# 第五章　项目要求

## 一、项目需求

### （一）采购内容

1.金融级 AI 技术中台系统一套，含开发实施。

2.代码助手、大模型知识库智能应用各一套，含开发实施。

3.后续或有服务价格参考，包括平台维保价格。

### （二）开发周期

本项目研发及投产周期预计 12 个月，分阶段实施，具体以项目实际进度为准。

### （三）业务功能要求

1.金融级 AI 技术中台

搭建大规模智能服务的基础设施，形成一套完整的智能模型全生命周期管理平台和服务配置体系，基于私有化部署方案，具备 AI 异构算力资源池管理能力，支持模型/算法库复用，支持主流 AI 框架，兼容机器学习、深度学习模型及大模型，面向行内提供从数据处理、模型开发、模型训练、模型评估到模型推理部署等功能模块的 AI 开发全流程支持，对前台业务提供智能服务的迅速构建能力。AI 技术中台包含但不限于以下四层能力，具体要求如下：

（1）算力资源层。

建立异构算力资源池，统一管理和弹性分配各种 AI 算力资源，包括 CPU、GPU 及 NPU 计算资源等，支持少于 1 卡的虚拟 GPU 切分，标准化接口和协议，实现对算力资源的自动化调度和优化，提高算力资源利用率和响应速度，并对算力资源具备优秀的监控预警能力。

（2）模型训练层。

实现模型的统一管理，拥有纳管各类原子化 AI 能力的模型仓库，具备模型转换、模型开发、模型训练、模型评估等功能模块；支持主流的开源机器学习和深度学习框架，支持多人协作开发，具备模型本地化训练精调能力，支持各类主流开源模型及我行选定的商用模型，支持 SFT、Lora、RLHF 等微调训练方式；对训练数据进行管理，支持丰富的结构化和非结构化训练数据源接入、支持数据加工及标注功能。

（3）模型推理层。

对模型推理应用提供统一管理服务，包括模型服务网关、模型的弹性伸缩及灰度发布、流量控制、模型服务监控模块；对模型部署进行管理，包括提供在线服务、边缘服务等功能模块。

（4）模型服务 Maas 层。

提供高效、灵活的 AI 服务编排组合能力。通过可视化应用编排模块对数据、AI 算法等组件进行灵活编排，实现复杂业务场景中的 AI 能力的快速搭建和迭代，提高 AI 技术在业务中的落地效率。

2.代码助手

基于 AI 中台，实现全栈智能化开发辅助，提高软件开发效率、质量和可靠性，规范代码推动最佳实践。代码助手应支持丰富的编程语言及 IDE 开发插件，能够结合行内代码及开发框架进行模型微调，并具备以下四项核心功能：

（1）代码补全。通过分析注释描述和上下文代码语法语意，自动生成相关的业务逻辑代码、函数和注释。

（2）技术对话。实现情境感知的对话交互。通过对用户问题与相关代码上下文的理解，它能提供推理性的解答回复，并自动生成可能的后续提问推荐。

（3）单元测试。具备自动生成单元测试代码功能，可以让研发人员快速完成繁琐的单元测试编码工作，提高软件自动化覆盖率。

（4）代码诊断。提升代码可靠性、健壮性和易维护性，发现代码异常问题，给出相应的修复方案或调优建议。

3.大模型知识库

基于 AI 中台，由大语言模型和检索增强生成（RAG）技术构建的知识管理平台，结合企业私域数据或知识，进行高效智能的知识库自动化扩充，优化传统的问答梳理与对话树设计等繁重人力工作，构建智能知识检索能力，提供精准的知识问答应用范式，实现智能化的信息管理。大模型知识库应具有但不限于以下功能：

（1）文档管理。支持用户构建私域文件或知识管理的能力，支持创建应用知识库，管理知识库导入的文件，将整理的数据上传后，系统将自动完成数据的清洗、切割与分段并存储到向量库中，支持 txt/pdf/doc/docx/md/xlsx 等多种格式的数据单个或批量导入。

（2）模型管理。平台内置多种大模型和向量化模型。提供针对金融场景优化的行业知识库大模型，并支持多种大模型的切换；提供企业优化过的向量化算法模型，并支持向量化算法的切换。

