**《这本书，让你比90%的人更懂ChatGPT背后的技术！》**

-by 田浩

在当今这个AI技术日新月异的时代，生成式AI以其强大的内容生成能力，正在逐步改变我们的生活方式和工作模式。不管是用于学习还是用于解决日常的办公事项，大模型已经开始逐渐代替了传统的搜索引擎（对于我个人来言目前大部分的方案搜索基本都使用大模型了）。

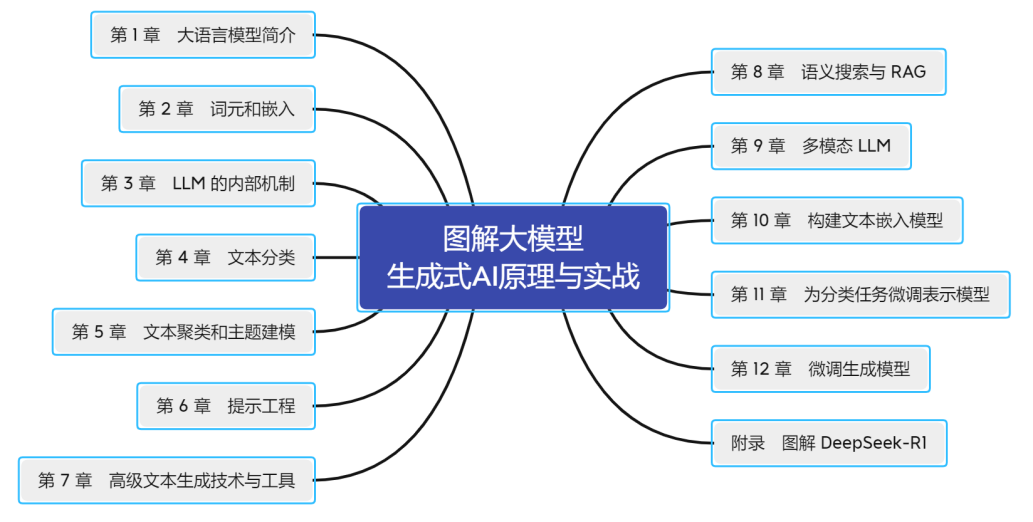
很多人觉得AI 或者大模型的相关技术很深奥，晦涩难懂。诚然，这些AI技术在一定程度上是有门槛的，但是也不是攻不可破的，Jay Alammar与Maarten Grootendorst强强联合，为我们读者打造了一本可以深入了解生成式AI原理的著作---《图解大模型：生成式AI原理与实战》。



**为什么要读这本书？**

**内容非常丰富**

我们先来了解一下这本书的内容部分:



**依据个人的理解，我来总结一下每一部分的内容：**

**第一部分：**第1章-第3章。

在这一部分作者先介绍语言AI的发展历程（词袋模型→嵌入→注意力机制→Transformer架构），然后对比开源与专有模型，并演示基础文本生成。接着介绍了词元与嵌入的相关基础的知识，解析分词技术（词级、子词级、字符级）及不同LLM的分词器差异性，通俗易懂。然后讲述了LLM内部机制，详解了Transformer架构（输入输出、前向传播、注意力机制、位置编码等）剖析了一些优化技术（如RoPE位置编码、高效注意力机制）。

**第二部分：**第4章-第9章。

在这一部分，主要讲述如何使用预训练语言模型。作者从文本分类出发，介绍了基于表示模型（如BERT）和生成模型（如T5、ChatGPT）的分类方法，涵盖监督与无监督两大常见的场景，同时讲述了使用嵌入降维（如UMAP）和聚类（如HDBSCAN）实现主题建模的方法和技术要点。在提示工程方面，系统性的分析了基础提示到高级技术（思维链、自洽性、思维树）的技术特点和演变之路。在核心的文本生成上，讲述了如何利用LangChain构建链式流程、记忆机制和智能体系统，扩展LLM强大的推理功能。语义搜索与RAG方面则实现稠密检索、重排序及检索增强生成（RAG）。在多模态LLM的技术上，作者介绍了整合视觉与文本模型（如CLIP、BLIP-2），用于高效处理图像描述、多模态对话等任务。

