跨平台智能算法模型转化工具采购项目

报 价 书

供应商： 中南大学

2025年 09 月 17 日

目 录

[一、 报 价 函 1](#_Toc23)

[二、 授权委托书 2](#_Toc2303)

[三、 资质文件 3](#_Toc25314)

[3.1 事业单位法人证书复印件 3](#_Toc30484)

[3.2 开户行出具的银行资信证明 4](#_Toc5597)

[3.3 2022年至2024年资产负债表 6](#_Toc9367)

[3.4 2022年至2024年收入支出表 7](#_Toc32659)

[3.5 履行合同所必须的设备和专业技术能力证明材料 8](#_Toc16338)

[3.6 ISO质量管理体系证书 11](#_Toc5999)

[3.7 近一年中任意一个月依法纳税的凭证 12](#_Toc25532)

[3.8 近一年中任意一个月缴纳社会保障资金的凭证（复印件加盖公章） 13](#_Toc28888)

[3.9 “信用中国”信息查询截图 17](#_Toc28971)

[3.10 主要负责人、股东或出资人信息 18](#_Toc16435)

# 报 价 函

中国船舶集团有限公司第七一六研究所：

1．我方已仔细研究了 跨平台智能算法模型转化工具 询价文件的全部内容，愿意以人民币（大写） 陆拾伍万元 （¥ 650000 ）的报价（其中不含税价为： ¥ 631067 ；增值税税额为： ¥ 18933 ）完成服务，并按照合同预定履行义务。

2. 我方的报价书包括下列内容：

（1）报价函；

（2）授权委托书；

（3）报价一览表；

（4）报价方案；

（5）资质文件。

……

报价书的上述组成部分如存在内容不一致的，以报价函为准。

3．我方承诺除商务和技术偏差表列出的偏差外，我方响应询价文件的全部要求。

4．我方承诺在询价文件规定的报价书有效期内不撤销报价书。

5．如我方成交，我方承诺：

（1）响应询价文件的全部要求

（2）在收到成交通知书后，在成交通知书规定的期限内与你方签订合同；

（3）在签订合同时不向你方提出附加条件；

（4）按照询价文件要求提交履约保证金；

（5）在合同约定的期限内完成合同规定的全部义务。

6．我方在此声明，所递交的报价书及有关资料内容完整、真实和准确，且不存在第一章“询价采购公告/询价采购邀请书”中规定的供应商不得存在的情形。

7． 无 （其他补充说明）。

供 应 商：中南大学 （盖单位章）

法定代表人（单位负责人）或其委托代理人：（盖校长签字章）

地 址：湖南长沙岳麓区麓山南路 932 号

网 址：https://www.csu.edu.cn/

电 话：0731-88836754

传 真：0731-88710591

邮政编码：410083

2025 年 09 月 18 日

# 授权委托书

本人 李建成 系 中南大学 的法定代表人（单位负责人），现委托 李双琳 为我方代理人。代理人根据授权，以我方名义签署、澄清确认、递交、撤回、修改 跨平台智能算法模型转化工具 询价采购项目报价书、商定合同和处理有关事宜，其法律后果由我方承担。

委托期限：自本委托书签署之日起至 跨平台智能算法模型转化工具 询价采购项目签订采购合同之日止。代理人无转委托权。

附：法定代表人（单位负责人）身份证复印件及委托代理人身份证复印件

法定代表人身份证复印件

|  |  |
| --- | --- |
|  |  |

委托代理人身份证复印件

|  |  |
| --- | --- |
|  |  |

供 应 商：中南大学

法定代表人（单位负责人）：

身份证号码：420106196412274833

委托代理人： （签字）

身份证号码：430521198608164739

2025 年 09 月 18 日

# 报价一览表

|  |  |  |
| --- | --- | --- |
| **预算科目** | **金额（万元）** | **说明** |
| 一、直接费用 | 21.00 |  |
| （一）设备费 | 8.00 |  |
| 1．设备购置费 | 0.00 | 指对外购买设备。 |
| 2．设备试制费 | 0.00 | 现有仪器设备无法满足项目（课题）检测、实验、验证或示范等研究任务需要而试制专用仪器设备发生的费用。 |
| 3．设备改造与租赁费 | 8.00 | 研发测试过程中租用云服务器和存储的费用。 |
| （二）业务费 | 8.00 |  |
| 1.材料费 | 0.00 |  |
| 2.测试化验加工费 | 0.00 | 一般不得超过总预算的30%。 |
| 3.差旅费 | 6.00 |  |
| 4.会议费 | 0.00 |  |
| 5．出版/文献/信息传播/知识产权事务费 | 0.40 | 用于购买科研文献、技术标准手册，均为项目必需的参考资料 |
| 6.其他费用 | 1.60 | 邀请 5 名行业专家（高校教授、企业高级工程师）参与需求论证与验收评审，按 4 号文 “技术咨询类专家人均 1600 元 / 天” 标准，共 2 天。 |
| （三）人力成本 | 5.00 |  |
| 1．劳务费（外部派遣或聘用人员） | 3.50 | 委托专业团队开展模型适配测试。 |
| 1. 临时聘用技术人员费 | 1.50 |  |
| 二、间接费 | 35.00 |  |
| 1．在职员工人力成本摊销 | 32.00 | 在本单位入职人员成本。 |
| 2. 使用已有设备成本摊销 | 3.00 |  |
| 三、管理费 | 4.00 | 按总预算 6.5% 计提 |
| 四、其他支出 | 5.00 | 预留紧急支出风险费和缴税费用 |
| 五、收益 | 0.00 |  |
| **总计** | **65.00** |  |

