会议纪要

|  |  |
| --- | --- |
| **项目名称** | 端边云产业赋能系统 |
| **会议时间** | 2022-11-30 16:00-17:10 |
| **会议地点** | 腾讯会议 |
| **参会人** | SMIL团队：谭明奎老师、朱金辉老师、陈岑老师、陈耀佛同学、许守恺同学等  南京先维：葛俊海、胡亮、卞成明、鲍全卫 |
| **会议议题** | 大模型部署平台功能演示及讨论 |
| **会议内容** | |
| 本次会议通过对大模型部署平台现有功能进行展示与探讨，从而明确平台未来的完善内容及完善方向。具体来说，各老师与同学探讨时对平台目标用户进行需求分析，并从功能的展示、规范与扩展三个角度提出如下意见：   1. 目标用户需求分析   大模型应用到下游任务的场景中，主要面向公司、企业、单位，在开发、部署、迁移上提供便利。   * 1. 甲方用户   提出场景, 场景里有需求, 需求落实到数据和指标. 需求要上传到系统里进行对接。   * 需求支撑：提供一个功能模块规范需求分析流程，方便公司填写需求，算力、芯片、任务要求等 * 管理支撑：通过平台把需求管理起来，方便沟通下游任务，从而降低沟通成本   1. 工程师用户（乙方）   用系统解决甲方任务, 根据需求提出方案。   * 平台目标：i. 能帮助用户方便快捷地生成可用场景模型，看到效果，解决甲方需求 ii. 帮助用户向甲方/上级进行展示与沟通，方便讨论 * 存在问题：用户之前的套路与经验没办法共享公开   1. 大模型开发团队   将大模型部署到平台上，提供模型与算法等的接口规范约束并将这些约束展示在平台中，从而把千变万化的大模型规范地放到系统中。   1. 功能的展示   拟完善各功能的展示内容，从而让用户更好地梳理功能间关系，掌握功能的使用方法，了解数据的详情。   * 梳理功能间关系 * 提供具体场景下功能的流程图 * 概览根据用户角色分类展示不同的使用流程 * 大模型左边目录设置逻辑是否不够准确，部署的业务、管理类的业务杂糅在一起，分不清，不同角色的人用这个平台，不能准确的找到功能范围 * 掌握功能的使用方法 * 数据录入的地方要展示规范约束，避免用户上传错误代码或文件，并在上传数据有错时立即反馈 * Pipeline启动命令给出合理提示，或者提供案例文件，让用户知道怎么使用平台 * 部署到实际应用的物理拓扑不清楚，部署是怎么样的一个具体概念，是产生一个场景模型？还是到应用终端 * 了解数据的详情 * 训练提供样本展示 * 将不同模型、不同数据集，训练评估结果保存下来，通过条件筛选，横向比对评估结果 * 大模型展示的资料信息要全一些，包括大模型类型、架构、架构图、接口、输入输出、训练数据集、模型来源（github，其他公司，自研）、参考资料（论文，专利）。可借鉴paper with code * 场景模型功能展示提供来源（添加类型（训练、转换）），增加其他模型信息的展示 * 展示模型迁移时的输入，输出，框架、是否需要数据、约束、参数、限制等信息 * 展示模型部署关联硬件输入，输出清楚，框架、是否需要数据、约束、参数、限制等信息  1. 功能的规范   拟规范各功能的接口，从而保证算法与模型等的可用性。   * 模型迁移: 参考大模型的资料信息完整，把metadata定义清楚。特别的, 场景约束(模型大小,参数大小,功耗约束,平台限制) * 入仓: 单独把入仓拿出来。把metadata定义清楚，越详细后面对接的工作量越少。把模型入仓和算法入仓放在一个层面。入仓后还有审批过程，要有严谨详细的流程。提供使用仓外模型的接口, 进行临时使用 * 输入输出的格式需要规范，metadata要足够多，能够规范的地方都规范下来，有些地方可能规范不到，是否可以通过灵活的方式提供一个解决方式，在规范和自由度之间权衡好 * 数据的录入要定义规范约束 * 算法入仓生成镜像区分构建镜像和非构建镜像两种入仓方式  1. 功能的扩展   拟对功能进行扩展，从而简化用户的使用或支撑更多使用场景。   * 在用户填写上传东西时提供提示与例子 * 目前配置支持ARM 等，后续增加其他模板配置 * 部署过程中补充增量学习与版本管理功能 | |
|  | |