**Weekly Reports**

|  |  |  |
| --- | --- | --- |
| Summary of  2024/03/04  -2024/03/09 | Task 1: 本周工作 | 备注 |
| 1. 绪论：整理了NLP中与大语言模型相关的基本概念（如N-gram，NLP四范式等），梳理了生成式语言模型发展的历程（从GPT-1到当前的LLM），简单描述了语言模型训练的四个流程（预训练，有监督微调，奖励建模，强化学习），由语言模型催生的11大研究方向。 2. 大语言模型的数据构建：总结了常见的语言模型的数据来源，预处理（冗余消除，隐私去除，词元切分）策略以及数据质量和数量对模型的影响。设计了代码示例理解Token概念和对LLM产生的影响。 3. 大语言模型的全流程构建：详述语言模型的建模方法，包含高效指令微调，模型上下文窗口扩展（例如Gemini 1.5 pro支持200k上下文输入，Claude2 / 3支持200k上下文窗口输入等），以及相关数据集的构建方法和格式。设计了代码示例对LLaMA2模型在Colab的T4进行监督微调训练。   文档地址：<https://jwr8w7hhd3h.feishu.cn/wiki/space/7341385633325318146?ccm_open_type=lark_wiki_spaceLink&open_tab_from=wiki_home> |  |
| Task 2: 存在问题 | 备注 |
| 暂无 |  |

|  |  |  |
| --- | --- | --- |
| To-do-list of  2024/03/11  - 2024/03/16 | Task 1: 下周计划 | 备注 |
| 1.整理大语言模型技术的基础知识，包括Transformer模型、生成式预训练语言模型GPT以及大语言模型LLaMA。  2.整理语言模型的评估策略与评估指标。  数据处理工作：  1.获取广州 / 深圳地铁线路的经纬度信息，以标准的json文件输出（师兄要求）。  2.代码封装为pipeline，方便获取全国任意城市地铁线路经纬度信息。 |  |