学习周报

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | | 田雨桐 | 时间范围 | 20241209-20241215 |
| 周次 | | 第十五周 | 研究方向 | 大模型压缩 |
| 本周完成工作 | 1.将OceanGPT挂在了ChatGLM的前端上，但是最后没用这版  2.阅读了OceanGPT: A Large Language Model for Ocean Science Tasks  3.看了两节MIT线代课和部分李宏毅机器学习 | | | |
| 本周  问题汇报 | OceanGPT 数据集的构建过程主要包括：  1.原始数据收集：从开放获取文献中收集了 67,633 篇海洋科学相关的文献，涵盖最新研究和历史重要文献。  2.数据清理与预处理：使用 pdfminer 转换为纯文本，去除图表、页码、URL、空格等无关内容，并通过哈希方法去重。  3.领域指令生成：采用 DOINSTRUCT 框架，利用多代理合作生成数据。每个代理作为某一领域专家，生成并扩展海洋科学指令数据。  4数据生成策略：  种子数据集：通过海洋科学专家手动编写代表性样本，结合 LLM 自动生成更多样本。  演化代理：使用 GPT-3.5-turbo对种子数据进行扩展和深入分析。  文献提取代理：从海洋文献中提取高质量文本，自动构建 (inst, output) 数据对。  5.质量控制：通过训练专家手动验证数据，并随机抽查 10% 样本，确保数据的准确性。最终IAA评分为 0.82。 | | | |
| 下周工作计划 | 阅读一篇小模型的综述A Survey of Small Language Models  ，了解一下相关技术 | | | |

学习内容笔记