（3）知识问答。支持根据不同场景创建各类知识库问答应用，能准确理解自然语言提问，能正确分析各类知识，整合生成答案并提供溯源数据链接，支持对文档中命中内容定位，能够设置知识回答的准确度及回答知识范围功能；能够设置提示词及开场白，支持基于文档知识的基本问答和基于任务流程式的问答，支持大模型上下文，支持过滤敏感词、违禁词；提供知识库调优和数据回流能力，对不满意回复修正知识库答案，设置文档召回提高分词准确率，能够有效减少幻觉问题。

（4）权限管理。支持对知识库进行分权限的方式维护，支持多租户多角色的权限管理，支持文档生效的范围，支持对文档进行多标签标识的权限管理，支持基于权限的 API 接口调用。

（5）统计分析。支持对知识库应用服务多种可视化监控，能够提供按用户、应用、满意度等多维度知识库使用情况数据统计分析。

**（四）技术要求**

| AI 中台-代码助手项目技术要求 | | | |
|---|---|---|---|
| | **序号** | **技术功能模块** | **要求描述** |
| **通用要求** | 1 | 信创支持 | 1、平台所有组件支持信创服务器、操作系统，适配不同的 CPU 体系架构，如 X86 和 ARM。<br>2、产品本身是信创产品。<br>3、支持信创计算芯片。 |
| | 2 | 产品架构 | 1、平台各组件低耦合，可维护性高。<br>2、平台各组件可独立升级和维护，不相互影响。<br>3、平台台组件支持两地三中心的高可用部署。 |
| | 3 | 异构计算芯片 | 1、支持主流的信创 GPU、NPU 等芯片。<br>2、支持多种型号的芯片。 |
| | 4 | 安全审计 | 1、支持平台所有产品模块管理操作的审计事件记录。<br>2、支持审计事件查询。 |
| | 5 | 用户及权限管理 | 1、为所有产品提供统一的用户体系。<br>2、支持用户的创建、 删除及权限配置。<br>3、支持细粒度的接口、按钮、菜单、以及数据权限控制。 |
| | 6 | API 接口 | 1、所有的页面功能提供 API 接口供行内调用。<br>2、所有 API 接口支持鉴权。 |
| | 7 | 监控告警 | 1、支持所有产品模块及集群提供统一的监控系统。<br>2、支持自定义告警指标。<br>3、支持多种告警推送方式，包括回调、消息推送等。<br>4、支持接入行内的智能预警平台及统一告警中心。 |

| | | | |
|---|---|---|---|
| | 8 | 日志 | 1、支持日志搜索。<br>2、提供可视化日志采集和投递配置。<br>3、支持接入行内的日志平台。 |
| **AI 中台** | 9 | 平台兼容性 | 兼容开源的 Kubernetes。 |
| | 10 | 算力虚拟化 | 1、支持算力虚拟化，对单卡算力进行更细粒度的切分。<br>2、实现任务间的资源限制、资源隔离。<br>3、支持可视化的算力虚拟化管理。 |
| | 11 | 算力池化 | 1、支持多种异构算力池共存。<br>2、根据用户需求，动态地分配和调度资源，具备弹性伸缩的能力。<br>3、支持可视化的算力池化管理。 |
| | 12 | 数据接入 | 1、支持多种数据集类型，不限于文本、图片、视频、语音等。<br>2、支持数据集本地导入导出。<br>3、支持多种类型远程数据集，不限于 MySQL、TDSQL、TBASE、orcale、hdfs、minio、ceph、hive 等。<br>4、支持对数据集进行版本管理，同一个数据集可以发布多个版本，并支持历史版本的管理。<br>5、支持数据分析可视化。 |
| | 13 | 数据标注 | 1、支持对图像标注、文本标注、音频标注和视频标注等。<br>2、支持在线标注、多人标注等多种标注方式。<br>3、支持从三方数据标注平台获取标注数据。 |
| | 14 | Notebook 建模 | 1、使用 Notebook 新建多个文件并进行交互式的编程。<br>2、Notebook 环境支持包括代码自动补全，代码跳转，断点调试等功能。<br>3、预置 PyTorch、Tensorflow、Spark、PySpark 等丰富的框架，支持框架版本选择及新增。<br>4、Notebook 中的开发脚本支持持久化，训练的模型支持推送到模型管理中心。<br>5、支持内置 git 及配置、支持多人协作开发。 |

