The Data Sessions: Generating High-Quality Data with Prompt Engineering for Data Scientists

Generating High-Quality Data with Prompt Engineering for Data Scientists

By now, you've probably already typed something ridiculous into ChatGPT or Bard or Claude 2 just to see how far you can stretch these conversational chatbots. Which is better, crunchy or creamy peanut butter? That's what I asked, and ChatGPT's nonanswer was actually impressive. It chose no side and rather started making a case for both, letting me know that this is a matter of taste and there's room in the world for both preferences. I continued this conversation much longer than I care to admit. At one point, I even asked it to pretend that it was arguing the issue before the United States Congress and was trying to eradicate crunchy peanut butter from the country. The case it made was pretty great and pretty entertaining, probably good for a congressional filibuster. And if any of you ever find yourself in that forum, I suggest that you make some entertaining and compelling arguments about the virtues of one peanut butter over the other with the help of GenAI. While it's all silly and it's fun to do, GenAI tools are actually way more useful than that, of course. Getting a specific desired output from these models, aside from just mere entertainment, requires some skill and a specific mindset. The capabilities embodied in the gamut of GenAI products to generate text, images, audio, video, datasets, and even full analysis is staggering, especially if you start using the paid-for versions of these products that come with even more impressive token limits and capabilities. There is a bit of an art to creating prompts, and if your prompts have characteristics for data science purposes, you can generate all sorts of useful data. Keep in mind this is all relatively new in the mainstream market, so there are always more than one way to do some of these tasks. Prompt engineering may not always give you what you need for every experiment, but I think here we can show that you can get a really good start generating high-quality data with what's out in the market today. There are definitely GenAI products out there that are managed services that take care of all the work for you. You could just upload a dataset, set some parameters, and get back a great generated synthetic dataset in return. We're going to focus a little bit on some of the conversational chatbots. The first type of data that we can generate easily is what many people call dummy data. This data is data that is completely fictional. It's extremely easy to generate with prompt engineering, and your outputs are used for things like skills practice, exercises, teaching, educational settings. This data doesn't have to be tied to real events. So let me show you using the paid version of GPT-4+. I'm going to enable advanced data analytics because I think it will do well for us to generate data. Many people suggest you should specify a role for ChatGPT. This is something like you're a data scientist. We can do that here, even though it may or may not make a difference, but you could also get more specific by adding in what you wish to accomplish. You're a data scientist who needs to synthesize a dataset of text-based answers to a question. It's still kind of vague, and we haven't really asked to do any work. We've just told it to assume a role and hinted at a task. Let's add some more specificity. I'm going to describe the question for which we're creating answers. That would sound something like the question is, what additional retail item are you most likely to buy together with shoelaces? Of course, this is hypothetical, but now we're pretending to be engaged in product market research and analysis just for the illustration today. Say we're trying to identify our next product to import for resale in the U.S. Details go a long way in prompting GenAI. I'm going to specify why I want this information. You're working on identifying what product your company should import next for resale in the U.S. that complements your existing line of shoelaces based on user responses. Now, I need to specify exactly what I need as an output. What type of data do I need? For illustrative purposes, let's just ask for some very, very specific data. The synthesized dataset should contain absolutely no PII, or personally identifiable information. It should contain 20 records. The fields should be user ID, response, time, date, location. The cities and states can be duplicated, but they need to be locations all over the United States. And there can be duplicate responses from customers. Then I give it a command, generate the new data. The response here will be a dataset that I could copy and use elsewhere. What if I already had a dataset though, and I wanted to synthesize additional data based on the dataset? Pretend that this set we just created was not actually fictional. I'm going to start a new chat, so we aren't using any of the previous context. And I'll modify the previous prompt with the introduction of an original dataset. I can modify the previous prompt to include something like using the uploaded dataset in CSV format as your original data, synthesize high-quality data of 100 additional customer responses. Make sure the statistical properties of the original data match the new data. Let's try it. Here's the prompt again. That's a really long specific prompt. GenAI chatbots give better and better results with clearer details, whether you put it in the original prompt or whether you continue the conversation asking it to be refined. I recognize this is a hypothetical situation, and there are some holes in the story here. But the anatomy of the prompt isn't that bad, and there are a lot of follow up instructions I could give GPT-4 in the conversation. After examining the output, I may ask GPT-4 to limit the responses to less records or provide even more records. I can even ask it to do an analysis for me and return graphs and recommendations to prove that it is the same statistical properties of the original dataset. I could then ask it to suggest clever product names for my new imported product. I could ask GPT-4 to explain the process it took to generate the data as if it were explaining it to my grandma. It's pretty fun to do this and to experiment. Yes, I do know there is a knowledge cutoff for ChatGPT that would affect the accuracy of my results, but all of this is just to illustrate how conversational the interaction can be in prompt engineering and how to include the specificity needed to generate a required output. Valuable data is typically information tied to a real event, item, or person. We often talk about data as something we collect, but where there are barriers to collecting original data in a useful quantity, what can be done? This is where synthetic data really comes in handy. So your company possesses real customer and user data that portrays user behavior or demographic information or private company