data. The algorithms will instead find patterns and uncover what the data is about. This usually requires huge amounts of data. Unsupervised learning may also uncover patterns that are too complex for humans to comprehend.

This approach to machine learning is at the cutting-edge. It’s where you generally need the expertise of a data scientist.

Yet these models can still underperform. It can be extremely challenging to measure the accuracy of the models since there are no labels to compare against. These models are often massive as well. For example, unsupervised learning is at the heart of LLMs.

Next, let’s take a look at several of the methods for unsupervised learning:

- Clustering: By far, this is the most pervasive algorithm. As the name implies, clustering involves grouping similar data items and the AI will then find the underlying patterns. To do this, there is a function that measures the closeness. There are various functions that can accomplish this. For example, centroid-based clustering uses a KNN algorithm and involves iterations of data items. Then there is the density-based clustering system. For the most part, this detects concentrations in the datasets. And then there is the hierarchical-based clustering method, which involves a random forest structure. This is where the dataset has a taxonomy.
- Anomaly detection: This involves the analysis of huge amounts of data and often this includes logs or web traffic. Anomaly detection will attempt to find quirky occurrences within the dataset. However, this can give off false signals. The pattern may ultimately be just noise. However, when done properly, anomaly detection can be quite powerful. Some of the use cases include fraud and malware detection. There are different types of anomalies the

algorithm can focus on. One is a contextual anomaly, which evaluates the environment, such as the time of day. For example, if a person is making a transaction at 5:23 AM, then this could be a red flag. There are also point anomalies, which is essentially an outlier, say when an invoice has an unusually large amount. But anomalies do not have to be negative. They can show encouraging trends, say when customers are showing interest in something new.

- Association: This applies computer logic to a dataset like an IF/THEN structure. It can also be a way to apply probabilities to the relationships. For those that have higher frequency, they have higher weight in the model – which is known as support. There is also a confidence measure that provides a sense of how predictive the variables are. Association is often used for recommendation engines as they can find unexpected patterns that can lead to higher sales.

The autoencoder is another useful technique for unsupervised learning. It's commonly used for generative AI applications.

With an autoencoder, it's possible to reduce the noise in the data and this makes it possible to find the features that have a real impact. The first step for an autoencoder is determining where the data is compressed. This is done by using a sophisticated neural network or deep learning model.

After this, there is the use of the decoder phase. Again, this uses a neural network or deep learning system to restructure the data.

Here's how a Nvidia blog describes this:

> *[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.*[4]

There is then the Variational Autoencoder (VAE), which can create interesting and unique images. The model takes samples of the images – based on probabilities – which allows for creativity.

Reinforcement Learning

Reinforcement learning is a form of unsupervised learning. The system is based on the concept of reward and punishment. That is, as it interacts with a dataset – or even a physical environment with robots – it will learn what to avoid and what to do more. It’s essentially how humans learn by trial and error.

The origins of reinforcement learning go back to the late 1930s. The pioneer of this concept was B. F. Skinner, who was a top psychologist. He helped to create behaviorism, which focuses on analyzing observable behaviors to get a sense of a person’s mental state. One of his experiments was the Skinner Box, which evaluated conditioning of animals. They would press a lever to get food, which was positive reinforcement, and avoid negative stimulus.

There was continued evolution of these theories. In the 1950s and 1960s, Richard Bellman and other academics provided a mathematical foundation for reinforcement learning. These would be key for AI models.

But it was after 2010 that this would gain momentum, accelerated with the developments in deep learning. DeepMind saw the huge potential of reinforcement learning and devoted substantial resources to this. The company focused initially on games, such as those from Atari. Reinforcement learning could quickly be trained to be super players. This was done primarily by analyzing pixel movements and the scores.

Games are good constructs to test AI. They have a consistent underlying structure with the game board as well as a set of rules. For the most part, a game is a self-contained environment that can allow for sophisticated simulations and learning.

However, the breakthrough came with Go, which is an ancient Chinese game. It consists of 19x19 lines on a board. While fairly small, it still allows for a seemingly endless number of moves. They are estimated to be more than the number of atoms in the observable universe.

At the time, the consensus was that it would be impossible to use a computer to beat a world Go champion. There would not be enough compute power and memory.

But DeepMind saw this as a tempting challenge. In 2014, the researchers set out to create the AlphaGo system, which relied mostly on reinforcement learning. The system was trained on the games from top players. Within a year, AlphaGo would defeat the European Go champion, Fan Hui.

But DeepMind knew it needed a more sophisticated system. To do this, there was the use of millions of simulated games. The result was an acceleration in the learning process.

To test out the new system, DeepMind setup a competition with Lee Sedo, a world Go champion. Millions of people watched the event and in the end, AlphaGo won four to one.

Sedo said: "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]

DeepMind saw that the system could be applicable beyond games. The company would go on to use it to diagnose disease and reduce energy in data centers.

Deep Learning

As noted earlier in the chapter, deep learning is a subset of machine learning. For the most part, it's about mimicking how the human brain operates: a network of complex learning systems.

Of course, deep learning can be challenging to understand. Even data scientists have difficulties with the subject as the underlying concepts of deep learning generally involve advanced math.

There is also lots of jargon. But it's important to have a high-level understanding of some of these terms. Let's take a look:

- Node or neuron: This is an algorithm that receives input from other nodes or external sources. Each node has a weight that is learned in the training process. An activation function will sum up the inputs, which helps the model learn.
- Parameters: These are the variables for a deep learning model. They represent what is learned from the training data. There are two main types of parameters. First, there are weights, which measure the strength and direction of the nodes in the neural network. Next, there are biases. These help provide flexibility in fitting the data. With the weights and biases, the model will attempt to minimize the differences between the predictions and the actual results.
- Scalar: This is a single number. It has magnitude but not direction. For a deep learning model, a scalar can be used for the learning rate, a loss function, and bias.
- Vectors: This is a one-dimensional array of numbers. Often vectors represent a large set of them. With these, it’s possible for a computer to understand the data for the processing of the model. For example, words or images can be converted into vectors.
- Word embeddings: These are vectors for words that have similar meanings, and this is a key part of generative AI models. It essentially allows for the predicting of the next best word.
- Matrix: This is a two-dimensional array of numbers. Matrices are common with deep learning models, as they can represent complex relationships in massive datasets.
- Tensor: This is a multi-dimensional array of numbers. It’s often used as a way to generalize scalars, vectors, and matrices. This makes it easier to work with large datasets, such as for images and videos.

Then how does a deep learning model work? Well, it is actually similar to regression analysis or logistic regression. The reason is that you will come up with random variables to train the model.

Then there will be many iterations to find the strengths and weights of the parameters for the input layer. This will continue until the errors are minimized.

However, compared to machine learning models, the processing is more extensive and involves the use of hidden layers, such as with activation functions. The output of each of the nodes will become input values for the next layers and so on. By doing this, the model will start to learn. The “deep” part of deep learning refers to the number of hidden layers.

Does this mean that having more hidden layers is better? This is not necessarily the case. Having a large number of them can result in bad results. Thus, a key role of a data scientist is knowing how to optimize the hidden layers and this often involves much trial and error.

Another key element of deep learning is backpropagation. With this, the output of the model will be sent back to the initial input layer and this can lead to much better results.

Flavors of Deep Learning

Deep learning is not monolithic. There are different categories of this technology and this is a testament to the significant innovation. But the variations of deep learning have also been to meet the unique needs of certain use cases.

One of the main types is the convolutional neural network (CNN). This is primarily for the processing of images. A CNN can learn about the different features of an image and the spatial aspects, so as to allow for accurate classification.

The origins of the CNN actually go back to neuroscience. This technology is generally a representation of the visual cortex of the brain, which is for the eyes to process the environment.

Research on CNNs began in earnest in the 1980s and the leader in this field was Yann LeCun. But of course, he had help from other academics like Yoshua Bengio and Patrick Haffner.

The breakthrough in CNNs came with the publication of “Gradient-Based Learning Applied to Document Recognition” in

1998. The authors included LeCun, Bengio, Haffner, and Léon Bottou.

What they demonstrated was that a CNN could successfully recognize handwritten digits, such as for zip codes, check numbers, and so on – which proved to be transformative for the financial services industry.

Then in 2012, CNNs showed an ability to recognize more sophisticated images. This was demonstrated with the ImageNet challenge, which we saw earlier in this chapter. Hinton's group built a model that had an error rate for the image recognition of only 16%. By comparison, the runner up was 26.2%.

If anything, the ImageNet results were a major catalyst that propelled deep learning. It showed that there were real-world applications for this technology.

Since the first ImageNet challenge, there was continued progress with CNN technology, such as with the development of new systems like VGGNet, GoogLeNet, and ResNet.

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.*[6]

Over the years, the errors would drop to only about 2%. To put this into perspective, it's roughly 5% for humans. Because of this, the ImageNet challenge was ended.

So how does a CNN work? It is known as "convolutional" because the algorithm is applied with many iterations. It will start with a small part of an image, say a few pixels, and then work its way through the rest. By doing this, the CNN will start to recognize the underlying patterns. It can then compare these with the rest of the images and begin to identify what they represent.

Before the introduction of the CNN, the process for image recognition was labor intensive. Data scientists would write equations to describe angles and shapes of objects.

Now another major type of deep learning system is the recurrent neural network (RNN). This uses nodes and hidden layers but the processing is done over a sequence of processes or tasks.

The RNN is based on decades of research that go back to the 1980s. Some of the early researchers in this category include John Hopfield, Rumelhart, Hinton, and Ronald J. Williams. In fact, it was Hopfield who made the initial breakthrough with RNNs, which happened in 1982 with the introduction of Hopfield networks. What this provided was a certain type of memory to allow for the system to better understand and learn from the data.

But it was not until 1990 that Jeffrey L. Elman came up with the modern RNN. It was known as a simple RNN, which included hidden layers and the use of backpropagation.

The memory of RNNs, though, was quite limited. This made it difficult for certain types of applications like NLP. In 1997, Sepp Hochreiter and Jürgen Schmidhuber published their research on a concept called Long Short-Term Memory (LSTM). This would allow for higher levels of memory, and it meant that NLP applications got much better.

For example, RNNs allowed for the creation of massive apps like Siri and Alexa. It powered the capability of text completion. If you enter something like “Pl” in a chat app, there will be a list that shows options like “Play” and “Please.”

Google also would use RNNs extensively. One use case was with its highly popular Translate app, which can interpret over 100 languages and handle billions of interactions per day.

The Issues with Deep Learning

The impact of deep learning is certainly huge. The technology has been one of the driving forces for the commercialization of AI.

But the technology is not without its limits. There are some serious drawbacks to consider.

Here’s a look at some of them:

- Data: You generally need a huge amount of data to build an effective model. This means you will need a sophisticated infrastructure, such as with storage systems and GPUs. There will also be a need of data scientists who can wrangle the datasets.
- Black box: Because of the complexities of deep learning, the models are extremely difficult to understand. This can mean less trust in the AI and can make it challenging to meet regulatory requirements.
- Overfitting and underfitting: These are common problems with deep learning models. Because of the complexities of the models, the results can skew the results. With overfitting, this is when the algorithms are not understanding or learning from the datasets. It’s just memorizing certain patterns and this can give a false sense of accuracy. In fact, if the accuracy rate is over 90%, then this is a telltale sign of overfitting. Underfitting, on the other hand, means that the model is not reflecting the real world. This often means there are not enough differentiated parameters.
- Conceptual thinking: Deep learning really does not understand concepts like justice, philosophy, politics, and so on. The systems usually do not have common sense either. Rather, deep learning systems are merely processing huge datasets to find patterns.
- Butterfly effect: Even a small change in a model can have a major impact. This sensitivity can make it difficult to calibrate the system.

Generative AI

Generative AI is a relatively new category. But of course, it is one of the most important in AI.

The origins of generative AI go back to 2014. It actually started at a pub in Montreal. Ian Goodfellow had an engaging conversation with his friends about how deep learning could create unique images.

At the time, he was already a respected data scientist. He received his Ph.D. in machine learning at the Université de Montréal and worked with luminaries like Yoshua Bengio and Andrew Ng. Goodfellow even wrote a book called *Deep Learning*, and as for his day job, he worked at Google.

OK, as for his discussion at the pub, there were various ideas proposed. But Goodfellow came up with an interesting concept. He thought that game theory could be a way to create images. This would involve deep learning models fighting each other.

