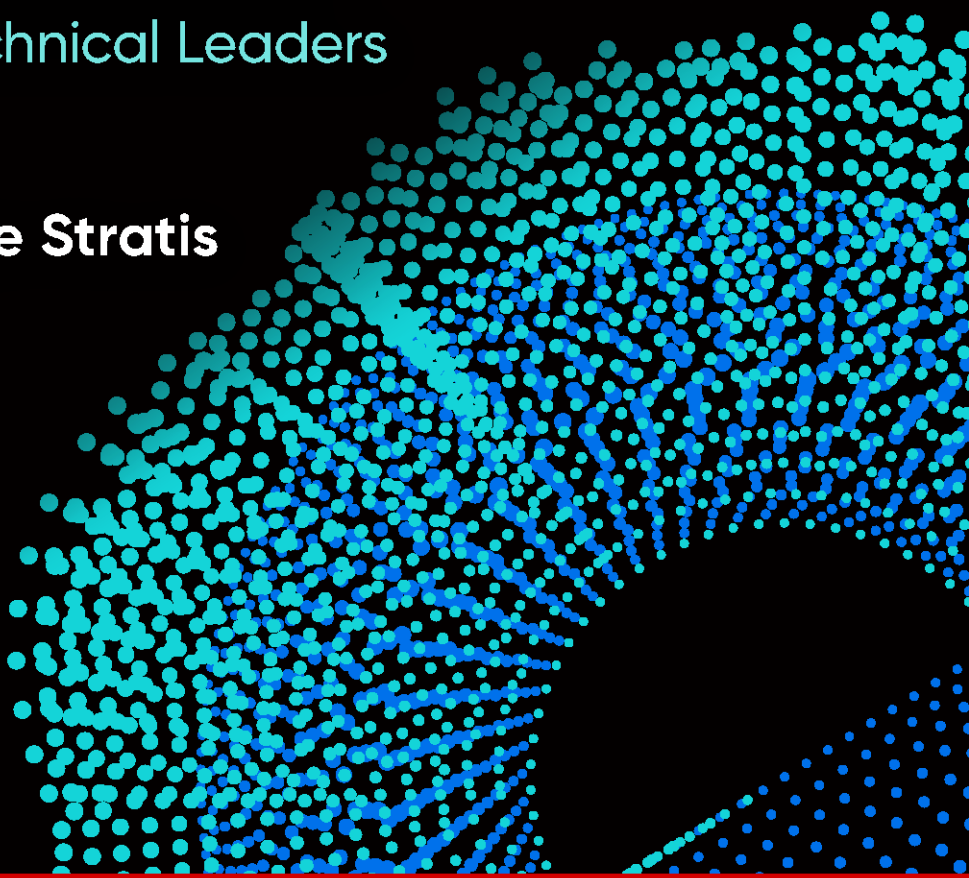
O'REILLY®

Compliments of Couchbase

What Is Generative AI?

A Generative AI Primer for Business and Technical Leaders

Kyle Stratis



REPORT

Couchbase

The Multipurpose Database for AI-Powered Applications

You can deliver hyper-personalized customer experiences, reduce risk and management efforts, and save your business money.

Learn More

What Is Generative AI?

A Generative AI Primer for Business and Technical Leaders

Kyle Stratis

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

What Is Generative AI?
by Kyle Stratis

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Nicole Butterfield
Development Editor: Jeff Bleiel
Production Editor: Gregory Hyman
Copyeditor: Audrey Doyle
Interior Designer: David Futato
Cover Designer: Ellie Volckhausen
Illustrator: Kate Dullea

December 2023: First Edition

Revision History for the First Edition
2023-11-30: First Release

This work is part of a collaboration between O'Reilly and Couchbase. See our statement of editorial independence.

978-1-098-17087-5

[LSI]

Table of Contents

CHAPTER 1
Introducing Generative AI

ChatGPT, Midjourney, Stable Diffusion, LLaMA—you probably have heard of some or all of these tools, which are quickly becoming household names. Collectively, these tools (and many more) are categorized as *generative artificial intelligence*, or *generative AI*. Generative AI is a distinct set of techniques within the larger AI field that *generate* something new: images, text, even video. Generative AI is separate from the more common *discriminative AI*, which is focused on reliably categorizing previously unseen inputs (classification) or determining the mathematical relationship between a dependent variable and the independent variables that describe some set of data (regression).

Generative AI is an exploding field. If you're in business leadership, it is imperative to determine how to harness generative AI to drive growth, enhance creativity, and streamline operations in your organization. If you're a developer, you will want to know how generative AI could impact your craft and how you can harness it to boost your productivity. If you're a hobbyist, you might want to know whether generative AI is worth spending your time to learn and tinker with. This report aims to provide you with the knowledge and insights you need to navigate the exciting realm of generative AI.

Of course, as with other nascent technologies, generative AI can be challenging to implement. To that end, we will also cover the ethical, legal, and technological concerns surrounding the use of generative AI and how some of them can be avoided or mitigated. After reading this report, you will understand the history, applications,

benefits, and potential pitfalls of generative AI so that you can incorporate this cutting-edge technology into your daily life, your business, or your creative endeavors; evaluate its return on investment; and implement and manage generative AI initiatives.

