The Rise of AI Engineering

The scaling up of AI models has two major consequences. First, AI models are becoming more powerful and capable of more tasks, enabling more applications. More people and teams leverage AI to increase productivity, create economic value, and improve quality of life. Second, training large language models (LLMs) requires data, compute resources, and specialized talent that only a few organizations can afford. This has led to the emergence of *model as a service*: models developed by these few organizations are made available for others to use as a service. Anyone who wishes to leverage AI to build applications can now use these models to do so without having to invest up front in building a model. In short, the demand for AI applications has increased while the barrier to entry for building AI applications has decreased. This has turned *AI engineering*—the process of building applications on top of readily available models—into one of the fastest growing engineering disciplines. This chapter begins with an overview of foundation models, the key catalyst behind the explosion of AI engineering. I'll then discuss a range of successful AI use cases, each illustrating what AI is good and not yet good at. As AI's capabilities expand daily, predicting its future possibilities becomes increasingly challenging. However, existing application patterns can help uncover opportunities today and offer clues about how AI may continue to be used in the future.

From Language Models to Large Language Models

The statistical nature of languages was discovered centuries ago. In the 1905 story "The Adventure of the Dancing Men", Sherlock Holmes leveraged simple statistical information of English to decode sequences of mysterious stick figures. Since the most common letter in English is *E*, Holmes deduced that the most common stick figure must stand for *E*.

The set of all tokens a model can work with is the model's vocabulary. You can use a small number of tokens to construct a large number of distinct words, similar to how you can use a few letters in the alphabet to construct many words. The Mixtral 8x7B model has a vocabulary size of 32,000. GPT-4's vocabulary size is 100,256. The tokenization method and vocabulary size are decided by model developers.

From Foundation Models to AI Engineering

AI engineering refers to the process of building applications on top of foundation models. People have been building AI applications for over a decade—a process often known as ML engineering or MLOps (short for ML operations). Why do we talk about AI engineering now? If traditional ML engineering involves developing ML models, AI engineering leverages existing ones. The availability and accessibility of powerful foundation models lead to three factors that, together, create ideal conditions for the rapid growth of AI engineering as a discipline: *Factor 1: General-purpose AI capabilities* Foundation models are powerful not just because they can do existing tasks better. They are also powerful because they can do more tasks. Applications previously thought impossible are now possible, and applications not thought of before are emerging. Even applications not thought possible today might be possible tomorrow. This makes AI more useful for more aspects of life, vastly increasing both the user base and the demand for AI applications. For example, since AI can now write as well as humans, sometimes even better, AI can automate or partially automate every task that requires communication, which is pretty much everything. AI is used to write emails, respond to customer requests, and explain complex contracts. Anyone with a computer has access to tools that can instantly generate customized, high-quality images and videos to help create marketing materials, edit professional headshots, visualize art concepts, illustrate books, and so on. AI can even be used to synthesize training data, develop algorithms, and write code, all of which will help train even more powerful models in the future. Because of the resources it takes to develop foundation models, this process is possible only for big corporations (Google, Meta, Microsoft, Baidu, Tencent), governments (Japan, the UAE), and ambitious, well-funded startups (OpenAI, Anthropic, Mistral). In a September 2022 interview, Sam Altman, CEO of OpenAI, said that the biggest opportunity for the vast majority of people will be to adapt these models for specific applications. The world is quick to embrace this opportunity. AI engineering has rapidly emerged as one of the fastest, and quite possibly the fastest-growing, engineering discipline. Tools for AI engineering are gaining traction faster than any previous software engineering tools. Within just two years, four open source AI engineering tools (AutoGPT, Stable Diffusion eb UI, LangChain, Ollama) have already garnered more stars on GitHub than Bitcoin. They are on track to surpass even the most popular web development frameworks, including React and Vue, in star count. Figure 1-6 shows the GitHub star growth of AI engineering tools compared to Bitcoin, Vue, and React. A LinkedIn survey from August 2023 shows that the number of professionals adding terms like "Generative AI," "ChatGPT," "Prompt Engineering," and "Prompt Crafting to their profile increased on

Foundation Model Use Cases

The number of potential applications that you could build with foundation models seems endless. Whatever use case you think of, there’ s probably an AI for that.10 It’s impossible to list all potential use cases for AI.