| | | | |
|---|---|---|---|
| | | | 6、支持数据集挂载。<br>7、内置 Tensorboard 等可视化库。<br>8、支持 Notebook 选择多种容器镜像环境，支持用户自定义的 notebook 实例可导出为自定义镜像，并进行管理和复用。<br>9、支持 Notebook 调试过程中的资源监控。 |
| | 15 | 任务建模 | 1、支持添加多个数据集、添加超参数、添加预训练模型。<br>2、支持查看任务详情、监控、日志。<br>3、支持任务生命周期管理，支持任务按优先级或预定时间顺序执行。<br>4、支持常见的训练加速方法。<br>5、支持常见的指令微调，偏好对齐，继续预训练等模型优化方式。 |
| | 16 | 工作流建模（可视化建模） | 1、支持通过工作流的方式建立模型。<br>2、支持可视化的拖拉拽方式构建工作流。<br>3、内置丰富算子和算法，如数据预处理、数据转换、特征提取等。<br>4、内置多种框架的常用版本，如 Tensorflow，Pytorch 等。<br>5、支持流程导入导出。<br>6、支持历史版本管理。<br>7、支持导入自定义算子。<br>8、自定义组件算子可基于 Notebook 环境编写和调试脚本。 |
| | 17 | 模型评测 | 1、支持对比评测，具备多种评价指标的计算，可视化输出评测结果。<br>2、支持自定义评估指标，可使用自定义的评估指标脚本进行评估。<br>3、支持多维度的内置评估指标，效果评估指标支持准确率、精确率、召回率等；性能评估指标支持计算时延、CPU/GPU 占用率、内 |

| | | | |
|---|---|---|---|
| | | | 存/显存占用率等。<br>4、支持评估报告自动生成和下载。 |
| | 18 | 模型优化 | 1、支持在不降低模型推理准确性或降低可承受准确性的情况下，对模型进行优化，包含模型压缩、模型加速等。<br>2、支持数据回流，可以收集发布任务的回馈信息用于模型的迭代。 |
| | 19 | 模型仓库 | 1、支持模型中心统一管理。<br>2、支持导入各种格式的本地模型，包含 pt、pkl、bin、safetensor、ckpt、guf、pth 等。<br>3、支持第三方的开源或商用模型接入。<br>4、支持训练任务和模型仓库的互通。<br>5、支持各种格式模型的训练和推理，包括大模型、专用模型等。<br>6、支持模型版本管理。<br>7、支持一键发布模型服务。<br>8、支持将第三方模型文件导入模型中心统一管理。<br>9、支持镜像文件类型的第三方模型导入。<br>10、支持对接行内镜像仓库。 |
| | 20 | 服务管理 | 1、对平台纳管的所有模型，具备模型服务治理的能力，例如灰度发布、弹性扩缩容、统一网关、流量控制等。<br>2、对平台纳管的所有模型，具备滚动更新、平滑重启的能力。<br>3、对平台纳管的所有模型，支持请求日志、服务日志查看。<br>4、支持对平台纳管的所有模型进行服务事件查看，包括容器启动、重启、销毁记录等。<br>5、支持第三方服务注册（包括后端服务请求地址与服务协议）、查询、修改、删除等基本管理功能。<br>6、支持对模型服务进行性能评估，支持对模型进行压测，并查看压测指标，包括包括成功率、平均 RT、TPS、CPU/GPU 使用率、内 |

| | | | |
|---|---|---|---|
| | | | 存使用率和网络 IO 等。<br>7、支持模型服务编排。<br>8、支持模型服务鉴权。<br>9、支持模型服务接口级管理，包括创建接口，编辑接口，删除接口、屏蔽接口、接口限流等。<br>10、支持模型服务负载均衡能力。 |
| | 21 | 模型应用 | 1、支持可视化的模型应用及 Agent 智能体编排。<br>2、模型应用编排支持知识库、三方服务接口、函数组件等。<br>3、支持模型应用的编辑、查看、删除、更新、发布、下线、回滚、扩缩容等。<br>4、支持通过 API 的方式提供模型应用服务，并提供鉴权能力。 |
| | 22 | 服务监控 | 1、支持服务质量的可视化监控告警，包括 tps、响应耗时、交易量等。<br>2、支持对服务资源的可视化监控以及告警。<br>3、支持对模型服务的监控，可监控的颗粒度包括：应用、单个服务、POD 层。<br>4、服务与应用的监控的纬度包括：调用量、QPS、总流量、平均流量、总处理时间、平均处理时间等。<br>5、POD 监控纬度包括：CPU 平均使用率、GPU 算力平均使用率、内存平均使用率、加速卡显存平均使用率等。 |
| | 23 | 监控告警 | 支持基于集群的多维度监控，包括 CPU、GPU、内存、IO、网络流量等。 |
| **代码助手** | 24 | IDE 插件集成 | 1、支持全系 Jetbrain 产品，包括 IDEA、Goland、PyCharm、WebStorm 等。<br>2、支持 Visual Studio Code。 |

