Now, in this next lesson in your introduction to prompt engineering, we are going to talk about what goes in your toolkit.

And very important to note, this is an evolving toolkit. There are some pieces and parts where the manufacturer of the tool may change, but it's going to continue to be something that you reach to. We'll talk about this in a way that you should have some general principles that will last as the technology evolves, which is one of my big goa goals in teaching you how to think about prompt engineering. Okay.

So the first tool in your toolkit is the foundation model. So a foundation model is a very capital intensive piece of technology. It's a mathematical object that exists on a computer, weights and a matrix basically that allow you to give it a bunch of input. So in the case of a Large Language Model, those would be tokenized text inputs. And it provides the outputs, right?

So in ChatGPT, you give it a bunch of text. You get the text back. The literal thing that's doing that, there are layers on top of it where there's infrastructure and stuff going on but the core thing is the foundation model. And so when you're using ChatGPT or hopefully ChatGPT Plus, you're going to be accessing GPT 4, which is at least at the time of recording the most advanced publicly available foundation model.

There are some other ones. Most likely because of how expensive it is to create these things from scratch, both in terms of computation, which can be hundreds of millions of dollars, I think for GPT 4. And there's even some rumors and speculation because of how these scaling laws work, where just more is more. The more you stuff in and the more computation. The more cycles you run through to create the weights of the model, the higher quality the results are. The smarter it gets. It's entirely possible that it could cost billions of dollars just to create future iterations, GPT 5, 6, 7, 8.

So, this is something where there are going to be very few, at least initially, organizations that have the capabilities and the capital to create the models, but then we all benefit from it. So it's like a power company. Or something in that sense where you have these centralized sources of electricity generation, but then it's wired into all of our houses. I really like the electricity metaphor. I think it works pretty well.

Okay. This is a very important and necessary, but probably not sufficient tool in your toolkit. Yeah. GPT4 from OpenAI. OpenAI is one of the leading organizations that's creating these models. Then you've got Gemini from Google. That's an upcoming one. That's supposed to be a competitor to GPT4.

Claude from Anthropic is a spin off of OpenAI. So there's some OpenAI employees who started the thing. You've got Llama from Meta, formerly known as Facebook. They've gone the open source route, meaning they make the model way it's available to developers who want to host it and use it like on their own machines. Very different strategy than what you have with the other labs where they're keeping their models proprietary and making them available as a service.

And you've got GPT 3.5, Bison. These are models that are just of lesser capability than what you have at the time. Now there's the model itself, which is accessible through an API or in the case of Llama, you can literally get the literal weights, like the physical object in some sense, and put it on your computer. And then tinker with it and do whatever you want. But most people are going to be experiencing or taking advantage of these models through products, a consumer product like ChatGPT or enterprise software, like Microsoft 365 that has GPT4 hooked into it and things like that.

So you'll see these models, they're hooked into these different products, but behind the scenes, this is really what you're taking advantage of. It's the core intelligence makes these capabilities possible. All right. So that's important. It's necessary, but for you to take full advantage of these capabilities and become a good prompt engineer, there's a lot of other complimentary technology and tools that you probably want to consider incorporating into your toolkit.

Some you'll use more than others. The first one is a prompt repository. This is if you developed a prompt and you engineer the prompt for high reliability because you're going to have a lot of throughput. Meaning you're going to want to use that prompt over and over again to perform the same task.

Something I've said before, a CRM entry based on the transcript of a conversation with a prospect or something like that. Having a template, like a Notion page, or it doesn't really matter the format that you put it in. But having that stored somewhere so at the very least you can do an easy copy and paste is not a bad idea.

Some of the consumer applications are starting to make this easier as well. So ChatGPT has a little like share link so you can share the beginning of a conversation or whatever it gets to with somebody else and then they can continue it. That's another way to share some of your prompts.

And then in OpenAI Playground, which is really for developers to come up with prompts and and test some of the API capabilities through web interface. You can actually save those prompts like you can with chats and ChatGPT and use it that way.

So not a bad idea to start at least having a list of these things that you can come back to and use yourself. Or you can share with other people. Ultimately, some of those can be incorporated into full blown software, but that's outside the scope of this lesson.

Then there is the matter of dictation. Okay, I think this is just so incredibly important. I've talked before about how more is more. How one of the advantages of this technology is it allows you to interface a lot more naturally with your computer and just speak in natural language. Whisper is an open source model from OpenAI which is state of the art, meaning it's really the best speech to text model there is. And you can download it on your computer. You can run it locally. I've done it. I asked ChatGPT to actually tell me how to do it and run the code, so that was wild. But it's also incorporated increasingly into other products that you can use on your phone, on your computer for dictation.

