Deloitte.

Deloitte AI Academy™

Generative AI Fluency
Part 3: Generative AI and Beyond

Consolidated Transcript

Table of Contents

Welcome 3
Course Overview 3
Learning Objectives 3
Meet the Team 3
Topic 1: Identifying Applications 3
What Makes a Use Case Generative? 3
Generative AI Benefits 4
Accelerate with Text Generation 5
Accelerate Actions like Writing Basic Code 5
Generative AI Benefits 6
Automate Chatbot Responses for a More Natural Response 6
Automate Time-Consuming Interactions into Actions 7
Personalize Offerings by Generating New Recipes Based on Ingredient Availability 7
Simulate Artificial Environments with Generative Image Editing 8
Create the Unexpected with Generative Multimodal Approaches 9
Applications 9
Topic 2: Life Cycle Example 10
Generative AI Solution Life Cycle 10
Phase One: Identification of Need 10
Phase Two: Algorithm Selection and Development 10
Phase Three: Evaluating the Quality of the Generated Output 11
Phase Four: Deployment and Implementation of the Selected Output 11
Topic 3: Risks, Limitations, and Costs 12
Risks 12
Limitations and Costs 13
Ignorance 13
Compute 14
Intent 15
Topic 4: Conclusions 15
A Human-Centric Future 15
The Future is Generative 16
Summary 16

Welcome

Hi, thank you for joining the third of the Generative AI lecture series. This is part of the Deloitte AI Academy's Generative AI Fluency series and produced in collaboration with the Deloitte Technology Academy.

Course Overview

Like I mentioned, this is the third of a three-part lecture series. In the first lecture, we went through the foundational models of Generative AI and kind of an introduction to Generative AI, got a little bit under the hood there. In the second one, we went into a lot more of the technical mechanics of the Generative AI models, a lot more under the hood. And went into some of the different data modalities like natural language processing, computer vision and how these fit into the space.

Today, we're going to be looking more at the applications of Generative AI. We'll go through a full end-to-end example and the hope is that this will give you more of an intuition on how to determine that something is a Generative AI use case and give you a little bit of an understanding to start to guide some scoping questions. We'll end with a touch on the future trends that we're looking at in general in the Generative AI space and hopefully get you really psyched to start to use these in your own projects.

Learning Objectives

So, we have different components in today's talk. The first is going to be in identifying applications. This is something that will take probably the meat of our talk today. And then, we'll go through one end-to-end life cycle example. So, one specific example and all the different components of it. We'll talk a little bit on the risks, limitations, and costs of these models and then we'll conclude with kind of where the future of the space is going and give you a quick summary of all three of these lectures.

Meet the Team

My name is Celia Ludwinski. I'm the Head of Technical Operations here. Within SFL Scientific, we are an AI division within Deloitte that works on applied R&D projects for our clients. What I do is a lot of managing in the technical learning journeys, trainings, improving on how we deliver projects and ensuring that we have great client delivery.

Topic 1: Identifying Applications

What Makes a Use Case Generative?

So, we're going to be talking a lot about what makes a use case generative. Traditional AI versus Generative AI, the real key word that I want you to focus on here is "Generative."

So traditional AI tends to be more prescriptive, tends to be more rules-based or tends to have defined inputs and outputs, whereas generative is really creating something novel or synthetic or some sort of brand new interpolation, interpretation of information and data that's been provided so far.

Within the traditional AI space, there's a ton of different modalities, there's a ton of different examples. But to just go through a couple within the image space, one example is traditional image classification,

and that's where you provide some images, and you try to get different classes of known categories out. You seen a lot of this in the very basic cat/dog or dog/blueberry type of examples that have come about so far.

In the Generative AI space, instead of identifying or classifying, what you're really doing is creating novel image synthesis. So, this has been become very popular in recent months where people are able to do a multimodal text-based input and then get an image-based output that's brand new and novel, and it's democratizing the capability there of generating new images. In the text-based space within traditional AI, language translation is an example that you can consider. This is basically trying to move from one language to another one.

And traditionally this is done in segment-based, rules-based, one-to-one translations, whereas now within Gen AI, you can start to see language translation, maybe not necessarily between languages, though probably there as well. But really, you're starting to see language translation that adheres to style and tone, so I can translate something from Shakespearean style to modern hip-hop style and back and forth because of Gen AI and the understanding of those different pattern- recognition spaces.

