(3) Imagine Austin | Prompt engineering for automation developers -

Transcript:

(00:07) hello everyone quick hey well thank you for coming to today's session around AI prompt engineering for automation developers so my name is Max Cassidy I'm a senior program manager here at automation anywhere here is over my name is YY Falon I'm from Google cloud and I work with automation anywhere on some of the um generative AI capabilities that they're bringing into their platform please to meet you all now in today's session we are going to be going over the fundamentals of prompt engineering

(00:41) when it comes to integrating this within automation anyway but before we jump in I just want you all to have a quick look at our Safe Harbor if you can all take a look through this this is basically just to make sure that any purchases made with automation anywhere are made by your decision and that we haven't pushed you in order to be able to make these decisions moving forward I'll leave that up there just for a few more seconds in order to be able to go through that and read that before we move on to the main

(01:14) context now automation anywhere is Bridging the Gap between Automation and generative AI by offering the fun the functionality to be able to integrate with generative Ai and large language models straight from the platform itself to be able to create create costeffective automations streamlining processes and start tackling Advanced use cases that before were not possible or required massive amount of human intervention now the agenda for today's session we will Begin by taking a look at how large language models can advance

(01:51) and enhance your automations we're going to take a look at the benefits of generative AI tools and what it is that they offer as well as how these generative AI tools can enhance customer interactions while providing support and even tailoring error handling for very efficient ways of managing your automations within production so enhancing automations with large language models now in this era of Automation and Innovation large language models are are an amazing technology that we can all get on board with

(02:30) automation anywhere is Bridging the Gap between enhancing automations using generative AI now when it comes down to these large language models these are sophisticated AI tools that are trained on vast amount of data they allow for human interaction to be able to take place talking asking questions being able to generate responses and being able to obtain a response from this regarding of the prompt that you are going to be going in with now we're going to go through some use cases that's going to bridge this

(03:04) Gap later on down the line but when it comes to utilizing large language models with automation they can assist with aspects like your doc your document structuring if you are now working with unstructured data that's captured from documents before it would require a massive amount of human interaction in order to be able to work with this unstructured data now by leveraging large language models we can actually pass the data through these large language models to receive the fields that we need we can do query analysis to

(03:36) be able to make sure that if we let's say we are working with responses or we are working with feedback received by certain products that we've got we are able to go through this provide sentiment analysis within seconds around the product or the review that we're working with and specific responses if you're working within customer care and you have responses or questions that are coming in from your customers by leveraging Automation and large language models you are able to provide specific

(04:05) and tailored answers unique to the customer that is asking the question based off historic data that you're going to be working with now in order to be able to achieve the strongest responses from your large language models it's all about how you can tailor your response and your prompt to be able to make sure that you are Conta receiving all of the information correctly it's vital that your prompt is tailored correctly and we're going to go through this later on in this presentation but by integrating this

(04:41) with automation anyway you can receive direct responses that are very contextually correct around what it is that you are asking the large language model okay so when we start looking at the anatomy of a prompt lety over to you for this one yeah sure so when you look at anatomy of a prompt what what are you doing you're really just asking the language model to provide a response and when you do that you're trying to give it a context you know to say I want you to be for example if some of you have been into the um Bot games challenge

(05:19) that we have going on it's a product review option where you get to review credit cards you can say you are a credit card reviewer and I want you to review the following comment as a context then you can say you know from a review point of view I want you to just review the comment and give me whether it's a positive comment whether it's a negative comment or whether it's a neutral comment if that's one of the options that you want right then you can then also give it examples if you want

(05:44) to do that as well so you can give it a you know maybe two or three because you know you have three options positive neutral and negative you can give it three examples and then and then say this one is a positive this one's a negative this one's a neutral as part of that as well then maybe one of the things you want to do is you know as part of your automation you actually wanted to to provide a sing format so folks are familiar with Json for example you know because you want to use it forther Downstream to do some

