附件2

|  |  |  |
| --- | --- | --- |
| **大模型管理平台功能需求** | | |
| 序号 | 功能名称 | 需求描述 |
| 1 | 资源管理 | 资源管理模块主要负责对大规模的CPU、GPU、存储等资源进行管理和调度，资源调度引擎基于集群的硬件和组网特点，实现对算力资源的统一管理、调度和监控，进行细粒度的资源实时分配，支持多种不同厂商的GPU机器纳管，支持构建算力集群，支持用户资源共享与隔离。 |
| 2 | 模型管理 | 模型管理模块提供模型上传/下载、模型分享、模型部署上线以及纳管第三方模型服务。支持一键部署为API能力，支持按照用户分配调用权限、限制调用QPS、调用量，提供高可用、可扩展的模型推理服务。可实时监控模型运行状态、性能指标和资源消耗情况，及时发现和解决模型运行问题。支持版本管理和灰度发布，确保模型更新过程的平滑稳定。 |
| 3 | 模型开发 | 支持SFT全参、Lora等多种模型训练方式，可创建微调任务，对于模型微调任务，支持新建、删除、搜索等基础操作，支持查看详情、查看日志、停止任务、训练模型可视化、故障恢复、模型上架等功能。 |
| 4 | 模型评估 | 模型评估模块支持对部署在平台上的模型进行评估，以了解模型在不同场景下的表现，验证模型的泛化能力，从而更好的优化模型，主要包含：评估模板的创建与管理、评估任务的创建与管理、评估结果的展示。 |
| 5 | 知识库管理 | 提供从结构化数据、非结构化文本等多种数据中抽取知识，构建领域知识库的能力，支持自定义分割策略、预处理策略、知识增强策略，支持单文件导入和批量导入两种导入方式，提供高效可靠的知识存储和管理，保障数据安全，提供知识检索、推理、问答入口，底层支持多种检索模式，保证更好的知识检索结果。支持针对检索策略进行召回测试，判断检索系统的召回结果是否符合预期。 |
| 6 | 智能体管理 | 涵盖智能体的可视化开发、部署、监控和维护等功能。通过直观的图形界面，用户可以轻松设计和配置智能体的行为逻辑，减少对编程技能的依赖。平台支持一键部署，将开发完成的智能体快速应用于生产环境，并提供实时监控工具，跟踪其运行状态和性能指标，提供流量控制功能。 |
| 7 | 工作流 | 提供可视化流程编排工具，支持将大模型、知识库、工具、自定义代码等算子按照顺序进行串联，构建自动化工作流，知识基础的开发、发布和上线管理，支持定时任务、实践触发等多种任务调度方式，确保工作流高效稳定运行，实时监控工作流运行状态，及时发现和处理异常情况。 |
| 8 | 语料管理 | 提供语料的获取、清洗、加工、治理、应用和管理的全生命周期，提供基础的语料清洗、语料切分和语料评测工具。 |
| 9 | 权限管理 | 平台具备完善、清晰的的权限架构：用户根据所分配到的角色，继承该角色的权限来进行模型开发、应用开发和管理，管理员可对用户、用户组、角色权限进行自由地创建和修改，并且针对不同部门的和职责的人员，设置差异化的权限，实现权限的集中管控。支持数据、模型、应用、资源层面的隔离，满足个农商行独立开发的需求。涉及关键的流程步骤，支持触发审批流程，实现操作留痕。支持按照调用量、调用速度控制用户访问权限。 |
| 10 | 运维管理 | 平台提供对专属资源池的监控，用户可查看整个集群的CPU、GPU、内存等监控信息，查看已上线模型、模型应用的调用量、失败率、token、QPS、TPM等指标，提供统一的日志收集、查询、存储、下载和配置，帮助运维人员轻松应对日志采集、查询等运维场景。提供监控告警管理，运维人员通过告警监控来监控、管理系统自身或者对象上报的告警或者事件，告警提供了丰富的监控和处理规则，可以将故障通知给运维人员，帮助高效监控、快速定位和处理网络故障，保障业务的正常运行。 |