**Generative AI Foundations: Getting Started**

Introduction and Overview

Okay, there's a bunch of different ways I could begin a course about generative AI. I could begin with hype. Isn't technology wonderful? Generative AI can take care of everything now. But I could begin with criticism and tell you to beware of generative AI for it is biased. It is unreliable. It is wrong. It is deceptive. I could show you news headlines about all the jobs it's going to destroy, but I could show you other headlines about the jobs it's going to create. And okay, I'm going to talk a little bit about all of this because there are valid reasons for all the hype and for the criticism, reasons to be anxious and reasons to be optimistic about this technology. But whether you are personally enthusiastic or skeptical or somewhere in between, this is important. Generative AI is a new technology already having a massive impact on the world, and I guarantee you it is useful to understand it better, know how to use it, and just feel more comfortable with it. But this term generative AI describes a variety of different software tools and applications including ChatGPT, Google Bard, Midjourney, Anthropic Claude, DALL‑E, GitHub Copilot, Microsoft Copilot. There are dozens of others and new ones appearing all the time. Some of these are websites you go to; others are apps you install on your mobile phone or your desktop or laptop. We're seeing generative AI features being added into older existing applications like Microsoft Office and Adobe Photoshop, and other generative AI is designed to be integrated into your own custom software and applications. But the point of any generative AI is that you provide some instructions, often in the form of a prompt, which can be a simple sentence to describe what you want, and the generative AI then generates brand new content. And that's what the word generative means here, that when you ask that AI for something, it's not just finding some existing file and handing it over. It actually generates or creates brand new results. Many of them generate text, all kinds of text from emails to poems to packing lists, stories to presentations, summaries, business strategies, conference schedules, hiring plans. But generating text also means some of them can generate specialized kinds of text like computer code. ‑Write a Python function to convert Fahrenheit to Celsius. ‑Now other generative AI tools can generate images from abstract pencil sketches to icons, diagrams, all the way to photo‑realistic cinematic images. Gen AI can also generate audio. We can have new music, new sound effects, or we can have new generated voices. It can create slide decks for presentations. We can do video processing like I showed a moment ago to change the style of an existing video. We can also make video avatars and make them say our words instead of us. And okay, it might seem like all these tools must be incredibly different from each other. But the way that generative AI works is fundamentally the same, whether it's being used to generate text or generate images or generate code or anything else. And that's why here, we're not just going to talk about using one generative AI tool. This is how you can get better results from any of them. And that's not always obvious because working with generative AI is often counterintuitive. What I mean by that is sometimes you will give one of these AIs a complex request. You might even secretly think, oh, there's no way it's going to do this, but it does. It instantly gives you exactly what you wanted. So you think, well, if it was good at that complicated thing, it must be really good at this easy thing. And then you give it something simple to do, and it fails miserably. We're going to talk about why this happens. Along the way, we'll cover several important terms like LLM. We'll talk about prompt engineering, context windows, foundation models, hallucinations, and more. But this is a short course. So if you are looking for a deep technical dive into the details of the algorithms running behind the scenes, this is not that course. And unlike what some people would have you think, you do not need to be any kind of artificial intelligence expert to just use generative AI successfully. However, there are a few things that's worth knowing about how this works because that's how you can understand what it's good at and what it's not good at. So whether you just want to keep up with current technology, whether you want to understand which tools you should be looking at that would work the best for you or just because this is a fast moving area of technology, what you should expect in the future, welcome. Let's dive right in.