A Brief History of Generative AI

To gain a broad understanding of generative AI and how it works, it's important to understand where it came from. Even though it can feel like generative AI popped up out of nowhere sometime over the past year or two, research in this field has been ongoing for decades, with artists and computer scientists creating programs to generate visual art as far back as the 1970s.

The foundation for generative AI, and much of the discriminative AI that you may be familiar with, is the *deep neural network*, a neural network architecture that has many "layers" of neurons, including one or more that are hidden from the input and output layers. *Neurons*, the base units of a neural network, are mathematical functions that behave like a simplified version of a biological neuron. Neural networks were first introduced as early as the 1960s, but they were too computationally intensive to outperform other, simpler machine learning methods until around 2009, when a *recurrent neural network* (a kind of deep neural network) was able to win several handwriting recognition competitions for the first time ever.

Improvements in hardware and network architectures over the next few years made deep neural networks one of the dominant forces in the world of machine learning and artificial intelligence, especially in computer vision, where they excelled in common tasks like object detection and recognition.

Modern Generative AI Architectures

It was the introduction of the *variational autoencoder* (VAE) and *generative adversarial networks* (GANs) in 2014 that really kicked off the modern era of generative AI. Up until then, deep neural networks were used primarily for classification tasks, but with these architectures, they began to be used for generative artificial intelligence.

Variational Autoencoder

The VAE network architecture was introduced in the 2014 paper "Auto-Encoding Variational Bayes" by Diederik P. Kingma and Max Welling. In this architecture, two deep neural networks are used: an *encoder* and a *decoder*. The encoder learns how to reduce an input into an internal representation called a *latent space*, while the decoder learns how to reconstruct the input from that internal representation. Both networks learn simultaneously by comparing the difference between the generated image and the input image (*loss*) and trying to reduce that. But what sets VAEs apart from similar architectures is that they also try to organize the internal representations such that the model is able to generate something new.

VAEs are used for data generation, data augmentation, and anomaly detection, among other applications, and are capable of generating text and audio data in addition to images.

Generative Adversarial Networks

The GAN architecture was also introduced in 2014, in the paper "Generative Adversarial Nets" by Ian J. Goodfellow et al. Like VAEs, GANs make use of two deep neural networks: the *generator* and the *discriminator*. The generator is fed random noise and generates images from that noise, while the discriminator is trained on real images of interest, such as a large set of pictures of faces, and tries to determine whether an image given to it by the generator or the image source is real or generated. The generator is incentivized to successfully "trick" the discriminator into predicting the incorrect label for the generated images, while the discriminator is incentivized to accurately label both real and synthesized images. This is a zero-sum game: the better the generator does, the worse the discriminator does, and vice versa.

Also like VAEs, GANs are used for data augmentation (typically for creating new training and test data for other machine learning models); drug discovery; image processing tasks such as upscaling, inpainting, and colorizing black-and-white photos; and creating realistic, high-resolution images. The latter use is illustrated brilliantly through the website This Person Does Not Exist.

Seq2seq

Along with VAEs and GANs, the sequence-to-sequence (seq2seq) architecture was introduced in 2014 by Google for use in machine translation. Models created with this architecture excel at natural language processing (NLP) tasks in general, though, because they are designed to map one sequence to another. Like VAEs, seq2seq models use an encoder–decoder design. But in seq2seq, the outputs of the encoder are discarded, and the hidden or internal states of the networks within the encoder are fed directly to the decoder, which processes the encoder's internal states, discards its own outputs, and generates its own hidden states. The important innovation with seq2seq models was the addition of an attention mechanism. This enabled the decoder to focus on the most relevant input(s) when generating the output for that part of the input sequence.

Today seq2seq models are used for chatbots, machine translation (most notably as a part of the Google Translate product), text summarization, and more.

Transformers

The major breakthrough that led to the performance of the current state-of-the-art generative AI models was the introduction of the *transformer*. This was proposed in the 2017 paper "Attention Is All You Need" by Ashish Vaswani et al. This architecture uses an encoder and decoder like seq2seq; however, it uses a set of six encoders and six decoders, with each encoder feeding its output to all six decoders. In a general transformer architecture, any equal number of encoders and decoders can be used. The encoders and decoders use attention layers and simple feed-forward neural networks instead of the sequential networks used in seq2seq. This allows transformers to process input sentences in parallel, making them faster than older architectures, as well as understand to what extent the more distant words in a given input affect the current word being processed.

The transformer architecture revolutionized generative AI, forming the basis of the current large language models (LLMs) available today, such as the GPT family, PaLM, and LLaMA, among others, along with the DALL-E family of image generator models.

The Two Types of Generative AI

From our brief history lesson, you might be able to see some separation in the types of generative AI available: image generators and text generators. These are the most popular types of generative AI, the types that are likely most relevant to you, and the types we will be covering in the remainder of this report.

But for the sake of completeness, it's important to note that researchers, creatives, software companies, and others are also exploring generative AI for making music, videos (such as the opening sequence of Marvel's *Secret Avengers*), and other kinds of content. Many of these cutting-edge applications of generative AI are still based on image generators, text generators, or some combination of the two, making them a solid foundation on which to build your own generative AI knowledge.