In traditional AI, there's one other space that I want to highlight, and this is a space that's typically just not been well-developed because there hasn't been sufficient data or for some other reason. And in one example there you can see fraud detection. And typically, that's been done using things like statistical models, and anomaly detection, and things like that. And that's because with fraud there aren't a superhigh volume of fraud detection cases. Rather, there's a microcosm of fraud compared to the high volume of good actors.

And so being able to use Generative AI, you can actually start to create synthetic data to mirror or to mimic what these fraudulent actors would look like. And that really helps with the entire training pipeline and the way that we work with machine learning models today. So this idea that synthetic data can be generated using Generative AI and then can be used to feedback into a system where data isn't already prevalently available can be applied to a wide variety of industries and is a great location for Generative AI to live.

Generative AI Benefits

So, with that, I know this is a little bit of a review, but I want to quickly highlight a slide that you've probably already seen before, and that is the progression of maturity of Gen AI use cases, so this has been presented in lecture one. We talked a little bit about this, but generally from left to right you can see a increase in complexity of category and therefore an increase in the difficulty and generally a decrease of maturity in the Gen AI industry.

So, on the left you see this "Accelerate" category. That's what's actually happening a lot in industry today. You're seeing it with things like ChatGPT that's come out, with Midjourney that's come out, and a ton of other AI tools that have come out that are allowing users to provide first drafts of something and then iterate very quickly and provide their own subject matter expertise and accelerate work that they're already doing.

Slightly more complex is the "Automate" industry. You can see a lot of automation already happening traditionally with chatbots, and this is something that generally is automating low-level type work. So it's

removing that human necessity component, but isn't something that's automating really high-risk or high-complexity components at the moment.

You also probably see some instances or at least discussions of "Personalize," "Stimulate," and "Create", and there's a lot of gray zones and overlaps and boundaries here within these three categories. Some of these are being pursued using some traditional methods. Some of these are being pursued using Gen AI methods. Some of them are still quite immature. But really the long-reaching scope of each of these have not fully been realized yet. And so, we'll talk a little bit about what that means.

I'm going to start and dive right into some use cases for each of them. So, you guys have a very clear intuition of what it means to be in each of them and then we'll go from there. "Accelerate" is one of the categories that I wanted to talk about here, and really this is about trying to alleviate some tedious tasks that require a large amount of time that operate at a high volume in the space. So, happen pretty frequently by a number, a large number of different users, but that don't require really intense subject matter expertise. So, we're seeing a lot of this acceleration starts to happen right now because of the high volume and the high volume of data that's available and the democratization of certain AI tools that are out there in the world.

Accelerate with Text Generation

So, we've heard about large language models performing text generation. This is something that can be used for writing emails as can be seen on the right-hand side, writing proposals, writing paper reports, writing posts, things like this. Given an input of the known info in a concise but disorganized manner, a large language model can really very rapidly create a first draft that is concise, cohesive, and ready for a human-in-the-loop to perform their set of review. It's actually been studied now, and this can boost productivity and is boosting productivity by nearly 40% in realistic writing tasks at high volume by low subject matter expertise. This type of thing can be improved in niche areas still.

So, the public large language models can be worked into niche areas to do similar types of productivity benefits. But in terms of what's already available out in the world, things like writing these types of emails can really be rapidly sped up using large language models, and there's a decreased barrier of entry for these types of skills. So, it can save time for that for a large population.

Accelerate Actions like Writing Basic Code

The language use cases that can be accelerated using large language models do not have to be limited. I want you to expand your mind a little bit here. The goal is to make you think creatively, and really have an intuition or an understanding of where these use cases can be applied. So, English to English is not the only type of use case here. You can think about any other human language. You can think about a combination of human languages, English to Japanese, or vice versa. You can also think about using things like coding languages. These are also text-based language models that can be used for use cases.

So, I'm going to go through an example to really highlight what that means. We had a task at hand that a coder was meant to develop, so the goal is to take in a series of documents, be able to scroll through them and create a recording of what that scrolling would look like and create a video output. Now, if I gave this to a coder today to traditionally code by hand from scratch, assuming that they understand the

Python language that you're trying to develop and assuming that they understand generally some of the functionality that's being looked for, but maybe don't understand 100% of the specifics of the video encoding would probably take them a few hours in order to hit some of that learning curve, actually type up all of the concepts they had in their head and understanding that approach and then implement and push forward. So, let's say three-ish hours.