performance data or even proprietary research data. There could be any number of valuable data points available. But let's say you, as a data scientist, wish to conduct experiments on this data. Why might you choose to use synthetic data rather than just the truly original consumer data? Can synthetic data be just as useful or helpful as the truly original data? First, let's quickly cover what synthetic data is before we answer that question. The simplest definition is that it is data generated by some artificial means other than an actual real event, person, or item, or phenomenon. Using machine learning and artificial intelligence, we can create datasets in a number of ways. We could manually prescribe a set of rules and then let machine learning models render labels for items such as that we would do in a computer vision task. This rule-based data creation is a common use case for AI. The rules you prescribe then allow a machine learning task to categorize or label items. And this works really well on simple tasks, and rule-based synthesis requires manual involvement each time the rules need to be changed. It also requires that you have a robust set of data. Generative AI can be used to generate code that labels your data. I searched on Google and there were about five examples of people generating Python code that would classify a statement by the sentiment detected therein. Creating labeling rules and code to run the labeling rules on a set of data is something you can employ generative AI for now. It can write that code. Generative AI can help us to synthesize data as well in situations that are far more complex. Lets look closely at generative AI and what is happening when we use it to synthesize data. We can train a model on an original dataset, which can then produce synthetic data that represents the same trends, shape, and statistical properties of the original data. For this type of synthetic data to be helpful, it isn't simply random and arbitrary though. It's not dummy data. It can't be dummy data. Nor is it exactly rule based. It's representative and simulated and authentic because it's the output of the model that was trained on the original dataset. The model can capture the complex relationships within the data and then generate similar data to the existing original data. So back to my original questions, can synthetic data be useful, and why would I choose to use generated synthetic data rather than original data from my company and its customers? There's a few reasons why. There may be strict privacy barriers to the open use or sharing of customer data inside and outside of your organization. There are even laws like the European Union's General Data Protection Regulation, or GDPR, that prohibit the use of data for purposes other than those for which a user has explicitly given consent. This is very common as a security and privacy issue for data scientists who wish to design meaningful experiments without breaking the law or misusing data. Synthetic data can be generated to replace the PII and still preserve the integrity and statistical properties of the original data. Using a synthetic dataset that mimics the original data can be a great alternative because this allows you to capture the complexity and the nuance and the relationships in the original data without breaking any of the laws that would put you in jail or get you in trouble or even just harm your customers trust in your company. Another situation could be that you may have imbalanced data with some populations being underrepresented entirely. Often with survey results or other voluntary data, biases can be overrepresented based on the populations of the respondents. You'll need to employ your expertise to make sure that you aren't departing from the meaning of the original data. But you might have a need to amplify a certain voice in your experiment with synthetic data to be more representative of the actual customer base or population, not just representative of those who responded, especially if you find that you have a skewed sample of respondents. Perhaps you're dealing with data that is from rare events. Depending upon how rare the event is, data may be pretty hard to round up. This might be some sort of a rare weather phenomenon or a rare chemical reaction or even occurrences of a rare disease. If you trying to use AI or ML to simulate a very rare adverse drug reaction, but the drug is brand new and you're seeking approval for it, you don't have a lot of data from testing it. So you might need to have some synthetic data to work with to model some of the possible results. Perhaps you have enough data to train a model, but not enough to run a broad experiment of a much larger scale. Generating synthetic data can easily augment your dataset without losing the properties that existed in the original data. Synthetic data can also come in handy when you're trying to test an algorithm or a model. It can be used in testing the performance of these without the time and cost of procuring original data as well. As I've detailed, there are a number of use cases for synthetic data that are meaningful. It is not a panacea. It will not solve all of your data-related issues. It works in some situations as long as you've got good synthetic data, but there are some limitations. Sure, you can create all sorts of random data to practice your skills on, but the only data that yields meaningful insights is data that's rooted in reality and based on original data. So you can't just manufacture data from thin air and have the same meaningful results. You really do need some original data to start with, and it needs to be of a high enough quality that you could train a model using it. Further, in regards to quantity, you'll have a hard time if you're only working with a couple 100 records. The more data you have, the more accuracy you can get. So all of this is great, but how exactly do you synthesize meaningful data that isn't just text based? So let's step away from prompt engineering for a minute, and let's look at a specific scenario to illustrate how you would do this with images. There are many possible industries that could benefit from synthesized data. But here I'd like to walk through how to create high-quality data for medical research. You wouldn't use ChatGPT, which is text based, nor would you really be able to use something like DALL-E or Midjourney because the scale of the experiment that I'm going to describe and the image-based data results. Imagine we're building an artificial intelligence workflow that can analyze medical images to detect a rare disease, but there's not enough real patient data to properly train our model. First, we collect a small dataset of real medical images to use. These images are our original dataset. Next, we need to find or build a generative model that can synthesize new images. Generative adversarial networks, or GANs, work well in this situation because they are really good at fine details like those in