(06:09) postprocessing so you might say yes I want uh positive neutral and negative but I actually want you to give me a Json with a status code with it so that then you can use it on in following um uh processing so those are really the the key the key pieces so it's it's the context it's the um example it's the format and also the instructions and one of the key things to realize here is that there are multiple things you can do with a language model but you always want to simplify your prompt the simpler

(06:40) the better and if that means having to break your prompts down into multiple steps even that because you know when I'm sure you guys are building um automation um Bots you always simplify your Bots to do simple things and put them all together same thing with a language model you want to do simple prompts even if you do multiple prompts so that that's the uh that's the essence of a good prompt design and here we have the anatomy of a prompt which is going over exactly what you was just saying there so here is a

(07:10) perfect example of how we would want to structure a prompt when working with automation anywhere to pass it through in order to be able to receive the response that we want to begin with we have the context and instruction and this builds the scene this is acting as somewhat like a story really making sure that the large language model is understanding in the context of what it is that we are asking in this circumstance we can see here that we're asking it to clearly outline each activity's duration and intensity which

(07:40) this are two pieces of the instruction that we're asking for we're tailing The Prompt exactly what it is that we're wanting to create which is a concise routine for a morning workout that does include stretching when we start looking at the format modifier this is a sample routine this means that the prompt is going to use this as an example of the format of how we want to receive that output back as you can see here we've given it an example and from the example within the format modifier itself that means that

(08:12) the response we're going to get from that is going to match this area here when we go down and we actually have a look at the example this is an example rather than being an example of how we want the data to come back to us this is an example of what it is that we're wanting to receive so as you can see we want a well-rounded routine that ensures that you start your day right so this is showing that we need to make sure that the routine that we get back is not just focused on one specific key area but is

(08:42) to be well-rounded now whenever it comes to a case of being able to tailor your prompt it's always vital that you test multiple outputs now with generative AI it's not always the case depending on your prompt that your response that you retrieve back will be the same each time I've seen circumstances with generative AI where you're going through it and you can get a consistent response eight times but without including the context in context and instruction your format modifiers and your examples you still

(09:15) leaving that bridge in order to be able to allow that large language model to interpret a different response and maybe output something that actually isn't contextually correct to what you are after so whenever you are going into your writing of a prompt you you always want to make sure that these key aspects are carried out and included and even if it means breaking it down like this just like Yi said you can build your prompt in sections asking for the feedback in different areas and even when it comes

(09:44) to automation these responses been can be captured in variables so if you're breaking this response here down into three separate areas them three areas can easily be stored into three separate variables and then strung together at the end to get your perfect response response likewise if you are asking your prompt let's say you have 10 separate questions that you want to ask you do not want to overflow the large language model with these questions so always take a step back with your prompts make sure you're asking are you asking the

(10:15) correct context and instruction are you including that format modifier and are you given an example also taking a note have I got multiple questions here that can be broken down or can it all go through correctly testing is key when it comes to this to be able to make sure that the response that you get back is correct and especially if you're integrating this within your automation Solutions now your tokenization so the way that these large language models work is via tokenization to be able to monitor each word

(10:51) sentence or punctuation and each one classes as a token now when it comes to leveraging generative Ai and large language models this is an area where you want to make sure you are keeping on top of because this is also that what can save you cost down the line now the more the more questions that you're asking your prompt and the more feedback and the more data that you're receiving from this means that the more tokens that this is going to use each token has a value of cost depending on the usage that you're going to be working with it

(11:23) as you can see on the example at the left here each word is assigned a single token what is one token time is another token it is is another token and so on and this is what allows the generative Ai and the prompt solution in order to be able to generate the response that is going to be cost effective to your solution so again when it comes to being able to manage your prompts and work within generative AI the tokenization aspect is key to keeping that cost down and streamlining the operation on a whole there is a tokenizer tool that is

(11:58) also embedded with in chat GPT this will allow you in order to be able to input your prompts before integrating them with automation anywhere and receive the response back in terms of how many tokens are being used and how many characters are being used now why is this important this is important because if you are integrating automation with generative AI you want to keep the cost as low as possible that means that your prompts want to be as low tokens as possible and limit the responses that you're retrieving back if you are

(12:32) extracting specific pieces of data from a document all you want as a response are the pieces of data that you are after you don't need filler words to say I have extracted this data for you because all of that is additional tokens that you don't need and this can all be stripped out by your effective prompt writing techniques again whenever you're going to be integrating automation with large language models utilizing a tool like this is going to make sure that you have a cost effective solution perfect now the temperature the

(13:14) temperature is what allows you to be able to make your prompt either give you a direct response back or get creative by tailoring the temperature if we have our temperature set to 0.2 you are going to have a direct response back it is not going to open it up for creativity in the response itself so an example of this if I have my out my temperature at 0.

