FOUNDATION OF DIGITAL BUSINESS

Surojit Mookherjee

Vinod Gupta School of Management

Indian Institute of Technology Kharagpur

Week 07

Lecture 33

Lecture 33 : Prompt Engineering

Good morning, continuing on my module 9 tools and techniques for AI, I take up a very interesting topic called prompt engineering. So, prompt engineering techniques and some generative AI resources. So, this is a very commonly used technology nowadays everybody is using it from school students, college students, university students and of course, everybody in their workplace also. So, it looks very apparently simple because you just type a query and you are expected But when you are one is a simple query like when is this flight or when I what I want to do or what is meant by something you do not understand your say some

topic in physics or chemistry or anything any subject you can give that or when was the you know what happened in the battle of panipat blah blah. So, these are all very simple educational queries questions which are your tools are very adapted responding to. But when you come to business application, then you need to take this up more. seriously as to prompt it is a technique. So, that is why it has got a name, but prompt engineering altogether I mean it is like a new subject like mechanical engineering, electrical engineering.

So, that it has been proved to be so useful, it has become so important and it is becoming so ubiquitous that in everyday life including our business life that it has been now classified as an engineering, prompt engineering as if it is a subject in which you can probably get a degree on. So, you are a prompt engineer qualifies. So, days are coming very soon when you will have like a software programmer or data analyst or business analyst you will have somebody who is titled as prompt engineer. So, all he is doing is evolving various methods of prompting and whole purpose is to get the right thing.

The challenge is you have a host of information available on the internet. that all of us know or these models have been trained. So, they can get it retrieved for you, but are they firstly the let us forget the hallucination part the information is coming out is correct or whatever that is fine you understand that, but is it the right one is it exactly what we want or is it we want something else something better or we meant this it understood that. So, that will always happens that is the contextual

thing which we always talk about. So, my context was this. However, the computer did not really understand what I wanted and gave me this. So, I keep on trying. So, it becomes an iterative method.

So, I can keep on trying playing with my prompt till the time I get a right response. It might happen the other way I get worse and worse and worse response. So, that is where this engineering thing is coming into or this is where some training is required. So, knowledge is required or it is not just simple straightforward playing with English words and English language I am assuming we are only using English for this business communication. So, it is not as simple as that because the English knowledge

also varies from individual to individual. So, I can express certain thing if the same thing is given to a group of people may be 10 or 20 in a company organization and each of them are been expressed to identify the problem or bring out the use case in one sentence or two sentence you will have may be 10, 15 different sentences and if you feed all of them separately to any GPT model. will get different outputs.

So, which is the right one? So, which is the best one? Again you need a human intervention because it is all augmenting the human knowledge with the outputs from the tool from the GPT tool and then we use our human experience and intelligence to find out identify which one is the most appropriate or it could be that none of them are appropriate enough. So, I will have to again change.

So, this is in a sense the background for the prompt engineering. So, is the process of creating effective prompts that enable AI models to generate responses based on given inputs. So, the key is effective prompts. Prompt engineering essentially means writing prompts intelligently for text based AI tasks, more specifically natural language processing tasks. They help in generating accurate responses by adding on some additional guidance for the model.

not generalizing a prompt too much make sure that the information added is not too much as that can confuse the model. So, as a input we are giving some data some information do not give to try to give too much of information in one prompt single prompt because that will confuse the model it may not be able to handle so many pieces of information. do not generalize a prompt too much, do not make it too general because then it is open to various interpretation and the model will jolly well go do all those multiple interpretations and make you confused.

Making the user intent and purpose clear for the model to generate content in the relevant context only, obvious that this is the most fundamental thing we want to do achieve, but it is not so easy as it is looks like. to make the user intent and purpose clear for the model, not for me I am clear what I want, but is my model is the computer is it on the same page with me or not that is the question or can I make bring it to same page. As an analogy imagine you are instructing a very talented, but inexperienced assistant you have hired some brilliant fellow.

come from an Ivy League or big school, very talented, but no experience. And then you want them or him to complete a task effectively. So, you need to provide clear instructions to him because he does not have any experience. So, prompt engineering is something similar. It is about crafting the right instruction called prompts to get the desired results from a large language model.

Like it is a student who has come, you have hired somebody who has lots of knowledge because he has come from the best university and he was the topper, ranker, etc. He knows everything, but then now the problem you have to put it in the right context so that he can give you the best answer. This is exactly parallel of how prompt engineering works with LLMs. So, the working of prompt engineering involves crafting the prompt, you design a prompt that specifies what you want the LLM to do. This can be a question, a statement or even an example.

The wording, phrasing and context you include all play a role in guiding the LLM's response. Understanding the LLM, different prompts work better with different LLMs. some techniques involving giving the LLM minimal instructions which is known as zero shot prompting in the language of prompting engineering while others provide more contextual examples which is few shot prompting. The point here is that different models behave differently.