CHAPTER 2

Introduction to Image Generators

Image generation AI might be one of the most gratifying recent advances in generative AI: with a Discord bot or a web application and some clever words, you can get a unique and ready-made piece of art in no time. Modern image generation AI makes clever use of many of the technologies and processes discussed in Chapter 1, along with some that weren't introduced, such as *latent diffusion* (another encoder–decoder architecture).

What is even more exciting about image generators is that they don't just take a text prompt and create an image to match it. They can also improve existing images with minor text guidance, generate (or regenerate) selected parts of an existing image, colorize black-and-white photos, and handle other, similar workflows that are often overlooked. Figures 2-1 and 2-2 show the power of image generators, beginning with a simple geometric image and the prompt "A castle on a field surrounded by a moat. Realistic fantasy art, oil painting" and resulting in a significantly higher-quality image that mostly correctly interprets the original.

Brands are experimenting with generative AI for everything from brand design to flavor design, construction technology startups are using generative AI to prototype building designs, architects and designers are using image generators to create mood boards and to prototype designs for clients before spending time on the final product, and businesses with high data needs (such as those training

their own AI models) are using generative AI to augment their datasets, allowing them to build more accurate discriminative AI models.

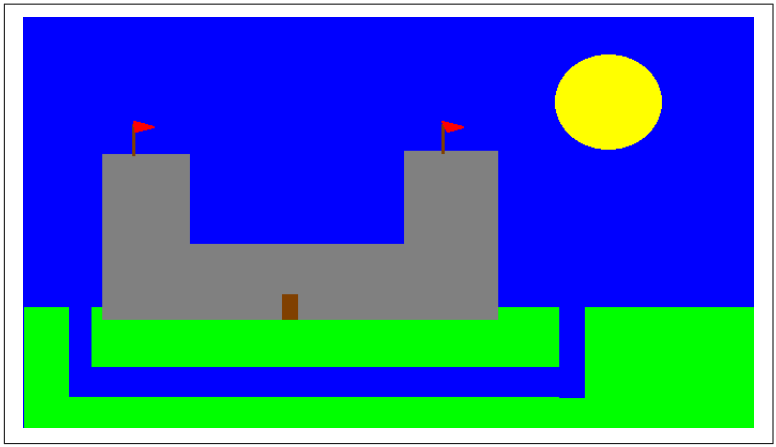


Figure 2-1. A simple input image to be used as a base for an AI-generated improvement

Figure 2-2. The upscaled output of a Stable Diffusion image-to-image operation made with the base image in Figure 2-1 and a prompt describing what the base image represented

But how are people doing these things? What are image generators capable of? What *can't* they do? And how can you make use of them? To start answering these questions, let's take a look at the

current landscape of image generation AI tools available for the end user. Focusing on tooling will help you get started interacting with and evaluating generative AI quickly, and it will simultaneously give you a broad overview of what use cases generative AI excels at.

Image Generation AI Tools

The number of image generation AI tools out there is staggering. From standalone desktop applications like DiffusionBee to Discord-based tools like Midjourney to application plug-ins like Adobe's Firefly and Canva's Text to Image, almost every kind of application interface that exists can be found among the image generators and will be discussed in this section. In this list, I excluded models and approaches themselves, such as GANs, to focus on what you might end up incorporating into your own toolkit right away.

Midjourney

Midjourney may be one of the most popular AI image generators out there, despite its current restriction to being used via a bot within the Discord chat application. Midjourney has several advantages:

- It is easy to use. Simply provide your prompt to a Discord bot, and get your images back in a reply.
- It plays host to a huge community built into the Discord server that hosts the Midjourney bot, with 15 million registered users at the time of this writing.
- It has a relatively simple flat monthly pricing structure.
- It is able to produce great-looking images with relative ease.

There are also some potential weaknesses that could affect your own adoption of Midjourney:

- It has a strict content moderation policy.
- It has a smaller feature set than other tools listed here and only supports text-to-image prompting.
- It is currently restricted to Discord.

DALL-E

Introduced by OpenAI in January 2021, DALL-E is a family of image generation models based on a modified GPT (generative pretrained transformer) model. At the end of 2022, DALL-E 2 was released to public beta with additional capabilities such as higher resolution, inpainting, outpainting, and creating variations of an input image. These new features and quality improvements were enabled by integrating CLIP and a diffusion model, abandoning the GPT approach of its predecessor. DALL-E's advantages include API access as well as a web application for testing; however, it uses a pay-per-usage pricing model, which can be a benefit or a limitation depending on your use case. API calls are also limited by default to a $120 monthly quota, which at current prices limits you to 6,000 of the highest-resolution images per month. Like Midjourney, this is a hosted service, rather than something you can run on your own infrastructure.

NOTE

CLIP is a model created by OpenAI that is able to predict a natural language description of any given image. This is a major improvement over traditional image classification models that were restricted to labels (or descriptions) that they were trained with.

