Welcome to Lesson 2 of Module 3: An Introduction to Prompt Engineering.

This time I'm going to talk to you about what prompt engineering actually is. We're going to first dive in by giving you a precise and concise definition of prompt engineering. So what is prompt engineering? Said simply, it is the art and definitely in parentheses the approximate science of f talking to a computer, specifically one that's hooked into a Large Language Model (LLM), and especially in the context of a particular goal. In short, you're trying to do something and you are leveraging your computer powered by AI to do it.

Now the art and the science piece, this is a key point I want to hone in on. Working with AI, talking to your computer, doing prompt engineering, so to speak, is very much an art and it's not a science in the sense that there are highly specific and precise tricks that are going to stand the test of time that are going to allow you to use these language models the way you are today in the future. So I say it's an approximate science.

There are some techniques that probably will stand the test of time. Many won't as the capabilities increase and the technology evolves. I'm going to unpack the prompt and the engineering. We're going to break that out.

But first I want to talk about actually the third component – the problem. You can prompt, you can engineer your prompt, but what's even more important is getting very clear on your goal. What it is you're trying to do when you're working with this intelligent machine? Another way to say this is you're trying to go from point A where you are now to point B. There is some better future that you're trying to get to. Maybe it's over the long run. We're not at that point with AI right now where you're like, “Oh, I want to have a billion dollar business.” And then you just hit enter and it does the thing for you. Maybe it'll get to that point at some point. But at that point, I'm not sure how much money is going to matter.

But at the very least, getting from point A to point B is something that's more proximal. I want this particular prospect to respond to my email, to some kind of outreach. I want to get their attention. I want to attract their attention. I want to facilitate the beginning of some kind of authentic relationship that is ultimately going to be mutually beneficial for the purposes of expanding my business.

That's a pretty good point B. And the point here. Why would you use AI to accomplish this particular goal?

There are a few reasons. Certainly it could be less expensive to do it this way. You're spending a few cents or a few dollars on a couple of calls to AI versus hiring a BDR or a salesperson or something like that, in the case of the example I just described.

You could be improving quality in some cases. It definitely happens that there are certain types of tasks, administrative work, the creation of summaries and reports and things like that nobody really wants to do. It just might be the case that your AI system will be able to create something of higher quality than you would have created manually.

But there's a third benefit or dimension which is perhaps the most important, which is reducing the time to the result that you're trying to get to. The cycle time from going from point A to point B. Maybe it would have taken you an entire hour to write this email, to do all the research that you would need to do about the person, to watch a 30-minute interview with them to get a feeling for their vibe, or like interesting things that they would talk about.

You can decrease that time to result or the output that you're trying to create to get the result by an order of magnitude. Again, it would have taken you an hour to do the research and then write this email. You could, with AI assistance, bring that down to five or six minutes. Truly.

Prompting simply means working with the computer. Maybe you're giving it a command. You're telling it to do something. Or it could be that it's more of this discourse, this symbiotic relationship that you're cultivating to try to figure out how to get to point B.

Once that's clear, then it turns into more of a rote command with higher degrees of precision and you can start moving things into the engineering phase. And this is more advanced, but if there's a very wide cone of uncertainty, so to speak, around what point B is or where you need to be looking, you can certainly use AI to try to bring some clarity to that through a technique that I call “thought partnership” as well. We'll talk about that later.

So point A to point B and figuring out where and how to focus the limited attention that you and your AI assistant have. Single prompt, multiple prompt depends on the situation, but you're just sending these commands. You're talking to your computer in natural language to get a particular result.

So let me define prompting. You are figuring out how to use AI to generate outputs of sufficient quality. So there's a quality threshold that needs to get crossed. They get you from point A to point B significantly faster and at least commensurate cost with the status quo.

Quality is about the same or it's at least cost is about the same, or it could be a lot less depending on the situation. But the key point is you're doing it significantly faster. You're reducing the time to result, the time to capability.

If we switch over to the engineering side, once you figured out the prompt, the way in which you want to prompt and talk to the computer, work with the computer, then there's a question of, “Okay, am I going to be doing this again? Am I going to be applying the same technique to multiple problems that are close enough in terms of the characteristics that they have? Am I going from sending one prospecting email to 10 or a hundred or a thousand?” And depending on how you want to scale this up, just like with traditional software, you want to start focusing on reliability. You want to start focusing on generalizability, like will the prompt continue to work across these different cases?

And that's where the engineering part of it comes into play. A very specific definition on engineering. As you're figuring out how to use AI to generate particular outputs, given sufficiently specified inputs at higher levels of reliability. This is an awesome image that I pulled from Twitter from Andrej Karpathy who's a researcher over at OpenAI. I think formerly of Tesla as well. Really interesting guy. Super genius type who's a machine learning researcher.

And he shared this image on Twitter a while ago to flesh out the concept of prompt engineering.

Let me break it down. You'll see, going from zero to a hundred percent, in this case, he says task accuracy, but that would be similar to what I would say is quality. So there's a certain threshold where you would find output to be acceptable when you, you delegate the creation of that output to a machine.

And the question is, how do you get there? Now, the first step. This is simply coming up with the correct prompt or perhaps series of prompts. If you're breaking this down into steps, like I talked about in the previous slide, there are some specific ingredients and inputs. But you're really just trying to tell the computer, like in plain language, give it enough information such that then given that input, it's going to be able to produce the output. And this would be called a zero shot prompt.

Then you have few shot prompts. You basically figure out, “All right, here's some examples that I can show it of the output sort of gold standard, high quality outputs of the same type. That it can refer to when creating additional iterations of the same output.”

And then you start getting into more advanced stuff – retrieval augmented view shot prompting. Very complicated way of saying. Basically pulling more relevant contextual information into the conversation as you're working with a computer along with examples to get quality up a little higher.

And then we get to the stuff at the end. We're not even really talking about it all with fine tuning. And there's like other really bleeding edge, very expensive in terms of engineering labor and all this kind of stuff you can do with the models to eke out those last few points of reliability.

Now, just to just unpack this a little bit further. One of the points that he was making in the original tweet was there's a lot of discussion right now using small models, open source versus closed source, all that kind of stuff. If there's one thing to take away from this is anecdotally true. I experienced, there's also a lot of research that points to this.

Being the case, bigger is better. The more cutting edge, the bigger the model is. And frankly, this is going to be the stuff that's not open source that the AI labs and different big tech companies are going to be providing. Almost always, it's going to be the best thing to reach for. And instead of trying to train your own model and do this all in house by simply using the big model and putting a lot less effort in all the way upstream. And just getting that initial prompt, that's going to take you extremely far.

All right, so we talked about prompting. We talked about engineering your prompts. The idea of getting as fast as you can from point A to point B. You can't get from point A to point B immediately. Maybe you break it up into steps and then you follow that same process.

But that's essentially how this works. You're trying to get from point A to point B using a computer which you engage in a sort of discourse or dialogue using natural language. Giving it commands and then working through that process like I described. And that's pretty much it. And that's something that will continue to be true as the technology gets better and better.

Of course, let us know if you have any questions. I will see you in the next lesson.