| | | | |
|---|---|---|---|
| | 25 | 编程语言 | 1、支持多种后端语言，包括 C、C++、C#、Java、Python、Go、PHP 等。<br>2、支持多种前端语言和前端框架，包括 Html、JavaScript、Typescript，Css、React、Next、Vue 等。<br>3、支持多种 App 端开发，包括 Swift、ObjectiveC、Kotlin 等。<br>4、支持数据库 SQL 语句。 |
| | 26 | 代码补全 | 1、支持根据注释生成代码、SQL 语句。<br>2、支持根据上下文补全代码、SQL 语句。<br>3、支持多种补全方式，包括单行补全、多行补全。<br>4、支持跨文件的代码补全。 |
| | 27 | 代码辅助 | 1、支持代码优化。<br>2、支持自动生成单元测试。<br>3、支持根据代码生成注释。<br>4、支持代码解释。 |
| | 28 | 编程交互 | 1、支持在 IDE 工具中进行编码知识对话，并给出合理的问答结果。<br>2、支持进行多轮对话。<br>3、支持对话历史保存，重启或关闭后依然可以继续历史对话。<br>4、支持对问答结果进行好坏的反馈。<br>5、支持直接将问答结果插入到代码区域。 |
| | 29 | 平台监控 | 1、支持对代码补全场景的数据进行多维度的统计监控，维度包括总数和日均，指标包括生成率、接受率、接受行、响应时间、每日使用人数、每日触发补全次数等。<br>2、支持对编程交互场景的数据进行多维度的统计监控，维度包括总数和日均，指标包括接受率、每日使用人数、每日对话次数等。 |
| | 30 | 模型服务 | 平台的自带的大模型不小于 7B。 |
| | 31 | 增量训练 | 支持行内根据自身代码、文档等进行代码模型的增量微调及训练。 |
| | 32 | 模型切换 | 模型与应用低耦合，可进行不同代码生成模型的切换。 |

| | | | |
|---|---|---|---|
| **知识库** | 33 | 知识库应用 | 1、支持知识库应用管理，包括创建、删除、编辑、发布。<br>2、支持可视化知识库应用编排能力,支持提示词、变量、上下文、下一步问题建议、模型幻觉等配置。<br>3、支持创建基于文档知识和问答知识的知识库应用。<br>4、支持导入多种不同格式的文档知识，包括 txt、pdf、doc、docx、markdown、xlsx 等。<br>5、支持一次导入多个文档知识或问答知识。<br>6、支持选择不同的向量化模型。<br>7、支持选择不同的大模型。<br>8、支持配置知识库应用的 CPU、算力、存储、内存等资源规格。<br>9、支持配置召回数量，包括文档最大召回数量、问答最大召回数量。<br>10、支持对文档知识进行分段处理并存储至向量知识库。<br>11、支持采用重排序模型进行结果召回。<br>12、支持提供追溯原始数据并展示原文,支持对文档命中内容定位。<br>13、支持对问答进行敏感词、违禁词的配置。<br>14、支持知识库应用参数设置，包括检索的置信度、匹配文档的最大数量、是否精确匹配知识库等影响知识回答准确度的参数。 |
| | 34 | 知识管理 | 1、支持进行文档知识管理，包括导入、导出、删除等。<br>2、支持进行对话知识管理，包括手动录入、批量导入、导出等。<br>3、支持对知识进行标签化管理，包括标签新增、标签设置、按标签检索知识。<br>4、支持对文档知识和对话知识进行权限控制。 |

| | | | |
|---|---|---|---|
| | 35 | 知识服务管理 | 1、支持服务治理的能力，包括弹性扩缩容、流量控制等。<br>2、支持通过 API 调用和 web 端访问的方式提供知识服务，并具备鉴权能力。<br>3、支持服务质量的可视化监控告警，包括 tps、响应耗时、交易量等。<br>4、支持对服务资源的可视化监控以及告警。<br>5、支持对知识服务的监控，可监控的颗粒度包括：应用、单个服务、POD 层。<br>6、服务与应用的监控的纬度包括：调用量、QPS、总流量、平均流量、总处理时间、平均处理时间等。<br>7、POD 监控纬度包括：CPU 平均使用率、GPU 算力平均使用率、内存平均使用率、加速卡显存平均使用率等。<br>8、支持知识服务命中信息监控，包括命中率、命中关键字统计等。<br>9、知识服务支持 API 方式供三方调用，并提供鉴权功能。<br>10、支持按调用方进行知识服务的监控，监控的维度包括：调用量、QPS、总流量、平均流量、总处理时间、平均处理时间等。 |
| | 36 | 知识服务调优 | 1、支持对知识服务进行验证测试。<br>2、支持对知识服务进行效果调优。<br>3、支持收集知识服务的满意度反馈。 |
| | 37 | 模型服务 | 1、平台自带多种向量化模型。<br>2、平台自带多种大模型，模型大小不小于 7B。 |