medical images and variability, which would be present when you have a full set of images. We would train, again, on our real dataset so it learns to generate realistic medical images. Remember to split your real dataset into a training set and a validation set. Only train the GAN on the training set, then test it on the validation set. This prevents overfitting of the model, which is when your statistical model fits so exactly to the original dataset that it can't accurately process any new unseen pieces of data, in this instance, a new image. There isn't just one GAN option either though, so depending on the work you're doing, you'll need to find the one that fits well for you. In our situation dealing with medical imagery, what I found by researching online was that we would do well to use a deep convolutional GAN, or a DCGAN, because they're known widely for efficiency and accuracy with image generation. Data is typically dirty though, so it takes some cleaning. What does that mean with images? Hypothetically, we probably are dealing with some x-rays, MRI scans, even maybe histology slides. Make sure they're formatted correctly for the GAN. This means some resizing and normalization might be necessary. So next, you load or create the GAN model in a machine learning framework like TensorFlow or PyTorch The model will consist of two neural networks, the generator and the discriminator. The generator takes in random noise or variables as inputs and outputs a synthetic image. The discriminator takes in the created image and outputs a probability of it being real, usually as a score. So we train the GAN by alternating between updating the generator and the discriminator. The generator tries to fool the discriminator, while the discriminator tries to detect real or fake images. Over many training iterations, the generator gets better at creating realistic medical images that can pass as real by the discriminator. And this is what the term adversarial means in generative adversarial networks. The two models are competing as if they are adversaries in a game. The key is starting with a small, but representative real image dataset. This gives the GAN a basis to learn from before it can generate new additional data. After training the GAN, we can sample from it to generate unlimited amounts of new synthesized images. But how do we know these images are of high quality? This is where the validation set comes in. We can use it to test if the GAN's outputs look real. If you're new to this, like we all have been at some point, don't tune out here. Metrics like the Frechet inception distance, or FID, can compare synthesized images to real ones. Basically, it compares statistical properties like means or averages and covariances between your original data and the synthesized data. For images, it reduces some of the characteristics of the image down to assigned numbers and then compares the difference between those number values of the synthesized image and an original image. A lower FID means that the synthesized data is very close to the original data. If your FID is too high, you have to adjust the GAN. We now have a GAN that can generate highly realistic medical images, but diversity is also important. We have to make sure the GAN outputs varied images, not just copies. So to promote diversity, we can feed the GAN random input vectors. If you're in TensorFlow, for instance, you can just dial the input vector from the program's presets. Essentially, this is a collection of numbers sampled from a simple distribution that promotes higher quality and varied results. There are a number of common practices if you have to generate your own input vectors, but this is potentially also a space in which you could use AI to generate some of these vectors. The last step is annotation. Use clinicians, real people like pathologists and physicians to add realistic labels like disease type and location to the synthesized images. This is a manual task. And for large synthesized datasets, you might need to have many clinicians to get work done at scale. Sure, it could be possible to go back to using a rule-based AI method to label the images, but it's risky because it's important to have some sort of human validation of the synthesized data at this point in the process. Generative AI can be a little bit unpredictable and inaccurate sometimes. These are known as hallucinations. Things like hands with six fingers or people with three eyes happen in generated images. However, the output of this annotation process, discarding the images that are unusable, is a high quality and complete dataset of medical images that we could use. We've kind of gone through one example of what synthesizing image-based data might look like in practice. And in this instance, we used medical images as the story in health care research. Other examples are prevalent all over the market. Amazon's teams that develop and maintain Alexa's artificial intelligence use synthetic data to improve Alexa's understanding of languages where they have less user input to benefit from. Waymo has been training self-driving cars with simulated driving data, and these driving observations otherwise would take an incredible amount of time and effort to collect. Facebook AI research published a paper titled Recipes for Building an Open Domain Chatbot, in 2021, describing how generated synthetic data helps train an AI chatbot. So right now, we're experiencing a massive AI boom in the market, and the sheer number of generative AI projects and products and experiments that have been released in the last 12 months is staggering. I see that as only increasing moving forward in the future, which is a good thing because everything will improve. A quick internet search for generative AI for data synthesis reveals products that will do the work for you. These are plug-and-play options that exist as SaaS solutions for your work. Some that come up in my search are Gretel.ai, MOSTLY.ai, and Tonic.ai. This is where we get back to prompt engineering. These products are designed to synthesize data and are specialized to do so. Going with a general or foundational model is like asking a jack of all trades to come and do some advanced carpentry on your house. Depending upon your use case, you might want to find a product specifically designed to accept prompts that do a limited thing. I am confident that the number of options will only grow and grow as this space is really exciting and new, and each product seems to improve upon the last. It's important to note that the use of synthetic data should be carefully validated to ensure that it accurately represents the characteristics of the real data. Synthetic data should not be used as a replacement for real data where data is readily available. Instead, it should be considered as a complementary tool to address limitations in data availability, in privacy regulations, and limitations in diversity. Hopefully, some of this comes in handy in your future experiments.