# 报价方案

## （一）技术响应偏离表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 条目 | 招标文件要求 | 投标文件响应 | 偏离情况 |
| 1 | 功能指标 | 可支持模型的跨平台、跨底层架构的格式转化 | 支持模型的跨平台、跨底层架构的格式转化 | 无偏离 |
| 2 | 可根据模型格式进行模型可视化 | 具备模型可视化功能，可根据模型格式选择 | 无偏离 |
| 3 | 性能指标 | 可支持不少于20 类主流目标检测、目标分类、目标跟踪、自然语言大模型等智能体模型的网络结构可视化 | 支持目标检测、目标分类、目标跟踪、自然语言大模型等20 类主流智能体模型的网络结构可视化 | 无偏离 |
| 4 | 可支持不少于20 类主流目标检测、目标分类、目标跟踪、自然语言大模型等智能体模型的格式转化 | 支持目标检测、目标分类、目标跟踪、自然语言大模型等20 类主流智能体模型的格式转化 | 无偏离 |
| 5 | 支持不少于4 种包括华为、瑞星微、英伟达、寒武纪平台模型的转化 | 支持华为、瑞星微、英伟达、寒武纪平台模型的转化 | 无偏离 |
| 6 | 硬件运行环境 | 可适配英特尔、鲲鹏CPU，英伟达、华为、寒武纪GPU进行部署 | 适配英特尔、鲲鹏CPU，英伟达、华为、寒武纪GPU进行部 | 无偏离 |
| 7 | 软件运行环境 | 操作系统支持Linux（Ubuntu20.04 及以上）、麒麟v10 或Windows Server 2019 及以上 | 支持Linux（Ubuntu20.04 及以上）、麒麟v10 或Windows Server 2019 及以上操作系统 | 无偏离 |

## （二）技术方案

本项目技术方案聚焦“跨平台智能模型转化工具”在试验场内场调试机的内嵌应用需求，以 “解决模型跨架构格式壁垒、降低模型理解门槛” 为核心目标，构建 “B/S 架构支撑 + 核心功能模块 + 全流程管控” 的技术体系，严格匹配 “20 类模型适配、4 类平台兼容” 的性能指标与 “跨平台转换、模型可视化” 的功能指标。

系统采用分层架构设计，自下而上分为硬件资源层、数据存储层、应用服务层与用户层：硬件资源层部署华为昇腾、瑞芯微、英伟达、寒武纪 4 类平台真机测试节点，通过 Docker 容器化统一管理运行环境；数据存储层采用 MySQL+MinIO+Redis 混合架构，分别存储结构化任务数据、非结构化模型文件与高频缓存数据；应用服务层拆分为转换系统、可视化系统与支撑服务，实现微服务化协同；用户层支持主流浏览器与响应式界面，确保调试机操作便捷。

核心功能模块围绕两大指标展开：一是跨平台格式转换模块，通过 “原生格式→ONNX 中间格式→目标平台格式” 两步流程，集成 4 类平台专用转换工具（华为 ATC、瑞芯微 RKNN Toolkit 等），开发 20 类模型专属适配模板，解决算子不兼容与精度损失问题，转换成功率≥98%；二是模型可视化模块，基于 Netron 二次开发，支持网络结构拓扑图展示、层参数详情查看与跨平台结构对比，针对计算机视觉与 NLP 模型提供差异化视图，可视化响应时间≤2 秒。

同时，配套开发跨平台推理测试模块与任务监控模块：推理测试模块从功能、性能、精度三维度验证模型适配效果，自动生成可视化报告；任务监控模块实现全流程任务跟踪与异常告警，确保短周期内高效交付。整体技术方案兼顾兼容性、稳定性与易用性，可直接支撑试验场智能模型部署测试需求。

### 1.系统设计

#### (1)轻量化 B/S 架构支撑系统

作为工具运行的基础骨架，该系统以 “资源精简、响应高效” 为设计原则，适配调试机内网环境与有限硬件资源：

用户层：仅支持 Chrome、Edge 两款主流浏览器（减少兼容性开发成本），界面采用 “功能分区 + 快捷工具栏” 设计，核心操作（模型上传、转换配置、结果查看）均在首页完成，无需多页跳转。支持单文件拖拽上传（最大支持 3GB 模型，适配调试机存储容量）、转换参数一键保存（常用配置可预设为模板），进度展示采用 “环形进度条 + 关键节点提示”，直观呈现任务状态。

应用服务层：采用 “核心服务 + 插件扩展” 架构，仅保留模型解析、转换调度、可视化渲染 3 个核心服务，其余功能（如日志管理、权限控制）以轻量化插件形式加载，减少调试机内存占用。前端基于 Vue3+Vite 开发（构建速度提升 60%），界面元素采用按需加载；后端基于 Flask 框架（轻量型 Python 框架），接口响应时间≤300ms，适配调试机低算力环境。