Goodfellow couldn't stop thinking about this idea. When he returned home, he created an initial model and referred to it as a Generative Adversarial Network (GAN). Later on, he co-wrote a paper – titled "Generative Adversarial Nets" – and presented it at the Neural Information Processing Systems conference (NeurIPS) in Montreal. It was a bombshell. Yann LeCun boasted that the GAN was "the most interesting idea in the last 10 years in Machine Learning."

There would be an explosion of research. But savvy developers would create their own GANs to spin up immersive images – many of which spread across Twitter.

Yet this technology had a dark side. Some people would use GANs to create deepfakes, which could spread disinformation or even defame people.

Regardless, the proverbial "genie" could not be put back into the bottle. Going forward, there would be new ways to deal with the new world of generative AI.

What Makes Generative AI Different

Until the emergence of generative AI, most forms of AI were built on discriminative models. As we've seen in this chapter, this is about grouping and classifying datasets. By doing this, it’s possible to detect complex patterns, which can lead to insights.

Granted, generative AI is similar to discriminative modeling. They both use sophisticated deep learning models and process huge amounts of data.

But the main difference is that the output of a generative AI system is new data, which is based on advanced probabilities. This allows for variations and creativity. But the output still shares the same types of features of the original data.

Yet generative AI models are mostly black boxes. They also rely mostly on unstructured data, which makes it difficult to measure the accuracy and performance of the models. This helps explain why ChatGPT and Bard can sometimes generate output that is misleading, false, or even ridiculous.

There are a myriad of generative models. In the next few sections in this chapter, we’ll take a look at some of them, including GANs, Variational Autoencoders (VAE), diffusion models, and the transformer.

GANs

For the most part, GANs are for applications with images, videos, illustrations, and audio. Again, the technology involves a competition between deep learning models:

- Generator network: This takes in the training data. But the generator will use this to make fake images, which is called noise.
- Discriminator network: The focus of this model is to differentiate the fake images from the real ones. Usually, this is done by using loss functions. At first, the discriminator network has a low rate of success in detecting the fake images. But with more iterations, it gets better.

Typically, a GAN uses a CNN as the deep learning models. The reason is that these work better with image data.

What about the drawbacks with GANs? A key problem is the "vanishing gradient." This is where the discriminator network is too successful, which can stall the model. The result is that there may not be much variation in the output.

VAEs

To understand a Variational Autoencoder (VAE), you need to first see how an autoencoder works. This is a deep learning model for unsupervised learning. What this often accomplishes is denoising the dataset, which means stripping out random and inconsistent information. This process allows for a better representation of images, audio, and video.

This is the basic structure of an autoencoder:

- Encoder: This is where the initial dataset is fed into an algorithm and the encoder will generate a condensed version. This is known as the latent space. The encoder will attempt to optimize minimal data without sacrificing the quality of the model.
- Decoder: This is a deep learning system that takes the latent space and tries to reconstruct the original dataset.

As for a VAE, it has a similar structure as an autoencoder. But a key difference is that it uses the latent space as a distribution of probabilities. In other words, there are many more vectors, which allows for more interesting representations. The VAE then takes samples of the data, which is called reparameterization. There are also several loss functions that help smooth the data. Through this process, it is possible to create compelling images, audio, and video.

Diffusion Models

Diffusion models are essentially next-generation VAEs. This new approach came about in 2020 with the publication of "Denoising Diffusion Probabilistic Models." The authors included AI researchers Jonathan Ho, Ajay Jain, and Pieter Abbeel.

This is how the diffusion model works:

- Add noise: This is the diffusion process. The initial sample of the dataset will include random noise, and this is done as a distribution, such as with a Gaussian statistical calculation. This means that the results are according to the bell curve, which has 68% of the observations within one standard deviation, 95% within two standard deviations, and 99.7% within three standard deviations. This has been shown to generally represent many phenomena in the real world like IQ scores, height, or weight for populations. The adding of the noise is also done in multiple steps.
- Denoising: This is a complex function that reduces the noise in the data but this is done as an iterative process. The output will be a sample of the data distribution.
- Training: The model will attempt to mitigate the differences between the initial dataset and the samples.

With diffusion, the ultimate output can have high quality. Images, for example, can look unique and realistic. There is also more control over the process.

But there are some drawbacks. Diffusion assumes that relationships are linear and stationary and this can skew or distort the images. For example, this can mean that an image of a human can have more than two hands or legs!

Yet there is considerable research on diffusion models. The result is that the systems are getting better and better.

As for the types of applications for diffusion models, they are generally for creating images. The user will enter a prompt – usually a sentence or two – and the system will often create four images. This is done by using a CLIP (Constrastive Language-Image Pre-training) model, which tries to find the best match between the text and the image.

You can then prompt the system to make edits and changes to the image you are interested in. Some of the leading applications include OpenA's DALL-E 2, Adobe's FireFly, Stable Diffusion, and Midjourney.

The Transformer

Generative AI is a relatively new category. Some of the initial research started in 2014 with the focus on “attention.” This is about analyzing data sequences and finding the relevant parts. The research was mostly on text information, which generally has a logical structure.

At first, researchers used attention with RNNs but this presented challenges. Again, a big problem was with the memory.

There had to be a new way – something very different. And this emerged in 2017 with a paper from Google researchers. Titled “Attention Is All You Need,” it set forth the transformer model. They used eight GPUs for the model and trained it for three and a half days. The results were standout, as there were high levels of accuracy for English-to-French translations.

The transformer model was a huge breakthrough in the AI world. The irony is that Google did not aggressively pursue the opportunity. Instead, it was OpenAI that saw the huge potential for this technology.

As for how the transformer works, it involves something known as self-attention. This is where the algorithm finds the relationships among all the words.

For example, suppose we have the following two sentences:

Jack has a strong interest in artificial intelligence.

Jack earned 5% interest on his bank deposit.

The transformer will find the associations of the word “interest.” It will determine that this word means something different when it is related to a bank deposit versus the topic of artificial intelligence. In other words, the generative AI is gauging the context of the language. However, with a traditional RNN, this would be nearly impossible to do.

There are other aspects of a transformer model to note:

- Parallel processing: This is required in finding the relationships for the whole text. Otherwise, it would take too long for the analysis. But with a GPU, it has been shown to work extremely well with handling the parallel processing.
- Pretrained model: This means the model has already been created and can be used easily, such as by accessing an API. This is why companies can quickly add generative AI to their own systems. But the models can be trained and fine tuned, such as for other domains like medicine, security, financial services, and so on. There is no need to develop a model from scratch. This can be a huge time and money saver.

Figure 7-2 shows an illustration of how transformer works, which is from the 2017 paper. No doubt, it is complicated but let's take a look at the main parts.

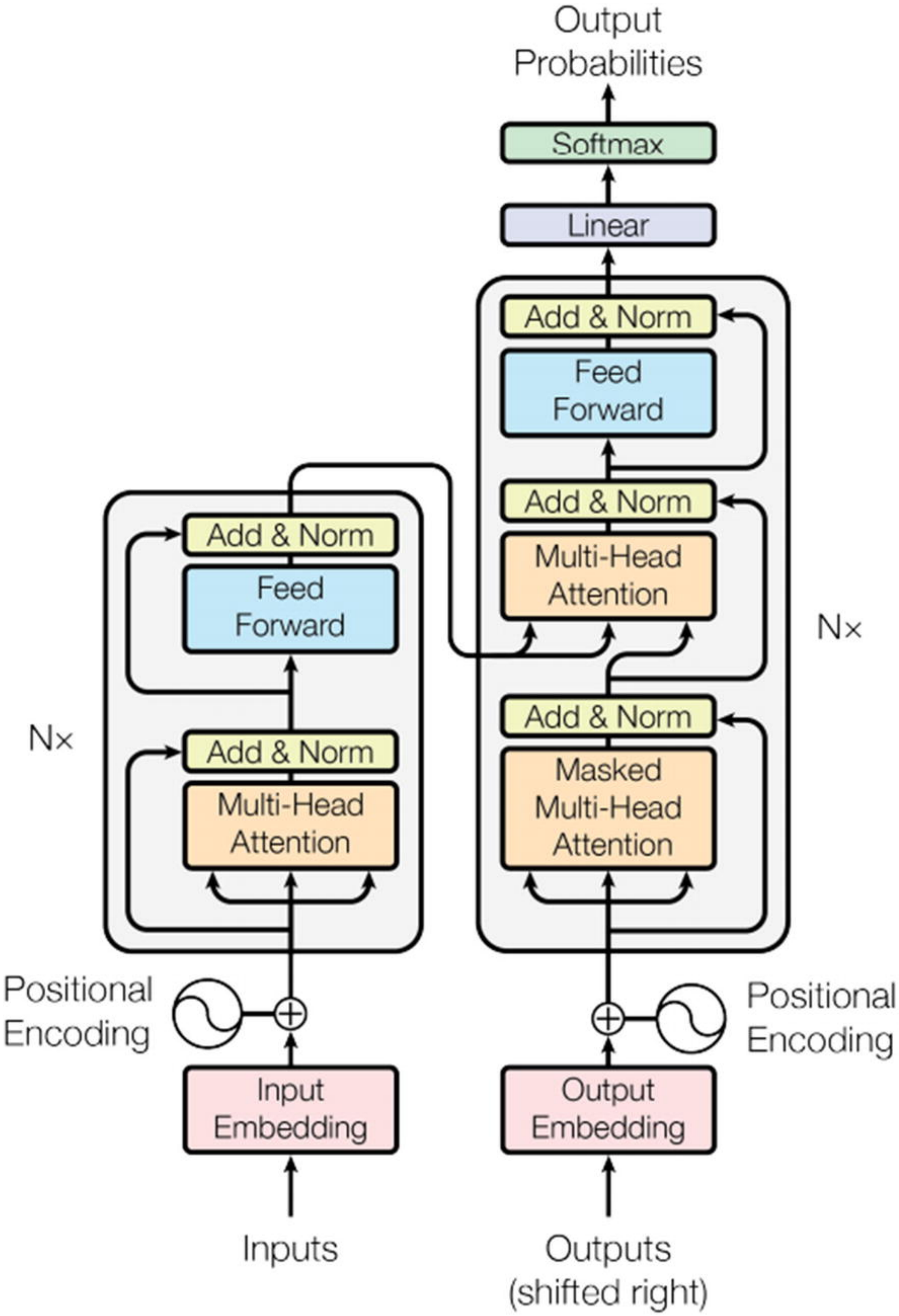

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

Figure 7-2 *This is the diagram of the transformer model from the "Attention Is All You Need" research paper*

As you can see, it has two parts. There is the encoder, which takes in the data and converts it into word embeddings. Then there is a process initiated called positional encoding, which determines the relationships among the tokens. This is accomplished with a stack of encoders that iterate across the dataset.

Next, there is a decoder, which will take the output from the encoder. There will also be many iterations of the processing.

But a model may not need to have both of these systems. There are encoder-only models, such as for BERT, as well as decoder-only models like GPT-3.

Conclusion

In this chapter, we looked at the core foundations of AI. We started out by looking at the rich history of this technology, which goes back to the 1950s with the monumental Dartmouth conference. We then covered some of the early successes of the "Golden Age" of AI. But unfortunately, this would turn into the "AI Winter" in the 1970s, when funding dried up.

Yet there would remain some researchers who pushed the boundaries of AI, such as Hinton. They introduced important concepts like backpropagation. By 2010, AI would see explosive growth as deep learning became practical because of cloud computing, GPUs, venture funding, and open source libraries.

Next, we covered the different types of AI. For example, machine learning relies on labeled training data to make predictions through approaches like supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning. Then there is deep learning. This utilizes neural networks with multiple hidden layers to uncover intricate patterns, especially in image, text, speech, and time series data. Finally, we looked at generative AI. Some of the key systems include GANs, VAEs, diffusion models, and transformers.

Of course, AI still faces challenges related to data needs, interpretability, overfitting, and performance constraints. But rapid progress continues thanks to global AI research collaborations and game-changing innovations.

In the next chapter, we'll look at data for AI applications.

Footnotes

1

www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/

2

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

3

https://squarkai.com/machine-learning-faq/#:~:text=%E2%80%9CMachine%20Learning%20is%20a%20field,to%20computer%20pioneer%20Arthur%20L.