Your First Five Minutes with Generative AI

I expect many people watching this have already used generative AI. The problem is the way a lot of people are introduced to this is using trivial examples, like create a bedtime story, write a poem, plan a fishing trip. And when you begin that way, it can often be tough to make the leap into actually being productive, being useful with this. So even if you have used generative AI already, for a moment, let's pretend you hadn't and just kind of begin again. What should your first 5 minutes with this have looked like? Now I could demonstrate generative AI using many different tools. We are going to begin with ChatGPT because it's the most popular, most well known of the generative AI chatbots. ChatGPT is at the website chat.openai.com. You'll need to create an account if you haven't already. But let's just jump right in and start with a prompt. Write a one‑paragraph email to my boss asking for Friday off, and here's what ChatGPT comes back with. ‑I hope this message finds you well, am writing to formally request time off this coming Friday. ‑But watch what happens when I ask it to rewrite it to be less formal. ‑Hope you're doing great. I wanted to ask if it's okay for me to take this Friday off. ‑When you use a chatbot, it keeps track of context, meaning the history of chat, the relevant details that were mentioned earlier. So here, I just said rewrite it to be less formal. The AI knows that it means the email to the boss I just asked you to write. Now this is very different from say using a search engine where every interaction you have just begins again. Now I could keep this chat, this conversation going, but I'll start a new one for a different kind of question. Let's say I just got a message asking me to fill in for someone on a technical interview that's going to begin in 5 minutes. I need some prep. So, write some interview questions for an entry‑level web developer. Now an important point. If you're following along using ChatGPT or any other AI chatbot and you typed in exactly the same prompt I did, you should expect the results to be similar but different. If I entered the exact same prompt again and again and again, I would expect a different result every time. Generative AI does not produce identical responses. There's a level of randomness and unpredictability. Just as if you'd asked 10 people to do this, you would expect 10 slightly different results, it's the same here. But let's switch gears to another example. Let's say a colleague just sent me a link to a research paper. At the time, I'm recording this late 2023, if I just have a basic ChatGPT account, I can't upload a PDF. But if I have a basic account at Anthropic Claude, I can. So let's switch over to a different chatbot. It's a PDF, but I know this is publicly available on the web so I don't mind uploading it. And I tell Claude to summarize the main points in this research paper. And this is one of the things generative AI is great at, the idea of summarizing longer amounts of content. But let's say I looked at that summary and I saw there's something I'd like to know more about. So I'll refine it. Expand on what the paper says about .2. It can be a great productivity tool right from the start to digest long documents or long emails into concise insights. And here's the thing. Your prompts don't have to be perfect. Most AI chatbots are pretty forgiving about punctuation and grammar. This is not like programming languages where you have to write statements within an exact perfect syntax. I could have written these requests in many different ways. However, a little later, we're going to talk about the different kinds of things you'd say in your prompts to help guide the AI to better results. But when you're starting out with generative AI, do not get hung up on crafting the perfect prompt right away. There's a lot of people out there that tell you they have the perfect prompt for some business situation. They're usually trying to sell you something. The key is to have this attitude of continuous experimentation and revision. Treat the initial prompt as a starting point. Ask the AI to build on those ideas and rewrite its responses. Try different prompts from different angles. Tweak the prompts that don't work well, and sometimes just give up and start again. There is an art to prompting, and it develops through actively engaging with this. So rather than expect instant perfection, just embrace this exploration idea. See what works, see what doesn't, and improve along the way. So, we've covered helping write emails and then rephrasing those emails, creating interview questions and then refining those questions, asking for summaries, then expanding on those, and we're just starting to scratch the surface.