数据存储层：采用 “精简混合存储” 方案，MySQL 8.0 仅存储任务基础信息（任务 ID、模型名称、转换状态）与用户配置，剔除冗余字段；MinIO 采用单节点部署（初期无需分布式扩展），仅存储原始模型、转换后模型两类核心文件，日志文件按 “7 天自动清理” 策略管理，避免调试机存储空间溢出；Redis 仅缓存任务实时状态与高频访问的模型元数据（输入输出维度、算子数量），缓存有效期缩短至 12 小时，释放内存资源。

硬件资源层：每类平台仅部署 1 台核心测试节点（华为昇腾 Atlas 200I、瑞芯微 RK3588、英伟达 Jetson AGX Orin、寒武纪思元 370），通过 Conda 虚拟环境替代 Docker 容器（减少调试机虚拟化资源消耗），为每类平台预置最小化运行环境（仅含驱动、核心 SDK 与依赖库），环境部署时间≤2 小时，适配调试机快速调试需求。

#### (2)高效跨平台模型转换系统

聚焦 “低精度损失、高转换效率”，针对调试机场景优化转换流程，核心包含四大模块：

精简转换工具集：以 ONNX Runtime 为基础框架，仅集成 4 类平台必需的转换工具，剔除冗余功能。华为昇腾采用 ATC 工具轻量版（仅保留格式转换与基础算子融合功能）；瑞芯微选用 RKNN Toolkit Lite（内存占用降低 40%），仅支持 Min-Max 量化（满足调试机快速量化需求）；英伟达采用 TensorRT 简化版（关闭动态形状优化，减少计算资源消耗）；寒武纪使用 MagicMind 基础版，仅保留核心算子映射。同时集成 ONNX Simplifier 轻量插件，仅执行冗余算子删除、常量节点合并两项简化操作，简化耗时缩短至原流程的 50%。

模型适配模板库：针对 20 类主流模型（目标检测：YOLOv5/v8、Faster R-CNN；目标分类：ResNet50、MobileNetV3；目标跟踪：SORT、DeepSORT；NLP 大模型：BERT-base、GPT-2），开发 “最小化适配模板”，每个模板仅保留模型结构解析规则与关键算子映射表，剔除非必要优化策略。例如 YOLO 模板仅聚焦检测头 Anchor 信息保留与 backbone 算子适配，BERT 模板仅优化 Transformer 层注意力计算逻辑，模板文件大小均控制在 100KB 以内，加载速度提升 3 倍。

两步式转换逻辑模块：优化 “原生格式→ONNX→目标平台格式” 流程，第一步仅解析模型拓扑结构与核心参数（忽略权重细节校验），PyTorch/TensorFlow 模型解析耗时≤1 分钟；第二步根据目标平台自动选择 “快速转换模式”，华为昇腾跳过算子性能优化、瑞芯微关闭量化校准数据生成，转换耗时较常规流程缩短 40%。转换过程仅记录关键日志（解析开始 / 结束、转换成功 / 失败、错误信息），日志文件大小控制在 10MB 以内，适配调试机存储需求。

轻量任务调度模块：采用 “本地任务队列 + 单节点调度” 替代分布式架构，用户提交的转换任务直接进入调试机本地队列，按 “提交时间先后” 排序（无需负载均衡，减少调度开销）。任务状态仅包含 “等待中、处理中、成功、失败”4 种，无中间过渡状态，前端通过定时轮询（轮询间隔 5 秒）获取状态更新，减少调试机网络通信压力；转换完成后仅校验输入输出维度一致性，跳过权重哈希值对比，校验耗时≤10 秒。

#### (3)直观化模型可视化系统

以 “简洁清晰、重点突出” 为原则，针对调试机小屏显示优化可视化效果，核心包含三大模块：

基础网络结构可视化模块：基于 Netron 核心解析能力开发精简版渲染引擎，仅提取模型关键层（输入层、卷积层、全连接层、输出层）与层间连接关系，忽略 BN 层、激活层等次要结构，拓扑图节点数量减少 60%。不同层类型仅用 “颜色 + 图标” 简单区分（卷积层蓝色方块、全连接层绿色圆形），支持鼠标滚轮缩放（缩放范围 50%-200%）与节点单击查看基础参数（仅显示算子类型、输入输出维度），避免界面信息过载。

场景化可视化适配模块：仅针对计算机视觉、NLP 两类模型开发专用视图，计算机视觉模型采用 “纵向分层” 展示（输入层→backbone→输出层），标注每层输出特征图尺寸；NLP 模型采用 “横向链式” 展示（词嵌入层→Transformer 层→分类层），仅突出注意力头数量与隐藏层维度两个关键参数。不支持自定义配置功能，仅保留 “默认视图 + 黑白主题” 两种展示模式，减少调试机渲染压力。