(13:41) 2 and I say what is the current time where I'm at it's going to give me the current time for the location of where it is that I'm positioned however if I open up the creativity here and change my temperature let's say 0.8 or1 now the response that I get may also include other time zones your time zone is here which is 1 hour behind X place 2 hours behind y place and it allows for that elaboration on the data that you're receiving again this can be create if you're creating them specific responses to let's say customer reviews if

(14:11) somebody leaves a review for a product that you're wanting to sell and you're wanting it to make sure that response that you give leveraging a large language model and automation anywhere is creative friendly engaging by adding them keywords and increasing the temperature you giving your large language model flexib ility in the response that it can generate back to you which means that it's not going to seem like a robot is actually crafting this response and giving it back to the customer but it can actually start

(14:39) adding some human touches and even if you start threading it through with previous examples of chat logs that's already been received from the customer through to the client itself it'll be able to start matching the language patterns while still living that little bit of Leverage to get creative with the responses and generate that conversation in place one thing to remember with regards to this though is with that higher temperature as it says at the bottom you are less likely to be able to predict

(15:09) that response so if you are after a very direct response without leveraging too many tokens you want to keep that temperature at a low Point don't stay this yep there go so I think of this stuff we've already covered but um it's just giving you some guidelines on when you are structuring a prompt it's really been consigned to precise it just being I know I want you to do this one thing and do it if I need to do multiple things break it up into multiple prompts and then you can have a chain of prompts and obviously that

(15:46) allows you to also control uh some of the um parameters that Max just talked about around the tokens that come back or the number of characters and also you can then apply different temperatures as well to the different prompts depending on how predictive you want to be versus just being a more um more uh general for lack for better word as well um use action verbs say you know develop create um you are a AI assistant doing a credit card review just kind of place kind of place the uh llm into the world you want

(16:18) it to be would kind like the actor you want it to be as part of your prompt um design as well and then with constraint I talked about um the the bot game so we gave it three constraints we just want you to return positive neutral and negative so give it constraints if you know you actually wanted to return certain things give it a constraint if you know you wanted to return a stting format give it the format output you may even have to go as far as giving it an example of the output that you wanted to produce as well so the llm understands

(16:47) the actual final output you're looking for when you send the prompt in okay now when we look at the integration between automation anywhere and the generative AI tool that it is that you want to utilize we are very adaptable we're supporting trat GPT we're supporting as your GPT we have Google vertex Ai and there is Amazon Bedrock there as well now remember whenever you're using any of these platforms to the right hand side you will see these three key areas intelligence efficiency and Innovation now these large language

(17:28) models are extremely intelligent and that's why it's vital in order to be able to craft your prompt to receive only the data that you want to receive remember you want to make sure that the response is relevant to what it is that you're asking so be very intelligent when it comes to being able to craft that response and provide it all of the data needed in order to be able to generate the output that you are requiring the efficiency around large language models when streamlining this Within current use cases for automation

(18:01) we're going to cover a few aspects of this shortly down the line with a few examples that we've got but when it starts to come in your merg in automation with these large language models you can really start to streamline processes that before would have taken a large amount of time a great example for where this would actually come into play would be around that customer service or even a marketing campaign if you are launching a new marketing campaign you've gone through the prerequisites of knowing exactly okay

(18:33) this is the context that I want to include within my marketing campaign these are the graphics around this idea of what it is that I'm wanting to deliver large language models can be leveraged in order to be able to generate a lot of the marketing content for you in a very quick amount of time when passed over to a human this would then go to human review and human validation in order to be able to check that response and then start utilizing it the purpose of the large langage language models for areas like this