4

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

5

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

6

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

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_8

8. Data

A Critical Element for AI

Tom Taulli[1]
(1)
Monrovia, CA, USA

Neal Patterson, Paul Gorup, and Cliff Illig worked at Arthur Andersen in the 1970s. But they saw an opportunity to create a company to leverage computer technology for the healthcare industry.

The result was the founding of PGI & Associates, which would be renamed to Cerner in 1984. This is when the company launched its platform and the growth was brisk. Within a few years, Cerner pulled off a successful IPO.

Over the years, the system evolved and included more mission-critical features. The focus was primarily on helping healthcare organizations to better manage their IT systems, especially with the sharing of patient data. Along the way, the company acquired various companies like Siemens Health Services.

In late 2021, Cerner announced that it agreed to Oracle's $28.3 billion acquisition offer, which was one of the biggest deals of the year and the largest ever for Oracle. It also highlighted the value of data.

With Cerner, Oracle would be able to bolster its lagging cloud business. Access to huge amounts of data would also allow for many opportunities with creating sophisticated AI applications. Keep in mind that healthcare expenditures account for roughly 20% of the GDP of the United States.

Larry Ellison, the cofounder and chief technology officer of Oracle, had this to say about the deal:

> *With this acquisition, Oracle's corporate mission expands to assume the responsibility to provide our overworked medical professionals with a new generation of easier-to-use digital tools that enable access to information via a hands-free voice interface to secure cloud applications.*[1]

Besides the deal for Cerner, there was another large acquisition for a data-rich company – that is, Microsoft shelled out $19.7 billion for Nuance Communications.[2] The company, which was founded in 1999, built a platform for speech applications based on advanced analytics and AI. The focus has been primarily on the healthcare industry. In 2022, more than 55% of physicians and 75% of radiologists in the United States used Nuance's applications.

Such mega deals will likely continue. As data becomes more essential for AI, there will be a scramble for these valuable assets.

In this chapter, we'll take a look at the importance of data. Although, this is a category that is something often overlooked. But this is a big mistake. If you want an effective AI strategy, you will also need a solid plan for data.

> **Note** At the time of Oracle's acquisition, Cerner was the No. 2 player in the electronic medical records industry. The leader? It was Epic Systems Corp, whose software manages the data for 78% of the patients in the United States and over 3% for the rest of the world. Judith Faulkner founded the company in 1979 and she got only a $70,000 investment. From there, she grew the company into a powerhouse. Faulkner has never taken the company public – and she says she will never do so. In 2023, her net worth was estimated at over $7 billion.

Types of Data

No doubt, large-language models (LLMs) require huge amounts of data. Some of these systems actually scrape much of the Internet, which can come to petabyte-levels (a petabyte is 1,000 terabytes).

In light of this, data creation is one of the fastest growing categories. Consider IDC's Global DataSphere report. In it, the research firm projects that data growth will be about 21.2% per year by 2026, exceeding 221,000 exabytes (an exabyte is 1,000 petabytes).[3]

There are a myriad forms of data. The most common – which is composed of 90% of data – is unstructured data. As the name implies, this is information that is not standardized like images, emails, videos, text files, satellite images, and social media messages.

For an AI project, unstructured data is usually critical. It is at the heart of systems like ChatGPT and Bard.

Then there is structured data. This includes formatted repositories like databases and spreadsheets. The information is often stored in rows and columns, such as to track phone numbers, addresses, credit cards, and inventory items.

Structured data is generally easier to work with. For example, you can develop a program to parse the data and then use it for analysis. But as for unstructured data, there needs to be sophisticated systems like deep learning to recognize the complex patterns.

Another type of data is semi-structured data, which is composed of structured and unstructured sources. Some examples include JSON (JavaScript Object Notation), which allows for creating APIs (application programming interfaces), as well as XML (Extensible Markup Language). These provide underlying guides or maps of the unstructured data.

And are there other types of data? Yes, there are. Here's a list:

- Metadata: This describes data. This can help provide a better understanding of a certain item. For example, a document may have metadata for the file name, size, and time for the latest update.

- Time-series data: This is a stream or sequence of data for a certain duration. This can be helpful to gauge trends. Time-series data is also a common form for AI applications.
- Real-time data: This is data that is collected and analyzed when it is generated, say for traffic or weather information. Real-time data for AI is for cutting-edge applications. They are found in apps like Uber, where time is essential. But developing and maintaining these types of systems can be expensive and complicated, requiring a team of experienced data scientists.
- Synthetic data: This is artificial data. In fact, generative AI creates this type of information, say for blogs, tweets, and emails. Synthetic data can also be used as training sets for AI models. This is common for self-driving systems since it can be difficult to get information about rare edge cases.
- Dark data: This is data an organization has but does not realize that it can be useful. Some examples are sensor data, log files, and social media mentions. But this data can be useful for developing advanced AI models.
- Genomic data: This is information about DNA. This has significant potential for creating novel therapeutics.

Understanding Big Data

As British mathematician Clive Humby noted in 2006, data is the new oil. It was a forward-thinking statement that has turned out to be spot-on. Today’s most valuable companies like Microsoft, Google, Meta, and Apple have mastered how to leverage data to create compelling systems.

A term that aptly describes this is Big Data. Gartner analyst Doug Laney coined this term in 2001 and it quickly caught on. But his conception of Big Data was more than just about large amounts of data. Laney crafted an interesting framework – called the three V’s – which includes the following:

- Velocity: This highlights the acceleration of the creation of data. There are so many sources of this, whether mobile apps, cloud computing, streaming, social media, and IoT (Internet-of-Things). This has also been hugely beneficial for AI, which relies heavily on training data for creating better models.
- Volume: When Laney published his framework, the amount of data creation was relatively small. But nowadays, the sheer volumes are off the charts. In fact, there is no agreed-upon standard for how much data qualifies for Big Data. But when it comes to cloud computing, these systems essentially provide access to data storage that is seemingly infinite – at reasonable costs for customers.
- Variety: This shows the different types of data, which we described earlier.

Laney's model has held up fairly well. But given the dynamism of technology, there are gaps. To fill these, analysts have added other V's to Big Data:

- Variability: The content of data evolves. This is due to a variety of factors, such as a change in society or scientific knowledge. This is important for AI since models need to adapt. At the same time, they need to be adjusted for inherent problems with datasets like bias.
- Value: This shows the inherent usefulness or effectiveness of data. It's often just a part of a dataset that has the most value. The rest of the information is noise. This is where having data science expertise can be helpful.
- Veracity: This shows the accuracy of data, which is a major challenge with AI. The datasets often have false or misleading statements. Thus, there needs to be ways to detect the problems.
- Visualizations: This is turning data into graphs and charts, which helps users better understand the trends and insights.

Database Systems

All AI projects need a database. This is technology that stores and manages data. Databases are one of the earliest systems when computers emerged in the 1950s. Although, they were fairly primitive and complex. Yet they were able to accomplish major feats like the Apollo space program.

But the major breakthrough in databases came during the early 1970s, with the emergence of relational database management systems. The pioneer of this technology was Edger Codd, an IBM researcher. He wrote a paper titled "A Relational Model of Data for Large Shared Data Banks."

He showed the basic structure of a relational database. It would be composed of different tables, such as for categories like payroll, employees and customers. A table would have rows and columns to store the data.

With this structure, it was much easier to search the database and manipulate the information. This was made possible with a scripting language called SQL (Structured Query Language). The commands were in English and easy for non-technical people to understand. An example of an instruction is: SELECT * FROM EmployeeTable (the use of * is for "all").

Another key benefit of the relational databases is the ability to connect tables. This allows for creating sophisticated relationships, which are done using primary and foreign keys. A primary key is an identifier – usually a number – that represents a piece of data. The foreign key allows for connecting the primary key to another table.

With these keys, you can create three types of relationships:

- One-to-one: This is where a row in a table is joined to a row in another table. The value must be unique, say with a customer number.
- One-to-many: This is a value in a row that is connected to two or more rows in another table. This could be a customer table that is linked to multiple purchase orders.

- Many-to-many: This is where more than one row is linked to more than one row in another table. These relationships can get complicated. But then again, they can create powerful structures.

The irony is that Codd's framework drew tepid interest from senior managers at IBM. It's not clear why this was the case. But it could have been due to the fact that IBM already had a highly successful database, called IMS (Information Management System). In other words, there could have been fears of cannibalization of existing revenue streams.

But this was not a concern for Larry Ellison, who was an engineer in Silicon Valley. When he read Codd's research, he saw that this was the future of database technology and he wasted little time in cofounding Oracle.

IBM would eventually create its own relational database, which was Db2. But it was too late as Oracle would eventually dominate the market. Consider that the company's market value is currently much more than IBM's.

Relational databases have proven to be quite durable. They are still widely used for mission-critical operations.

But the technology is not without its drawbacks, especially for modern applications like AI. Here is a look at some of the main ones:

- Costs: These can be hefty. Even to run a relatively small AI project, the licensing or subscription fees can weigh on the ROI. But of course, many modern AI models require huge amounts of data.
- Capabilities: Relational databases are ideal for structured data. But with AI, this is not enough. A database needs to handle semi-structured and unstructured data.
- Data sprawl: This is a major problem with relational databases. If not managed properly, the data can easily be scattered across an organization. This can make it difficult to consolidate the datasets for AI projects.

Because of these problems, there have been many innovations in the database market. A major driver for this has been the need to manage the enormous requirements for cloud systems, large mobile apps and social networks.

A key source of the innovation came from open source projects. These allowed for creating thriving ecosystems and evolving sophisticated platforms.

Some of the initial projects – like Hadoop – focused on handling data at scale and in real-time. These would become known as data warehouses. Companies like Facebook, Yahoo! and Amazon would adopt these types of open source systems. This would significantly accelerate adoption and encourage more innovation.

As AI started to gain momentum in 2012 – with advances in deep learning for image recognition – there was the emergence of new data systems to facilitate projects. One of the most important platforms was Apache Spark, which was the creation of a team of students and researchers at Berkley. They would then cofound Databricks to commercialize the system. The timing was perfect and the growth was robust. And this has continued until today. For the fiscal year end January 31, 2023, the company reported revenues over $1 billion and the top-line soared by more than 60%.[4]

At the heart of Databricks is a modern data warehouse, which is called a lakehouse. This is a sophisticated cloud-based platform that seamlessly integrates any data silo across an organization. The lakehouse can also be built on a cloud system like AWS, Microsoft Azure or Google cloud – which is a must-have for enterprise customers. The result is that AI projects can get the maximum impact with data, allowing for much more effective models.

Besides data warehouses, there are other new forms of databases that have emerged. Consider NoSQL systems. These are databases that store data in a free-form document model, which makes it easier to integrate data sources into complex AI models. A key is that NoSQL systems are built to handle unstructured and semi-structured data.

The dominant company in the space is MongoDB. In 2007, tech veterans Dwight Merriman, Eliot Horowitz and Kevin Ryan cofounded the venture. The inspiration for the company came when they were at DoubleClick, a leading ad network. They had challenges creating custom data stores that could handle the enormous scale, such as with processing up to 400,000 ads per second.

The limiting factor was the traditional database. To deal with this, the founders of MongoDB created an open source application and then a company to manage it for customers.

Currently, MongoDB has over 40,000 customers across more than 100 countries. From 2018 to 2023, revenues have gone from $166 million to $1.28 billion.[5] Since the launch of the company, users have downloaded over 325 million copies of the database.

Another major player in the database market is Snowflake. The company's cloud-based system has become critical for a wide variety of AI projects.

The company got its start in 2012. The cofounders were Thierry Cruanes, Benoit Dageville, and Marcin Zukowski, who had stellar careers at companies like IBM, Google and Oracle. They had first-hand experiences with the challenges of relational databases and wanted to build something better – that was based on modern approaches.

While Snowflake is incredibly powerful, it is also easy to use. It can take a minute to fill out a form to spin up a sophisticated database.

Snowflake has built a thriving ecosystem, which has been bolstered with an Appstore. There are more than 1,800 apps available on the platform.

As for the business model, this is attractive to customers as well. There are no subscriptions or licenses. Rather, the fees are based on consumption.