快速渲染优化模块：采用 “预渲染 + 局部更新” 策略，模型上传后先预渲染整体拓扑结构，用户点击节点时仅局部加载该节点参数，初始渲染时间≤2 秒；采用 Canvas 绘图替代 SVG（减少 DOM 元素数量），大模型（500 层以上）可视化界面帧率保持在 30fps 以上；权重数据不支持可视化展示，仅提供 “参数范围文本提示”（如 “权重值范围：-1.2~1.5”），平衡展示需求与调试机性能。

#### (4)实用化硬件集成验证系统

聚焦调试机 “快速验证、问题定位” 需求，简化验证流程，核心包含三大模块：

精简硬件测试集群：每类平台仅保留 1 台测试节点，节点状态仅监测 “在线 / 离线” 与 “CPU/NPU 利用率” 两项核心指标，通过简单指示灯（绿色在线、红色离线）在界面展示，无需复杂监控面板。节点访问采用 SSH 直连方式，避免部署额外监控服务，减少调试机资源消耗。

基础验证模块：仅开展功能验证与简易性能测试，功能验证仅检查模型是否能正常推理（输出非空结果），不校验结果格式合法性；性能测试仅采集单轮推理延迟（取 10 轮平均值），不计算吞吐量与资源占用峰值，测试数据直接以文本形式展示（如 “推理延迟：250ms”），无需生成复杂图表。精度对比仅针对目标检测模型（YOLO 系列）与分类模型（ResNet 系列），采用 mAP 与 Top-1 Accuracy 两个核心指标，以英伟达 Jetson AGX Orin（FP32 精度）为基准，精度损失阈值放宽至≤8%，减少调试机测试时间。

简易报告生成模块：验证完成后生成 TXT 格式简易报告，仅包含测试模型名称、目标平台、测试时间、功能验证结果（成功 / 失败）、推理延迟、精度损失 6 项核心信息，报告大小≤100KB，支持直接在调试机打开查看，无需额外安装 PDF/HTML 阅读器，适配调试机简化需求。

### **2.架构设计**

本项目架构突破传统分层架构局限，针对 “试验场内场调试机内嵌” 场景，采用 “核心引擎 + 场景化插件 + 轻量化支撑” 的三维架构设计，以 “低资源占用、高适配效率、快响应速度” 为核心目标，实现 20 类模型与 4 类平台的精准适配，同时满足调试机轻量化运行需求，架构分为 “核心引擎层、场景化插件层、轻量化支撑层” 三层，各层职责明确且协同高效。

#### (1)核心引擎层：架构核心驱动力

作为整个工具的技术底座，聚焦 “模型解析、格式转换、结构可视化” 三大核心能力，提供标准化接口供上层调用，所有引擎均做轻量化优化，适配调试机有限算力：

多格式模型解析引擎：整合 ONNX Runtime 基础解析能力与自定义格式适配逻辑，仅保留 PyTorch（.pth/.pt）、TensorFlow（.pb/.saved\_model）、ONNX（.onnx）三类核心原生格式解析功能，剔除小众格式支持，解析时仅提取模型拓扑结构、输入输出维度、核心算子类型三类关键信息，解析耗时≤90 秒（1GB 模型），内存占用控制在 512MB 以内。

跨平台转换核心引擎：以 ONNX 为中间格式枢纽，内置 “原生格式→ONNX” 统一转换逻辑与 “ONNX 优化器”（仅保留算子融合、常量折叠两项核心优化），对外提供标准化 “ONNX→目标平台格式” 转换接口，支持 4 类平台插件接入，引擎自身不包含平台专用逻辑，确保轻量可扩展。

模型可视化渲染引擎：基于 Netron 精简版内核开发，仅保留网络结构拓扑图渲染、层参数基础展示两项核心功能，采用 “静态预渲染 + 局部交互” 模式，加载模型时先生成整体拓扑图静态画面，用户点击节点时再动态加载该节点参数，渲染响应时间≤1.5 秒（500 层模型），不支持复杂动画效果，降低调试机 GPU 消耗。

#### (2)场景化插件层：功能扩展载体

围绕 “4 类平台适配、20 类模型优化” 场景需求，以插件形式实现差异化功能，插件可按需加载 / 卸载，避免资源浪费，核心插件包括：

平台专属转换插件：为华为昇腾、瑞芯微、英伟达、寒武纪 4 类平台各开发 1 个轻量化插件，仅包含该平台专用转换工具调用逻辑（如华为插件调用 ATC 工具、瑞芯微插件调用 RKNN Toolkit）与基础算子映射表，插件大小均控制在 50MB 以内，加载时间≤10 秒，支持独立更新（无需重启核心引擎）。

模型适配优化插件：针对 20 类主流模型（目标检测 6 类、目标分类 5 类、目标跟踪 4 类、NLP 大模型 5 类），每类模型开发 1 个适配插件，仅包含该模型专属结构解析规则（如 YOLO 插件保留检测头 Anchor 信息、BERT 插件优化 Transformer 层逻辑）与精度校准参数，插件可通过 “模板导入” 快速新增，无需修改核心代码。

推理测试简化插件：仅包含功能验证（检查推理输出非空）、基础性能测试（采集单轮推理延迟）两项测试逻辑，支持 4 类平台测试节点接入，测试时不生成复杂图表，仅输出文本化测试结果，插件内存占用≤256MB，适配调试机快速验证需求。