So, if it means that if you give the same prompt to say chat GPT, perplexity or Gemini or Gork you will get different responses may or may not. Defining the prompt it is often a trial and error as I was telling you process you might the prompt based on the LLM output to get the kind of response you are looking for. So, you have to tweak it to get the best possible result or find in the prompt. Some of the applications of prompt engineering are language translation is the process of translating a piece of text from one language to another using relevant language models.

So, relevant prompts carefully engineered with information like the required script, dialect and other features of source and target text can help in better response from the model. So, you give it a particular script to translate and then you mention some features like what is the source, what is the region or area etcetera. So, all that will help a particular dialect if you mention this script is from this dialect will help the model to give a better translation output. It is just not simple translate this and give a text and you translate that.

So, that will be what we have seen might be something or quite better, but if you want to improve the color translation also add some metadata from this region or this dialect etcetera. Questions answering chatbots, a QA bot is one of the most popular NLP categories to work on these days. Forms on which an AI chatbot model is trained can largely affect the kind of obviously, responses it gets. So, whatever chatbot you are using behind that there is a. LLM model and your prompt is nothing but your chat whatever you type is nothing but a prompt.

So, this is one of the most commonly used prompt application which most of us are using it nowadays. Text generation, task can have multitude of application. Hence, it become again becomes critical to understand the exact dimension of the query. The text is generated for what purpose can largely change in tone, vocabulary as well as formation of the text. So, when you are trying to generate a text you have to be very careful about the prompt you are giving.

So, the tone everything can express a context the way you write the prompt. So, what are some of the prompt engineering techniques? So, as I told you prompt is an engineering topic, now you are calling it as prompt engineering. So, you are a prompt engineer. So, it is a playground that has all the tools to adjust your way of working with the big language models with specific purpose in mind.

So, it is just not a drafting of a prompt you have to think it is a playground that has all the tools you need to adjust your way of working with the models such with a particular purpose in mind. So, that you must be very first step one is the purpose should be clear in your mind what I want. So, some of the foundational techniques information retrieval which is where prompt is being used it is entails creation of a prompt so that the LLM can get its knowledge base and give out what is relevant. So, it is again something like a rag for example.

So, what we were talking in the previous session. Context application sorry context amplification gives supplementary context to the prompt in order to direct the understanding and attention of the LLM to its output. So, I want to augment anything something topic. So, I ask the tool the LLM to do it for me. So, add something more to whatever I know.

Summarization is induce the LLM to generate write the summaries about complex themes which we have discussed earlier. Reframing, rephrase your reminder to the LLM to consider a specific style or format for the output. it has given an output you do not like the formatting. So, you ask him to reframe this in or maybe shorten it in. So, maybe it has given a you can open things summarized and did it in 100 sentences or whatever long it is too long can you make it shorter or

can you use simpler English or maybe shorter sentences or can you break it up in 3-4 paragraphs depending on the different topics. So, that is reformatting reframing of the output and then iterative prompting is of course, the common one breakdown the complex tasks into smaller parts and then instruct the LLM sequentially in how to achieve the end result. So, it is a complex thing I can break it up first this then that sequence I can do I can I know this topic. So, I can break it up in the sequence and I want the prompt to respond in that sequence whatever I want.

So, the input to the LLM should also be in that particular sequence. So, what are the best practices of prompt engineering? It can largely be tuned by using a correct prompt, the question arises how can we make sure our prompt is right for the task at hand. So, this is what I want that this is the prompt I am giving whether it is the best prompt or right prompt etcetera. So, how do I find out that or what are the steps to ensure that the best practices.

So, begin with an objectives and goals. So, whenever interacting with a model the goal of the conversation and the objectives to be achieved via it should be absolutely clear even before one begins. So, step 1 is your goal clear what you want. So, go over that first to make it very this is exactly what you want. Relevant and specific data identification and usage, like every prompt and its objective should be described clearly.

Similarly, only absolutely relevant data should be used to train the model. One should make sure there is no irrelevant or unnecessary data in the training or the data you are using part of the prompt it should be absolutely relevant data. Do not use any unnecessary other extra data. If you think I give more data it will probably able to give me a better output not necessarily so. Focus on finding the relevant keywords that is very crucial relevant keywords make it is a huge difference in the type of response it generates.

The keyword used correctly in the right place can lead to a much different results altogether. So, this is very crucial thing could I identify the keywords for the prompt of course, based on the goals or objectives which I want. Make sure your prompts are very simple and clear, when crafting prompts it is important to keep them simple and clear by using plain language and avoiding complex sentence structure. Do not use any complicated English words, nobody is testing your English language capability like you are not a Shashi Tharoor for example, just to take a name reference. So, it is important to use very plain simple English.