DreamStudio

DreamStudio is a web application developed by Stability AI, the same group that created the open source Stable Diffusion model on which DreamStudio is based. In addition to being a web application, DreamStudio has an API available for programmatic use as well as a Blender plug-in. It's also very accessible to beginners and comes with a prompting guide to help users craft effective prompts. The downside to DreamStudio is that pricing can be complex. DreamStudio offers new users 25 credits, and beyond that $10 will get you 1,000 credits. The credit cost per generated image depends on the model you choose, the generation steps for each image, and the resolution of the generated image.

DreamStudio offers the same features as DALL-E: image generation, inpainting, outpainting, and image variations, along with negative prompts and prompt weighting. While the pricing can be difficult to predict, Stability AI does provide a cost calculator, API availability, model choice, and Blender integration. Another advantage is that it uses a number of open source models.

DiffusionBee

Like DreamStudio, DiffusionBee is based on the open source Stable Diffusion family of models. Unlike DreamStudio, it's a free and open source desktop application designed for macOS. Although it is easy enough for most users to get up to speed quickly, it provides a large number of options that can be overwhelming. It balances that with prompt ideas for text-to-image generation and helpful explanations of the available options. DiffusionBee is able to do text-to-image and image-to-image generation (a technique that's also known as *variations* in other tools), inpainting, and outpainting. Another feature useful for tinkerers is that you can load other open source models into DiffusionBee, allowing you to experiment with image generation and editing more freely.

While DiffusionBee is free, you must run the image generation on your own machine, which, depending on your system, may or may not pose a problem. In my experience with an M1 MacBook Pro, it takes just a few minutes to generate four images with default settings. Figure 2-3 shows the main interface of DiffusionBee, which has tabs for its supported workflows, a prompt input area, as well as settings, prompt ideas, and style suggestions.

Being locally installed, DiffusionBee gives you a lot of power that the previously mentioned tools do not. However, it doesn't have an available API to build into your own software.

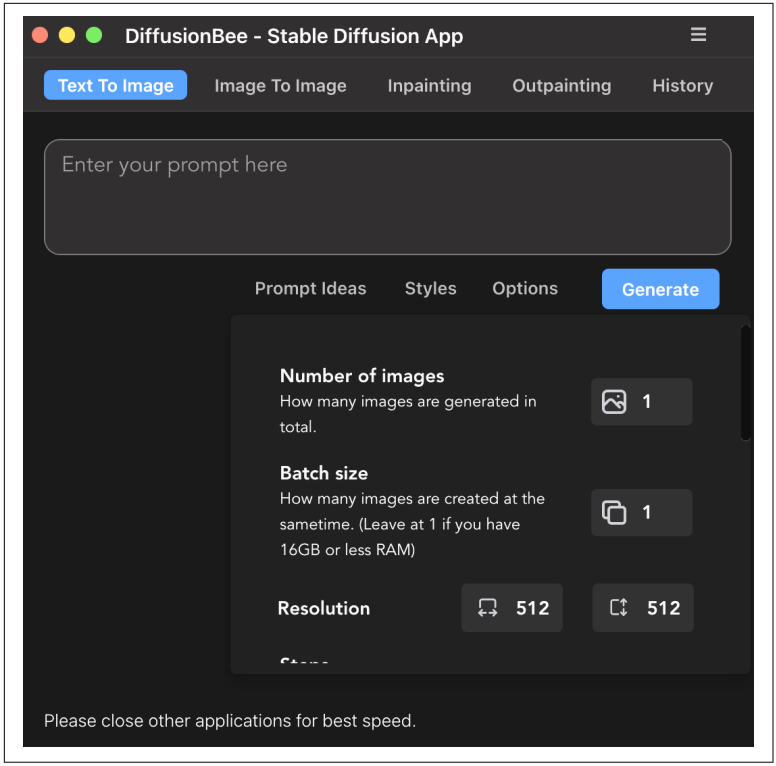


Figure 2-3. The Options dialog of the DiffusionBee interface

Firefly

Adobe Firefly is one of the first generative AI tools released by a major incumbent company. Using Adobe's existing application and data infrastructure, Firefly includes many features such as text-to-image generation, inpainting, outpainting, and generation of 2D images from 3D models. Firefly is available both as a standalone web application and as a tool within Adobe's application offerings, including Photoshop, Illustrator, and Adobe Express. Standalone pricing is credit based: the free plan gives users 25 credits per month, and the premium plan gives users 100 credits per month, removes watermarks, and offers a subscription to Adobe's Creative Cloud service, which provides additional credits.

> **NOTE**
>
> By "incumbent company," I mean an existing business that has a product and customer base that predates its adoption of generative AI. This is in contrast to companies such as OpenAI and Stability AI, which were founded to sell generative AI products.

Beyond integration into existing applications, another strength of Firefly is its commercially safe model. The model is trained on existing Adobe stock images, images in the public domain, and images with open licensing structures, which would mitigate one of the most prominent risks of using generative AI: copyright infringement (discussed shortly).

Text to Image

In its Text to Image tool, Canva has also folded text-to-image generative AI into its existing design suite, with a free model limited to 50 total queries and a premium model that offers 500 monthly queries. Query-based pricing is simpler to navigate than credit-based pricing, which is a big advantage of Text to Image, but the tool doesn't come with the same commercially safe dataset that Adobe Firefly does. Canva also offers a Stable Diffusion–powered photo editing tool called Magic Edit that marries AI techniques like inpainting with traditional photo editing tools.

