# 2.预训练

## 1. 为什么要增量预训练?

**预训练学知识**，**指令微调学格式**，**强化学习对齐人类偏好**，所以要想大模型有领域知识，得增量预训练（靠指令微调记知识不靠谱，不是几十w条数据能做到的）。

## 2. 进行增量预训练需要做哪些准备工作?

1. **选取底座模型**：可以根据自己的项目需求和硬件基础来选择合适的底座模型及模型参数量的大小。
2. **收集数据**：一般来说需要收集大量的文本数据，包含各个领域，主要从互联网上获取，一般预训练数据的大小都是 TB 级别的。
3. **数据清洗**：所有的信息都能够在互联网信息中被找到，只是**信息密度**相比「人工精选数据集」要更低。例如「明星信息」、「如何写代码」这些信息都能在新闻网站、或是问答网站中找到，只不过「维基百科」或是「Github」则是将这些信息给「高密度」且「结构化」地进行了存储。这使得我们在使用维基百科作为训练语料的时候，模型能够更快的学习到这些高密度信息（人物的经历、年龄、性别、职业等等），而这些内容在互联网信息（如新闻）中的信息密度则较低，即很少会有一条新闻完整的介绍一个艺人的过往经历。只要我们**对互联网信息进行严格的处理**（去除冗余信息，提高有用信息的密度），就能够加快模型的学习速度。

## 3. 增量预训练所用训练框架?

- **超大规模训练**：选用 3D 并行，Megatron-Deepspeed拥有多个成功案例
- **少量节点训练**：选用张量并行，但张量并行只有在 nvlink 环境下才会起正向作用，但提升也不会太明显。
- **少量卡训练**：如果资源特别少，显存怎么也不够，可以使用 LoRA 进行增量预训练。

## 4. 增量预训练数据选取思路有哪些?

垂直领域预训练有三种思路：

- 先用大规模通用语料预训练，再用小规模领域语料二次训练
- 直接进行大规模领域语料预训练
- 通用语料比例混合领域语料同时训练