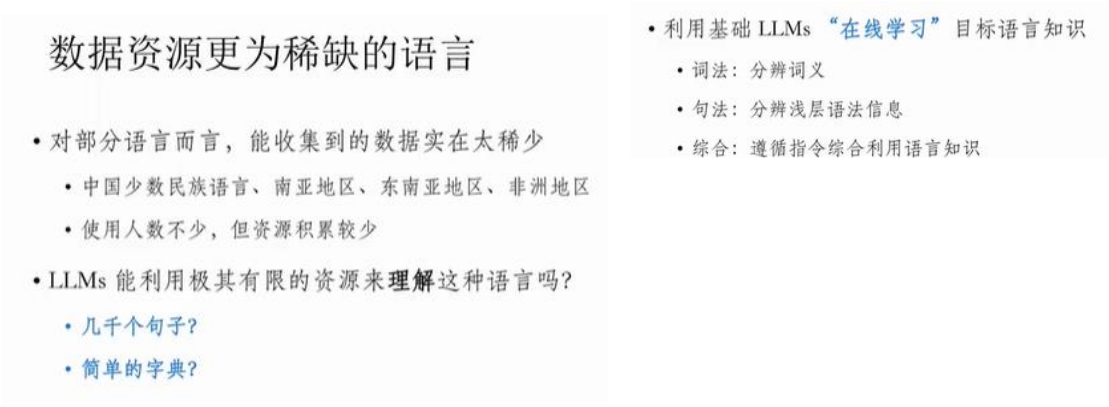
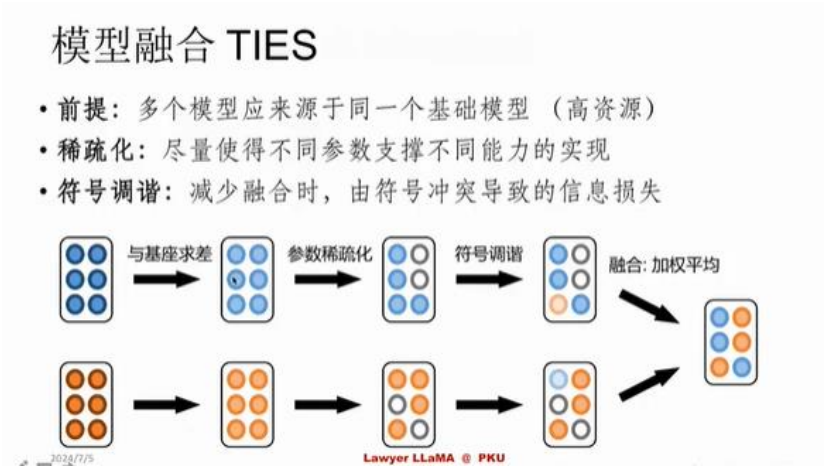
讲座 3：大语言模型如何学习新语言

大语言模型：

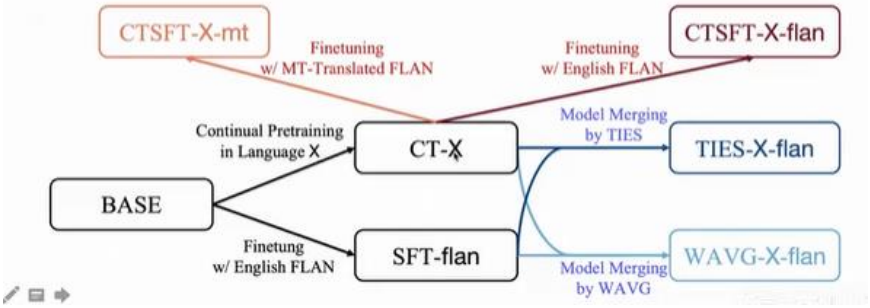
1. 数据资源是影响 LLMs 性能的主要因素之一
2. 主要由英语等少数资源丰富语言所主导
3. 很多语种在当今主流的 LLMs 中并没有很好的体现，编码效率低
4. 很多语言在 NLP 与 LLMs 资源中不受重视，恶性循环
5. 现有的处理流程对资源丰富的语种更有利，可能导致数据污染





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