Generative AI That Isn't ChatGPT

The generative AI marketplace is evolving rapidly. There are new tools, new applications, and new businesses appearing all the time. ChatGPT often gets the most attention, but there are many other powerful gen AI tools out there. Now we can't possibly cover them all here, and we don't need to. But it is useful to have kind of an overview of the different areas of generative AI because that's the first question. If the point of any generative AI tool or application is that it generates something, well, what does it generate? Text, images, audio, video, something else? So let's start with text because even this has multiple different aspects to it. We have the chatbots designed for us to have these back‑and‑forth conversations. ChatGPT from OpenAI is the most well known, but there's also Google Bard, Anthropic Claude, Microsoft Bing. Inflection has the Pi chatbot. Quora has the Poe chatbot. Now, chatbots are broadly similar in the way you interact with them. If you know how to use one, you know how to use any of them. But they do have different intended use cases, different features, different capabilities. For example, at the time I'm recording this, the end of 2023, the current version of ChatGPT only knows about things that happened before April 2023 because the way it was built, and we'll talk more about this later, it has a knowledge cut‑off point. Now, a quick sidebar. This cut‑off point is periodically updated. So at the time you're watching, it may be more recent. Now how do you know? Well, it's a chatbot. Just ask it. What's your knowledge cut‑off point? Some chatbots are designed in such a way where if you ask them about something more recent, they can go out to the web and find up‑to‑the‑minute information. Now there are differences. For example, some chatbots allow you to upload documents and provide links; others don't. Now all of these features change and improve regularly, but they're chatbots. If it's not obvious what they can and can't do, ask them. But beyond that, their intended use case can be different. So ChatGPT is designed to be as general purpose as possible. You can ask it to write a bedtime story or create a poem or summarize a complex business document, write a business case. But Inflection's Pie chatbot, now that's designed more to be a personal companion rather than a productivity and research assistant. It's something you use for talking through personal decisions and act as a sounding board. But we also have text‑focused generative AI that are more specialized to a specific business area like marketing or content creation. Examples here would include Jasper and copy.ai. Now while they do generate text, they are much more structured. They provide templates and guidance around specific uses. And for software development and code generation, you can use many of the general purpose chatbots and just ask them to generate code or also kind of as an educational tool. You can paste in some code and ask them to explain it to you. But there are also specialized applications like GitHub Copilot, which integrates into a programmer's development environment to suggest code as they type. But another major category of these tools deal with images like DALL‑E, also from OpenAI. This is an AI designed to create images from any text prompt. Similar tools include Midjourney and Stable Diffusion. Like the chatbot, we use prompts here, but the prompts look a little different. There's sometimes less of a regular sentence and more of a list of qualities and characteristics you want from the new image. Because designers and artists often have their applications they've been using for years, we're also seeing generative AI features being added into popular existing design tools like Adobe Photoshop. Photoshop now has a tool called generative fill. I can type a description of what I prefer to see in this section of the image, and it will give me several different options for this. Now over to video. There are tools like RunwayML and Kaiber that can generate video from text prompts. Now right now, the video tools aren't at the point where you could just provide any prompt you want and get production‑ready results. Some of the results are a little strange, but there are some more specific targeted areas of video creation that it's useful for. One of the things we can do is work with video avatars. Organizations like HeyGen, Synthesia, and D‑ID all provide options where you can provide a script, and they will generate a virtual person speaking that script. There's also options to use generative AI for audio. We can have it generate music or sound effects or voices. Companies like ElevenLabs specialize in voices. I can choose from a selection of precreated ones. I could even clone an existing voice by providing some samples. ‑And unlike the computer voices of old, the delivery here is often very realistic. Now, a lot of the current generative AI products focus on one kind of thing. They only generate images or only video or only text, what's often called a single modality. And the reason I mention this word is we also have this growing idea of multimodal generative AI where the same AI can deal with more than one modality. It could deal with text and images and video. But as these tools develop, we're also seeing more and more useful combinations of them. For example, I can ask ChatGPT for an image. And while ChatGPT doesn't directly create the image, it could take my basic prompt and then pass those over to DALL‑E to actually create those images. So the power becomes in tying several of these together. Another example. Because we have generative AI that can do text translation from one language to other and we also have the ability to do audio voice cloning and we also have the ability to do video processing, if you combine these three things together, that allows us to have video translation where companies like HeyGen and Synthesia will take my video that I've recorded in English, translate it into other languages, then also clone my voice and process the video on my face to make it sound like and look like I'm actually speaking that foreign language. If you combine these three things together, (speaking in French). So first, think about the kind of output you need and whether you want a more general use or a more specialized use. And new generative AI tools are appearing all the time, and they often have a lot of hype around them. So do take the time to properly evaluate your options.