The Windows and Apple speech to text capabilities are, frankly, at the time of recording, not that good. So I would highly recommend using Whisper. It's just gonna be less of a headache. You don't have to correct all these mistakes. And it works really well.

If you're on a Mac, I would recommend purchasing SuperWhisper. I think it's a five bucks a month and it allows you to do dictation. It's super accurate. And again, you can just provide more context. You can talk at length to provide the relevant information and give clear direction and natural language.

And then on the ChatGPT app on both iOS and Android, Whisper is incorporated. There's a little symbol to just speak into it. And it'll turn everything you say into text and you can input.

There's also an audio to audio feature coming soon as well. You can simply just talk naturally and the app will actually talk back to you, not just provide text. So, that's interesting. But anyway, across the board, anything that you can incorporate to make the experience more fluid and provide lots more context is going to be good.

Dictation is the first step in this direction. There's some other crazy hardware tools – ambient computing. These different consumer products that are coming to market that'll expand the scope of this where you're just automatically be collecting contextual information and feeding it into AI. But that's still probably a few years away from mainstream adoption.

So in the interim, having these tools for dictation is going to be super useful. Okay. Then another tool is Otter. This is one many. meeting AI note taker transcription tools that you can use. It'll join your internal meetings and your external meetings. It automatically creates summaries with AI assistance. Has notes you can share with people. You can make highlights. You can do all kinds of interesting things with it.

The reason why I like Otter versus all the other ones and the all the other ones are quite good as well, is it provides you with a live transcript. So if you're having a conversation with somebody either remotely or in person, you can just open the Otter app. Or have the note taker join your meeting. And then you can actually go to the conversation in the Otter interface and copy and paste what people are saying live, and then feed that into AI to do things with it.

So just as an example of how this is useful. I've had a few folks who I work with where I'm a subcontractor. They have a relationship with a bigger company and they want to bring me in to do some of the AI implementation work. We'll have a conversation over the course of 30, 40, 50 minutes. Work out all the details of the proposal that we'd like to get to the client, even role play it a little bit. And literally in real-time, I'll command A and command C, copy and paste the entire transcript from that conversation into an AI tool like Claude or ChatGPT or whatever it'll fit into and just say, “Hey, can you help us whip up the first version of this proposal?” And it does a really good job. And then work with it using the prompt engineering techniques that I shared before to get that really into shape.

So Otter is fantastic for collecting not just the meeting transcript, but making available to you live. So again, you can reduce the time to capability of getting the result that you want by feeding that data directly into AI.

And then, two other tools I'll share. One is called Descript which I'm using to record this presentation. Descript is pretty cool. And it's video and audio editing, but it automatically transcribes everything. So you always have the transcript as soon as you've done a screen recording or a recording in general.

And what I'll often do is if I'm trying to explain a process, or something long form and I need to show and I'm going to speak naturally to try to work through the the steps, I will make the recording. And then what I'll do is I'll take the transcript of what I've said, feed it through AI, create summaries, say, “Hey, I want you to turn this into a Slack message I can send. Include a TLDR, Too Long Didn't Read headline at the top with some bullet points of action items or, you know, key takeaways or something like that.”

And I will send that text that it produces, which is a summary of what I've said, to somebody in an email or a Slack message or whatever the appropriate medium is, and then include a link to the original video if they want to revisit the very source and see for themselves and go deep.

Descript is pretty fantastic for that. Really streamlines the process of communicating effectively and documenting processes in a lot of cases.

Okay, and then Rewind AI is pretty interesting. I'll just show you this quickly, but it's hard to see. But here you can see I'm actually like going. There you go. I'm editing stuff.

It's a digital memory, so it records everything you say, hear, do on your computer. They're expanding to other platforms. It's on a Mac only at the moment, but it provides you with a digital memory that you can query in natural language.

Again, I've talked about more is more. Context is the most important complement to commoditized artificial intelligence. And so just recording everything, you want to make sure you take privacy and security extremely seriously and really get that locked down. But if you can create this digital repository, you're just going to have a lot more material that you can work with and use AI more effectively.

So, those are basically the tools. You've got the foundation models. You've got the complementary tools that help you use those capabilities a lot better.

If you have any questions, feel free to reach out and ask, and I'll see you in the next lesson.