

讲座 3: 大语言模型如何学习新语言

大语言模型:

1. 数据资源是影响 LLMs 性能的主要因素之一
2. 主要由英语等少数资源丰富语言所主导
3. 很多语种在当今主流的 LLMs 中并没有很好的体现, 编码效率低
4. 很多语言在 NLP 与 LLMs 资源中不受重视, 恶性循环
5. 现有的处理流程对资源丰富的语种更有利, 可能导致数据污染

大语言模型 + 低资源语言

- 充分利用大语言模型自身的语言能力
- 需要大语言模型学习目标语言知识

NO MAGIC → 质量优先

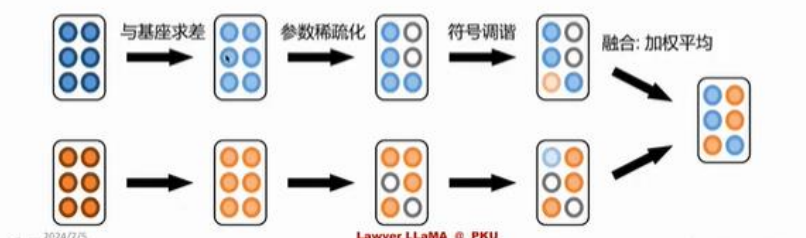
- 数据来源白名单
 - 尽量避免相似语言之间的影响
 - Github Copilot 生成网页内容主题抽取代码
- 充分利用现有资源
 - CulturaX, Wikipedia, NLGIW 2023 Shared Task
 - 二次确定语种, 降低识别错误
 - 约占 15%
- 精细化去重、过滤
 - URL、精辟去重、模糊去重策略相结合
 - 去噪、隐私数据
- 利用基础 LLMs “**在线学习**” 目标语言知识
 - 词法: 分辨词义
 - 句法: 分辨浅层语法信息
 - 综合: 遵循指令综合利用语言知识

数据资源更为稀缺的语言

- 对部分语言而言, 能收集到的数据实在太稀少
 - 中国少数民族语言、南亚地区、东南亚地区、非洲地区
 - 使用人数不少, 但资源积累较少
- LLMs 能利用极其有限的资源来**理解**这种语言吗?
 - 几千个句子?
 - 简单的字典?

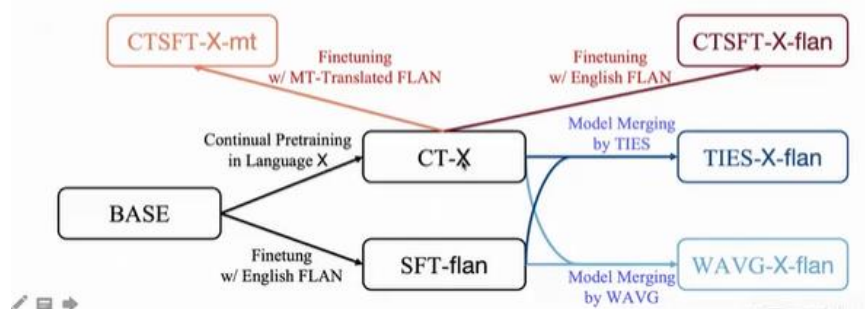
模型融合 TIES

- **前提**: 多个模型应来源于同一个基础模型 (高资源)
- **稀疏化**: 尽量使得不同参数支撑不同能力的实现
- **符号调谐**: 减少融合时, 由符号冲突导致的信息损失



模型融合

- 高资源语言基础模型、低资源语言持续预训练
- 高资源语言指令微调数据、低资源语言机器翻译指令微调数据



总结

- 质量优先策略对于资源稀缺语言至关重要
 - 在高质量数据基础上通过预训练可以帮助LLMs更好的支持资源稀缺语言
 - 数据收集中仍存在很多问题：
 - 政府公告、新闻占比过高、生活数据（论坛、社交媒体）偏少
 - 多样性、代表性；指令数据匮乏
- 基础语言能力强大的LLMs 潜能巨大
 - 在极其困难的场景下，仍可以有令人惊艳的表现
 - 新构架、新技术可以缓解资源稀缺语言所面临的困境