大模型数据标注流程

1.项目创建与数据导入

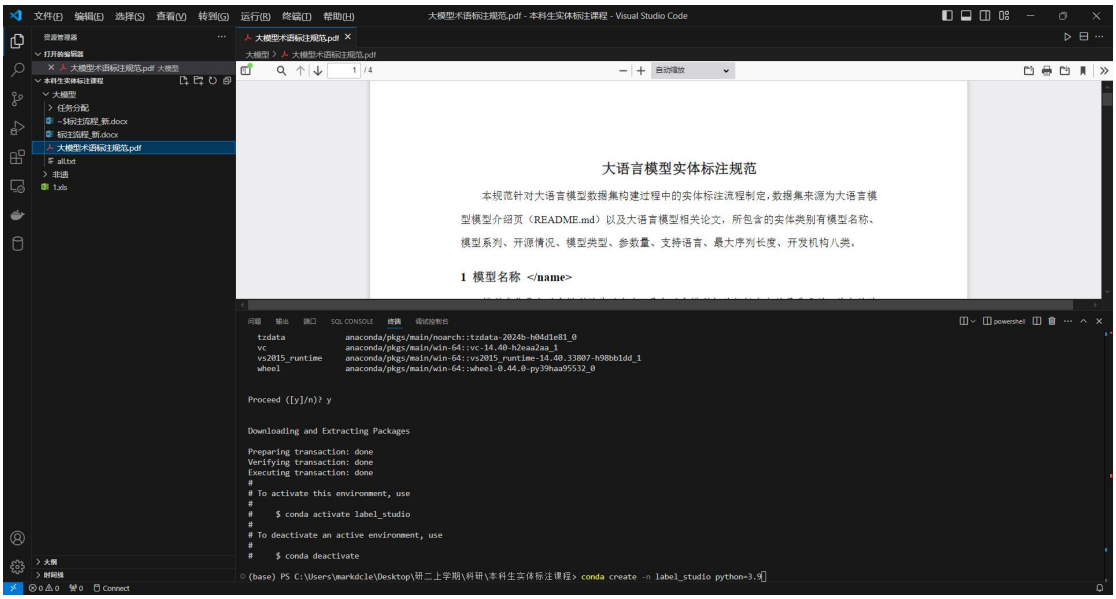
本次数据标注使用 label studio 平台进行，本地访问部署（VSC/pycharm终端):

