

模型训练

- 在开始实际训练之前，确保代码、环境和数据都是正确的。
- 检查训练环境的最大运行超时时间。检查云平台的余额。确保没有任何因素会中断训练。
- 确保如果发生中断，我们能够恢复任务。中间结果不会丢失。
- 如果代码有问题，训练过程中肯定会暴露出来。
- 如果数据不干净，肯定会影响后期阶段。
- 采用迭代方法。按数量级扩展。先训练数百万个标记，然后是数十亿个标记，接着是数万亿个标记。
- 如果训练结束，会发生什么？训练的最终处理是否正确完成？
- 从他人的错误中学习，比如 Llama2 所犯的 error。
- 不要只修复结果，要解决问题的根本原因。