**（五）售后服务要求**

批注 [v1]:

1.项目人员要求。项目经理应具有 6 年（含）以上工作经验，3 年（含）以上项目管理经验，具有丰富的同类项目管理和实施经验；核心人员需具有 4 年（含）以上工作经验，具有丰富的同类项目实施经验。

2.落地实施。需驻场实施，按照采购业务需求及技术要求等开展系统建设，积极配合我行接口调试，确保适配我行相应的业务要求，并接受系统开发实施监理。

批注 [v2]:

3.知识转移。为规范系统运维和用户操作使用，需对系统功能、技术架构和日常运维等关键内

容进行解释及必要培训，提供操作手册及相关技术文档、源代码等必要的文档。培训内容涉及 AI 系统基本功能操作使用，模型开发、模型训练，模型部署等，两个垂直领域大模型的微调和优化，确保行内人员具备在采购平台上进行模型开发、管理和维护的能力，而且能够独立操作。

批注 [v3]:

4.试运行期不低于 6 个月。若规定试运行期满，不能达到验收标准的，视其情况，可延长试运行期。

5.提供原厂维保，免费维保期不低于 1 年，自项目验收合格日起开始计算免费维保期。在维保期内，维保方应：

（1）提供免费技术支持服务，服务内容包括不限于采购软件的技术咨询、系统恢复、系统功能故障处理、软件版本的维护（模型和系统的升级）、模型和系统调优支持、保障业务系统流畅运行等。

（2）提供免费现场系统巡检，每年至少 1 次，检测软件系统的运行的健康状况，并在每次检测后提供检测报告和相关专业技术意见，供招标人掌握并优化中标软件系统的使用。

（3）对本合同中中标软件提供：一是指定专业技术人员提供 7*24 小时免费技术支持服务，为招标人提供与本项目有关的各种技术咨询、操作指导、问题解答等技术支持。二是发生故障，中标人必须在用户发出技术求援请求后的 30 分钟内响应，组织相关技术人员在约定时间内给予解决。在必要的情况下，专业技术人员应在 24 小时内到达指定现场，并立即修复或排除故障。

6.供应商工作实施需满足我行质量评价管理要求。供应商需制定科学合理的工作推进计划，经采购人审核通过后实施，各阶段产出成果均应经采购人审核通过，若未通过审核，采购人有权要求整改完善，直至审核通过。

7.其他要求

（1）供应商提供的产品及产出成果，不得侵犯第三方知识产权及其他合法权益，并自行承担相关责任。根据采购人业务需求所做的改造部分的开发成果知识产权归采购人所有。

批注 [v4]:

（2）供应商对获知的采购人未通过正式渠道公开的客户信息、经营信息、产品信息等商业秘密以及本项目相关信息、产出成果应严格保密，由此给采购人造成损害的，需赔偿采购人损失。

（3）提供代码助手在生产环境使用时的代码采纳率案例（用百分比表示），需提供真实佐证；提供大模型知识库在生产环境中的问答准确率案例（用百分比表示），需提供真实佐证。供应商需承诺对提供数据真实性负责，后续过程招标方将对该数值进行核对，如果实际出现较大偏离，招标方有权要求中标方承担相应责任。

（4）项目实施完成后，如相关后续事项需供应商协助处理的，供应商应提供必要的协助工作。

**（六）验收交付**

系统上线并稳定运行 6 个月，我行与供应商双方对软件开发成果进行整体项目验收。验收时由我行与供应商双方（或我行指定的第三方）人员共同组成验收小组开展验收工作。验收依据为《项目工作说明书》，软件开发成果达到《项目工作说明书》规定的功能和性能要求的，验收小组共同签署《项目验收确认书》，视为通过验收。对规定试运行期满，不能达到验收标准的，视其情况，延长试运行期。