Test and refine your prompts, the final step is to use a variety of test cases to evaluate the performance of the generated prompts and make accordingly you make adjustments to the prompt and so that you get a better output. The future of prompt engineering. It is a very recently developing an upcoming technology and hence it can actually serve to a very crucial part of most AI and NLP tasks and other areas as well. So, that is why people are treating this as a subject prompt engineering.

It becomes so useful, so common and so widely used. In AI and LLP domain technologies advance one expects to see significant improvements in the accuracy and effectiveness of prompts. With more sophisticated algorithms and machine learning models prompts will advance and be more particular to the specific use cases. Simple thing you can ask write me a program to do this software and it writes that. So, all of these applications will increase more and more once these tools become better and better.

Integration with other technologies. Prompt engineering is likely to become increasingly integrated with other technologies such as virtual assistants, chatbots and voice enabled devices. This is quite understandable like we are talking about all these assistance etcetera you give a prompt and it does something. One it tells you gives you information, but then said not only that it also executes some actions tasks completes some tasks like buying a plane ticket for booking a hotel for example. So, this will enable users to interact with technology more seamlessly and effectively improving the overall user experience.

So, more I can use it for doing simple task or whatever business task or personal task for me, the more I will be using it. Increased automation and efficiency, you can also expect to see increased automation and efficiency in the process along with more advanced prompts and streamlining the development of prompts and therefore, improving the output. So, these are all quite linked together. So, one of these things will lead to the other. So, the more powerful tool it becomes, more actions it can do, more tasks it can complete, more I will use in prompts and prompt engineering.

Some prompt engineering techniques. So, this is I just took one reference it is a 12, you will have another reference will give 20 techniques. It is a huge area and many people are just coming up with different ideas. Some of the names are quite generic commonly used shared. So, I have just pointed out some of them and I will describe one or two of them few of them for your understanding, but it is a huge area you can search and you will get any number of prompt techniques etcetera by different authors.

So, it is a definition engineering can be described as an art form creating input request for large language models that will lead to a n researched output. So, this can be kind of a high level definition for what is prompt engineering. So, these are the here they have listed some 12 prompt techniques. So, one of them is least to most prompting. So, I just try to explain this thing least to most prompting.

It decomposes a complex problem into series of simpler sub problems step by step we go and subsequently solving for each of these sub questions. So, you give small small break it up into small. Hence least to most prompting is a technique of using a progressive sequence of prompts to reach a final conclusion. Now to give you a very simple example, this example of course is very simple straightforward so easy to understand. So, I using a prompt to generate a specific type of story.

So, the least starting prompt could be an open ended request tell me a story. So, the AI tool will generate some story whatever it feels like to do does. Now you check add slightly more prompting adding a context tell me a story about an animal or it can be more specific context tell me a story about a squirrel or next stage could be you can add some key elements tell me a story about a talking squirrel, a squirrel which talks in human language tell me a story about that.

So, write a story for children. or adding a tone or genre, tell me a short humorous story about a talking squirrel. So, now I am adding not only a talking squirrel, make the story short and make it humorous. So, adding the tone of the story, most intrusive prompting. So, providing constraints and detailed instructions, adding specific plot, it could be like write

200-word story about a squirrel named Squeaky who can talk. The story should involve Squeaky trying to convince a human that he can talk, but the human thinks they are imagining it. End the story with Squeaky sighing and saying that humans are so dense. So, we start with that simple tone and then finally end with this, and finally, you will probably get a good story, hopefully. Self-ask prompting: this approach allows an LLM to give answers to a question it was not explicitly trained on.

The model might not have the direct answers to a question, but answers to sub-questions will exist in the LLM datasets. So, the initial question that answer probably is not part of the training data set is not there, but then you can break up sub questions and then gets small answers for each of those I mean answers for each of questions—and then finally come to a conclusion. Now, direct prompting: one example. If I take this one in the chain of thought, who lived longer, Theodore Hacker or Harry Vaughan Watkins?

This is the question. The answer is: Theodore Hacker was 65 years old when he died; Harry Watkins was 69 years old when he died. But it still did not answer who lived longer. You can gauge, of course—we know that 65 and 69. And he says, so the final answer: the name of the person is Harry Vaughan Watkins. So, who lived longer? So, it breaks it up like this.

Now, who was president of the US when superconductivity was So, it says superconductivity was discovered in 1911 by Heike Kamerlingh Onnes. Woodrow Wilson was president of the United States from 1913 to 1921. So, the final answer for the name of the president is Woodrow Wilson, which is a wrong answer. But if you ask the self-ask thing, if it was a self-ask, this was a chain-of-thought model prompting technique. Now, if you go to self-ask prompting, what it will do for the same question—who lived longer, Theodore Hacker or Harry Vaughan Watkins—it will break it up like this.