I have asked someone to do this with a large language model equivalent, basically consolidate the concept to the task, put it into a large language model as a prompt and request the code as output. And what this is able to provide was you can see on the right-hand side some code-based output that was not immediately able to achieve what we were looking for. And that's because the large language model didn't understand the environment. But with the help of the coder to enable libraries and installations and things like that from start to finish of coder putting in prompt, coder getting out the large language model output, coder fixing the environment variables using their own subject matter expertise, we were able to complete this task end-to-end within about 15 minutes.

So, what I mean by time savings or acceleration here is also applicable in this type of world as well. So, from 3 hours down to 15 minutes and I do want to call out the limitations here. Firstly, both input and output required subject matter expertise. Model doesn't know the environments that you're living in and model needs that human-in-the-loop to really synthesize a concise prompt up at the beginning. But it was able to then using that human loop, accelerate that subject matter expert's workflow and get them to move much faster.

Generative AI Benefits

So, the point I want to call out in both of those use cases, we just went over in the "Accelerate" space is that they still require that human subject matter expert to provide inputs, to provide outputs, to review, to send out, to perform the actual actions. But Gen AI is really supplementing the internal workflow of the individual users in order to create an increase in productivity, that acceleration component. If you wanted to take some of those really low-level tasks and remove the human judgment entirely from the space, and this is something that's already being pursued by various different corporations and industries on things that are kind of low-risk but high-volume in their industry, then you would want to move into this "Automate" space.

Because of these types of considerations, this is a little bit more along the progression of complexity and therefore a little bit lower on the maturity level, which is why I'm highlighting that this is most commonly seen in things like low-risk use cases. So, let's dive into an example with a chatbot.

Automate Chatbot Responses for a More Natural Response

Chatbots are not new to the world of automation. In fact, these have existed for quite a long time historically so far. But there is a lot of frustration with the current capabilities of these types of chatbots, and that comes in part from essentially two different reasons.

First is that actually getting to a resolution at this point requires a ton of different hierarchical, multistep follow-ups, questions to really get at the intention of the consumer using a very detailed, essentially flow state staging of the chatbot. It's very rules-based. Some of it can be interpreted, but it's usually a more direct word-for-word, one-for-one, step-by-step interpretation.

The second reason that we see a lot of frustration with traditional chatbots is the lack of this kind of empathy or human-type response or understanding that makes it more organic and more like a conversation between humans. So, both of these things are actually things that Generative AI can and has started to overcome and can be addressed in this chatbot or this automation-type world.

Automate Time-Consuming Interactions into Actions

So, if we're talking about the ability to remove some of those hierarchical questions and those hierarchical meanings, Generative AI is actually a little bit better at understanding the intention of what's coming behind the consumers' goals. And I'm not going to say it's great because I'm going to talk about that in limitations later. But it is one of those things where when you say I'm upset because the watch I ordered has arrived broken the Gen AI tools will be able to identify a little bit better what order you're talking about, what item you're talking about, because it can contextualize better rather than demanding that you provide very specific contextual details in a step-by-step manner. So that can be done a little bit better.

Secondly, the idea of trying to make the conversation more empathetic, more organic, and more natural. This is something that language models are getting really good at because they're getting really good at imitating the way that humans speak to each other or are typing to each other. So, using the right type of human training data set, the right type of empathy language to really provide this background to this chatbot means that it will start to provide that type of interaction a little bit more organically to the consumers and alleviate some of that frustration to get to the end solutions in a much faster continuing to be automated, time-consuming way for the end user.

I want to pause here to talk a little bit about the differences between some very similar and often interchanged words in this space. So, when we say "personalize," a lot of times you'll hear the words "customize" and "recommendations" coming out. And I want to differentiate between personalization and customization a little bit more clearly. And we'll also do that with an example. But in the space, you've seen tons of different examples of recommendation engines, customization options, things like that.

Amazon has a huge catalog of products that they make recommendations to you for based on your past-purchasing behaviors and how they align with other users. Past-purchasing behaviors. If you were to, for instance, log on to a large shoe brand website and try to create a new shoe based on a series of different color options, and lace options, and sole options, you're actually creating what's called a customized output rather than a personalized output. And the difference here, or at least personalized with Generative AI. But the difference here is a finite, concise set of options and responses for customization rather than essentially an infinite and really uniquely adapted to individual personalized outputs in the sense of Gen AI.