In the fiscal first quarter of 2023, the company reported revenues of $623.6 million, up 48% on a year-over-year basis. It had 373 customers with trailing 12-month product revenues over $1 million. According to Snowflake CEO, Frank Slootman: "Data has gravitational pull, and given the vast universe of data Snowflake manages, it is no surprise that interest in data science, AI and machine learning is escalating while its uses are rapidly evolving."[6]

As the data workloads have continued to grow, there have been more innovations in scaling database systems. Take a look data lakes. These are built to handle enormous amounts of data – whether structured or unstructured. Data lakes have become a common aspect for large AI projects. These may be connected to platforms like Apache Spark or Google's TensorFlow, which is for developing deep learning models.

Then there is another type of database that has even more scale. It's the data store. The pioneer of this technology is Uber. In the early days of the company's hypergrowth, it needed a new type of database and the company invested heavily in building its own, which was Michealangelo. It was created mostly for AI workloads, such as for the training, deployment and monitoring of advanced models. Michealangelo proved to be highly successful and was critical for Uber's success.

Another type of modern database is the graph database. It is for storing and handling the complex relationships for applications like social networks. This technology has morphed into graph neural networks, which are for AI applications. In a database, there are nodes that are connected to lines or edges, which can be helpful in discerning intricate relationships. For example, graph neural networks have been effective with recommendation engines, fraud detection and drug discovery.

Finally, there are vector databases. These store and manage embeddings, which are long lists of numbers. These can be features of an AI model and vectors can represent many types of data, whether text, images, audio and video. Vector databases have become a common way to build generative AI systems.

A leading player in the market is Pinecone, which Edo Liberty founded in 2019. Before this, he was a director of research at AWS and was the lead of Amazon AI Labs.[7]

It was through these experiences that he saw there was an opportunity to create a database tailored for working with vectors. But it was the emergence of LLMs that Liberty saw that the timing was right for a product like Pinecone.[8]

In April 2023, the company announced its Series B funding for $100 million. The lead investor for the round was Andreessen Horowitz and the valuation was set at $750 million.

According to Liberty:

> *Pinecone's mission is to provide the Long Term Memory for AI, and we are well on our way to achieving that. Our self-onboarding product, efficient growth engine, and world-class team are here to power the next wave of AI-powered innovation. We are excited to support the fundamental shifts in the way companies approach data management in the era of AI.*[9]

The Cloud

When it comes to AI projects, the cloud is often the preferred way for the creation and deployment of the models. A key reason for this is that the data is centralized, which allows for much more robust models. But the cloud also provides huge amounts of scale. With a cloud platform, you can easily create virtual machines to host the applications and this can also be supported with sophisticated systems like GPUs and TPUs. Finally, the costs are generally reasonable. After all, they are based mostly on the usage of the cloud service.

However, the cloud has various flavors. Here's a look at the main ones:

- Private cloud: This is a cloud system that has only one user or tenant. A company will pay for the infrastructure, such as the data center, and the facility can be rented or owned by the company. However, it will have complete control of its capabilities. A private cloud is a default option for many regulated business, such as in healthcare and finance.

- Public cloud: This is a multi-tenant option. That is, many companies can use the same public cloud. The biggest players in this market include AWS (Amazon Web Services), Microsoft Azure and Google Cloud. Because of the sharing of the same infrastructure, the public cloud is subject to more security risks. Then again, there are often many guardrails in place like firewalls, identity management and encryption.
- Hybrid cloud: Yes, this is a blend of the private and public cloud. Often, this is where a company will house its most important processes in a private cloud and the rest will be offloaded to the public cloud. This helps to lower the overall costs and maintenance requirements. In fact, the hybrid cloud is becoming a popular approach for many large enterprises.

Cloud providers like AWS, Azure and Google cloud have dominant leads. Yet it's actually common for customers to use two or more of these platforms – which is known as the multi-cloud strategy. This can allow for more redundancies and access to better functionality.

But this does not mean that customers should not look at other cloud providers. Companies like Oracle and IBM are making great progress with their own offerings.

There are various startups that focus on providing hosting for sophisticated AI workloads. One is CoreWeave, which was founded in 2017. The original focus was on helping provide cloud infrastructure to help with crypto applications like mining. This was about making it easier for customers to access GPUs.

But the founders of CoreWeave saw that its platform would be ideal for hosting AI systems. To this end, the company has made it possible to easily spin-up and manage sophisticated GPUs, such as Nvidia's H100s, A100s, A40s and RTX A6000s. Some of the clients include fast-growing AI companies like EleutherAI and Stability AI.

A differentiator for CoreWeave is its business model. The company only charges its customers for the usage of resources. There is no fee for when existing capacity is idle. The upshot is that CoreWeave can provide a cost-competitive service.

In April 2023, CoreWeave announced a round of funding for $200 million at a $2 billion valuation.[10] The lead investor was Magnetar Capital.[11]

> **Note** To power OpenAI's ChatGPT, Microsoft spent "hundreds of millions of dollars" to develop a state-of-the-art supercomputer. This involved connecting thousands of Nvidia GPUs on the Azure platform. According to Eric Boyd, who is the corporate vice president of Azure AI:
>
> > *We saw that we would need to build special purpose clusters focusing on enabling large training workloads and OpenAI was one of the early proof points for that. We worked closely with them to learn what are the key things they were looking for as they built out their training environments and what were the key things they needed.*[12]

Data Strategy for AI

To bolster your AI efforts, it's critical to have a data strategy. This does not necessarily have to be highly detailed. But it should cover some main areas. There also should be the use of best practices.

In the next few sections of this chapter, we'll take a look at how to do this.

Data Collection

On its face, the concept of data collection seems straightforward. But as is the case with many parts of IT, this is far from the reality. Data collection can be a tough process – and it can be easy to make mistakes.

First of all, you need to inventory your organization's existing datasets. What are the contractual rights to these? How are these datasets being used? What needs to be improved with them? Can they be combined with other datasets? Or will there need to be some new ones added from outside sources?

Interestingly enough, there may be datasets that could be overlooked. But they can still be quite essential for AI success. Just some of these are

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

There are a myriad of third-party sources for data but the fees can be high. They may also not necessarily be effective. It's important to do some due diligence before making a purchase.

Then there are free datasets. These are often from the government, say the Department of Commerce, the Labor Department and the Federal Reserve. These can be rich sources and useful for creating sophisticated models.

The following are some popular free data sources:

- Data.gov: This is an amazing resource. It has about 250,000 datasets from the federal, state, city and local governments in the United States. The topics are wide-ranging, from education to healthcare to agriculture to climate.
- U.S. Healthcare Data: This includes data on environmental health, social services, substance abuse, medical devices, diseases and drugs. The sources for the datasets are from organizations like the US Department of Health and Human Services, Centers for Medicare and Medicaid Services, Centers for Disease Control and Prevention, and Food and Drug Administration
- IMF Data: This is a portal that provides access to a large amount of data about economies across the world. Some of

the topics include exchange rates, interest rates, prices, trade of goods and GDP.

- MNIST: This is an extensive dataset for handwritten digits. It's been around since 1999 and is managed by top AI researchers like Yann LeCun. The dataset is useful for applications like pattern recognition.
- IMDB Movie Reviews: This dataset has 50,000 reviews, which are roughly evenly divided by a training and testing set. They are polar, which means there are only positive and negative reviews. This dataset is useful for AI applications for sentiment analysis.
- Google's Open Images: This is a massive dataset of more than 15.8 million images. This is popular for creating computer vision AI systems.
- Project Gutenberg: Author Michael S. Hart launched this in 1971. It's the world's first digital library of books. Currently, it has over 70,000 freely downloaded titles.
- BookCorpus: This is another repository of books but this is focused on fiction. There are more than 11,000 titles across 16 genres, such as romance and historicals.
- OpenWebText: This includes millions of shared URLs on Reddit. Each has at least three upvotes from users.
- CommonCrawl News: This is a dataset of news stories from websites across the globe.

Another option is to use webscrapting. This involves using tools to take data from the Internet. While this can provide useful data, there are issues with copyright infringement. There may also be risks of bias, which can negatively impact a model.

Data Evaluation

When you have one or more datasets, you should do some evaluation of the content. You do not want to quickly start an AI project to only realize later that the dataset is not the right fit.

For the evaluation process, here are some factors to keep in mind:

- Relevance: True, this can be difficult to determine in the early stages of an AI project. This is why it can be good to do a small experiment to test out the dataset. A data expert or data scientist will definitely be helpful for this process.
- Timeliness: Data can have a finite usefulness. The information can easily become stale and may not reflect current reality. The result is that an AI model will provide suboptimal outcomes and this is why you should get a dataset that is updated.
- Legal: Do you have the right to use the data for the purpose of your AI model? This is something to determine early in the process. To this end, you should get the legal department involved.
- Target audience: You want the dataset to appropriately reflect your customers or users. Thus, if you are creating an AI system primarily for women, then make sure the dataset accounts for this.
- IT infrastructure: You may not have the right tools and systems for the AI project. This is quite common as many organizations have legacy systems. Early in the AI process, try to estimate the IT requirements.

Data Wrangling

No dataset is perfect. There are always problems like gaps, outliers, and bias. This is why there needs to be data wrangling or cleaning.

This is a critical part of the AI process but often it does not get enough attention. After all, the process is tedious and time consuming. It can also be daunting when the datasets are massive.

There are many techniques for data wrangling. Some are fairly straightforward while other approaches need the assistance of a data scientist.

Here’s a look at some of the common techniques:

- Outliers: This is where some of the data is outside the general ranges. This can be an indication that the data is not accurate or it is noise. Yet outliers should not be dismissed out of hand. The reason is that sometimes this data can be extremely valuable. This is the case where the data may indicate a signal, say for of a hack or fraud. It's typical for cybersecurity systems to look for outliers.

- Nonsensical data: This can be where there is a negative number for an age or a piece of data where a person is, say, 15 feet tall! Obviously, these are mistakes and they are easy to correct. But there still needs to be a scan of the data.
- Duplications: This is common with datasets. There may be mistakes in data entry but they need to be corrected – or else the model is likely to suffer.
- Standardization: This involves setting up conventions for how certain data is expressed. For example, should states in the United States be abbreviated or spelled out? There is no definitive answer. But what's important is that you are consistent. Another approach is to have input forms that have drop-down menus for this type of data. This can greatly reduce problems with standardization.
- Conversion table: Some of the data may have alternative measurements. An example is the difference with the metric system. To deal with this, an algorithm to make the conversion to one measurement may be needed.
- Missing data: This is common for most datasets. If only a small part is missing, then this should not impact the effectiveness. But there are some approaches to help when the gaps are more prevalent. You can use an average of the numbers of a column to fill in for missing data, which can smooth the dataset. Generative AI algorithms can also help with missing data. That is, it can simulate it, which is based on the underlying patterns.
- Binning: This is for how to categorize data. Often, this does not have to be highly specific. Having age groups of 30–32, 33–35, and so on is probably not useful. A better approach would likely be to have a longer period, say 10–20 years.

- One-hot encoding: A good way to understand this concept is to use an example. Suppose that you have a dataset of iPhones, which include iPhone 14 Pro Max, iPhone 14 Pro, and iPhone 14 Plus. You can classify these as numbers. This could be setting iPhone 14 Pro Max to 1, iPhone 14 Pro to 2, and iPhone 14 Plus to 3. By doing this, the company can better process the information. However, there is a problem: an AI model may think that 1 is more important than 2 and 3. There may be an assumption that there is a ranking. But one-hot encoding can resolve the problem because you will have a column for each of the devices. You could call them iPhone_14_Pro_Max, iPhone_14_Pro, and iPhone_14_Plus. If the iPhone 14 Pro is being evaluated, then there will be 1 in the column and 0 for the others.

Note With a dataset, you might be able to create data based on the other data. How so? An example is with birth dates. If you have a column for a person’s age, then you can create a new column that subtracts this from the current date.

There are many more techniques, and some of them can be highly complicated. But there are various tools to help automate the process. They can also do this at scale – saving considerable time and expense.

A leader in data management systems is Alteryx. The company’s Trifacta system can handle all of the approaches that we have discussed, like standardization. But here are some others:

- Predictive transformation: This is a visual designer tool that helps to resolve data transformation problems. The user does not have to be a data expert to use it.
- Adaptive data quality: This will scan the dataset and provide actionable suggestions to improve the quality of the data. It will evaluate for column dependencies, integrity constraints, and formatting.

- Active data profiling: This is another sophisticated technique. But this one will look at the validity, accuracy, and completeness of a dataset.