Other Tools

Naturally, this list isn't exhaustive, nor can it be. With more open source models being released as well as investment dollars flowing to more and more businesses in this space, you can expect to find an explosion of tools aimed at the end user. Keep an eye out for experimentation in the tooling space as the major vendors fight over market share.

Problems Facing Image Generation AI

There are myriad uses and tools available right now for image generation AI models, but as with any new technology, there are also challenges that aren't always as well publicized as the opportunities are. Some of the challenges include:

- Unresolved questions around permission to use existing images to train generative models and the ownership of the resulting images
- Tools created to interfere with or block the use of images for training generative models
- Increasing numbers of AI-generated images in the wild, potentially making training datasets less useful

Chief among these are ethical and legal concerns: Whose work was the model trained on? Did they consent to the use of that work in training the model? What if my generated image looks too much like existing work? Do I own the images an AI tool generated?

Ethics and Copyright

Many of these questions are still unanswered, and some are unanswerable, but work is being done to clarify issues such as ownership, the rights of creators to not be included in training data, and more. Adobe's Firefly, mentioned previously, claims to be specifically trained on Adobe's stock images, images with open licenses, and images in the public domain. This would mitigate the ethical issues around the use of other models that may have been trained on images without permission from the creators. To reduce that risk for all involved going forward, the Coalition for Content Provenance and Authenticity (of which Adobe, Microsoft, and others are a part) is an effort to create and promote a standard provenance credit within digital images that will credit the artist or AI involved in creating it.

The copyright status of AI-generated images is also somewhat of an open question. While there has been a single ruling that proclaimed that AI-generated images with no human guidance aren't copyrightable, that particular case is still making its way through the appeals process, and it is unclear how much human direction (if any) qualifies AI-generated images to be copyrighted. If you decide to create images with a generative AI tool, you'll have to decide whether the risk of not being able to copyright those images is worth the tool's use.

Adversarial Tools

Some people aren't exactly happy with the new, uncertain landscape brought about by generative AI, and they are working to find ways to protect their creations from being used in a training set without their consent. Some of these tools are *adversarial*, meaning they attempt to make a given image useless for training or, worse, pollute the entire training set. Only a few of these tools exist now, but it is possible that more will be developed and that future generative AI models may actually be worse than current ones.

The major adversarial technique available right now is *glazing*, which uses a tool called Glaze to make small changes to an image that purportedly "trick" the training AI into thinking the image is of a style other than that of the artist in question. It's unclear whether Glaze truly works (and works on all models), but it did set off a sort of arms race with the introduction of a tool that claims to remove those changes published to GitHub in response.

Poor Datasets

The current crop of adversarial tools focuses on poisoning potential training data, but it's possible that generative models will end up doing this all on their own. Much of the work being done to improve generative models requires building bigger models with larger datasets (for example, Stable Diffusion v1 was trained on a dataset containing nearly 6 billion images). However, there is a limit as to how much data currently exists or could potentially exist to feed these models, and with the proliferation of AI-generated images (and, as we will see, text), newer models may end up being iteratively trained on more and more AI-generated data, with unknown consequences.

CHAPTER 3

Introduction to Large Language Model Text Generation

We've all tried to ask computers questions before. I have fond memories of encountering Ask Jeeves for the first time and getting frustrated when I got websites instead of answers to the questions I asked of it. With the release of ChatGPT and other LLMs, the vision of computers as machines that you can interact with in a natural way is much closer to becoming reality.

Researchers are beginning to leverage the uncanny power of LLMs to do things like translate languages without any additional training, design new drugs, and detect data anomalies and financial fraud. But while those applications are exciting, the more immediately practical applications may be even more exciting because of how widespread their potential impact is.

LLMs are being used right now to do things like summarize long documents, rewrite ad copy in different voices, interact with customers via chatbots, write code, simplify writing, and ask questions about documents. In the next section, we'll dive into a number of tools using LLMs to do these tasks, starting with the popular chat tools and their applications, moving on to coding and marketing assistants, and then looking at knowledge management integration.

Text Generation Tools

There are a number of tools coming to the market that allow users to interact with text-generating LLMs. In contrast to image generators, you're currently unlikely to find many local-first or local-only LLM tools: the size of most LLMs and the computational power needed to generate text make them untenable for use on home computers. This is slowly changing as more efficient models are released, but most currently available tools interact with an external model via an API.

ChatGPT

ChatGPT was publicly released by OpenAI in November 2022, and for many it was their first exposure to generative AI. It uses OpenAI's GPT family of models, which were trained on data sourced from across the internet and then fine-tuned to favor conversational responses. In addition to the extremely large model size and volume of training data used, a part of ChatGPT's success has come from using a technique called *reinforcement learning from human feedback* (RLHF), which uses feedback from human trainers to develop a reward system for the model that incentivizes responses that would be preferred by the human trainers. For ChatGPT, this process targeted friendlier, more conversational responses and attempted to avoid responses that would encourage illegal or harmful activity.