conda create -n label_studio python=3.9

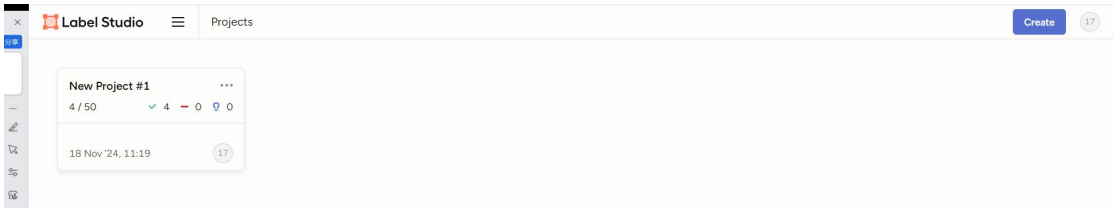
conda activate label_studio

pip install label-studio`-i https://pypi.tuna.tsinghua.edu.cn/simple`

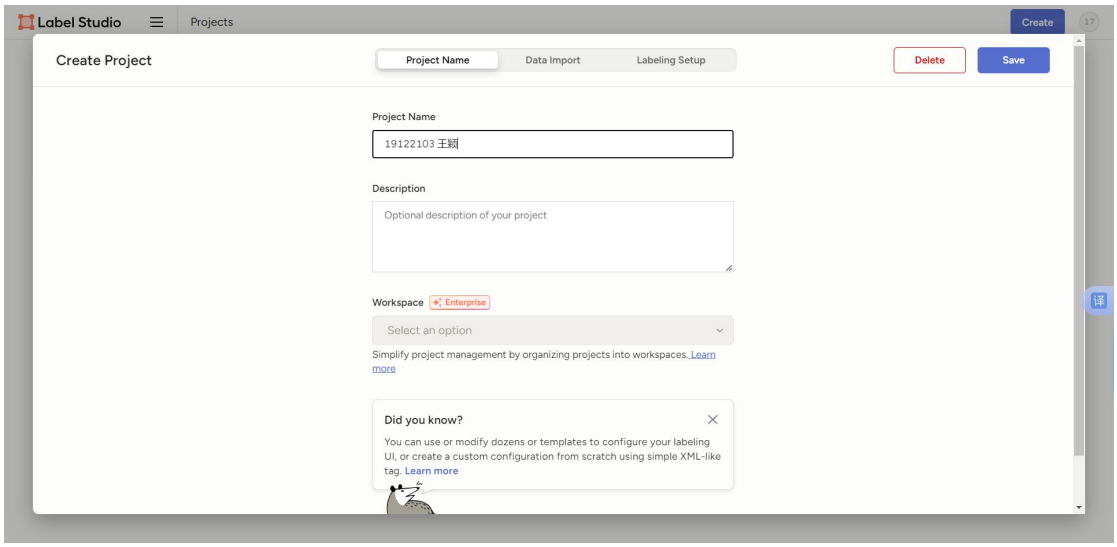
label-studio start



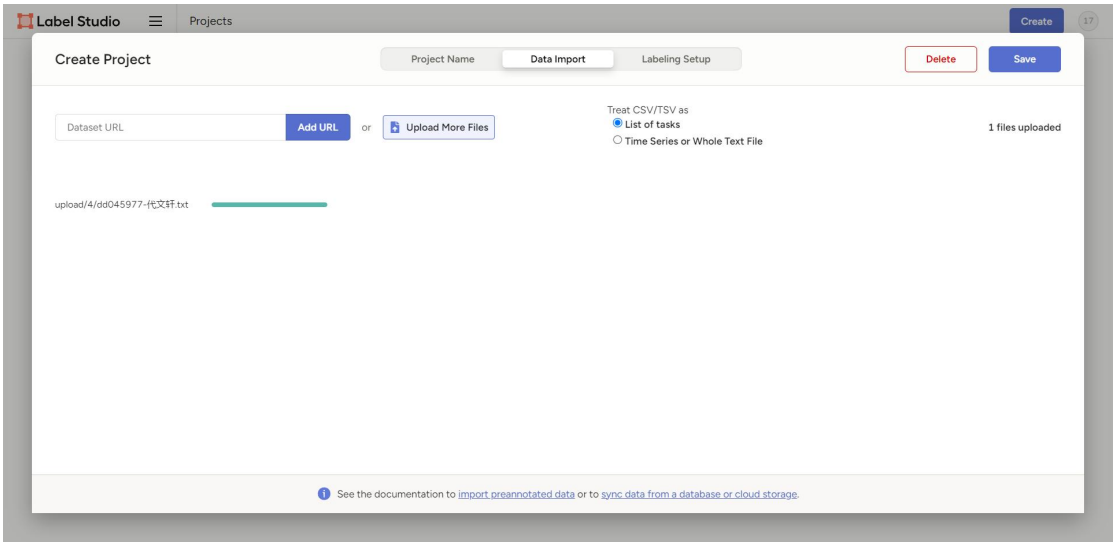
2. 注册输入邮箱、密码即可注册，注册完成后可直接登录，进入项目主页：



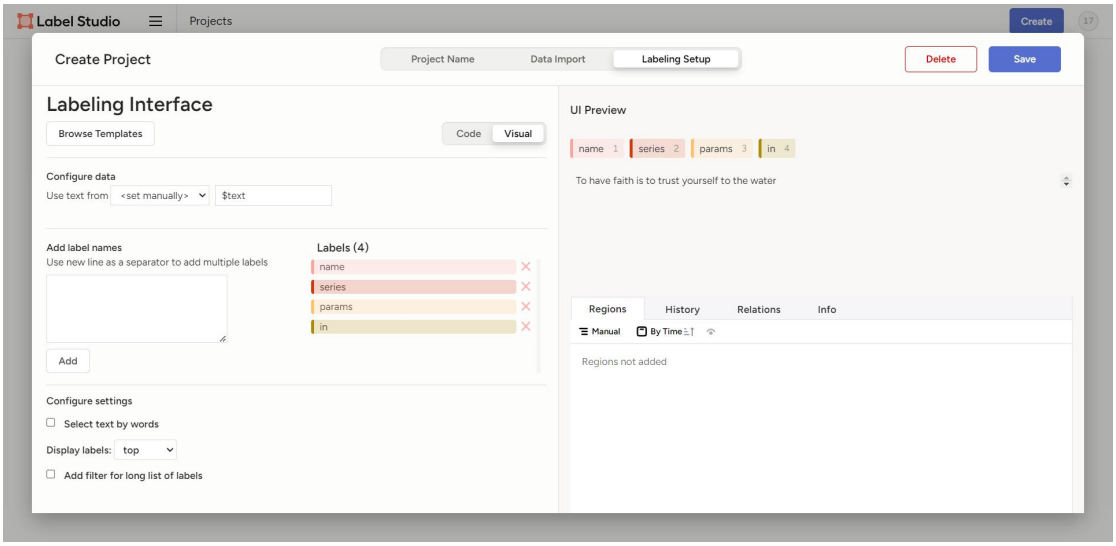
3.进入项目主页后，点击 create 进行创建项目，首先设置项目名称，为便于后续对标注结果进行整理与组织，项目名称设置为学号+姓名，例如“ 19122103 王颖”



