项目需求书

本项目采购标的所属行业为：软件和信息技术服务业。

（一）项目概况

在数字经济时代，采购人作为知识创新与人才培养的核心阵地，算力资源已成为支撑教育数字化转型的战略性基础设施。本项目目标聚焦于构建基于智能计算能力的智慧化知识服务体系，通过整合高性能算力资源与知识管理平台，全面提升师生对海量学术资源的获取、分析与应用能力，推动教学科研范式革新。

本项目为支撑采购人教学、科研、管理与服务场景中的大模型应用建设，拟采购大模型 MaaS 服务（模型即服务，Model as a Service，MaaS），形成如下三项服务能力。

1、平台可面向校内教职工提供国内顶尖大模型调用 API 服务；

2、平台具有较完备的权限控制与模型用量检测，有效的控制大模型的使用成本；

3、建设一套面向软件工程的 AI 编码助手工具，提高软件编码效率。

（二）技术要求

序号	分项名称	技术参数要求
一、大模型调用接口		
1	大模型调用接口规格	提供以下大模型调用接口服务： 大语言模型提供至少 9 种模型。整体包括模型调用服务、技术支持、售后保障、安全运维等相关配套服务。具体模型需求如下： 1、通义系列包括： ※qwen3.8-max； qwen3.7-max； qwen3.7-plus； ※qwen3.8-flash；

		qwen3.8-27b； ※qwen3.8-omni-flash； 2、Deepseek 系列包括： ※deepseek-v4-pro-0813； 3、GLM 系列包括： glm-5.3； 4、Kimi 系列包括： Kimi K3； 5、另外还可提供图片生成与编辑、视频生成与编辑、语音合成、语音生成、全模态等大模型，满足不同业务场景下的特色需求。 ※其中文生视频模型要求单次生成时长最长 30 秒、输出帧率不低于 30fps，支持首帧/首尾帧控制和视频编辑与延长，提供模型的官网截图。 需提供以上每一个模型在官网或控制台截图并加盖供应商公章（用于证明以上模型上架情况）。
2		通过 Maas 服务平台提供多地域部署的模型推理 API 服务,支持如下服务方式。 ※1、OpenAI 兼容接口：平台模型支持 OpenAI 兼容接口，只需调整 API Key、BASE_URL 和模型名称，即可将原有 OpenAI 代码迁移至本平台的模型服务。 ※2、批量推理：支持上传包含多个请求的 JSONL 文件，创建批量推理任务。系统在后台队列中处理任务。可通过控制台或 API 查询任务进度和状态。任务完成后，系统生成结果文件和错误文件，提供官网或控制台截图并加盖供应商公章。 ※3、支持每分钟 Token（TPM）吞吐性能模式：为指定模型锁

		定专属推理容量，预留容量内的调用不受公共资源限流影响。支持预留的 TPM 容量业务专属，不与其他用户共享。该模式下，支持的模型包括但不限于千问 3.8-max、glm5.3、deepseek-v4-pro。提供官网或控制台截图并加盖供应商公章。 ※4、平台支持以独占 GPU 资源为指定模型提供性能和吞吐保障，支持基础模型与自定义模型部署，该模式下，支持的模型包括但不限于 qwen3.7-plus-2026-05-26 、 glm5.2 、deepseek-v4-flash。提供官网或控制台截图证明材料，并加盖供应商公章。
3		API 采用 API Key 方式进行身份认证，具体机制如下： 1、认证方式：通过 HTTP 请求头 Authorization: Bearer <API_KEY> 传递凭证。 2、API Key 获取：平台支持创建 API Key。 3、权限体系：支持主账号、RAM 子账号、业务空间多级权限管理，可对子空间设置模型调用权限。 4、API Key IP 地址白名单：支持通过 IP 地址白名单的方式，控制大模型接口仅响应白名单内的访问请求。提供官网或控制台截图并加盖供应商公章。 5、API Key 使用模型管理：支持针对 API Key 的模型授权管理，有效限制单个 API Key 调用模型服务的范围，提供官网或控制台截图并加盖供应商公章。
4		※需支持 token 用量统计，支持按月、天、小时、产品/模型类型、接入点维度统计输入、输出 token 用量。支持查看日账单、月账单、消费明细。支持按模型或 API Key 查询消费费用。提供官网或控制台截图并加盖供应商公章。
5		对于本项目采购的模型接口服务，如果模型服务发生变化，包