(19:00) isn't to streamline that entire process it's to be able to make a chunk of it a lot quicker and more efficient to be able to streamline your workflow and make your lives easier I always picture these large language models as a personal assistant imagine having a personal assistant that is proficient within language if you ask them in order to be able to create a document that outlines a summary of x amount of data within your records large language models can do that for you within seconds if you are utilizing different

(19:33) types of code JavaScript Python and you go through code reviews in order to be able to check that it follows best practice techniques large language models can help assist with that by passing forward the code they're enable to perform a code analysis likewise if you need a complex segment of code for something that you are currently building by giving the prompt the data in order to be able to tell it exactly what it is that you're wanting to achieve what language you're wanting to write that code in large language models

(20:04) can generate that code for you allowing you to then work on that and tailor it knowing that it's going to follow best practice techniques especially if you did provide a format modifier of how you are currently coding your inhouse technology and again Innovation with large language models you want to think outside of the box and you want to make sure that you're crafting that perfect prompt in order to be able to receive the information that you need but likewise in certain circumstances you may not want to limit what it is that

(20:35) you're receiving with regards to the output from that prompt so always think and always test whenever it comes around to actually receive in that output and the prompt that you put in so this is one of the use cases that we've got in play so GP services at the moment especially in the United States when you go and you tell your GP that something's wrong you've got a cough you've got a flu your symptoms actually all go down to a code so after you've been the that GP will have to go through and they're

(21:09) going to have to turn the notes that they've generated from your interview into specific pieces of code and load that through into specific documents like you see here CPT icd10 and hcpcs in order to be able to assign it for insurance purposes and to make sure that that money is generated now as a GP they are trained in this they have to go through extensive training in order to be able to understand and interpret these codes whereas this data is available within large language models so what we now see is GPS rather than

(21:43) having to spend hours after seeing each patient creating the documents adding the codes to the forms and being able to make sure that everything is correct for insurance purposes we can leverage large language models in order to be able to turn the notes into a structured format with all of the codes noted on there which means now the GP is only validating and going over a specific document rather than having to create it all from scratch after each patient as you can imagine with a fair few patients this is going to be a very timeconsuming

(22:14) job by leveraging generative AI this can all be done at the end of the day and streamline the process allowing the GP to potentially see more patients that single day now as I said we can really enhance automation with these large language models if we wanted to previously start taking automations to The Next Step by adopting complimentary technology integrating with python code using JavaScript or vbscript or leveraging microsoft.

(22:47) net you would have to have a complex software developer or some sort of developer in place in order to be able to generate this now you don't now we can leverage large language models again like I said to be able to generate this code in order to be able to use within your automations you can use large language models in order to make sure that if you have five or so developers all developing code these can make sure that all of them code follow best practice techniques and they can actually tell the developer exactly what part of the

(23:17) code either needs to be looked at and the issue itself so any developers in here that are that are used to see in them syntax errors that are popping up by leveraging large language models and able to put the code into these large language models stating the error message that we're receiving these large language models will be be able to tell us the issue at hand and even sometimes fix that problem for us again making troubleshooting a lot more efficient now when it comes to testing prompts for your automation consumption

(23:55) this part is key now when you're passing your prompts through from automation into your large language model you're going to be utilizing apis Now API calls when leveraging this technology is different to if you are going to be doing it manually so it's vital that for example what we see here is from chat GPT and this part is called the playground and what this will allow you to do is actually input your prompt exactly how you would be inputting it within automation the response that is then given here is an example response

(24:27) of exactly what you would receive by passing it through an API now again if you are leveraging your large language models and using automation you want to make sure that all of your prompts are tested before being put into production using the API tester that is all included this makes sure that your prompts are concise straight to the point and you are receiving the data that you want to receive every single time now human in the loop validation by leveraging large language models and Automation and working alongside a human we can really start to