Personalize Offerings by Generating New Recipes Based on Ingredient Availability

So, I'll go through a little bit of an example of what that means. Historically, if you were trying to get a recipe recommender out, you would actually start by using your past behavior recommendations, maybe the ingredients that you have in your fridge, and maybe a few different preferences about flavors, and allergies, and things like that. And what would pop out of this type of recommendation

engine is from a large database, a seemingly massive database, you would get a specific recipe that best matched those types of preferences. So more of a matching recommendation algorithm.

What we're seeing in the Gen AI space is that instead we can now create personalized recipes, and these are generated novel recipes. They may seem very similar to something that already exists in the Suzy O's cookbook, but they have been personalized to you and are completely novel recipes based on something like your preferences. So, if you have a hatred of citrus and a love of spicy food, then this Tuscan roasted chicken recipe might have been adjusted with various components of the red bell peppers and other types of components in order to make it most applicable to your flavor palate, as well as what types of ingredients you actually have in the fridge.

So just really want to call out the difference here. What we're seeing in this example is personalization, because it's novel, because there are individual components and line items that have been and can be changed out rather than customization, which is more of a matching and/or more of a finite list of options that can be selected between.

Next, I'm going to talk about simulation. So, a lot of the concepts that exist within the world of simulation are a little bit more abstract concepts. It's the idea of digital twins, or synthetic data, or data augmentation. These are all things that we're abstracting, conceptualizing and are sometimes hard to understand what use cases could also be generated from in the simulation space. So, I'm going to try and give you a practical example of what a simulation use case might look like.

Simulate Artificial Environments with Generative Image Editing

In the traditional retail world, shelf stocking is something that happens all the time. But and we do know that stocking of shelves is something that has a high impact on the overall profits of the store and the purchases of the customers in satisfaction, even in the idea of being able to find something with ease.

So, this type of high-volume task that is not usually done with a ton of intention, at least not a ton of intention on a high-frequency basis at the time of shelf stocking, but maybe more as an overall strategic goal and infrequent basis is something that can actually be addressed using Generative AI and synthetic use cases.

So, given an input of layouts of the stores, things that already exist, performance of different shelf stockings of that store in the past, and an input of the current inventory of the store, as well as additional information about holidays and the goings-on and activities in the external environment as well. Different simulations of how shelves could be stocked can be generated in either an image or a virtual reality type display and evaluated using a very rapid flip through human-in-the-loop and understood understand where these types of items can start to be displayed. These can be tied in with a goal of trying to enhance something like cost savings, productivity, overall profit, margins, things like that. These types of factors can be tied in to create simulations that are more targeted towards that type of outcome and this is a space where there is an infinite number of combinations of all of these types of items. So, understanding visually how these types of items can be displayed on a shelf is a generative space and is a simulation space within that Generative AI world.

I have one last example to go through and that is the "Create" area of Generative AI. This has, at the moment, been inspiring a ton of the discussion that's been going on in the world on this topic, and it's something that we'd love to achieve in fullness going forward. But in reality we aren't quite at the full

level of maturity of what it means to be creative in the Generative AI space. And I will explain what that means in just a moment. But some of the examples that we can talk about that you might be familiar with and that might be incredibly achievable, are things like creating novel images

Create the Unexpected with Generative Multimodal Approaches

We've already talked a little bit about creating first drafts of text but being able to go one-step further and create entire works things like ghostwriting papers, a full-on logo and branding design, stuff like that are things that exist in the "Create" space as well.

Diving right into it, these are all probably familiar with the idea of creating images using text-based prompts and inputs. So, you can see on the left an image that has been generated using one of those Generative AI tools of a teddy bear on a skateboard in Times Square. And this is something that can be generated because of interpolation. A model has seen a teddy bear, a model has seen photos of Times Square models, seen photos of skateboards, and can easily put those all together in a way that starts to make sense in an image. So, this looks real, and that's the goal of these types of generative models right now, is being able to look real.

Similarly, on the left, images of corgi, fire, and trumpets have all been seen before, so you can get a combined approach to see those types of generated outputs. And if I were to ask a model to create an image of a spaceship that filtered through infrared light to allow for genetic mutation experiments of plants that we shot into orbit, I would bet that we would be able to get a solid image out of that type of concept.