		括但不限于模型升级、模型下线，需及时通过短信和邮件通知采购人，模型下线通知中需提供替代模型建议。 问题反馈与改进：建立问题反馈机制，定期收集采购人对模型服务、技术支持、售后保障等方面的意见和建议，根据采购人需求优化服务流程、提升服务质量。
6		1、数据安全：需严格遵守《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等相关法律法规，建立完善的数据安全管理制度，确保采购人的数据安全，严禁收集、存储采购人未授权的任何数据（包括输入文本、生成内容、业务数据等）用于本项目以外的其他用途。供应商需提供承诺书（格式自拟）并加盖公章。 采购人通过 API 调用服务时传入的输入数据和获得的输出结果，投标人需不得将其用于模型训练或改进。 2、模型安全：需确保提供的模型服务无安全漏洞、无恶意代码，不会对采购人的业务系统、服务器环境造成破坏；定期对模型进行安全检测、漏洞扫描，及时修复安全隐患。 3、接入安全：平台需支持以加密方式接入模型推理功能。
二、AI 编码助手工具		
1	使用功能要求	SaaS 化的部署方式： 无需采购人额外投入服务器、网络设备等硬件资源，供应商负责完成全部部署、调试工作，确保系统正常运行。
2		需支持以下 IDE，且插件安装简单、兼容稳定，无明显卡顿、崩溃问题： 1、基础 IDE：VS Code、IntelliJ、PyCharm； 2、补充 IDE：WebStorm、GoLand。
3		※支持多智能体专家协作机制：采购人提出需求后，系统自动

		进行任务拆解，组建包含方案设计、编码实现、测试验证与质量检查等角色的专家团队，并行执行并交付可落地的工程结果。多智能体协同支持增强提示词、skill、mcp 等功能。要求专家团调用不限次数。提供官网或产品功能截图并加盖供应商公章。
4		※支持对整个代码仓库进行结构化文档自动生成，并增量跟踪代码变更。提供命令行干预与配置化管理（wiki_plan.yaml）。生成结果支持团队级共享，新成员加入即获得完整工程上下文。提供官网或产品功能截图并加盖供应商公章。
5		※内置工程专家智能体，能够自主审视项目的工程规范（Harness），诊断工程短板并生成改进建议，支持一键启动修复任务，持续推动项目工程质量演进。提供官网或产品功能截图并加盖供应商公章。
6		具备以下核心编程辅助功能 1、代码补全：支持实时代码补全，可根据上下文智能推荐代码片段、函数、变量，支持跨文件补全； 2、智能调试：可识别代码语法错误、逻辑错误，提供精准的错误定位及修改建议，支持断点调试辅助； 3、跨文件分析：可分析多文件间的依赖关系、函数调用关系，提供代码导航、重构建议； 4、智能问答：支持自然语言提问，解答编程相关问题（如语法、框架使用、bug 排查等； 5、代码优化：可对现有代码进行优化建议，提升代码可读性、执行效率，支持代码规范检查。
7		企业级管理功能： 1、授权管理：支持统一创建、分配、回收用户授权，可按部

		门、岗位划分授权范围，实时查看授权使用情况； 2、审计日志：记录所有用户的操作行为（如登录、代码查询、调试等），日志留存≥1 年，可导出、查询； 3、安全控制：支持 IP 白名单设置，仅允许指定 IP 范围内的设备登录使用，防止未授权访问； 4、数据统计：可统计用户使用频率、功能使用情况、代码补全准确率等数据，生成企业级使用报表。
8		需全面支持以下编程语言，无功能缺失，适配各类开发场景： 1、后端语言：Java、Python、Go、PHP； 2、前端语言：JavaScript、TypeScript、HTML、CSS； 3、其他语言：C、C++、R、Ruby。
9	安全要求	1、数据加密：采用 HTTPS 加密传输，数据存储采用对称、非对称加密算法，防止数据泄露、篡改； 2、隐私保护：供应商不得收集、泄露、滥用采购人的代码数据、商业信息，合同终止后，需删除所有相关数据（经采购人书面同意留存的除外）；
10	售后服务要求	1、响应时效：提供 7×12 小时技术支持（工作日 9:00-21:00，节假日 10:00-18:00），电话、在线客服响应时间≤30 分钟，故障解决时间≤4 小时； 2、技术培训：提供不少于 1 次技术培训，覆盖全部授权采购人，培训内容包括功能使用、常见问题排查等，培训形式可选择线上或线下； 3、版本更新：服务期内，提供产品版本更新、功能升级服务，及时修复产品 bug； 4、售后保障：建立专属售后对接人制度，提供对接人姓名、联系电话、邮箱，确保沟通顺畅