Data Labeling

Data labeling is the process of providing descriptions to items in a dataset. This makes it possible to compare the results of an AI algorithm to determine the level of the accuracy. Data labeling is common for applications like image recognition.

A famous example of this is the research from AI expert, Fei-Fei Li. In 2006, she wanted to label the ImageNet dataset. But how could she do this cost-effectively? Keep in mind that the dataset had about 14 million images. She tried different techniques but wound up using a crowdsourcing strategy. Li used Amazon's Mechanical Turk, which involved paying thousands of people to do the labeling. This approach worked – and ImageNet became a standard for building AI models.

But a data label can be complex. It can be a detailed description of the data item, which can be particularly important for areas like self-driving cars. It's imperative to make sure that there is a complete understanding of a visual.

Yet there still needs to be some judgment. What details should be a part of the label? If there is too much description, it can degrade the quality of a model.

Given all this, it should be no surprise that data labeling is a challenging process and typically requires the help of data scientists.

But this does not mean you have to hire your own team. There are various companies that focus on providing services for data labeling. The costs are generally affordable. There is also the benefit of leveraging the deep expertise of the organization.

A top player in the space is Scale AI, which Alexandr Wang and Lucy Guo cofounded in 2016. At the time, they were students at the Massachusetts Institute for Technology and saw first-hand the difficulties with working with datasets.

With Scale AI, the founders built a platform that could efficiently label datasets for text, audio, video, maps, and sensors. Since the launch of the company, it has managed about 7.7 billion annotations and over 1 billion 2D and 3D scenes.[13] Some of the customers include Microsoft, Meta, Toyota, and Fox.

Consider a use case of this technology with SAP, which is one of the world's largest business software developers. The company had a massive trove of customer documents, but it was challenging to wrangle the data. There were also the issues of data ownership and privacy.

With Scale AI, SAP was able to create high-quality labeled data for three languages, various document types, and over 200 unique fields.[14] This also involved properly handling personally identifiable information like names, phone numbers, and emails.

The result was that the data extraction had accuracy of more than 95% – with processing speeds under 60 seconds.

How Much Data?

It's true that having enormous amounts of data helps create much more robust AI models. The effectiveness of LLMs demonstrates this.

But there needs to be caution. Massive models are expensive and time consuming to work with. It can also be difficult to find enough data. Granted, this can be alleviated with synthetic data. But this is far from ideal and can give off lower quality results.

Or a model may get bloated with data and parameters. This can lead to the "curse of dimensionality," which often means that the models underperform.

In other words, data quantity is something that is nuanced. Sometimes even very small amounts of data can be fine. This is the opinion of Andrew Ng, who is the CEO of Landing AI and the former head of Google Brain. He says some models can work with only 100 data points.

Conclusion

The phrase “garbage in, garbage out” is a cliché. But it is a valid warning for AI models. If the underlying dataset is faulty, then the results of the model will likely be off the mark.

In this chapter, we looked at the fundamentals of data management. This included understanding the types of data – such as with structured and unstructured sources – as well as the systems like relational databases, NoSQL databases, data lakes, and vector databases.

We also looked at putting together a data strategy. We covered some best practices, say with data collection, evaluation, wrangling, and labeling.

Yes, the process can be tedious and time consuming. But there are companies that can help out. Often, these are good options for creating useful AI models.

In the next chapter, we’ll take a look at the risks of generative AI.

Footnotes

[1] www.wsj.com/articles/oracle-to-pay-more-than-30-billion-for-cerner-11640006531

[2] https://news.microsoft.com/2021/04/12/microsoft-accelerates-industry-cloud-strategy-for-healthcare-with-the-acquisition-of-nuance/

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

[4] https://techcrunch.com/2023/06/21/databricks-is-proof-that-strong-unicorns-can-grow-their-way-out-of-a-market-correction/

5
https://investors.mongodb.com/static-files/090d8215-b0fa-4d74-ae26-d189c7dac830

6
https://investors.snowflake.com/news/news-details/2023/Snowflake-Reports-Financial-Results-for-the-First-Quarter-of-Fiscal-2024/default.aspx

7
https://medium.com/lightspeed-venture-partners/generative-nyc-is-back-with-edo-liberty-of-pinecone-d2ec635ae702

8
https://sacra.com/research/edo-liberty-pinecone-openai-vector-databases/

9
https://www.pinecone.io/learn/series-b/

10
https://techcrunch.com/2023/04/20/coreweave-a-gpu-focused-cloud-compute-provider-lands-221m-investment/

11
https://www.businesswire.com/news/home/20230531005132/en/CoreWeave-Secures-200-Million-Series-B-Extension-Bringing-Total-Round-to-421-Million

12
https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai

13

https://scale.com/about

14

https://scale.com/customers/sap

© The Author(s), under exclusive license to APress Media, LLC, part of Springer Nature 2023

T. Taulli*ChatGPT and Bard for Business Automation*
https://doi.org/10.1007/978-1-4842-9852-7_9

9. Risks

There Are Serious Downsides to AI

Tom Taulli[1]
(1)
Monrovia, CA, USA

For decades, there have been many warnings about the dangers of AI. These have often been from science fiction novels and movies like *2001: A Space Odyssey* and *Terminator*.

But since the launch of ChatGPT, the warnings have grown even more intense – and often from top AI researchers, entrepreneurs, and business leaders.

In March 2023, the Future of Life Institute published an open letter that recommended a pause of AI development for six months. This was focused on technologies that would exceed the capabilities of GPT-4. The letter noted: "Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable."[1]

The open letter garnered over 33,000 signatories, which included Elon Musk, Yoshua Bengio, Stuart Russell, Steve Wozniak, and Emad Mostaque.[2]

This created huge media buzz but there was no pause. The vigorous research and development seemed to continue at a faster pace.

Interestingly enough, some of the signatories said that a pause was unrealistic. But they believed that the open letter would be a good way to highlight the grave risks of AI.

The open letter would be followed up with several others. In May, the Center for AI Safety published a short statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[3]

This also had many notable signatories. Examples included Geoffrey Hinton, Demis Hassabis, Sam Altman, Dario Amodei, Bill Gates, and Ilya Sutskever.

With all this extreme pessimism, there was an impact on Corporate America. The Yale CEO Summit conducted a poll of 199 CEOs of large companies like Walmart and Coca-Cola. About 34% of the respondents said that AI could potentially destroy humanity within a decade, and 8% believed it could be within 5 years.[4]

Yet there was pushback to the doom and gloom. Some AI experts thought that the open letters were overly dramatic. There was even talk that this was a clever attempt to create more hype and excitement about the technology. The open letters could be a ploy to spur regulation, which could potentially help mega companies like Microsoft, Google, and Meta.

Yann LeCun, the chief AI scientist at Meta, had this to say:

> *Will A.I. take over the world? No, this is a projection of human nature on machines. It's still going to run on a data center somewhere with an off switch…And if you realize it's not safe, you just don't build it.*[5]

It's a spirited debate. The fact that the "end of humanity" is a mainstream topic is an indication of the importance of AI.

However, the focus on existential matters could be a distraction. The fact is that AI has many other risks – which may be nowhere as scary – but are important to understand and address.

That's what we'll do in this chapter.

Bias and Discrimination

On a high-level, the structure of an AI system is straightforward. It uses one or more algorithms to process data. From this, the AI will find patterns, which can lead to insights or the creation of new data.

But this opens up two big issues: the data and algorithms could be skewed or biased. This is especially the case with data. After all, the information is often a reflection of society, which has prejudices and false information and notions. In some cases, these problems can lead to discrimination or unfairness with the AI models.

Even synthetic data can have the same issues. For the most part, this is based on underlying patterns of existing data.

Here are some of the main areas where faulty data and algorithms can pose challenges:

- Hiring: An AI tool can be helpful in processing large amounts of resumes. But does the dataset adequately reflect the population? It may not. If a dataset mostly focuses on a certain category or demographic, then this can leave out many qualified employees.
- Facial recognition: This is a technology that has one of the highest risks. Just look at the case of a Black man who was jailed for nearly a week in a Georgia jail because a facial recognition program identified him as a criminal.[6] This was for a crime committed in New Orleans. But he had not been in the state. The computer system made the wrong identification. And this was not an outlier. There are various cases of facial recognition giving off false matches, and this has actually been more common for minorities. Because of these problems, companies like IBM, Microsoft, and Amazon have refrained from selling facial recognition technology for criminal matters. Some states and cities have also enacted restrictions.
- Criminal justice: Some courts have used AI to suggest the length and type of sentences for criminals as well as determining the amounts for bail. These are based on factors like the nature of the offense and recidivism. By

doing this, the judgments are expected to be less impacted by bias and discrimination. But unfortunately, the results have still been problematic. A study from ProPublica found that AI-based evaluations give the false impression of higher accuracy and impartiality.[7] This can mean that a judge may be tempted to use the results at face value and not look deeper at the matter. But the ProPublica study also found that AI came up with the wrong outcomes two times more often for Black people versus white people.

- Credit approval: In November 2019, Apple announced the launch of the Apple Card. This involved a strategic partnership with Goldman Sachs. But quickly, there were complaints popping up on Twitter. Notable tech entrepreneurs said that the Apple Card was biased against women. For example, Apple co-founder Steve Wozniak tweeted that his wife got a lower credit limit – even though she had a higher income. This spurred an investigation from New York regulators. Although, this resulted in the judgment that there was no violation of lending fairness laws. Yet this does not mean that there were no problems. The regulator noted that Apple and Goldman Sachs did not use data for gender, marital status, and race in their algorithms.[8] But when it comes to these types of AI systems, discrimination is often more subtle. There are "proxies" for data for gender, marital status, and race. For example, if a dataset includes information about professions, then this could mean bias for certain categories like caregiving, which have more women.

Another issue with data is that it is often siloed across an organization. This is particularly common for large organizations that span the globe. Over the years, the IT systems have been fragmented because of regulations, mergers and acquisitions, and internal policies. But this means that there will not be a cohesive view of the data.

But as seen with the case with Apple Card, it can be extremely difficult to determine whether a dataset is discriminatory or unfair. What are the standards for this? Let's face it, what may be fair in the United States may actually be illegal in China. The cultural differences can be significant.

Yet there will certainly be progress. There will be more understanding of the negative impacts of AI. But this could pose liability risks for companies. During the next decade, what if researchers have much more proof that AI discriminates and even violates state and federal laws?

This is why some companies will avoid the use of AI for certain use cases. This is the situation with PepsiCo. The company, for example, will not use AI when making hiring decisions.[9]

When it comes to mitigating the discrimination in datasets, a best practice is to start early in the AI process. This could be something similar to how cybersecurity works with software development. It should not be an afterthought or about checking the boxes.

Consider how Visa manages the AI process. The company has setup a team for model risk management.[10] Before the deployment of a model, they will evaluate it according to Visa's guidelines for responsible and ethical AI.

If there is not a proactive approach with AI, then the process of making corrections could be expensive. In some cases, the only option may be to start from scratch and redo the model. This could certainly have a disruptive impact, especially if the AI is mission critical.

Another problem with discrimination and AI is with the lack of diversity with data scientists. This is often called the "white male problem." Yes, for the most part, when it comes to the creation of AI models, it's usually from white males. They often have attended elite schools like Stanford and have advanced degrees.

Now this does not mean that they somehow intentionally create models to engage in discrimination. Rather, the problem is that their backgrounds have an influence on their actions and approaches.

The Women in Data Science (WiDS) conducted a survey – titled "Identifying and Removing Barriers for Women to Pursue Graduate Studies in Data Science and AI" – on gender representation in data science. The research showed that 80% of data scientists in the United States have graduate degrees.[11] However, only 7% of students in master's programs for Computer Science are women.

Data also shows low levels of representation for racial minority groups.

Lack of diversity may not only lead to discrimination and unfairness in AI models but also could mean missing out on the benefits of having different perspectives. If your systems do not represent the demographics of your customers, then there will likely be underwhelming results.

Note Columbia University researchers conducted a survey about algorithmic discrimination.[12] It involved roughly 400 AI engineers. They were tasked with creating systems to make predictions about a group of 20,000 people. The study found that the demographics of the AI engineers had an impact in creating bias in the outcomes.