**第三部分：**第10章-第12章。

第三部分的重点是训练与微调语言模型。作者介绍了基于对比学习（如SBERT）训练嵌入模型，涵盖监督、无监督（TSDAE）微调方法，基于这些方法我们可以很好的构建文本嵌入模型。在微调模型方面有微调表示模型、微调生成模型。微调表示模型针对分类任务微调BERT，涉及少样本学习（SetFit）和命名实体识别（NER）的相关任务。微调生成模型则介绍了分阶段训练（预训练→监督微调→偏好调优），使用QLoRA、DPO等技术优化生成质量与对齐。

通过我简单的总结可以知道，这本书的内容十分丰富，基本上每一章都是不可或缺的关于大模型的核心知识。每一章节都是必读内容。

也就是说学完这本书我们就可以掌握如下的知识点：

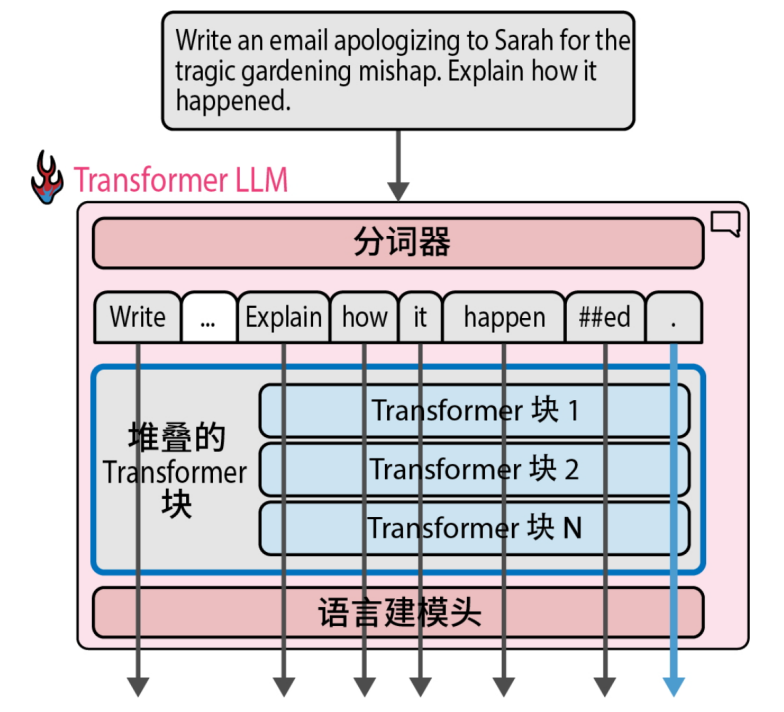
1. Transformer架构及其在文本生成与表示中的优势。
2. 构建高级大模型处理流程，用于文本聚类并挖掘其中的主题。
3. 打造语义搜索引擎，借助稠密检索、重排序等方法，实现超越关键词的搜索能力。
4. 探索生成模型的使用方式，从提示工程到检索增强生成。
5. 通过监督微调，对比学习和上下文学习，训练并优化适用于特定任务的大模型。