However, where the current capabilities of "Create" in the maturity spectrum versus the future capabilities lie is that we don't currently have the ability to take that image of that spaceship and connect it with logic and start to create engineering plans. So, I did say create engineering plans that can then be put into manufacturing. So, the ability to move into the next step of the "Create" space, firstly combining some of the logic and secondly, making sure that it's robust to various physics- based rules are things that still need to be incorporated into this world of "Create." And that's where the lack of maturity is something that we can strive to start to push against in our creation of new Gen AI use cases.

Applications

We've gone through a lot of different applications of Gen AI use cases so far, we've touched on a few different industry verticals, and we've touched on several different data modalities. We talked a little bit about some of the natural language processing ones, we've talked about a computer vision one, and you've even briefly, tangentially started to reference some graphs when we spoke a little bit about Amazon.

What I want you to take away from this concept, though, is that there can be conceived a Generative AI use case in almost every category of data modality and industry verticals. And this list that I have on the screen here is not a comprehensive list of those options. This is a beginning list and something that you guys should start to iterate on and think on and imagine for yourselves the way that we can use Generative AI in starting to expand in these different areas.

Topic 2: Life Cycle Example

Generative AI Solution Life Cycle

Hopefully this slide is familiar to people who've seen the previous two lectures where we have the four phases of the Generative life cycle solution. In the first phase, we have the identification of the need, and this is really about the business need, the problem is trying to be solved. I'm sure you guys are all familiar with identifying needs and problems, so I'm not going to go into this a ton.

But that's immediately followed by phase two, which is the algorithm and solution development. And it may seem strange that we're saying solution development at phase two. And then there are two additional phases beyond creating the solution. But I'll go into why that is in a minute. The algorithm solution development really involves preparing and cleaning, preprocessing any data, preparing any pretrained models, and/or creating your own models, creating various algorithms and testing them for their initial effectiveness.

Phase three is really about evaluation, and this is one of the areas that is going to require more heavy critical thinking, in particular in Generative AI use cases compared to some traditional use cases. And that's because in Generative AI there can be many correct answers. So, evaluating something to understand whether or not all of the correct answers belong there, whether or not some of the correct answers belong there, is really going to be tying back into the business need that was identified in that first phase. And some of these evaluations are going to be things that can be done using performance metrics, quantitatively trying to get our calculations. But some of them are going to be things that need some sort of human loop, human evaluation, and/or an indication of how it should be deployed in phase four with that type of human loop or iterative capabilities. So, phase three may not take the most amount of time in any Generative AI life cycle solution, but it is critical to have a lot of thought put into this in order to make sure that you're as successful as possible.

Now, phase four is about the real world productionization, understanding how you're going to start to monitor it and understanding how it's going to live and be implemented with the end users in its full life cycle. So, let's go through a quick example.

Phase One: Identification of Need

I'm going to expand upon the example that was talked about in one of the previous lectures where we went through a campaign name generator. Let's expand it and add a multimodal business development marketing tool where we're trying to create both campaigns and logo and product mock-ups and associated text content for captions and other types of marketing materials, all combined in a consistent style that allows for feedback and integration posting and then moving on from that. So we've identified our need here, our business need. It includes style consistency, it includes a few different components that need to come out from it. And obviously we want to achieve better sales through our marketing campaign here.

Phase Two: Algorithm Selection and Development

I'm not going to go through a ton of the details of what the technical model components are, obviously of a foundational model that can come out, and either through a combined approach or a serial approach, you can use it to start to create multi-outputs. So, after the realness discriminator, you can

either create a serial approach to create all of the images and then associate the captions or you could do them combined. This pipeline then is developed in phase two and evaluated to say that it is developing or it is outputting images and texts that seem like it's appropriate for the types of inputs that are coming out.

Phase Three: Evaluating the Quality of the Generated Output

Now, how do you evaluate that it's actually doing the job that it's supposed to be doing? And this is again where we talked about phase three is going to be really important. In our example, we talked about consistency and wanting to make sure that we had a cohesive connection between the multimodal components.

So, this in the top right corner, this relevancy and consistency bucket is where we would probably focus all of the evaluating the all heavier lift of the evaluation components for this use case. In this case, relevancy and consistency are kind of qualitative evaluations, right? It's not something you can't put a number on how relevant or consistent something is, at least not incredibly well.

So, this is an instance where you might want to include a human loop, either in evaluating the model itself as performing well before you put it into a client production state and/or considering including a human loop in the final production state. And in this case, that's what I think we're going to go with, is something that our pipeline has a creation of a several model outputs evaluation by the human loop marketer or whether or not it meets the needs that are looking for and then approval before being implemented into a marketing campaign itself.