Black Box

Highly complex AI systems are often referred to as "black boxes." The reason is that it is impossible for a person to understand how the algorithms come up with their conclusions. After all, an AI system – such as those that rely on deep learning and generative AI – can have hundreds of billions of parameters, many hidden layers and the use of advanced mathematics. The datasets can be massive because some models can have more than one trillion tokens.

This has been a nagging issue with AI. With little transparency, it is tougher to engender trust with a system. This can lead to a lack of adoption.

There are also regulatory challenges. Government agencies need to understand how an AI system works before it can be satisfied that there is compliance. In fact, it is common for “black box” algorithms to be banned.

Consider the new policy that the California Department of Insurance adopted in June 2022.[13] It set forth requirements for the use of AI, which include explainability. The regulation required that an insurer provides the reasons why an algorithm made decisions about matters like premium increases.

Focusing on explainability can be more than just about being compliant. This can also help with the bottom line. This is the conclusion of research from McKinsey & Co.[14] It showed that companies that use explainability can attribute at least 20% of their EBIT (earnings before interest and taxes) to their use of AI. By bolstering trust with customers, there is a better chance of growth of 10%+ for revenue and EBIT.

There is research to help deal with black boxes. This is usually called explainable artificial intelligence (XAI). This will provide a description of a model, the potential results and impacts, and the bias.

There are a variety of tools to help with this. However, the market category is still in the nascent stages. XAI tools are also being integrated in large AI platforms. They often do sophisticated sensitivity and feature importance analysis. They can also leverage surrogate models for comparisons. The tools often provide this information with visualizations like heatmaps.

A top developer of XAI tools is Fiddler Labs. Krishna Gade founded the company in 2018 because he saw the problems with AI models while he worked at companies like Facebook and Pinterest. Since the inception of the company, he has raised over $45 million. Some of his investors include Insight Partners, Lightspeed Venture Partners, Bloomberg Beta, Lockheed Martin, and The Alexa Fund.[15]

A case study of Fiddler Labs’ technology is with how.fm, which creates training systems for warehouse staffs. The company uses a variety of sophisticated AI systems, such as for natural language processing (NLP).

The company had a lean team that leveraged agile practices to speed up the process of development. How.fm evaluated building their own system for explainability of the models and monitoring. But this would have been too expensive. The company also did not have expertise in these areas.

The better approach was to use a third-party system. How.fm scrutinized various offerings and Fiddler Labs stood out.[16] The system offered the key benefits and fit the requirements for NLP. But implementing the technology was also fairly easy, taking only a day for onboarding.

With this strategy, how.fm's team could focus their efforts on those areas that drove the most value – say for creating new models. Before using the Fiddler Labs system, it was common for the data scientists to spend hours on tedious activities like analyzing logs to detect AI model issues.

Besides tools, some companies are creating governance committees. These groups help with crafting policies for responsible and ethical AI, including guidelines for explainability. A governance committee will usually be composed of managers, legal experts and AI professionals.

There can certainly be some tough dilemmas. For example, a model that is opaque may be highly accurate and useful. But when explainability is added, this can mean lower quality for the results. Often, this is because certain data is excluded or the algorithm is modified.

The decision then becomes how important explainability is. In some cases, it may be essential – again, say for regulatory purposes.

> **Note** The different approaches for dealing with black boxes can get confusing. Terms like transparency, explainability, and interpretability are often treated as the same. Yet there are clear differences. Transparency is where employees and customers have an understanding of how a model works. Transparency can also be segmented, such as with process transparency, data transparency, and system transparency. As for explainability, this shows the reasons for why the system has

made certain conclusions. Interprebility, on the other hand, gauges the predictability of the model's outputs.

Physical Safety

AI can potentially have real-world harms, say with physical safety. An area of major risk is with healthcare. If an AI model is based on faulty data or algorithms, there could be a misdiagnosis.

Consider a study from researchers at the University of Washington. In paper published in Nature Machine Intelligence, they analyzed models to detect COVID-19 by using X-rays. What the researchers found is that the predictions often used shortcuts – similar to what a human often does. The conclusion was that using sophisticated AI could result in errors for clinical purposes.

According to Alex DeGrave, a data scientist and one of the authors of the study:

> *A physician would generally expect a finding of COVID-19 from an X-ray to be based on specific patterns in the image that reflect disease processes. But rather than relying on those patterns, a system using shortcut learning might, for example, judge that someone is elderly and thus infer that they are more likely to have the disease because it is more common in older patients. The shortcut is not wrong per se, but the association is unexpected and not transparent. And that could lead to an inappropriate diagnosis. A model that relies on shortcuts will often only work in the hospital in which it was developed, so when you take the system to a new hospital, it fails—and that failure can point doctors toward the wrong diagnosis and improper treatment.*[17]

But there are other areas where physical safety has been a major problem. This has been the case with self-driving cars. This technology has proven extremely difficult to develop because of the challenges with anticipating the seemingly endless scenarios.

Because of this, there may ultimately be a need to require some level of human intervention. This could be with having a person monitor various vehicles from a remote location – like an air traffic controller. This would not only help reduce the error rates but also bolster peace-of-mind for passengers.

On the other hand, AI can help bolster physical safety as well. This is the case with work injuries, especially with industrial settings.

Based on the latest data – from the US Bureau of Labor Statistics – there were over 2.6 million nonfatal injuries and illnesses in the workplace during 2021. There were also over 5,100 deaths.

As for AI, it can help with the following:

- Use computer vision systems to monitor the workplace. This can be a way to detect if workers are using the necessary protective equipment as well as engaging in safe practices.
- Data from sensors, IoT (Internet-of-Things) systems, and other apps can be used to identify any harmful conditions.
- Drones and robots can be used to survey a worksite. This can be a way to evaluate the safety.
- Wearables can be useful to measure worker fatigue.

An interesting startup in the AI-based workplace safety category is Invisible AI, which focuses on factories. The company's technology uses many sophisticated cameras that do not require an Internet connection. Each of these systems has an AI chipset and 1 terabyte of storage. They are able to determine depth, which helps with finding dangers in a workplace.

In 2022, the company announced a Series A funding for $15 million, bringing the total amount invested to $21 million.[18]

Deepfakes

A deepfake is a form of AI that creates realistic images, audio, and video. There are different types. One replaces a part of an image, say with placing a head on an image. For the most part, these are usually easy to detect as deepfakes. They have a tendency to look exaggerated or unnatural.

But a much more sophisticated approach is to use generative AI. This involves processing training data to create the deepfake. This can create sound or video that looks real.

The following are some examples of notable deepfakes:

- In March 2022, a deepfake video of Ukrainian President Volodymyr Zelenskyy went viral.[19] In it, he urged the surrender to the Russians. The video actually had some flaws. The voice had issues with syncing and the voice seemed false. But detecting this would have taken some effort. In fact, the deepfake video got picked up with a local TV station in Ukraine.
- In May 2023, a video of an explosion at the Pentagon appeared on social media.[20] It actually led to a drop in the stock market. But yes, it was a deepfake and the Pentagon quickly put out a statement about it.
- In 2020, an employee at a Japanese company received a call from his boss. He said that he needed a wire for $35 million for an acquisition.[21] The employee carried it out. The problem? The boss's voice was a deepfake.

With cutting-edge technologies like generative AI, deepfakes are getting much more sophisticated. They are also becoming much easier to create. Some tools may take just a few minutes to develop a realistic deepfake.

The FBI and other federal agencies have rung the alarm bells about this and have put out warnings. There has also been a dramatic increase in deepfakes for nefarious purposes.

One terrible category is sextortion. This is where a deepfake impersonates a person and makes it appear they are engaged in salacious activities. According to the FBI, these types of videos saw a 322% increase in the United States from February 2022 to February 2023.[22] No fewer than 12 of the cases resulted in suicide.

Businesses are also becoming a common target for deepfakes. In fact, this is turning into one of the biggest threats.

Perhaps the area of highest risk is with business identity compromise (BIC), such as noted in the earlier example about the theft of $35 million. This type of deepfake can lead to other dangers like damage to a company's brand, release of confidential information, and harm to customer and partnership relationships.

Deepfakes are not illegal per se since there are many legitimate uses for the technology. But it is criminal if they are used for criminal purposes, such as fraud.

What can be done to mitigate the potential harm of deepfakes? First of all, there should be employee training. This will provide guidance on how to recognize a deepfake scam.

Here are some other guidelines in helping to detect deepfakes:

- Patches: A deepfake will often create synthetic skin, say around the eyes, hairline, nose, and chin. These may be slightly off-color. There could also be differences in skin tone because of the divergences with the lighting of the images.
- Eye movements: Are they natural? Is there actually any movement? Eye movement is one area where deepfakes usually fall short.
- Audio: The lip syncing could be off or the voice may be monotone or robotic. There could then be mispronunciations of certain words.
- Body movements: A deepfake could be jerky or stiff. Another telltale sign is if the body language is not consistent with the emotion being expressed. In some cases, the nose and eyes may be pointing in different directions.

Next, you can reevaluate the corporate processes for making major decisions, such as for authorizing a payment. This can include multiple ways to identify that the person is not an imposter.

Finally, cybersecurity companies are creating tools to help detect and ward off deepfakes. These systems could be quite helpful.

Cybersecurity

AI is emerging as a major cybersecurity threat. Just look at a research report from CyberArk, which is a top security firm. It was based on the feedback from about 2,300 cybersecurity professionals from private businesses and government institutions.[23] The survey found that 93% of the respondents believed that AI was a serious threat to their organization. The biggest worry was about AI systems that hackers use to create malware.

Another threat vector is phishing, which involves an email that includes links that capture confidential information. This can lead to accessing systems and enabling ransomware.

With generative AI, hackers can now make more effective phishing emails. This is because the language is more natural sounding and can be translated into many languages.

To deal with these threats, there needs to be a robust cybersecurity strategy. It needs to be updated frequently and backed up with employee training. There should also be rigorous backups and the use of encryption to protect valuable data.

Privacy Laws and Regulations

With the rapid advances in AI, there have been growing calls for regulation of this technology. Some of the issues include privacy, the unintended consequences of the technology, the thorny problems with ethics, and the impact on national security.

When it comes to regulation, the United States has historically been resistant. The general philosophy has been to allow the marketplace to provide the necessary guardrails.

This has certainly allowed for creating a powerful economy. The fact is that the United States remains the dominant player for technology and is the leader in AI. For US policymakers, there is a key focus on making sure that the innovation is not stifled.

Regulation often comes about when there is a crisis. This happened during the Great Depression during the 1930s. The federal government took swift actions to enact laws to regulate the financial system, such as setting up the Securities and Exchange Commission. Then there was the financial crisis of 2008 to 2009. In response to this, the federal government enacted various laws to regulate risky activities with complex derivatives.

So might this pattern continue with AI? This would not be surprising.

Granted, there is some bipartisan support for regulation. But there are widely differing opinions on what to focus on. In other words, it seems like a good bet that there will not be aggressive action with AI regulation.

During the past few years, some states like California, Virginia, Colorado, Maine, Nevada, and Connecticut have passed legislation for privacy rights to data. This has had an impact on companies that deploy AI models, such as for marketing purposes.

In 2022, cosmetics retailer Sephora announced that it would pay a $1.2 million penalty for alleged noncompliance with the California Consumer Privacy Act (CCPA).[24] The company did not adequately disclose that it was collecting and selling certain information. There was also no opt-out option.

This was the first time California brought a case under CCPA. But there will likely be many more. The enforcement will not be just for larger companies either. California has indicated that it will also monitor small companies for violations.

While the state privacy laws are helpful, they do have drawbacks. They have led to the fragmentation of regulations, which has made it difficult for companies to navigate them. This is why there are efforts – often from businesses – to have a federal privacy law.

As for AI regulation, there are different approaches. One is to create a separate federal agency. This would help to centralize the regulations and provide more uniformity.

But this could lead to some problems. For example, there could be more bureaucracy for companies to work with. Then there is the potential for "regulatory capture." This is where large companies ultimately have outsized influence on the federal agency. They can do this because of their enormous resources for lobbying and making campaign contributions.

Another approach is where various agencies will take on AI regulation. While this can help mitigate regulatory capture, this can mean even more fragmentation.

Thus, there is no perfect solution. Yet there are some examples to consider, such as with the proposed AI Act for the European Union. This is a comprehensive set of regulations, which uses a risk-based framework. For example, those parts of AI that pose the highest potential for danger – like real-time, remote biometric surveillance, scraping public data for facial recognition or predictive policy – are banned. Lesser forms of AI will have varying degrees of scrutiny. There will also be disclosure, such as for when generative AI is used for content creation.

