提示工程

日期：2023年3月15日 | 预计阅读时间：21分钟 | 作者：Lilian Weng

提示工程（Prompt Engineering），亦称**上下文提示**（In-Context Prompting），涉及如何与大语言模型（LLM）进行交流，通过不更新模型权重的方式引导其行为，以获得预期的结果。这是一门基于实验的科学，不同模型之间提示工程的效果可能大相径庭，因此需要大量的试验和启发式方法。

本文专注于自回归语言模型的提示工程，不包括完形填空测试、图像生成或多模态模型。提示工程的核心目标是实现模型的对齐和可控制性。详情可参阅我之前关于可控文本生成的文章。

[我个人的辛辣观点] 在我看来，一些关于提示工程的论文没必要长达8页，因为这些技巧可以简单地在一句话或几句话内阐明，剩下的大都是基准测试。一个易于使用且共享的基准测试设施对社区更有益。而设置迭代提示或使用外部工具并非易事，让整个研究界接受这一做法也同样困难。

基础提示

零样本（Zero-shot）和少样本（Few-shot）学习是两种基本的模型提示方法，这两种方法最早由多篇大语言模型论文提出，并广泛用于评估大语言模型的性能。

零样本

零样本学习指的是直接将任务文本输入模型并询问结果。

（所有情感分析示例均来源于SST-2）

```
文本：我敢打赌视频游戏比电影有趣多了。
情感：
```

少样本

少样本学习通过提供一系列高质量的示例，每个示例都包含目标任务的输入和期望输出，帮助模型更好地理解人类的意图和评价标准。因此，少样本学习通常比零样本学习有更好的表现。但这种方法会消耗更多的Token，并且当输入和输出文本较长时可能触及上下文长度的限制。

```
文本：（劳伦斯在舞台上）到处跳跃，跳舞，跑步，出汗，擦脸，充分展现了他初次成名时的惊人才华。
情感：正面

文本：尽管所有迹象都表明相反，这部糟糕的电影还是设法冒充了一部正式的商业影片，收取全价门票，在电视上大肆宣传，自称能够娱乐小孩和成人。
情感：负面

文本：多年来第一次，德尼罗在情感上有了深刻的挖掘，可能是因为受到了合作伙伴出色表现的启发。
情感：正面

文本：我敢打赌视频游戏比电影有趣多了。
情感：
```

许多研究尝试了如何构建上下文示例以最大化性能，并发现**提示格式、训练示例及其顺序的选择可以极大地影响模型的性能**，从几乎是随机猜测到接近最优状态的表现。

Zhao等人（2021年）研究了少样本分类，并发现几种与LLM（他们使用的是GPT-3）相关的偏见导致了性能的高变异性：（1）*多数标签偏见*，如果示例中的标签分布不平衡；（2）*近期偏见*，模型可能倾向于重复最后出现的标签；（3）*常见Token偏见*，LLM倾向于生成常见的Token而不是罕见的Token。为了克服这些偏见，他们提出了一个方法，即当输入字符串为`N/A`时，调整模型输出的标签概率为均匀。