You can obviously include other evaluations in this case. Accuracy and context are things that could maybe be evaluated quantitatively. You can do things like semantic understanding to evaluate whether or not the same meaning and content that was in the prompt or the request is also equivalently in the outputs that are coming out. You can obviously do some sort of any Euclidean distance matching to understand whether or not again you're matching things up between the prompt and the output. So definitely can be some quantitative components to these evaluations. But generally what we're seeing in Gen AI solutions is either a need for openness for things that aren't quite right, so that kind of democratized access, a democratized and low-risk access and/or you're going to need a little bit more of that human-loop component through reviews or through final deployment.

Phase Four: Deployment and Implementation of the Selected Output

So, speaking of final deployment, let's go through what the what our example output has created. So, I have three images here that our marketing campaign created and in the left one, I don't know if you can really see it on the screen here, but what we see are firstly a little bit of morphed, weird looking faces and apples that don't look like apples. They're really just combinations of a bunch of red blobs. So, this is something that maybe could be approved for a campaign depending on the resolution requirements, the speed that it needs to come out, something like that, but probably isn't really relevant in the case of this marketing campaign where we're trying to create an ad for an apple cider company.

In the second option you see here, this is actually a really common mistake that Gen AI image models are creating right now. And that's that it doesn't overlay text really well onto images. So, you see that firstly, some letters aren't even really letters. And then so the words just aren't really words at all either.

So, the idea that we can create this logo using text overlaid on top is something that is probably more of a combined approach with traditional and Gen AI, but in this case, a marketer would see this would determine that it is obviously not relevant for what they would want to put in a final product and remove it from the option list.

And the third option, we have the logo that's been created, and this is something that realistically could represent an apple cider company's logo and fits the need for trying to create a marketing campaign without having to do ad wrecked design of materials and is approved by the human marketing reviewer for usage in the end.

And this idea that we need to use the marketer for that type of deployment is at this point known as the last mile problem. You could also call it the 80-20 rule or the Pareto problem, but it's the idea that you do need to require that kind of human loop to ensure quality at this point, since what model, which Gen AI models are trained for right now is to sound right or to seem realistic. They aren't necessarily trained for factual accuracy. They don't have that underlying physics models behind them. They aren't trained to be right. So, I like to say they're not trained to be right but trained to sound right. You're talking about large language models, but it's kind of true of all Gen AI models. This concept that there is a last mile where we need to double-check things aren't completely wonky.

Topic 3: Risks, Limitations, and Costs

Risks

Let's talk about some of the risks and ethical implications. And these are by no means all of the risks and ethical implications. These are just some of the ones that have been popping up with a little bit more volume and concern recently within the Generative AI space.

So, in the category of "Fairness," one of the things that you should know about Generative AI models and just in general all AI models, is that they're trained on the data that already exists, and that data contains biases, usually human biases, and the models are just learning a lot of those biases. So that can mean that Generative AI outputs can start to reflect some of these biases. You may see something where a model's textual output is assuming that a doctor that is referring to is a man. You can also see this in things like resume checkers that are evaluating for different demographic biases as well. This can sort of be obviated and mitigated using some really, really good data cleaning, data alignment, classification equality, balancing some careful considerations, things like that. But it's something that you should be aware of, especially if there's a high risk of problems due to these types of biases popping up in Generative AI outputs.

Second one I want to cover here is "Privacy and IP." And technically, these are actually different concepts, but they do end up being talked about a lot together. So, I bring them up together. So, one of the topics that comes up here is where does the data for all of these model trainings come from? And there does seem to be a bit of an ethical concern or like murky area on the sourcing of these data, both in terms of proprietariness, privacy as well as just this idea almost of that. And it's a bit of a gray area here that hasn't been fully regulated yet.

So, if it's a concern for the use case, then it's something that should be considered when going into using pretrained models, for instance. Additionally, you might want to consider that some of these types of models, because they come from these types of data sets, may reveal private data. And if you're living in the world of Gen AI flubs and use cases you may have already seen some of these things come out. So, some of the ways you can start to address it is using data cleaning, some PII scrubbing, making sure you're understanding the sources of your models, and/or just moving into more privatized, personalized model development.