注：

1. 加注“※”号的参数为重点参数。

（三）需求服务清单

模型	服务项目	权重	基准价	基准价链接
模型 1	qwen3.8-max	15%	百万 Tokens 输入 12 元 百万 Tokens 输出 36 元 百万 Tokens 输入（缓存命中）1.5 元	https://help.aliyun.com/zh/model-studio/qwen3-8-max?spm=a2c4g.11186623.help-menu-2400256.d_0_15_0_0_0.446e92a7r7ikhZ
模型 2	qwen3.8-flash	10%	百万 Tokens 输入 0.8 元 百万 Tokens 输出 2.7 元 百万 Tokens 输入（缓存命中）0.1 元	https://help.aliyun.com/zh/model-studio/qwen3-8-flash?spm=a2c4g.11186623.help-menu-2400256.d_0_15_0_0_9.505192a732pIOM&scm=20140722.H_1000000054157._.OR_help-T_cn~zh-V_1
模型 3	qwen3.8-omni-flash	15%	百万 Tokens 输入 0.8 元 百万 Tokens 输出 2.7 元 百万 Tokens 输入（缓存命中）0.1 元	https://help.aliyun.com/zh/model-studio/model-pricing?spm=a2c4g.11186623.0.i3#4c2e910ce4pcq
模型 4	文生视频模型	20%	每秒视频生成（480P）0.3 元 每秒视频生成（720P）0.6 元 每秒视频生成（1080P）1.2 元	https://help.aliyun.com/zh/model-studio/wan3-0-video?spm=a2c4g.11186623.help-menu-2400256.d_0_15_0_4_8.7d4afe0073o4kM
模型 5	deepseek-v4-pro-0813	5%	百万 Tokens 输入（缓存未命中、空闲时段）4.5 元 百万 Tokens 输入（缓存未命中、高峰时段）9.0 元 百万 Tokens 输出（空闲时段）13.5 元 百万 Tokens 输出（高峰时段）27.0 元	https://api-docs.deepseek.com/zh-cn/quick_start/pricing/
模型 6	qwen3.7-m	10%	百万 Tokens 输入 12	https://help.aliyun.com/zh/

	ax		元 百万 Tokens 输出 36 元 百万 Tokens 输入（缓存命中）2.4 元	model-studio/qwen3-7-max?spm=a2c4g.11186623.help-menu-2400256.d_0_15_0_0_1.adb0556fttrTnu
模型 7	qwen3.7-plus	5%	输入<=256k 条件下： 百万 Tokens 输入 2 元 百万 Tokens 输出 8 元 百万 Tokens 输入（缓存命中）0.4 元	https://help.aliyun.com/zh/model-studio/qwen3-7-plus?spm=a2c4g.11186623.help-menu-2400256.d_0_15_0_0_5.167492a7sWNumr
模型 8	qwen3.8-27b	10%	百万 Tokens 输入 3 元 百万 Tokens 输出 12 元 百万 Tokens 输入（缓存命中）0.6 元	https://help.aliyun.com/zh/model-studio/qwen3-7-flash?spm=a2c4g.11186623.help-menu-2400256.d_0_15_0_0_10.617971c0iB7Vdy
模型 9	glm-5.3	5%	百万 Tokens 输入 8 元 百万 Tokens 输出 28 元	https://docs.bigmodel.cn/cn/guide/start/pricing
模型 10	kimi-k3	5%	百万 Tokens 输入 20 元 百万 Tokens 输出 100 元 百万 Tokens 输入（缓存命中）2 元	https://platform.kimi.com/docs/pricing/chat