#### (3)轻量化支撑层：运行保障基础

为核心引擎与插件提供基础支撑，所有组件均做精简设计，满足调试机内网环境与低资源需求：

资源调度组件：采用 “本地任务队列 + 单节点调度” 模式，不依赖分布式框架，用户提交的转换、测试任务直接进入调试机本地队列，按 “提交时间” 排序执行，仅监控 CPU 利用率（阈值≤80%）、内存占用（阈值≤85%）两项节点负载指标，避免调度逻辑复杂导致的资源消耗。

数据存储组件：采用 “极简混合存储” 方案，MySQL 仅存储任务基础信息（任务 ID、模型名称、状态）与用户配置，MinIO 单节点部署仅存储原始模型、转换后模型两类核心文件（日志文件 7 天自动清理），Redis 仅缓存任务实时状态（有效期 6 小时），三类存储组件总内存占用控制在 2GB 以内。

用户交互组件：开发轻量化 Web 前端，仅包含模型上传、转换配置、可视化查看、任务监控四大核心功能界面，采用 “单页应用” 设计，不支持多标签页操作，界面元素按需加载（滚动时加载可视区域内容），前端资源总大小≤2MB，首次加载时间≤3 秒，适配调试机内网低带宽环境。

### **3.核心引擎层：架构核心驱动力**

模型管理作为系统核心基础功能，覆盖智能模型从“接入-维护-版本控制-存储-检索”的全生命周期管理，需兼容主流光模块及交换机（网络传输基础能力不是本项目重点），同时满足 20 类智能模型（目标检测/分类/跟踪/NLP 大模型）的高效管理需求，为后续跨平台转换、可视化展示提供标准化模型数据源。

#### (1)功能描述

核心引擎层承担模型处理全流程的核心逻辑，不包含场景化适配细节（如平台专用算子映射、模型专属优化策略），仅提供通用化、可扩展的基础能力，具体功能如下：

多格式模型解析功能：支持 PyTorch（.pth/.pt）、TensorFlow（.pb/.saved\_model）、ONNX（.onnx）三类主流原生格式的解析，自动提取模型关键信息 —— 拓扑结构（层与层的连接关系）、输入输出维度（如 [1,3,640,640]）、核心算子类型（如 Conv2d、Linear、MultiHeadAttention），忽略权重数据校验、冗余算子详情等非必要信息，确保解析效率；解析结果以标准化 JSON 格式输出，包含 “layers”（层信息数组）、“connections”（层连接关系数组）、“io\_info”（输入输出信息）三类核心字段，供转换引擎与可视化引擎调用。

跨平台转换基础功能：实现 “原生格式→ONNX 中间格式” 的统一转换，支持自动处理不同框架的格式差异（如 TensorFlow 的 NHWC 数据排布转为 ONNX 标准的 NCHW 格式、PyTorch 的动态 BatchSize 标记为固定维度）；内置 ONNX 优化模块，仅保留 “算子融合”（如 Conv+BN+ReLU 合并为单个算子）、“常量折叠”（将常量节点与计算节点合并）两项核心优化，减少后续平台转换的复杂度；对外提供 “ONNX→目标平台格式” 的标准化接口，支持 4 类平台插件接入，引擎自身不存储任何平台专用逻辑，确保功能通用性。

模型可视化基础功能：支持模型网络结构的静态拓扑图渲染，自动按层类型分配差异化标识（卷积层用蓝色方块、全连接层用绿色圆形、Transformer 层用紫色菱形），清晰展示层间数据流向；提供基础交互能力，用户点击层节点时可查看该层的核心参数（算子类型、输入输出维度、卷积核大小 / 步长等关键配置），不支持复杂交互（如节点拖拽、多层级折叠）；可视化结果支持导出为 PNG 格式（分辨率默认 1920×1080，适配调试机显示），无 SVG、PDF 等复杂格式导出功能，降低资源消耗。​

#### (2)技术实现方案

围绕 “轻量化、高效化、可扩展” 原则，采用精简开发策略与模块化设计，确保引擎在调试机环境下稳定运行，具体技术方案如下：

多格式模型解析引擎实现：

基于 ONNX Runtime 1.15 基础解析库，整合 PyTorch、TensorFlow 的专用解析工具（torch.onnx.export、tf2onnx），通过 “格式判断→工具调用→信息提取→标准化输出” 四步流程实现解析；格式判断通过文件后缀与头部标识双重判定（如.pth 文件匹配 “PK” 头部标识），避免格式误判。

解析过程采用 “异步处理 + 内存限制” 策略，通过 Python 多线程异步执行解析任务，避免阻塞主线程；利用 memory\_profiler 工具监控内存占用，当解析 1GB 以上模型时自动触发 “分块解析”（按层结构拆分模型，逐块提取信息），将内存占用控制在 512MB 以内；解析耗时通过 “关键步骤计时” 优化，如 PyTorch 模型解析时跳过权重加载仅提取结构信息，解析耗时缩短至 90 秒以内（1GB 模型）。

