前言

亲爱的读者朋友：

您好！欢迎阅读这本《面向开发者的 LLM 入门教程》。

最近，以GPT-4为代表的大规模预训练语言模型备受关注。这些模型拥有数十亿到千亿参数，通过学习大规模文本语料库，获得了十分强大的语言理解和生成能力。与此同时，OpenAI等公司推出的API服务，使得访问这些模型变得前所未有的便捷。

那么如何运用这些强大的预训练模型开发实用的应用呢?本书汇聚了斯坦福大学的吴恩达老师与OpenAI合作打造的大语言模型（LLM）系列经典课程，从模型原理到应用落地，全方位介绍大模型的开发技能。

本书首先介绍 Prompt Engineering 的方法，提示是连接用户与模型的桥梁，优化提示对模型效果至关重要。通过案例，读者可以学习文本总结、推理、转换等基础NLP任务的Prompt设计技巧。

然后，本书指导读者基于 ChatGPT 提供的 API 开发一个完整的、全面的智能问答系统，包括使用大语言模型的基本规范，通过分类与监督评估输入，通过思维链推理及链式提示处理输入，检查并评估系统输出等，介绍了基于大模型开发的新范式，值得每一个有志于使用大模型开发应用程序的开发者学习。

通过对LLM或大型语言模型给出提示(prompt)，现在可以比以往更快地开发AI应用程序，但是一个应用程序可能需要进行多轮提示以及解析输出。在此过程有很多胶水代码需要编写，基于此需求，哈里森·蔡斯 (Harrison Chase) 创建了一个用于构建大模型应用程序的开源框架 LangChain ，使开发过程变得更加丝滑。

在Langchain部分，读者将会学习如何结合框架 LangChain 使用 ChatGPT API 来搭建基于 LLM 的应用程序，帮助开发者学习使用 LangChain 的一些技巧，包括：模型、提示和解析器，应用程序所需要用到的存储，搭建模型链，基于文档的问答系统，评估与代理等。

当前主流的大规模预训练语言模型，如ChatGPT（训练知识截止到2021年9月）等，主要依赖的是通用的训练数据集，而未能有效利用用户自身的数据。这成为模型回答问题的一个重要局限。具体来说，这类模型无法使用用户的私有数据，比如个人信息、公司内部数据等，来生成个性化的回复。它们也无法获得用户最新的实时数据，而只能停留在预训练数据集的时间点。这导致模型对许多需要结合用户情况的问题无法给出满意答案。如果能赋予语言模型直接访问用户自有数据的能力，并让模型能够实时吸收用户最新产生的数据，则其回答质量将能大幅提升。

最后，本书重点探讨了如何使用 LangChain 来整合自己的私有数据，包括：加载并切割本地文档；向量数据库与词向量；检索回答；基于私有数据的问答与聊天等。

可以说，本书涵盖大模型应用开发的方方面面，相信通过本书的学习，即便您没有丰富编程经验，也可以顺利入门大模型，开发出有实用价值的AI产品。让我们共同推进这一具有革命性的新兴技术领域吧！

如果你在学习中遇到任何问题，也欢迎随时与我们交流。

祝您的大模型之旅愉快而顺利！