We'll talk a little bit about how this is one of the cases that may wear you more towards creating your own model using your own data set. We've also seen a lot of what's recently been dubbed "hallucinations." So that's where a model creates some outputs that are incorrect or harmful, but very convincing and persuasive. So, I kind of mentioned this a little bit before when I said that models are trained to sound right, not to be right. This is one of the risks that you should be aware of when it comes to Gen AI. If there is a long-term downstream output problem that comes from these types of hallucinations existing, then just be very aware and cautious of how you're trying to approach this type of problem creation.

One of the things that can be implemented to mitigate this is, is that human-in-the-loop component that I was talking about. You can also implement some sort of guardrails to prevent those types of incorrect solutions coming out, but it has to be very well thought out in the beginning. And then the other components that may help here is things like interpretability or explainability of the models to give a little bit more of an understanding, working of the outputs, and then also can help with the concept of change management and adoption as well. So just something to think about overall.

Limitations and Costs

I'm also going to talk a little bit about the limitations and the boundaries of Generative AI as it exists at the moment. We've been touching on this a little bit throughout the entire session so far. So hopefully some of this is going to start to sound a little bit familiar. These are also limitations that somewhat inform some of the risks that you've already seen talked about in that previous slide. So again, hopefully things will start to really tie together and give you an intuition of how these types of models are starting to work.

Ignorance

The first one I'm going to talk about here is "Ignorance." And I touched on this a little bit in that previous slide when we talked a little bit about the risks. But models rely on training data and that means that they're essentially interpolation- based approaches to solving problems. They don't really have a logical understanding of the backgrounds behind the scenes and physics behind them, and they don't do great with out-of-domain areas.

So, we've seen a lot of problems arise and I don't really want to say problems, but mistakes arise because of model ignorance. So, in the case of the image on the left, what you can see a lot of people talking about when we're generating images of people is that it's really easy to identify some mistakes that models make because very often they're creating hands with way too many fingers or weird alignment of knuckles. And while this may seem photorealistic and like it could be an image that was

actually captured, it is not a realistic representation of a human hand, which we understand because we have an abstract concept that a human hand contains five fingers.

You see this also in things like people calling out that you can identify models are incorrect because they do ears incorrectly or because the background patterns are way too perfect to be a real photo. One of my favorite means of mistakes that have come out of the Generative AI modeling space and the images that are being created from it is the way that humans eat spaghetti. So we have a socialized norm of how you eat spaghetti. Obviously, it's typically with a fork, and spaghetti itself is already a fairly hard concept to abstract out. But there are a slew of hilarious pictures of photos that have been generated of people very incorrectly eating spaghetti. And it's because this social norm hasn't really been fully indoctrinated into the model. And it's indoctrinating just generally how humans eat some some foods, especially in types of stock photos where this type of enthusiasm.

So you see this woman eating a bowl of spaghetti almost like she would be eating a sub rather than traditionally how someone would be eating spaghetti. Similarly, on the left, you can see an example of ignorance. This is an example where someone definitely could have drawn this type of image. But the idea that a monk riding a a snail would work at this type of scale doesn't really make sense because of our known physics understanding of the scale of these two objects and how they would mesh up in this way. So these are examples where the the models are creating images, the images look like real images that people could have created, but they're fundamentally wrong in some way because of ignorance about the way that the world works.

Compute

Moving on to "Compute." So, large models are faced with cost, resource cost, and at an actual financial cost. So, in order to train large models, you usually need pretty heavy hardware computation and a lot of time to do it, which then is associated with a high cost, particularly for training something from scratch. Where we start, started to see democratization of these types of approaches is being able to use pretrained models, which I believe we covered in lecture two and then retrain them or add on a little bit of fine- tuning in order to get a focus in the niche. But that original training of a full model from scratch is something that's a pretty heavy cost. And so you can see that here in this image where we look at the difference in cost that the top one is what you would see if you were trying to create something from scratch.

For most use cases, unless there's a really specific need or a very obvious ROI above that type of cost level, usually that doesn't make sense as an approach to train a full, for instance, large language model from scratch. Rather, what usually makes sense is to go with that fine-tuning option and/or the open-source option that exists. So, whether or not you go with one of the two options will depend on whether or not you're going to cross that threshold of ROI cost and financing.

So, when you fine-tune a model, oftentimes you can deploy it and host it yourself, and that is cheaper at a per inference or per runtime request than if you were to use an API call like OpenAI or other open API or other open solutions out there for paying per request. How quickly it takes to see that threshold crossover will really depend on the frequency of usage or expected usage of the solution output that you're trying to develop.