解析结果采用 “最小化 JSON” 格式，剔除 “层描述备注”“算子版本号” 等冗余字段，仅保留 “layer\_id”“layer\_type”“input\_shape”“output\_shape”“key\_params”（核心参数）等必要信息，JSON 文件大小控制在 100KB 以内（500 层模型），便于后续引擎快速读取。

跨平台转换核心引擎实现：

采用 “中间格式解耦” 架构，“原生格式→ONNX” 转换逻辑通过统一封装的 Python 类实现（如 Pytorch2OnnxConverter、TensorFlow2OnnxConverter），类内包含格式差异处理函数（如 NHWC2NCHW ()、Dynamic2FixedBatch ()），确保转换结果符合 ONNX 1.8~1.12 版本规范。

ONNX 优化模块基于 onnxoptimizer 开源库，仅启用 “fuse\_conv\_bn”“fold\_constants” 两项优化 pass（移除 “eliminate\_unused\_initializers”“fuse\_pad\_into\_conv” 等非必要优化），优化耗时控制在 30 秒以内（1GB 模型）；优化后的 ONNX 模型自动生成 “优化报告”（TXT 格式），记录融合算子数量、常量折叠节点数，便于问题排查。

标准化接口设计采用 RESTful 风格，定义 “POST /api/engine/convert” 接口，接收 “onnx\_path”（ONNX 模型路径）、“platform\_type”（目标平台类型）、“plugin\_config”（插件配置）三类参数，返回 “conversion\_status”（转换状态）、“target\_model\_path”（目标模型路径）、“log\_path”（转换日志路径），接口响应时间≤500ms，支持插件通过接口注入平台专用转换逻辑。

模型可视化渲染引擎实现：

基于 Netron 7.0 精简版内核，移除 “3D 视图渲染”“动态特征图展示” 等复杂功能，仅保留 2D 拓扑图渲染核心代码；采用 “静态预渲染” 策略，加载模型时先通过内核生成 SVG 格式的整体拓扑图静态文件，存储于调试机本地临时目录，前端直接加载静态文件展示整体结构，减少实时渲染压力。

交互功能通过 “前端局部请求” 实现，用户点击层节点时，前端发起 “GET /api/engine/visual/layer” 请求，传入 “layer\_id” 参数，引擎从解析结果 JSON 中提取该层核心参数并返回，前端通过弹窗展示参数详情（不刷新整个页面），交互响应时间≤300ms；层标识采用 “颜色 + 图标” 极简设计，定义 4 类核心层类型的专属标识（卷积层：蓝色方块 +“Conv” 字样、全连接层：绿色圆形 +“FC” 字样、Transformer 层：紫色菱形 +“TF” 字样、输入 / 输出层：灰色三角形 +“IO” 字样），避免标识混乱。

资源占用优化通过 “渲染缓存” 与 “内存释放” 实现，渲染完成后将 SVG 静态文件缓存 12 小时（有效期内重复加载模型直接复用缓存），关闭可视化界面时自动调用内存释放函数（如 clear\_svg\_cache ()、free\_layer\_data ()），释放渲染过程中占用的内存，确保引擎整体内存占用≤768MB（1000 层模型）。

### **4.场景化插件层：功能扩展载体**

加载、独立扩展、低资源占用” 为设计原则，通过插件形式承接核心引擎无法覆盖的差异化需求，既避免核心引擎冗余，又适配试验场内场调试机的轻量化运行要求，所有插件均支持单独更新与卸载，不影响整体系统稳定性。

#### (1)功能描述

场景化插件层围绕项目核心性能指标（支持 20 类模型、4 类平台）与功能需求（跨平台转换、推理测试），提供三类专项插件，具体功能如下：

平台专属转换插件：为华为昇腾、瑞芯微、英伟达、寒武纪 4 类平台分别提供专属插件，核心功能是承接核心引擎输出的 ONNX 模型，调用对应平台的专用转换工具完成格式转换，并处理平台特有的算子兼容问题。例如华为昇腾插件支持调用 ATC 工具将 ONNX 模型转为 OM 格式，自动执行算子融合（如 Conv+BN+ReLU 合并）与数据排布优化（NCHW 转 NHWC）；瑞芯微插件支持调用 RKNN Toolkit 实现 ONNX 到 RKNN 格式转换，提供 Min-Max 量化算法选项，适配 RK3588/RK3599 的 NPU 算力特性。插件仅输出转换后的目标平台模型文件与简化日志（记录转换耗时、成功 / 失败状态），无额外冗余功能。

模型适配优化插件：针对 20 类主流智能模型（目标检测类：YOLOv5/v7/v8、Faster R-CNN、SSD、RetinaNet；目标分类类：ResNet50/101、EfficientNet-B0/B4、MobileNetV3；目标跟踪类：SORT、DeepSORT、ByteTrack、OC-SORT；自然语言大模型类：BERT-base、RoBERTa、GPT-2、T5-small、DistilBERT），每类模型开发 1 个适配插件。核心功能是向核心引擎注入该模型的专属适配规则，如 YOLO 系列插件预设检测头 Anchor 信息解析逻辑，确保转换时保留目标框预测关键参数；BERT 类插件优化 Transformer 层算子调用顺序，减少跨平台精度损失；插件同时支持加载该模型的测试基准参数（如输入维度、精度阈值），辅助后续推理测试。