4. 之后，导入数据文件，为便于后续处理，数据导入过程中采用单个文件进行导入，后续需对标注完成数据进行切分

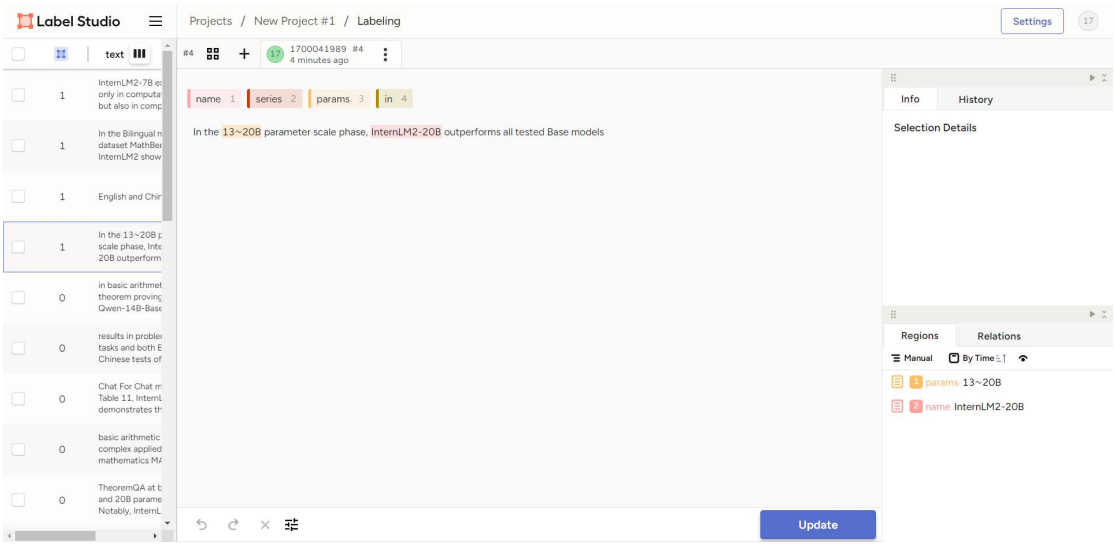


5. 最后，设置实体标签，选择左侧 natural language processing，点击 name entity recognitio n，设置实体标签，本次实体标注主要包括四大类实体，模型名称</name>、模型系列</series>、参数量</params>、开发机构</in>，详细标注规范见“大模型术语标注规范.pdf”。

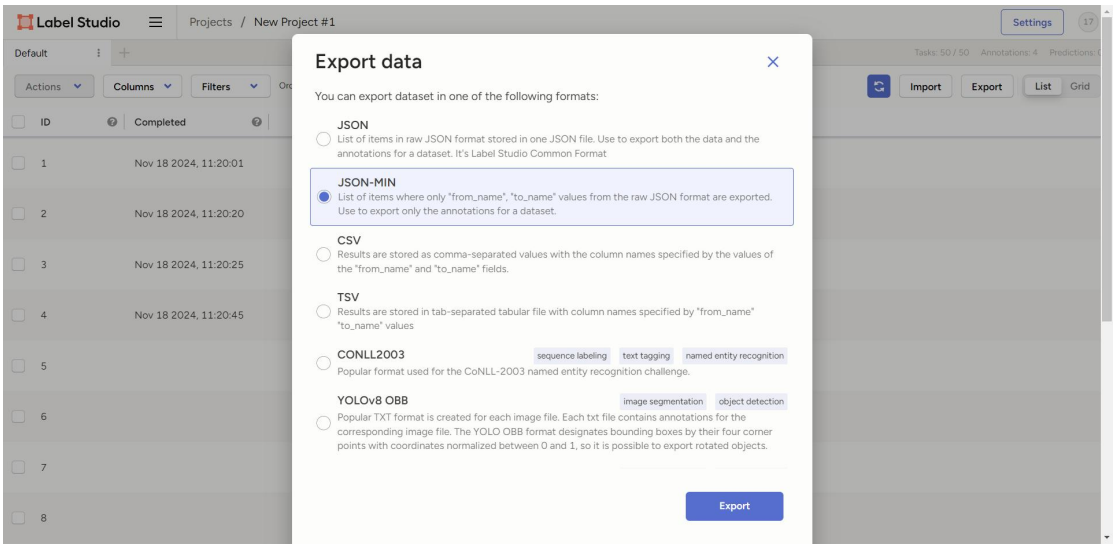


6. 标注：进入项目页面，即可看到导入的按行分割的数据，点击 label all tasks，进入标注页面，点击上方标签图标，选中文字，即可完成对实体的标注，右侧为标注结果，将一条文本标注完成后，点击 submit 进行提交。

PS：务必点击 submit 提交，即使没有可标注内容也要提交，否则会影响最终导出的结果数量



7. 标注结果导出：对一个项目标注完成后，进入项目页面，点击右上角 export，选择json-mini 格式进行导出，即可下载导出json 文件。



8. 提交结果：“姓名.json”