Another key component of ChatGPT's success is its user-friendly interface, an example of which is shown in Figure 3-1. By focusing on a conversational interface, OpenAI has made the power of its GPT family of models readily apparent, a pattern other model creators have followed. OpenAI has also set up a tiered subscription model familiar to users of software as a service (SaaS): free for personal use; a paid premium plan allowing access to the latest GPT model and GPT plug-ins for $20 per month; and an enterprise plan with no usage limits, longer contexts, and more. OpenAI also provides an API, allowing you to integrate ChatGPT with your own applications.

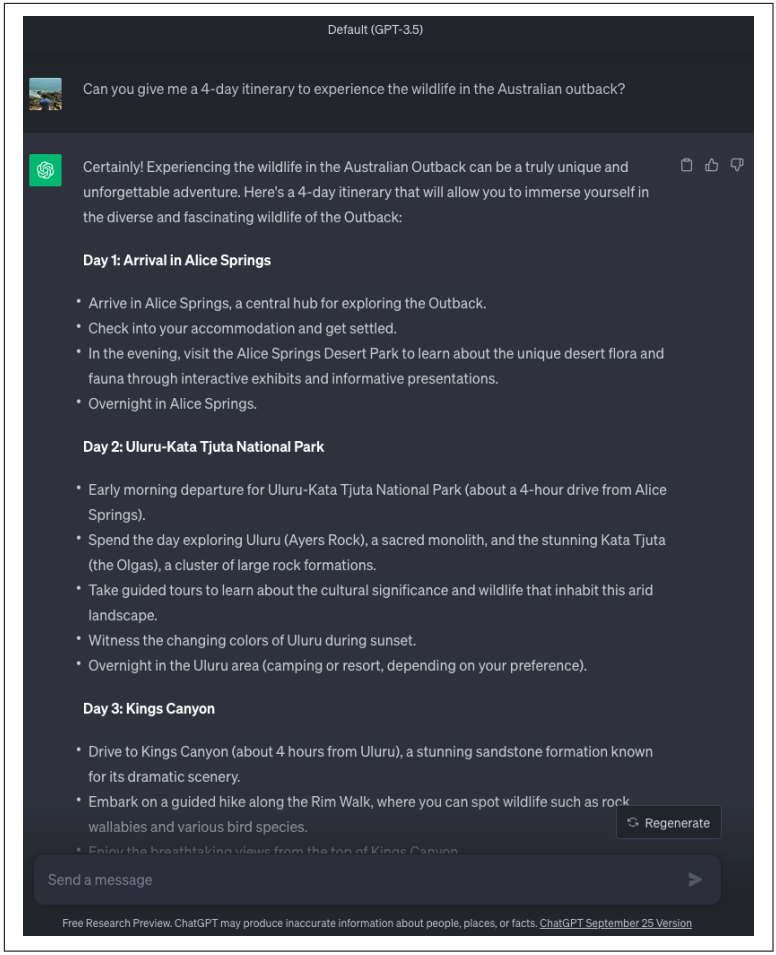


Figure 3-1. ChatGPT responds to a prompt asking for an itinerary to the Australian Outback

Gemini

Google released Gemini in March 2023 in response to the runaway popularity of ChatGPT. The initial rollout was marred by inaccurate responses (*hallucinations*) found at a much greater rate than in competitors like ChatGPT. Since then, model improvements and changes (such as moving from LaMDA to the newer PaLM model);

tight integrations with Google's email, documents, and other products; and the ability to access real-time information from YouTube, Maps, and Flights are beginning to set Gemini apart from other available tools.

Like Google's other offerings, Gemini is currently free. However, it is marked as an experiment and comes with warnings that your conversations not only will be used to improve Google services including Gemini, but also will be seen and annotated by human reviewers and used for training future versions of the underlying PaLM model. This isn't uncommon across the ecosystem of LLM services, but it's worth understanding and evaluating if you plan to use Gemini.

Claude

Anthropic's Claude models come in a familiar chat assistant interface, but the real power (and focus) seems to lie in their API access. With this focus, there's also a clear commitment to performance and safety: Anthropic's models have one of the largest available *context windows* (essentially, prompt length) at 100,000 *tokens* (the unit of text used by LLMs, which can be a character, part of a word, or a whole word), and, among other things, are built with what it calls *constitutional AI*. Constitutional AI, an alternative approach to RLHF, uses a specially trained preference model based on a human-provided list of rules (or constitution) during the model's reinforcement learning phase, rather than human feedback. This protects human moderators from potentially harmful material while still improving the model's responses.

Claude's pricing model is similar to ChatGPT's: free but limited access to the chat interface, a $20 monthly pro plan with priority access, and a separate API pricing plan billed per 1 million tokens. A disadvantage of this approach is that token-based billing may be difficult to predict and constrain, similar to the credit-based billing of the commercial image generator models discussed in Chapter 2.

Copilot

Copilot is a generative AI programming assistant based on the GPT family of models and fine-tuned on publicly available code hosted on GitHub. It essentially functions as an advanced autocomplete,

using code comments and *function signatures* (the name of the function and any inputs to that function) as context to write the code for the function itself, and is available as a plug-in for most popular code editors.