推理测试简化插件：聚焦调试机 “快速验证转换效果” 需求，提供轻量化推理测试功能，核心支持 4 类平台测试节点接入，可完成功能验证与基础性能测试。功能验证环节仅检查转换后模型在目标平台是否能正常推理（输出非空结果，如目标检测模型输出 bounding box、NLP 模型输出文本预测结果）；基础性能测试环节仅采集单轮推理延迟（取 10 轮平均值，单位 ms），不计算复杂指标（如吞吐量、资源占用峰值）；测试完成后输出 TXT 格式简易报告，包含模型名称、目标平台、测试时间、推理结果状态、平均延迟 5 项核心信息，适配调试机快速排查问题的需求。

#### (2)技术实现方案

围绕 “轻量化、可扩展、低耦合” 原则，采用 “插件注册 - 逻辑封装 - 接口对接” 的技术路径，确保插件与核心引擎高效协同，同时控制资源消耗，具体实现方案如下：

平台专属转换插件实现：

插件结构设计：采用 “配置文件 + 核心逻辑脚本” 的极简结构，配置文件（JSON 格式）记录平台名称、支持的转换工具版本（如华为插件配置 “tool\_version: ATC 7.0”）、算子映射表（如 “ONNX Conv2d→昇腾 Conv2D”），文件大小≤10KB；核心逻辑脚本（Python 语言）仅包含工具调用函数（如华为插件的 atc\_convert ()、瑞芯微插件的 rknn\_convert ()）与错误处理逻辑（如工具调用失败时重试 2 次，超时时间设为 300 秒），脚本代码量控制在 500 行以内，整体插件大小≤50MB。

核心引擎对接：通过核心引擎的 “插件注册接口”（POST /api/engine/plugin/register）完成接入，接口传入插件 ID、平台类型、配置文件路径三类参数，核心引擎验证通过后将插件纳入 “平台插件库”；转换时核心引擎通过 “调用接口”（POST /api/engine/plugin/execute）向插件传递 ONNX 模型路径与转换参数，插件调用对应平台工具执行转换后，通过接口返回目标模型路径与日志路径，接口响应时间≤100ms；插件支持独立更新，通过 “插件更新接口”（PUT /api/engine/plugin/update）上传新版本脚本与配置文件，无需重启核心引擎，更新耗时≤30 秒。

算子兼容处理：针对平台不支持的算子（如华为昇腾不支持 TensorFlow Unique 算子、瑞芯微不支持大 kernel 11×11 卷积），插件内置 “算子替换库”，华为插件将 Unique 算子替换为 “Sort+Deduplicate” 组合算子，瑞芯微插件将 11×11 卷积拆分为 3×3+3×3+5×5 组合卷积，替换逻辑封装为独立函数（如 replace\_unique\_op ()、split\_conv\_op ()），确保功能等价且性能损失≤10%；算子替换规则可通过配置文件动态更新，无需修改插件核心代码。

模型适配优化插件实现：

适配规则封装：采用 “模板化” 设计，每类模型的适配规则封装为 JSON 格式模板，包含 “模型结构特征”（如 YOLO 模板标注 “backbone: CSPDarknet, neck: FPN, head: Detect”）、“关键层解析规则”（如 BERT 模板标注 “Transformer 层需提取注意力头数量: 12”）、“精度优化参数”（如 MobileNetV3 模板配置 “量化精度: INT8, 校准数据集抽样数: 100”），模板文件大小≤5KB，支持用户通过前端 “模板导入” 功能新增模型适配（无需开发代码）。

插件加载逻辑：插件启动时自动读取模板文件，通过 “规则注入接口”（POST /api/engine/plugin/inject）将适配规则传入核心引擎的模型解析模块与转换模块；核心引擎解析对应模型时，调用插件注入的结构特征识别函数（如 is\_yolo\_model ()、is\_bert\_model ()），快速匹配模型类型并加载解析规则；转换时插件向核心引擎传递该模型的精度优化参数，如为 YOLO 检测头配置 FP32 精度、为 ResNet backbone 配置 FP16 精度，平衡精度与性能。

兼容性验证：插件内置 “模型兼容性校验函数”（如 validate\_yolo\_compatibility ()），加载模型后先校验模型结构是否与模板匹配（如检查 YOLO 模型是否包含 Detect 头），若不匹配则生成 “不兼容报告”（提示缺失的关键层）；同时支持 “版本适配”，如 YOLOv5 与 YOLOv8 模板共享基础解析规则，仅差异化配置检测头参数，减少模板冗余，20 类模型插件可共享 80% 基础代码，开发效率提升 60%。

推理测试简化插件实现：

测试逻辑封装：采用 “测试用例配置 + 核心测试脚本” 结构，测试用例配置文件（JSON 格式）记录测试数据路径（如 COCO 验证集图像路径、GLUE 文本数据路径）、测试次数（默认 10 次）、结果判定规则（如 “推理输出非空即判定为功能正常”），文件大小≤5KB；核心测试脚本（Python 语言）包含 “模型加载函数”（load\_model ()，调用平台专用 SDK 加载模型）、“推理执行函数”（run\_inference ()，执行推理并记录延迟）、“结果处理函数”（process\_result ()，判断结果是否非空并计算平均延迟），代码量控制在 300 行以内，插件整体内存占用≤256MB。