**内容极其细致**

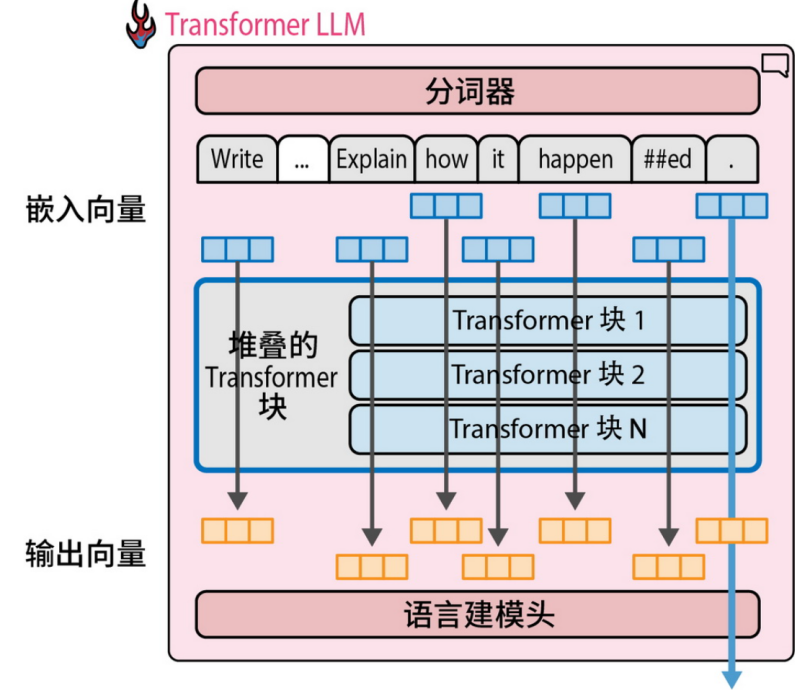
很多关于大模型的相关图书的内容也很丰富，不过一到讲解一些核心的知识点的时候却又显得“力不从心”，一种点到为止的感觉。不够具体也不够细致。而本书打破了这一尴尬点。

比如在第3章节讲述并行词元处理和上下文长度的时候，作者很细致的使用图解的方式来进行阐述：

1. 每个词元都通过自己的计算流进行处理：



1. 每条处理流接收一个向量作为输入：



这种使用递进式的图解讲解的方式，让读者能在第一时间了解模型处理文本的细节，能更好的理解模型的运行原理和机制。配上相关的公式、文字以及代码的讲解将LLM相关知识剖析得十分透彻。

**这本书特点如何？**

这本书不仅仅内容丰富，还有很多其他的特点：



关于这些特点，我觉得大家买到这本书，随便翻阅一下就能非常直观的感受到，无需多言。

**如何学习这本书？**

这本书的学习计划应该是这样的比较合理：



**第一阶段：理解语言模型（第1-3章）**

掌握大模型的基础概念，如分词、嵌入、Transformer架构等基础的知识。

（1）结合书中的300+幅图解，理解注意力机制、位置编码等核心概念。

（2）借助提供的代码示例（如分词器实验），加深对Tokenization的理解。

（3）使用Google Colab（科学上网）的环境运行示例。

**第二阶段：应用预训练模型（第4-9章）**

掌握使用大模型解决实际问题，如文本分类、聚类、提示工程等。这一章节在于实践。

（1）掌握BERT做文本分类（情感分析）。

（2）基于BERTopic的主题建模。

（3）LangChain构建对话机器人等工程实践。

**第三阶段：训练与微调模型（第10-12章）**

掌握如何优化大模型，如微调嵌入模型、分类任务优化等。

（1）基于QLoRA代码进行模型微调。

（2）尝试在Hugging Face上部署自己的微调模型。

**哪些人适合这本书？**

一本好的书的受众应该是广泛的，但是呢，也要有一定的针对性。那么这本书适合哪些读者呢？

1. 机器学习爱好者：机器学习爱好者借此机会提升大模型相关的技术
2. LLM 专业人士：快速提升开发与应用能力，掌握最新技术动向。

（3）授课老师：这本书作为大学里关于大模型的课程书籍也是非常不错的选择。

**个人感悟**

在阅读本书的过程中，我深刻感受到了作者对于生成式AI原理和应用的深入理解和独到见解。作者通过大量精美的图解和生动的案例，将复杂的原理和应用场景清晰地呈现在读者面前，使读者能够在轻松愉快的氛围中掌握生成式大模型的相关知识。作者还注重引导读者思考和实践，通过具体的代码实现和案例分析，激发了读者的动手能力和创新思维。

通过阅读《图解大模型：生成式AI原理与实战》，读者不仅能够掌握生成式AI的核心原理和技能，更能够在实践中不断提升自己的创新能力和解决问题的能力。因此强推这本书。

最后附上作译者的信息，以彰显本书的强大之处：



