1

What Are Transformers?

Transformers are industrialized, homogenized Large Language Models (LLMs) designed for parallel computing. A transformer model can carry out a wide range of tasks with no fine-tuning. Transformers can perform self-supervised learning on billions of records of raw unlabeled data with billions of parameters. From these billion-parameter models emerged multimodal architectures that can process text, images, audio, and videos.

ChatGPT popularized the usage of transformer architectures that have become general-purpose technologies like printing, electricity, and computers.

Applications are burgeoning everywhere! Google Cloud AI, Amazon Web Services (AWS), Microsoft Azure, OpenAI, Google Workspace, Microsoft 365, Google Colab Copilot, GitHub Copilot, Hugging Face, Meta, and myriad other offers are emerging.

The functionality of transformer models has pervaded every aspect of our workspaces with Generative AI for text, Generative AI for images, discriminative AI, task specific-models, unsupervised learning, supervised learning, prompt design, prompt engineering, text-to-code, code-to-text, and more. Sometimes, a GPT-like model will encompass all these concepts!

The societal impact is tremendous. Developing an application has become an educational exercise in many cases. A project manager can now go to OpenAI's cloud platform, sign up, obtain an API key, and get to work in a few minutes. Users can then enter a text, specify the NLP task as Google Workspace or Microsoft 365, and obtain a response created by a Google Vertex AI or a ChatGPT transformer model. Finally, users can go to Google's Gen App Builder and build applications without programming or machine learning knowledge.

The numbers are dizzying. Bommasani et al. (2023) created a Foundation Model ecosystem that lists 128 Foundation Models 70 applications, and 64 datasets. The paper also mentions Hugging Face's 150,000+ models and 20,000+ datasets! The list is growing weekly and will spread to every activity in society.

Where does that leave an AI professional or someone wanting to be one? Should a project manager choose to work locally? Or should the implementation be done directly on Google Cloud, Microsoft Azure, or AWS? Should a development team select Hugging Face, Google Trax, OpenAI, or AllenNLP? Should an AI professional use an API with practically no AI development? Should an end-user build a no-code AI application with no ML knowledge with Google's Gen App Builder?

The answer is yes to all of the above! You do not know what a future employer, customer, or user may want or specify. Therefore, you must be ready to adapt to any need that comes up at the dataset, model, and application levels. This book does not describe all the offers that exist on the market. You cannot learn every single model and platform on the market. If you try to learn everything, you'll remember nothing. You need to know where to start and when to stop. By the book's end, you will have acquired enough critical knowledge to adapt to this ever-moving market.