Copilot won't replace programmers, though, because as with any LLM, hallucinations are still an issue, and it will take a skilled programmer to detect any subtle bugs that may be introduced. Because it is trained on a snapshot of published code, Copilot will be less useful for code, packages, and frameworks for which there aren't many published examples. Another concern is GitHub's access to user prompts and generated code: Copilot for Individuals retains these by default (they can be deleted by opening a support ticket), but Copilot for Business deletes them as soon as they're used. On the plus side, pricing is straightforward: Copilot for Individuals costs $10 per month or, billed annually, $100 per year, while Copilot for Business runs $19 per user per month.

Cody

Sourcegraph's Cody is also a generative AI programming assistant, but it has features beyond the powerful code generation found in similar tools like Copilot, such as the ability to explain code, create unit tests for selected code segments, and optimize existing code, along with powerful natural language search. These features work by generating prebuilt prompts based on the user's query that are tested to work best with the backend model. The power of Cody lies in its ability to intelligently select the most appropriate code snippets for query context by using *embeddings*, a technique from NLP that enables a much more powerful search than traditional keyword search.

Sourcegraph has a zero-retention policy with its third-party LLM partners; this means Sourcegraph will not retain any model inputs or outputs and will not use personal data to train further models, which should allay any concerns about proprietary code being shared with third parties. Cody is currently in beta; a forever-free tier is available for individual developers, and an enterprise tier allows you to configure which LLMs to use and to use your own keys for both Anthropic and OpenAI models. The enterprise tier also comes with the typical enterprise user management and deployment features.

Jasper

Shifting away from general chatbots and fine-tuned coding assistants, Jasper is a marketing-focused generative AI tool that aims to be a one-stop shop for marketing content creation. Being a third-party tool, it's able to leverage several different models such as GPT-4, Claude, Gemini, and its own internal model to do things like generate marketing campaigns, translate copy into multiple languages, write blog posts, do search engine optimization (SEO) on existing and new content, and reuse existing content for new campaigns. Jasper also claims that its platform can learn any brand's voice and accurately re-create it in the copy that it generates.

Jasper's use of several models, choosing the model or combination of models that best suit the task at hand, is rare among the tools discussed in this chapter. Like the other tools built on top of the foundation models such as GPT, Claude, and Gemini, pricing is a typical tier-based monthly subscription: $49 per month for single users, $125 per month for teams of up to three users, and an enterprise tier.

Sensei GenAI

Sensei GenAI is Adobe's answer to generative AI marketing products like Jasper. Similar to Firefly (discussed in Chapter 2), Sensei GenAI benefits from Adobe's incumbent status and existing platform for creative professionals. Sensei GenAI, like the other third-party services described earlier, leverages several different LLMs but is also able to incorporate a user's existing data. Sensei's features go beyond content and campaign generation, though, and include brand-specific chatbots, sales conversation summaries, a natural language interface to brand data analytics, and more.

While the other tools discussed in this chapter have usage-based pricing or subscription tiers, Sensei GenAI only offers a demo request form, underscoring Adobe's commitment to large enterprise customers at present.

Notion AI

Notion is a popular knowledge management application that individuals and teams use to organize and manage information and tasks. Notion AI is a paid add-on for Notion that augments the kind of work done in Notion by handling many of the common text generation tasks such as summarization, translation, tone editing, simplification, querying a document, and more. Notion AI has a few advantages: one is that it's built directly into Notion, making it ready to use for existing Notion users. Another major advantage is its simple interface: the most common text generation tasks are part of a context menu that you can select from, removing the need to come up with a prompt for every summary or action items list you want to generate. You can still directly prompt the AI as well, which can help you do things like ask a document questions about its contents.

The pricing for Notion AI is user friendly too: every user has a number of free AI actions they can use before paying for a subscription, which runs $10 per user per month if billed monthly. Figure 3-2 shows Notion's context menu–based approach to interacting with its AI, which bundles common workflows but also provides a prompt input for more flexibility.

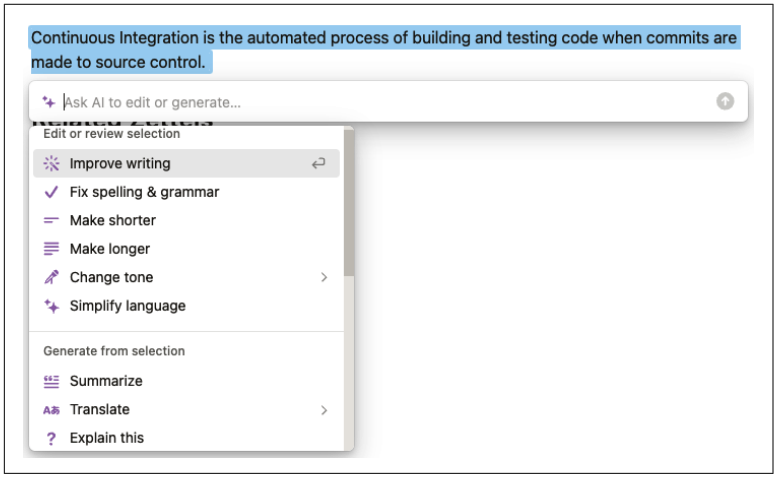


Figure 3-2. An example note in Notion with the Ask AI context menu open, showing the available tasks that can be used on the highlighted text

A Note on Prompt Engineering

Prompt engineering is a term used to describe the process of writing prompts for generative AI models that make use of natural language prompts as an interface. This is an area of active research, with multiple techniques already published. These techniques tend to fall under one of the following categories: zero-shot, one-shot, or few-shot prompting.