How Generative AI Works

Generative AI uses an approach called machine learning. And what that means is these systems all learn by example. And to do that, they need a lot of examples. So the companies developing any of these generative AIs have to begin by gathering massive, massive amounts of data to analyze and learn from, whether that's text or images or code. So, a text‑based generative AI is built by analyzing millions of text sources like web pages, books, articles, scientific papers; whereas an image‑based generative AI will have been trained on countless images, photographs, illustrations diagrams, and they need staggering amounts of data because they're trying to identify patterns and similarities in the data. And the combination of all the data that they're trained on, together with your suggestions or your prompts, allows them to generate that brand new content. Now, a quick sidebar. People often wonder what do you actually call this? I mean is it a program, is it an application, what is it? Well, we call it a model. The end result of this training process, which is very expensive and very time‑consuming, is a model. We say that we have trained the AI model on all of this data. There's even more specific terms. So when a generative AI is text‑based, meaning it's been trained on incredible amounts of text and language, we call it a large language model, or LLM. Now if the model was trained on some other kind of data like images or video, well, we're not going to call it a large language model because it isn't dealing with language. The term you'll often hear is foundation model. Training a generative AI model is very time‑consuming and extremely expensive. The large commercial generative AIs, the ones used by ChatGPT or Google Bard, they are estimated to cost tens of millions of dollars in computer costs alone just to train the model. They need these massive server farms and tons of equipment to do it, and it's one of the reasons why there is a cut‑off date with what ChatGPT knows about because that LLM took so long to train. It was trained up to a specific date. You can't just do it again the next day because it costs incredible amounts of money to train them. Now it's a common misconception to think that because these generative AIs are trained on millions of existing examples that they store all of that data they're trained on, but they don't. They try and recognize patterns in the original data, but they do not store all the original document. So a generative AI model is not like a database where you can just look something up. Here's what I mean. If I was a generative AI model, I've analyzed millions of documents including multiple documents that contain the phrase it was a dark and stormy night. Now I could recognize this pattern, recognize this occurrence of words one after the other. And after being trained, if you then ask the model, well, what comes after the words it was a dark and stormy, the model will say, well, night probably. I mean statistically, if somebody writes it was a dark and stormy, the next word is probably going to be night. However, if you now ask the model, okay, now tell me every single document that contained that phrase and the rest of the contents of those documents, it doesn't know because it didn't actually store all the documents it analyzed. It stored the results of the analysis. So while it's really good at recognizing patterns and making new predictions, it is not good at just looking something up. And while it might be an oversimplification to say that generative AI is like a big autocomplete system, there is an element of truth to that. Fundamentally, it is trying to figure out if you wrote these words or these sentences, what is the statistically most likely next word or sentence or paragraph? But also, because the model is trained on massive amounts of data, everything from scientific papers to novels, song lyrics, equipment manuals, the model doesn't actually care whether it's looking at the text for a washing machine manual or looking at the text for a fantasy novel. It's just text. And an LLM does not make a distinction between fiction and non fiction. It doesn't understand true and false. It is just trying to recognize patterns in the data. Now let's take another phrase like that was a massive waste of, and what comes next? Massive waste of time? Massive waste of money Massive waste of time and money? Massive waste of my time, your time, the company's time? All of these are possible, and they would all work as the next word or phrase after that beginning. But if you had previous context, let's say you also knew all the sentences prior to this one, it allows the predictions to be better and more accurate. And it's why a lot of LLMs will talk about having a larger and larger context length or context window, meaning what is the overall amount of things that they remember. What is in their working memory? What did they understand about this particular interaction? But here's the thing. This idea that there are often multiple possible answers is why generative AI can give you results where you understand why you got that result, but it's completely wrong.

When Generative AI Goes Wrong

Generative AI models are amazing at matching patterns and making new combinations, and they can often seem like they understand things, but that's a mistake. You see these AIs all do their best to generate new results that are convincing and believable. It doesn't mean that the results they generate are actually correct. There's even a term in generative AI called a hallucination where the AI basically makes up an answer. It might sound right and is often incredibly convincing. It's just not true. I've had multiple experiences using gen AI for research where it's made up a very realistic, convincing answer including citations. But if you actually looked up the publication mentioned in that citation, it doesn't exist. There's been news stories about attorneys using generative AI to build out their case files and it generating references to prior cases that never happened. And again, it's because of the way generative AI works. It recognizes patterns in data, and it's extremely good at using those patterns to generate statistically likely results. But just because an answer is statistically likely, it doesn't mean it's true. So while generative AI can give you amazing results, we're still at the point where you need to factcheck anything you get from them, particularly if you're using them for any kind of technical reference or technical explanation. Now sometimes it's a bit more useful to use the image tools to better demonstrate some of the issues with generative AI. So I'm going to give the same prompt to several of the image‑based generative AI tools, that an image I want is a glass of water with a lump of sodium in it. These are the results from DALL‑E, these are the results from Midjourney, and these are the results from Stable Diffusion. But those of you who remember your high school chemistry class may be itching to point out a piece of information, which is what happens if you drop a lump of sodium into a glass of water because the answer is it explodes. Pure sodium and water causes a massive exothermic reaction, something that is not being represented by any of these images. And that's because they're doing a great job of pattern matching. They're taking words like glass and water and sodium and combining them together. But most of the images they've seen with the word sodium were probably pictures of salt. I can understand why I got the results I did. It's still wrong because these generative AIs\ are not built with a fundamental understanding of the world. They don't understand consequences. They don't understand cause and effect. They're amazing at matching patterns and making new combinations. They can mimic, but it doesn't mean they actually understand anything. Now is this improving? Absolutely. And there are ways to mitigate against some of these situations. For example, if I use ChatGPT as a middleman to help me build that image prompt of sodium and water, the fact that ChatGPT is built on analyzing text, which often would have patterns talking about sodium and water will explode, then ChatGPT might not actually generate those images, but it could generate a more useful prompt and then pass that over to DALL‑E to generate the image. Now this might not be perfect, but it's more along the lines of what I wanted. Another common concern with generative AI is the idea of bias, and it's easy to see bias in generative AI. Use any of the image tools like Midjourney, DALL‑E, or Stable Diffusion and ask for an image that represents a job like I want a picture of a professor. You get multiple options, but do you see a certain similarity in the results? Let's try another. What does a software developer look like? And one more, how about a photo of a preschool teacher? Now we can understand why we get these results. These tools were trained on millions of existing images. Images with these keywords have historically skewed to a particular gender, age, and skin color. But the AI can up reinforcing this as if all professors have to be 60‑year‑old bearded white guys. Now, can you get generative AI to be more diverse? Absolutely. But you have to ask for it in your prompts. And to ask for it, you need to notice when it's happening. And while it's easy to see bias in a visual example like this, the same idea also applies to the way text‑based generative AI works, but it can be much harder to recognize. Generative AI is fantastic, but it does bring new issues. Not just hallucinations, but the potential effects of misinformation along with issues of data privacy and content ownership and many larger issues around the impact this is going to have on many professions.