The penalties are severe. The laws allow for fines of as much as 7% of a company's global revenues.[25]

Junk Content

One of the most compelling features of generative AI is the ability to write humanlike content. But this can cause lots of issues. When ChatGPT was launched, an early use case was with students who wrote their homework assignments with the service. Several school districts banned ChatGPT.

Some other teachers changed their approach to instruction. For example, they would have more in-classroom writing assignments.

But as generative AI has become more ubiquitous, it has led to something else: junk content. There has emerged a "get rich quick" cottage industry that sells tutorials on how to essentially create fake content sites that generate fees from Google ads. A quick search on YouTube or Amazon will highlight the many offerings available.

Even reputable media publications are experiencing challenges with this trend. Just look at *Clarkesworld*, which is a journal for science fiction and fantasy stories.

The site allows anyone to submit a story. Until the start of 2023, it would get about 25 spam stories a month. But this would explode to 500 by February.[26]

This situation got so bad that *Clarkesworld* had to temporarily shutdown its submission form. The editors had to rethink their existing approaches.

They did try various AI-content detection systems. But for the most part, they created too many false positives. Consider that many of the contributors to the magazine are from overseas and their English may be somewhat quirky. And yes, a detection system would flag the content as spam.

Interestingly enough, the editors realized that it was fairly easy for themselves to do the detection. The reason is that the stories were often formulaic and not interesting!

Yet this still meant they had to spend more time on something that was unproductive. There was less time to devote to legitimate authors.

Regardless, it seems that spam content will remain a problem. Many companies find this as an easy way to create blogs and white papers.

True, there is nothing wrong with this. Although, in terms of ethics, it's probably a good idea to disclose that the content is AI generated.

The AI content may also result in a long-term problem known as "model collapse" or "data pollution." This means that more and more of the content available used to train large-language models will be created by AI. This will result in an echo feedback loop, and the outcomes of the models could be less effective.

In May 2023, AI researchers from the UK and Canada published a paper on this topic, titled "The Curse of Recursion: Training on Generated Data Makes Models Forget." Their conclusions were not encouraging. The authors wrote:

> *We find that use of model-generated content in training causes irreversible defects in the resulting models, where tails of the original content distribution disappear… We demonstrate that it has to be taken seriously if we are to sustain the benefits of training from large-scale data*

> *scraped from the web. Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of content generated by LLMs in data crawled from the Internet.*[27]

The paper states that the model collapse can happen fairly quickly. It also shows that models can easily forget what they have learned from the original datasets.

Given this, researchers will need to be more circumspect when building their generative AI models. This may even be about using filters to exclude AI-generated content. This could be done if content is labeled or watermarked.

The Cloud

Cloud computing has been a critical driver for the growth of AI. It has allowed any company to cost-effectively utilize this technology – and scale it.

But the cloud has its own nagging issues. Keep in mind that there is considerable legacy infrastructure. The reason is that the cloud was generally not built for complex and data-heavy AI applications. The underlying computer systems are usually run on CPUs not GPUs.

True, the mega cloud providers are rushing to change their infrastructures. There are also various startups that are trying to fill the void.

Then there are some old-line companies that see this as an opportunity to catch-up on the cloud race. Take a look at Hewlett Packard Enterprise. In June 2023, the company launched its cloud service that’s focused on AI. This is not just for global companies but for companies of any size.

The platform, which is called HPE GreenLake, provides access to HPE Cray XD supercomputers. According to Justin Hotard, who is the company’s executive vice president and general manager of High Performance Computing, AI & Labs:

> *HPE GreenLake for Large Language Models allows our customers to rapidly train, tune, and deploy large language models on demand using a multi-tenant instance of our supercomputing platform – truly a supercomputing cloud combined with our AI software.*[28]

But it will take time to modernize existing cloud infrastructures. Ironically, companies are starting to now take more consideration of an on-premise strategy. This can provide more control over the system and security. The costs may even be lower as storage and networking systems are fairly cheap.

To capitalize on this, Dell Technologies has been aggressively developing its own line of AI servers called PowerEdge. These can be installed in or outside a data center.[29] They are also based on GPU systems.

Data Problem

Stuart Russell has remarked AI systems are "running out of text in the universe."

He should definitely know. Russell has a Ph.D. in computer science from Stanford and is a professor at Berkeley. His area of focus is on AI, and he has authored books like *Artificial Intelligence: A Modern Approach*.

Russell's fear about data availability is backed up from academic research.

Consider a paper published in 2022 titled "Will we run out of ML data? Evidence from projecting dataset size trends."[30] The authors predict that the world will run out of high-quality language data before 2026. Although, low-quality content will reach their limits from 2030 to 2050. This will be similar to vision data.

If these predictions prove to be true, then there will likely be a notable slowdown in progress with AI models.

Note Estimates show that generative AI models can be up to 100 times bigger than traditional AI models.[31] This means that they will consume much more resources, such as with storage, semiconductors, and electricity.

Intellectual Property

When an AI model creates data from training on existing content from the Internet, what are the rights to this? Is the new content the intellectual property owned by those who used the system generator? Or is there copyright protection for those who created the original content?

Well, there are no clear-cut answers. Those who operate AI systems argue that the "fair use" doctrine applies. This is where a copyright is not relevant because there is only a small part of the content used or it has been transformed.

However, the fair use laws did not anticipate technologies like generative AI. They are also somewhat vague and contradictory.

Because of all this, the Supreme Court may ultimately decide on this crucial issue of fair use and AI. But this could take some time. The Supreme Court is not known to act with much speed.

In the meantime, there has been a burst of litigation – and perhaps one of these cases will wind up being decided by the Supreme Court. Even so, this will not be a cure-all. After all, there is uncertainty across many other countries.

Two noted authors, Paul Tremblay (who wrote *The Cabin at the End of the World*) and Mona Awad (the author of *Bunny*), have alleged that ChatGPT has scraped their works.[32] Their lawsuit has claimed that the technology provides accurate summaries of some of their books. The authors are alleging violation of copyright laws.

Then there was a lawsuit from Sarah Silverman, who is a comedian and author. She has claimed that OpenAI and Meta have violated her copyright to her memoir, *The Bedwetter: Stories of Courage, Redemption, and Pee*.[33]

The litigation is also more than about books. There are also lawsuits emerging over the use of generative AI to create images. In 2023, Getty Images filed a lawsuit against Stability AI on this matter.

Getty Images has claimed that there was a “brazen infringement” of its intellectual property and that it was on a “staggering scale.”[34] The company said that Stability AI copied more than 12 million images.

With the swirl of litigation and uncertain legal liability exposure, companies are certainly worried about the implications. This is why some of them are taking a guarded approach to technologies like generative AI.

Although, some AI providers are being proactive. Just look at Adobe. The company has implemented a policy to provide compensation for the creators of its massive images database. Adobe has also agreed to provide some protections for its customers against potential litigation.

According to the company: “If a customer is sued for infringement, Adobe would take over legal defense and provide some monetary coverage for those claims.”[35]

This will certainly allow for much more adoption of its technology – especially from larger businesses.

An area that has become a flash point for intellectual property and generative AI is the entertainment industry. What if someone uses a generative AI system to create a character that is similar to say Ethan Hunt in the *Mission: Impossible* movie franchise? Well, it’s not clear.

But Hollywood attorneys are already amending contracts to deal with this brave new world of technology. There is definitely a lot at stake.

Consider that one of the biggest issues with the Hollywood strike for writers and actors is AI.[36] The writers believe that studios will be tempted to replace them with bots. This would be possible since there are huge repositories of scripts available as training data.

As for the actors, there is a similar threat. Film footage can be used as training data to create digital avatars.

Keep in mind that Hollywood already utilizes generative AI. An example is with the movie *Here*. Generative AI made it possible to de-age actors like Tom Hanks and Robin Wright.[37]

Or look at the latest *Star Wars* films. Disney leveraged generative AI to clone the voice of James Earl Jones for Darth Vader.

> **Note** In March 2023, the US Copyright Office made a major decision. It ruled that content created from generative AI could not have copyright protection. The reason is that it is not "the product of human authorship."[38]

The spate in litigation – or the threat of it – is likely to encourage settlements. This has already been the case with the Associated Press. In July 2023, the organization agreed to provide access to OpenAI for up to two years.[39] The financial arrangement was not disclosed. But OpenAI will have access to the full content archive that goes back to 1985.

Associated Press has been a leader in using automation technologies and AI. To this end, it has systems that generate stories for earnings reports, local sports and weather reports.

In the meantime, generative AI could represent an existential threat to industries like journalism. This is why some of the world's largest news and magazine publishers are discussing forming a partnership.[40] Some of them include *The New York Times*, *The Wall Street Journal*, Vox Media, Condé Nast (parent company Advance Publications Inc.), Politico, and Insider.

Such a partnership would be unique because these organizations are highly competitive. However, this does point to the perception that generative AI is a new type of threat that must be dealt with.

This is the strategy for Mathias Döpfner, the CEO of Axel Springer. The company is Europe's largest periodical publisher. In an interview with Fortune, Döpfner said:

> *I'm saying Axel Springer can be in three years gone, if we now give the wrong answers, or it can be at a completely different level. So I think that is the first and most important thing, that everybody has to rethink what she or he is doing.*[41]

But he is being proactive. He has restructured his newsroom at Bild, by cutting about 200 jobs. Part of this was due to the impact of AI.

Döpfner is also leveraging its resources to engage in active M&A (mergers and acquisitions) and venture investments to bolster its AI resources.

> **Note** In July 2023, the Washington Post reported that the Federal Trade Commission (FTC) was investigating the potential harm of OpenAI's ChatGPT.[42] The agency raised concerns that the technology was defaming or disparaging people because of the inaccuracies in the content.

Conclusion

As AI gets more powerful, there are certainly more risks. Some of the biggest are about how this technology works, which involves training models with existing data. This can lead to problems like unfairness, discrimination and bias.

Another problem with AI is that the models are getting much larger and complicated. They are essentially turning into black boxes. But this can make it more difficult for the models to pass muster with regulators. What's more, a model that lacks explainability may mean that people will not trust the results.

For businesses, some of the other risks include cybersecurity and deepfakes. AI can be a pernicious tool to allow hackers to enact serious damage.

Another interesting risk is the explosion of generated content. This can ultimately make it more difficult to train models. There is even the risk that the world will run out of useful data.

Even existing cloud infrastructures are a problem. They were not built for the intensive needs of modern-day AI models. This may even mean that on-premise environments could be more cost-effective.

Finally, intellectual property protection is another issue. They are extremely complicated – and the potential litigation exposure could be considerable. But it may take a while for there to be legal clarity.

Footnotes

1

www.brookings.edu/articles/the-problems-with-a-moratorium-on-training-large-ai-systems/

2

https://futureoflife.org/open-letter/pause-giant-ai-experiments/

3

www.safe.ai/statement-on-ai-risk

4

https://fortune.com/2023/06/15/yann-lecun-ai-godfather-destroy-humanity-threat/

5

https://fortune.com/2023/06/15/yann-lecun-ai-godfather-destroy-humanity-threat/

6

www.axios.com/2023/01/07/facial-recognition-issues-problems

7

www.bu.edu/articles/2023/do-algorithms-reduce-bias-in-criminal-justice/

8

https://techcrunch.com/2021/08/14/how-the-law-got-it-wrong-with-apple-card/

9

www.wsj.com/articles/rise-of-ai-puts-spotlight-on-bias-in-algorithms-26ee6cc9

10

www.wsj.com/articles/rise-of-ai-puts-spotlight-on-bias-in-algorithms-26ee6cc9

11

www.widsconference.org/blog_archive/new-wids-research-addresses-barriers-women-face-pursuing-graduate-studies-in-data-science-and-ai

12

www.cdotrends.com/story/17573/how-data-science-diversity-can-actually-create-better-products

13

www.debevoise.com/insights/publications/2022/07/california-restricts-insurers

14

www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it

15

www.prnewswire.com/news-releases/fiddler-raises-32-million-in-series-b-funding-as-it-leads-the-market-of-machine-learning-explainability-and-performance-management-301314395.html

16

www.fiddler.ai/customers/howfm

17

https://healthcare-in-europe.com/en/news/ai-shortcuts-could-lead-to-misdiagnosis-of.html

18

www.prnewswire.com/news-releases/invisible-ai-raises-15-million-series-a-to-rapidly-scale-innovative-computer-vision-platform-across-manufacturing-facilities-301635650.html

19

www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

20

www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai

21

www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=48fa0d9a7559

22

www.foxnews.com/media/fbi-warns-ai-deepfakes-being-create-sextortion-schemes

23

https://venturebeat.com/security/ai-powered-malware-is-a-growing-security-concern-cyberark-survey-finds/#:~:text=AI%2Denabled%20threats%20are%20a,by%20employee%20churn%20in%202023.