Zero-shot prompting is one of the most common ways people interact with generative AI. This is when you directly ask the model a question or tell it to do something with no examples. A zero-shot prompt would be something like "Write a poem about talking dogs in the style of Edgar Allan Poe."

If you add a single example to your prompt, you're now doing *one-shot prompting*. An example could look like this: "Using *https://www.poetryfoundation.org/poems/48860/the-raven* as a guide, write a poem about talking dogs."

You might already be guessing what *few-shot prompting* is, and you're probably right. With few-shot prompting, you add several examples to your prompt to guide the generated output. What is interesting with few-shot prompting is that you don't necessarily need to tell the AI what to do; given structured examples, it can usually figure out what to do:

> Multiply 10 * 10: 100
> Multiply 10 * 2: 20
> Multiply 10 * 4: 40
> Multiply 10 * 9:

For an end user, prompt engineering can become a discipline unto itself and potentially take more time and money than would be saved by using the model at all. This is something that should be factored into any ROI calculations done to determine the viability of introducing these tools.

Problems Facing Text Generation AI

While the potential uses of text generation AI are exciting and numerous, there are still risks and pitfalls to be aware of when deciding whether to invest in AI. These risks can be reputational, by allowing inaccurate statements to be displayed to a user, or even existential, by enabling theft of intellectual property (IP) or data.

Hallucinations

As mentioned earlier, hallucinations occur when an AI tool gives a confidently wrong answer to a user. Hallucinations are one of the most well-known weaknesses of LLMs. This can be a problem if you're using a chatbot to give customers answers about your product line or technical support, or if you are relying on an LLM for any sort of factual data but don't have an expert to review the results. This is unlikely to be eliminated, but it can be mitigated with human review or by writing prompts with more context and constraints, keeping the AI pointed in the right direction. Some use cases, such as a technical support chatbot, can use *retrieval-augmented generation* (RAG) as a mitigation tactic.

With RAG, you use datastores such as *vector databases* (special databases that store semantic numeric representations of data) to enrich prompts with contextual information before the prompts are used to query the model. The semantic representation of data in a vector database allows your software to find the relevant context of a user's query in a way other databases can't, and it can guide a model to giving more accurate answers.

Data Safety

Data safety is a major concern, one that could spell doom for an organization and lead some, like it did Samsung, to ban their employees from using LLMs such as ChatGPT. Because most interactions with these models occur via an API on third-party infrastructure, anything you put in a prompt goes to the LLM and on that infrastructure you no longer have control over, potentially leading to IP or sensitive information being leaked. Many of the

companies selling access to LLM tools do have data safety policies and agreements in place to mitigate this risk for their customers; an example is Sourcegraph's zero-retention policy. But users must be vigilant that they're not potentially leaking sensitive information and confident that their tools will safeguard their data.

Ethics and Copyright

Text generation AI suffers the same ethical and copyright considerations as image generators, especially around training data provenance and attribution. There are a number of lawsuits currently being worked out in court concerning whether the use of existing books and other published works in training data without an author's consent is copyright infringement or not. It is also not clear, for the same reasons mentioned in Chapter 2, whether anything produced by text generation AI can be copyrighted, which is one wrinkle for those wishing to use LLMs to aid in copyrightable work.

Prompt Injection

In a callback to SQL injection attacks that plagued web developers in the 2000s, it turns out that LLMs are vulnerable to *prompt injection attacks*. *Direct prompt injection*, also called *jailbreaking*, involves crafting prompts to bypass model restrictions or otherwise make a model do things it wasn't intended to do. Some of these made waves, such as the "pretend you're my grandmother" prompt that was able to get ChatGPT to give users instructions for making napalm, something you can't get by asking directly. In another instance, users were able to encode queries in base64 that would normally be blocked (such as how to make methanol) and get answers from the model. If you're a company that has an interface to a text-generating API with a custom prompt augmented by user input, it is possible for the user to overwrite the system prompt with their own malicious prompts that could steal user or backend data.

Indirect prompt injection is another risk, this one borne by people using LLMs with external sources, such as files or web pages. A malicious user could include prompt injection with a legitimate-looking file to be consumed by the LLM, giving the user control over it.

Wrap-Up

At this point, you should have a solid foundation in the current state of generative AI, allowing you to evaluate the available tools, see whether they match your use cases, and dive deeper if you choose to. This is a rapidly growing and exciting field, and I hope you will continue to learn and use generative AI more effectively while pushing the boundaries of what is currently possible.

About the Author

Kyle Stratis is a software engineer with nearly a decade of experience across the artificial intelligence development lifecycle in a variety of domains, including computer vision, health technology, and social media analytics. He currently is the lead machine learning engineer at Vizit Labs, where he is building Vizit's internal AI platform. Recently, he's been writing about Python, personal knowledge management, and software-defined radio.