(25:11) streamline specific actions in play so let's take for instance I won't go into the financial sector because we've got another example of that coming up but let's take for instance here a customer service representation so at the moment what that customer service rep is doing is they'll be receiving calls from the customer let's say managing specific complaints these complaints need to be managed within multiple systems that's coming into play and this complaints are all around a specific product now most

(25:39) of the time these complaints will come into either one centralized mailbox and they'll spend days going through each and every single one of them what we can start leveraging here is actually we can use large language models to be able to capture the complaint and generate that response the human in the loop would then rather than having to respond and contextually write a response back to that customer that's complaining now they're given a response they can tailor it they can make it more sincere they

(26:08) can make it more direct if they want to they can limit this just by utilizing a few clicks of a button now when we start leveraging large language models with complimentary Technologies let's say document automation or co-pilot then we get that more control we can integrate co-pilot into various Technologies and make make sure that the large language models that's running behind co-pilot are given the responses back to the human that is working on that project now when we start looking into the Realms of enhancing customer

(26:47) interactions and even how these large language models can support these people it's a case of efficiently speeding up the process of what the human is currently doing at that point in time you can see how we can intelligently classify Communications from information that's coming in so let's take for instance a live chat well a live a live Communication Center so rather than utilizing a chat bot you have that live chat feature that's open on a daily basis when somebody asks a question within that chat feature and you have a

(27:20) person on the other end of that that wants to give that human response the human would have to find the answer to the question that that person is asking now by leveraging co-pilot and large language models what we can now do is actually have the automation pick up the question that question can then be answered efficiently by the hu by the large language model and then they passed back to the human itself the human can then add a personal touch before responding and sending the response straight back to the person

(27:50) that's going to be asking that question we can reduce processing Times by making sure that the support all the questions that's in hand is linked through with large language models and generate your response within automation now a great area of this is around the financial sector again so handling customer complaints from various sources can be a very very tedious job and trying to get this into play compliance with f n r a rule 4530 you need to make sure that customer complaints are logged within 15 days

(28:27) before the the end of quarter depending on how many complaints that you've got in play that could be a very time consuming job and also a pricey job if you're actually needing to use humans in order to be able to focus and process these responses and complaints Now by leverag in again Automation and large language models automation can capture the complaints they can log them and make sure all of them in stored and large language models that can be leveraged for the actual response that they're going to be receiving and making

(28:58) sure that the complaint itself is contextually correct and it's actually accurate to what is being asked or complained about in that circumstance so when it comes to collaborating with generative AI large language models and automation anywhere as I said automation anywhere is bridging that Gap by providing you integrated packages within the platform to be able to directly link with these large language models so make the most of the automation anyway generative AI skills by following our boosters which is available on the Pathfinder Academy

(29:42) and this will be able to take you through the fundamentals of integrating your large language models in generative AI with automation anywhere again join the generative AI Pathfinder community group and this will allow you to be able to connect with other others within a forum and ask questions around your generative AI Solutions as you'll find that a lot of other customers and peir are going to be utilizing this for many many different operations and finally the collaboration is key this is what's going to present

(30:14) opportunities and it's going to allow you for efficiency to be able to grow your program even further now the Pathfinder group make sure you all experience Pathfinder today we start off by taking the Pathfinder assessment this assessment will be able to highlight key areas within your automation program that you can focus on in order to be able to scale once these areas have been identified we can start exploring the overall missions in order to be able to utilize and scale your automation program now the Pathfinder Academy will

(31:01) allow you to offer learning resources that's going to be able to advance your current developers and your automation program and show you how you can correctly integrate generative Ai and large language models with Automation and again that Pathfinder Community this is an overall forum and a community which allows everybody to be able to speak and ask questions to one another and work out how they can really start leveraging automation anywhere and generative AI in a way that we never saw was possible now I thank you all for your

(31:41) time today when it comes to going over generative AI large language models and integrating that within automation anywhere of course I'd definitely recommend taking a look at the automation anywhere platform especially under the generative AI packages and any information that you need on this can be found within our Pathfinder Community where we will be happy to answer any questions that you have

YouTube

https://www.youtube.com/watch?v=H9tb0n9xkbY