24

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

25
www.wsj.com/articles/ai-regulation-is-almost-here-in-europe-10b32c8f

26
https://techcrunch.com/2023/02/21/clarkesworld-ai-generated-submissions/

27
https://arxiv.org/abs/2305.17493v2

28
www.datacenterknowledge.com/cloud/hpe-unveils-ai-supercomputer-cloud-service

29
www.dell.com/en-us/blog/new-servers-deliver-major-leap-ai-performance/

30
https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

31
www.wsj.com/articles/rising-data-center-costs-linked-to-ai-demands-fc6adc0e?mod=hp_minor_pos4

32
-

33

https://venturebeat.com/ai/what-sarah-silvermans-lawsuit-against-openai-and-meta-really-means-the-ai-beat/

34

www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion

35

https://venturebeat.com/ai/what-sarah-silvermans-lawsuit-against-openai-and-meta-really-means-the-ai-beat/

36

https://uk.sports.yahoo.com/news/opinion-hollywood-actors-strike-over-134354030.html

37

www.wsj.com/articles/ai-chatgpt-hollywood-intellectual-property-spongebob-81fd5d15

38

www.wsj.com/articles/ai-chatgpt-hollywood-intellectual-property-spongebob-81fd5d15

39

www.axios.com/2023/07/13/ap-openai-news-sharing-tech-deal

40

www.wsj.com/articles/big-news-publishers-look-to-team-up-to-address-impact-of-ai-80c3b88e

41

https://fortune.com/2023/06/27/mathias-dopfner-axel-springer-ai-interview-cannes-lions-alyson-shontell/

42

www.wsj.com/articles/chatgpt-under-investigation-by-ftc-21e4b3ef?mod=hp_lead_pos2

Index

A

ABC Inc.
Active data profiling
Adaptive data quality
Adobe
AI
    bias
    Black boxes
    cybersecurity
    data problem
    data strategy
    deepfakes
    discrimination
    intellectual property
    physical safety
    privacy laws
    regulations
    risks
AI-based workplace safety category
AI code development tools
AI-content
AI-content detection systems
AI development team
AI-powered recruiting system
AI projects
AI text-to-image systems
Algorithmic discrimination
AlphaGo system
Amazon Web Services (AWS)
The American Institute of CPAs (AICPA)
ANALOGY
Anomaly detection
Apache Spark
Apple Card

Artificial general intelligence (AGI)
Artificial intelligence (AI)
    categories
    ChatGPT
    history
    machine learning
  See Machine learning
    strong AI
    weak or narrow AI
Artificial Intelligence: A Modern Approach (book)
Association of Change Management Professionals (ACMP)
Autoencoder

B

Backpropagation
Bard
    cases
    DeepMind
    description
    Duet AI system
    experiment
    gear icon
    images
    LaMDA technology
    launch of
    location data
    Med-PaLM 2
    modify the response
    PaLM 2
    pencil icon
    private/sensitive information
    prompt for new chat
    question-mark icon
    "regenerate drafts"
    resources
    response

  response/export
  revised tweet
  screen
  Sec-PaLM
  skillset
  at social media
  tables
  URL location
  "View other drafts" option
  your activity
BERT
Big Data
  description
  three V's
    variety
    velocity
    volume
  value
  variability
  veracity
Bing Chat
Binning
Black boxes
Boiling the ocean
BrowserOp
Businesses
Business identity compromise (BIC)

C

California Consumer Privacy Act (CCPA)
C-Corporation
Change advisory board (CAB)
Change fatigue
Change management
  AI
  boiling the ocean

  - CAB
  - change fatigue
  - Conner, Daryl
  - culture
  - definition
  - implementation
  - leadership
  - plan
  - roles
  - software tools
  - training
  - types
- Change Management Institute (CMI)
- ChatGPT
  - beta features
  - cases
  - chat interaction
  - chat sessions
  - custom instructions
  - data analytics
  - data controls option
  - definition of
  - Enterprise
  - generative AI
  - icons
  - image editing
  - launch of
  - LLM
  - main screen
  - models
  - new chat session
  - OCR
  - OpenAI
  - open plugin devtools
  - plugins
  - regenerate

  - resources
  - settings option
  - skillset
  - survey
  - technical aspects
  - technology adopted
  - translation
  - usage policies
  - warning message
- ChatGPT Sentiment Study
- ChatWithPDF
- *Clarkesworld*
- Classification models
- Claude 2
- The Cloud
- Cloud computing
- Clustering
- CoderPad
- Complaint handling
- Constitutional AI
- Constrastive Language-Image Pre-training (CLIP)
- Conversion table
- Convolutional neural network (CNN)
- CoreWeave
- COVID-19
- COVID-19 pandemic
- Credit approval
- Criminal justice
- CRM (Customer Relationship Management) systems
- CSV (comma-separated values) file
- Customer relationships
- Customer support
- Custom instructions
- Cybersecurity

D

Dark data
Data
    AI
    evaluation
    labeling
    quantity
    types
    wrangling/cleaning
Data analytics
Database systems
    Apache Spark
    Databricks
    graph database
    IMS
    MongoDB
    Pinecone
    relational database
    in 1950s
    Snowflake
    Uber
    vector databases
Databricks
Data collection
Data controls option
Data creation
Data extraction
Data lakes
Data pollution
Datasets
Data strategy
Deepfakes
Deep learning
    backpropagation
    CNN
    disadvantages
    hidden layers

  - matrix
  - node/neuron
  - parameters
  - RNN
  - scalar
  - tensor
  - vectors
  - word embeddings
- DeepMind
- Dell Technologies
- Diffusion models
- Domain-specific LLMs
- Drones
- Duet AI
- Duplications

E

- Earnings before interest and taxes (EBIT)
- EBITDA
- ELIZA
- Email marketing
- Equal Employment Opportunity Commission (EEOC)
- eXpert CONfigurer (XCON)
- Extensible Markup Language (XML)

F

- Facial recognition
- Feel-Felt-Found’ closing technique
- Few-shot prompting
- Fiddler Labs system
- Fiddler Labs’ technology
- Finance
- Financial management
- Forum clause

G

- Games

Gemini
Generative Adversarial Network (GAN)
Generative AI
    black boxes
    customer operations
    data
    discriminative models
    GAN
    marketing
    origins
Generative Pre-trained Transformer (GPT)
Genomic data
GitHub Copilot system
Gmail and Docs
Go games
Google
Google Chat
Google Cloud
Google Docs
Google Meet
Google Sheets
Google Slides
Google’s stock
GPT-3.5
GPT-4
GPUs
Graph database
Graphics processing units (GPUs)

H

Hackers
Healthcare industry
“Hearts and minds” concept
HPE GreenLake
HuggingChat
Human resources

Hybrid cloud

I

Ich liebe Schokolade
Image editing
ImageNet
ImageNet dataset
Incremental approach
Inflection AI
Information Management System (IMS)
Insurance
Intellectual property
Internet-of-Things (IoT)
Intuit Small Business

J

JavaScript
JavaScript Object Notation (JSON)
Junk Content

K

Kayak plugin
Key performance indicators (KPIs)
K-Nearest Neighbor (KNN)

L

Labeled data
LaMDA technology
Large-language models (LLMs)
Latent space
Legal concepts
Legal costs
Limited liability
LISP
Long Short-Term Memory (LSTM)

M

Machine learning

    categories
    deep learning
  See Deep learning
    history
    reinforcement learning
    supervised learning
    unsupervised learning
Marketing
Med-PaLM 2
Metadata
Meta’s LLaMa 2 LLM
Microsoft Azure
Missing data
Model collapse
Model-generated content
MongoDB
Multi-cloud strategy

N

Naïve Bayes classifier
Natural language processing (NLP)
Neural Information Processing Systems conference (NeurIPS)
New Chat
Nonsensical data
NoSQL databases
Nuance Communications
Nuance’s applications

O

One-hot encoding
OpenAI
    blog post
    “capped” for-profit organization
    deep learning
    develop policy
    founders

  history
  LLMs
  machine learning
  Microsoft investment
  non-profit artificial intelligence
  reward-punishment system
  strategy
  university
  usage policies
OpenAI Gym
OpenAI Playground
OpenAI Tokenizer
Open plugin devtools
Open source projects
Optical character recognition (OCR)
Oracle
Outliers

P, Q

Pathways Language Model (PaLM) 2
Patient data
PC revolution
Perceptrons (book)
Phishing
Physical safety
Pi
Pinecone
Plugins
PowerEdge
Predictive transformation
Principia Mathematica (book)
Private cloud
Prompt engineering
  benefits
  broad *vs.* specific
  challenges

  - context
  - description
  - factor
  - few-shot prompting
  - hire
  - instructions
  - OpenAI Playground
  - parameters
  - resources
  - zero-shot prompting
- Public cloud
- PwC HR Tech Survey
- Python
- Python language

R

- Random Forest
- Real-time data
- Recurrent neural network (RNN)
- Regenerate
- Regression model
- Reinforcement learning
- Reinforcement learning with human feedback (RLHF)
- Relational databases
  - categories
  - Db2
  - disadvantages
  - early 1970s
  - foreign key
  - mission-critical operations
  - primary key
- Remedial change management
- Reparameterization
- Return On Investment (ROI)
- Reward-punishment system
- Risk management

Robots

S

SAP
Scale AI
Sec-PaLM
Self-attention
Semi-structured data
Sentiment analysis
Sextortion
Small business
Snowflake
SOC 2 (System and Organization Controls 2) compliant
Social media
Standardization
Structured data
Structured Query Language (SQL)
Supervised learning
    classification models
    disadvantages
    KNN
    labeled data
    Naïve Bayes classifier
    Random Forest
    regression analysis
    regression model
    SVM
Support Vector Machines (SVMs)
Symbolic system
Synthetic data

T

Taxes
TextFX
Time-series data
TPUs

Trade secret
Transformational change management
Transformer model
Transparency
Tweaking
Twitter

U

Unplanned change management
Unstructured data
Unsupervised learning
  anomaly detection
  association
  autoencoder
  clustering
  LLMs
  uncover patterns
Use cases for ChatGPT and Bard
  AI code development tools
  AI text-to-image systems
  bias and fairness
  compliance
  customer support
  finance
    accounting and bookkeeping
    budgeting and costs
    and credit
    financial management
    Intuit Small Business
    time consuming
  HR
    diversity and inclusion
    EEOC
    hiring and termination
    offboarding
    onboarding

- recruiting
- retention

human resources
legal costs
- clarification of clauses and terms
- comparing clauses
- contract/legal document
- flagging issues
- general terms
- hourly fees
- insurance
- intellectual property
- legal concepts
- legal document creation
- missing clauses
- real estate
- responsibilities
- taxes

marketing
- content creation
- email marketing
- generative AI
- prompts
- SEO and SEM
- strategy

privacy
product development
- and best-selling author
- design and prototype
- feedback and iteration
- ideation and brainstorming
- market analysis
- marketing and launch
- video games

project management
- basics

        execution and monitoring
        leadership and team management
        planning and strategy
        project closure and review
        risk management
        skillset
    review the content
    sales
        closing techniques
        CRM systems
        email templates
        generative AI
        objection handling
        prospecting
        qualifying leads
        sales strategy
        survey
        tools
User interface

V

Variational Autoencoder (VAE)
Vector databases
Verbosity

W

Warehouses
Wearables
Web Browsing function
WeChat social network
White male problem
The Women in Data Science (WiDS)

X, Y

XAI tools
X-rays

Z

Zero-shot prompting

Table of Contents

Front Matter
1. New World of Automation
2. ChatGPT
3. Bard
4. Prompt Engineering
5. Use Cases
6. Change Management
7. Understanding AI
8. Data
9. Risks
Back Matter