The follow-up questions will be: Are follow-up questions needed here? It will ask the tool in the self-ask mode, and your answer should be yes. So, follow-up: How old was Theodore Hacker when he died? The intermediate answer will give: Theodore Hacker was 65 years old when he died. Follow-up question: How old was Harry Hogan when he died? Intermediate answer: Harry Hogan Watkins was 69 years old when he died.

So, the final answer is Harry Hogan Watkins. So, it goes in this method, finally arriving at the correct answer. The next question is about the US president: Who was the president of the US when superconductivity was discovered? He will ask: Are follow-up questions needed here? It will be yes. So, follow-up: When was superconductivity discovered?

Intermediate answer: Superconductivity was discovered in 1911, but that is not my question. So, follow-up: Who was the president of the US in 1911? Intermediate answer William Howard Taft. So, the final answer is William Howard Tuft and this is the right answer. So, you see the methods is being used the self ask technique how it is progressing to answer two of this question which one was asked by direct prompting method

and other was chain of thought method prompting method. So, you will get different answers if you are using different prompting techniques. This was a simple example etcetera, but you can imagine a more complex scenario in a business application. So, you have to probably find out which technique will be more appropriate. Chain of thought prompting is a technique is a very popular nowadays you will hear it in many places.

The technique used to improve the reasoning availabilities of large language models. Instead of just asking for the final answer you prompt the model to generate a series of intermediate reasoning steps that lead to the solution much like how a human would work. We always try to break down and then a complex problem given we break it down simple same chain of thought prompting something very human like. COT prompt chain of thought prompt for summarizing a legal document, the prompt is you are tasked with summarizing a legal document. So, identify the document type and purpose, what type of legal document is this?

What is the primary purpose or objective of the document? Who are the primary parties involved? So, this could be the chain of thoughts step by step by step. If you ask these are these prompts then you will get the proper output. Extract key provisions or clauses.

What are the main terms conditions or rulings outlined in the document? Are there any critical obligations rights or responsibilities assigned to the parties or identify any deadline penalties or specific conditions? Determine the scope and context, what is the scope of the document? Are there any specific legal principles, statutes or precedents referenced? What is the broader context or background?

Highlight key outcomes or implications, what are the primary outcomes or effects of the document? Are there any significant consequences for non-compliance or breach? Who benefits and who bears the obligations? So, you can see how details it can be this chain of thought for any complex assignment like analyzing a legal document which is a complex assignment. Simplify and synthesize, condense the information into a clear and concise summary avoiding legal jargon where possible, focus on the most critical elements, ensure that summary is neutral and fact

based reflecting documents intent without adding personal opinions. So, these are prompts. So, you have to make a note of how these prompts are made to ensure the quality of the output. Review for accuracy and clarity. Does the summary accurately reflect the documents key points?

Is the language clear and accessible to someone without a legal background? And have you avoided copying the original text verbatim while preserving the meaning? So, this will come in the case of sometimes you can use plagiarism for example. So, do not just copy the text, but make it a different reword reframe so that the at least the context is maintained the meaning is preserved. The output format you define provide a summary of no more than 200 words structured as follows.

Document type and purpose, briefly state the type of documents and its primary purpose, identify the main parties involved, summarize the critical terms rulings or obligations, highlight the primary outcomes or consequences and note any important background or scope details. So, once you give these things you expect this summary of 200 words and giving all these main features which you want in the output. So, that was the chain of prompt and this is one of the most commonly used prompting techniques. So, you can study that a bit more detail.

Here to finish in today's lecture. I just collated some generative AI resources for you for your reference. So, you can go through this try these out and find out and given what are the primary functions mentioned here some is talk to AI with voice, chat GPT is a alternative to chat GPT, email assistance for writing mails, transcribing YouTube videos, article summarization. The cloud AI for example, it is quite a powerful one, it can do many things I mean it is a general purpose NLM, but one of the specialization is for article summarization, book

summarization query, presentation creation, presentation is you go to pitch.com you want to present make some PowerPoint presentation etcetera documents you can do it using these tools. Resume writing, CV writing for example, you can use resume.io for writing this I am sure many of you will need this, it is a very useful tool that way. Code to website conversion, creating web pages, AI content generation, content idea generation, text based content creation, language translation to Spanish, this very specific Spanish translation tool, I do not know where I got it from. And then video editing this is again many of us are now making videos if you make videos you need to edit videos.

So, some of these tools you can use video editing, image creation, rap music creation creating a video from images etcetera. Then Sora.com also from open eye is creating videos from text prompts. Some of these tools I have not checked all of them. So, they could be giving you free access for certain limits and beyond that probably you might have to pay. So, all those things are there and they may not be completely free or open source.

So, you have to check that out, but I just enlisted that these here. So, that you have a quick reference for you and plus I have given the overall reference from where I got these sites. So, with that I will end this session on prompt engineering and Gen-AI resources. Thank you very much.