Intent

The last limitation I'm going to cover for now is “Intent.” And this is because AI and its current forms is a pattern matching or recognition type tool. It does not for the moment have any sort of intentionality associated with it at this time. That also means that interpreting human intent is purely based on pattern matching. So, if input prompts from our human developers are not things that perfectly match what the model's expecting, we get a little bit of a mismatch and a little bit of a wonky output from the model's goals.

So, you can see this in these types of examples. A generated prompt for an ad for a family fun pizza spot has created for Pepperoni Hug Spot, a motto that says, "Like family, but with more cheese." And while I personally find this incredibly hilarious, this may not be the best type of prompt that would go or the best type of output that goes with the intentionality of creating that ad for a family fun pizza spot.

So, this is because of the inability to understand the intentionality perfectly without a little bit of guidance and requires a little bit of iteration and/or some of that prompt engineering that we've been talking about. But in general, the things that you should be aware of, at least in the text world, is that models aren't super great at understanding human humor, detecting sarcasm, or responding to emotional language. This is also, again, I'm going to highlight back where the last mile problem fits in really well, including humans-in-the-loop to evaluate for and kind of massage and perform this type of iteration so that you can address the mismatch of intent.

Topic 4: Conclusions

A Human-Centric Future

I'm not going to go into the future of the technology. That alone is something that you can probably reference our previous lectures on, and it's something that is going to be incredibly rapid growing. So, it's not something that will be directly predictable and it's going to be influenced by you guys in your creativity, in the types of business use cases that we can start to bring to the Generative AI workspace and then drive the future development in that space. However, I do want to call out some considerations for the future usage and applications of Generative AI technology here. And these are things to really make it human-centric.

So Generative AI at its baseline, it's generating things for human consumption, for human needs, for human development, is really meant to help us in everything that we do. Looking forward through the lens of application of Generative AI use cases. There are a few considerations I want to call out, and these are all rooted in the concept that Generative AI is meant to be enabling and serving us and our needs.

So, there's a few areas that are really coming to light as we're starting to see implementations of these Generative AI applications. And one of them is considering speed versus accuracy. So, as Generative AI becomes more widespread, you're going to start to see a question of whether or not we actually need to get to that fully accurate value or whether or not getting to that 80% value makes sense. If you can do it incredibly quickly so that you can get it out to people and people can perform that 20% move faster.

Depending on the use case, you're going to land on one side versus the other. So that's definitely something to take in as one of those considerations.

This ties pretty heavily into this concept of democratization. What we're seeing a lot with Generative AI use cases is that this is very appealing to the everyday user because we're providing that speed without the 100% accuracy and we're giving access to things that are available at a high-volume but low-pulse, and kind of tedious, level of skill sets that can be enabled and replaced. So we're really seeing that productionization or that productivity improvement because of democratization and that type of thing is going to start to come out more and more as you're starting to see these use cases. So that's where you might want to consider pushing the bounds and/or where limitations start to push back.

The final one I want to call out here is interpretability or explainability. And this again ties into the idea of adoption and change management and giving access to people so that we can enable them to do their work. Just providing the output from a Generative AI model may start to really do that. But if you add on layers of understanding of why a Generative AI model thought that this was useful and highlight the different components that are relevant, that are most helpful, and use that as a way to guide iteration and/or to guide trust, you're going to start to see a lot of that improved adoption that you're looking for and a higher success value at the end of your use cases.

The Future is Generative

So, the last thing I want to leave off here with you guys is hopefully an inspiration or motivation to everyone. Generative AI use cases are not going to be solely conceptualized by your data scientists. They're going to be conceptualized by absolutely everyone. And this is because everyone has a different experience and can start to see where these types of capabilities will really enhance their own experiences, their own workflows, and things like that. So, I want to highlight that while this is a rapidly changing environment, your creativity, and your direction in trying to abstract and brainstorm and conceptualize these use cases is what's going to drive the entire movement and the direction of the technology going forward.

Summary

We've covered four different components of Generative AI applications, the project life cycle, the risks and limitations, and the future of Generative AI. This concludes the three-part series on introduction to Generative AI from the foundational view to details of the technology and generative data modalities and then the applications that you saw in today's workshop. Really appreciate you guys listening through this series and I really am excited to see all the things that come out of all of your inspiration and all the work that we'll be talking about going forward. Thank you, everyone.

Deloitte.