Getting Better Results from Generative AI

Getting better results from generative AI often comes down to the quality or creativity of your prompts, meaning the questions you ask, the commands you give. There's an idea called prompt engineering, and this can mean different things to different people. But let's just take it as the idea of applying some structure and formality and thought to our interactions with generative AI in order to steer it towards the results we want. And when we don't get the results we want, how we then interact with the system and how we refine and rephrase those prompts. There's this idea that with slight changes in the way you ask for things, you'll get very different results. Some of the most common suggestions to corporate in your prompts include the idea of a specific persona or profession where you can tell an LLM to act as if it's a type of person. It can be a job role. Act as a marketing copywriter, act as a physics professor, act as a financial advisor. You can even ask for specific people like historical figures or fictional characters. But there's one thing I personally find very useful here, which is the idea of flipping the roles. Here's what I mean. See, I see a lot of suggestions to use generative AI in a way that tries to replace the person writing the prompt. For example, I write and teach technical content, and a lot of people recommend that I might begin my prompts with words like act as an experienced instructor, act as a skilled trainer, act as an expert educator. I hate this recommendation. I am the experienced instructor. That's my job. I want gen AI to help me not replace me. And I don't like the internal message I'd be sending my own brain that I have a core competency, let me just give it to the machine. Now you can do this, but don't be annoyed at generative AI for taking your job if you're literally handing it over. But what's the alternative? Well, we can flip this around. Don't ask ChatGPT to act as an instructor. Instead, tell it to act as a puzzled learner and ask me three questions about joining tables in SQL or how variables in this language compare to variables in that language. I find it far more useful to have generative AI act in this role, and I often don't need the LLM to act in a specific role. I think that sometimes overused. Just tell it what you need. You can be specific about the audience. Explain as if I'm 10 years old. Write as if you're talking to a room full of software developers. You can also say what the desired tone should be. And I don't just mean formal or casual, but being extremely specific about how you want those results to feel. You can tell it what exact length you want, whether you want a short email or to write 5 tweets or you need 500 words or 3 paragraphs or 2 sentences. The more specific and creative you can be with the prompts, the more creative and useful your results. Now, in 2023, it was very common to see headlines like this, prompt engineering, the hot new skill employers are seeking. ‑Prompt engineering, the hottest job in tech. ‑Prompt engineering, the most in‑demand profession of the future. But I already believe this idea of everybody needing to be an expert prompt engineer is becoming less and less important. Now right now, it is important to be comfortable with writing generative AI prompts, but the tools themselves are quickly improving on this. They're taking even basic prompts and better reacting to what we're actually asking for. And it might sound like an odd thing to say, but the prompts were never the most important part of this. What's more important is your approach to them, not worrying about the perfect prompt, but keeping that attitude of experimentation and revision. And as the tools continue to change and improve, that is much more important. But the single most important thing to feel comfortable with generative AI, where it becomes just more intuitive to recognize what it's good at and what it isn't, is to use it again and again. Just start baking it into your life. So make ChatGPT or Google Bard or Anthropic Claude your home page. Get into the habit of using generative AI where in the past you might have used a search engine or Wikipedia or even just journaled in your own document. And if you want to go a little deeper, watch another topic in our Generative AI Foundation series. But thanks for joining me. I'll see you next time.