AI 大模型实战项目

一、个人任务

1. 任务一

1) 在hugging face等网站收集化学相关数据

- 数据集方向从下面两项中选择一项：

a) 国内外高中化学知识

b) 国内外本科各领域化学知识：有机化学、AI化学、化学应用、量子化学、无机化学、材料化学、分析化学、化学信息学等

- AI化学Key words (Including but not limited to)：

Machine Learning
Quantum Chemistry
Molecular Modeling
Drug Discovery
Predictive Analytics
Computational Chemistry
Chemical Informatics
Molecular Dynamics
Reactivity Prediction
Material Science
Deep Learning
Structure-Activity Relationship (SAR)
Cheminformatics
High-Throughput Screening
Property Prediction
Reaction Optimization
Virtual Screening
Chemical Space Exploration
AI-Driven Synthesis Planning
Data-Driven Chemistry
AI for chemistry

- 数据集类型：

a) 常规问答

b) 选择判断

c) 其他类型，大家集思广益

2) 收集相关数据集后，自己写一个 process function 对这些数据集进行处理

3) 处理后的格式要符合能够微调的格式，如llama-factory 中的一种

4) 标明处理后的微调数据集的用途，比如能够训练模型的答题思路或者回答某一类问题的模板

5) 提交形式：

- 将原始数据集，代码文件及处理后的数据集文件整合为压缩包
- 命名格式：姓名_数据集方向_数据集类型

✓ Ref：

LLaMA-Factory/data/README_zh.md at main · hiyouga/LLaMA-Factory · GitHub QwenFunctionCalling 的对话模板及训练方法总结 - 知乎 (zhihu.com)

https://github.com/THUDM/ChatGLM3

https://huggingface.co/datasets/Locutusque/function-calling-chatml?row=4

2. 任务二

1) 利用上课中敲过的微调代码或者其他模型自带的微调方法和任务一的数据集微调一个属于自己的大模型（自己选择自己设备能承受的最大预训练大模型）

2) 针对结果进行详细分析，如果效果好，说明为什么会表现得好（如果效果差说明为什么差 即使效果真的差也不用担心，方法论的验证也是有价值的，而且有的时候效果差也只是因为模型体量不足的原因）

3) 微调后模型的应用用途

- 常规问答类
- 判断类
- 能够激发使用者思考的类型
- 其他，大家集思广益

✓ Ref：

Qwen 的Github 项目等

4) 提交形式：

- 将训练数据集，测试数据集，验证数据集，代码文件，训练好的模型以及

对模型效果的分析报告pdf格式或者直接ipynb中的markdown格式都行整合为压缩包

- 命名格式：姓名_微调模型名称_用途

注意：

中期检查时间：2024年8月5日 18:00

终期提交时间：2024年8月11日 18:00

二、团队任务（2-3人一组）

1. 任务三（具体细节会在2024年8月8日前公布）

1) 自主收集目标需求文献形成资源库（pdf, txt 等各种类型）
2) 利用 RAG 及 Langchain 技术通过外挂知识库的形式，对知识库中的数据进行分类，按照具体类别分类最后将这些数据分类保存为特定的资源库，比如所有的反应类型、反应机理等，确保一一对应。
3) 需创建一个共同的 github 项目库，从而实现多人协同合作的目标

✓ Ref：

RAG 官方文档

Langchain 官方文档

4) 提交形式：

- Github连接和对应的资源库压缩包

注意：

提交时间：2024年8月21日 18:00

2. 个人毕业任务

1) 基于这一个月的实战，做一份海报（A2纸大小，支持垂直方向或者水平方向布局，中英文均可），海报中含两项个人任务及团队任务，海报中含有上面项目的github
2) 提交形式：

- 电子版和打印出来的海报

注意：

提交时间：2024年8月22日 18:00