CHAPTER 5
Prompt Engineering

Prompt engineering refers to the process of crafting an instruction that gets a model to generate the desired outcome. Prompt engineering is the easiest and most common model adaptation technique. Unlike finetuning, prompt engineering guides a model's behavior without changing the model's weights. Thanks to the strong base capabilities of foundation models, many people have successfully adapted them for applications using prompt engineering alone. You should make the most out of prompting before moving to more resource-intensive techniques like finetuning.

Prompt engineering's ease of use can mislead people into thinking that there's not much to it.[1] At first glance, prompt engineering looks like it's just fiddling with words until something works. While prompt engineering indeed involves a lot of fiddling, it also involves many interesting challenges and ingenious solutions. You can think of prompt engineering as human-to-AI communication: you communicate with AI models to get them to do what you want. Anyone can communicate, but not everyone can communicate effectively. Similarly, it's easy to write prompts but not easy to construct effective prompts.

Some people argue that "prompt engineering" lacks the rigor to qualify as an engineering discipline. However, this doesn't have to be the case. Prompt experiments should be conducted with the same rigor as any ML experiment, with systematic experimentation and evaluation.

The importance of prompt engineering is perfectly summarized by a research manager at OpenAI that I interviewed: "The problem is not with prompt engineering. It's

1 In its short existence, prompt engineering has managed to generate an incredible amount of animosity. Complaints about how prompt engineering is not a real thing have gathered thousands of supporting comments; see 1, 2, 3, 4. When I told people that my upcoming book has a chapter on prompt engineering, many rolled their eyes.