硬件节点对接：通过 “节点注册接口”（POST /api/plugin/test/node/register）接入 4 类平台测试节点，接口传入节点 IP、平台类型、SDK 版本，插件验证节点在线状态后将其纳入 “测试节点库”；测试时插件根据目标平台自动选择对应节点，通过 SSH 协议远程执行测试脚本，实时接收测试数据（如每轮推理延迟），测试完成后自动生成 TXT 格式报告，报告存储路径通过 “结果回调接口”（POST /api/plugin/test/result/callback）同步至核心引擎，供前端查看。

资源优化：测试过程中采用 “单节点单任务” 调度策略，避免多任务并发导致调试机资源过载；测试完成后自动清理临时文件（如测试日志、中间推理结果），调用内存释放函数（free\_test\_resources ()）释放 SDK 占用的内存；支持 “测试任务暂停 / 取消”，通过 “任务控制接口”（PUT /api/plugin/test/task/control）发送暂停 / 取消指令，插件立即停止测试并释放节点资源，响应时间≤500ms，适配调试机动态调整测试需求的场景。

### **5.轻量化支撑层：运行保障基础**

轻量化支撑层是整个系统稳定运行的基础保障，聚焦 “试验场内场调试机低算力、有限存储、内网环境” 三大约束，以 “资源精简、逻辑简化、适配高效” 为核心设计原则，通过极简的组件设计与协同逻辑，为核心引擎层、场景化插件层提供资源调度、数据存储、用户交互三大基础支撑，确保系统整体运行流畅且资源占用可控。

#### (1)功能描述

轻量化支撑层不承担业务逻辑，仅专注于基础运行保障，通过三类核心组件满足系统基础需求，具体功能如下：

资源调度组件：负责统筹核心引擎与插件的资源分配，以及转换、测试任务的有序执行。支持 “本地任务队列” 管理，用户提交的模型转换、推理测试任务直接进入调试机本地队列，按 “提交时间先后” 排序，避免分布式调度带来的资源消耗；实时监控调试机硬件负载（仅监测 CPU 利用率、内存占用率两项核心指标），当 CPU 利用率≥80% 或内存占用率≥85% 时，自动暂停新任务分配，待负载降至阈值以下再恢复执行；支持任务优先级手动调整（仅区分 “普通”“紧急” 两档，紧急任务可插入队列前端），以及任务暂停 / 取消操作（暂停后保留进度，取消后释放占用资源），适配调试机单节点运行场景。

数据存储组件：承担系统核心数据的存储与管理，仅保留 “任务数据、模型文件、高频缓存数据” 三类必要数据，剔除冗余存储。MySQL 数据库仅存储任务基础信息（任务 ID、模型名称、目标平台、转换状态、提交时间）与用户配置（常用转换参数模板），无复杂关联表设计；MinIO 存储仅保留原始模型文件、转换后目标平台模型文件两类核心非结构化数据，日志文件按 “7 天自动清理” 策略管理，避免调试机存储空间溢出；Redis 缓存仅存储任务实时状态（如 “处理中 - ONNX 转换阶段”）与高频访问的模型元数据（输入输出维度、算子数量），缓存有效期缩短至 6 小时，释放内存资源。

用户交互组件：提供调试机用户与系统的交互入口，界面设计聚焦 “核心功能快速触达”，仅包含模型上传、转换配置、可视化查看、任务监控四大核心功能模块。支持单文件拖拽上传（最大支持 3GB 模型，适配调试机存储容量），转换参数配置采用 “模板预设 + 简化选项”（如目标平台下拉选择、精度模式仅保留 FP32/FP16/INT8 三档），避免复杂参数配置；任务监控采用 “列表 + 环形进度条” 极简展示，仅显示任务 ID、模型名称、状态、进度四项关键信息，无多余数据可视化；所有交互操作均在单页完成，无需多页跳转，适配调试机快速操作需求。

#### (2)技术实现方案

围绕 “低资源消耗、高响应速度、适配调试机环境” 目标，采用 “精简组件选型 + 轻量化开发” 技术路径，具体实现方案如下：

资源调度组件实现：

任务队列设计：基于 Python 内置的queue模块开发本地任务队列，不依赖 RabbitMQ、Kafka 等分布式消息中间件，队列容量设为 50（满足调试机单节点任务并发需求），任务入队时自动生成唯一任务 ID（UUID 格式），并关联模型路径、目标平台、参数配置等基础信息；队列采用 “FIFO + 优先级插入” 调度策略，普通任务按入队顺序执行，紧急任务通过insert(0, task)方法插入队列头部，调度逻辑代码量控制在 200 行以内，内存占用≤64MB。

负载监控实现：通过psutil库实时采集调试机硬件数据，每 5 秒获取一次 CPU 利用率（取所有核心平均值）、内存占用率（已用内存 / 总内存），数据采集逻辑封装为get\_hardware\_load()函数，避免频繁采集导致资源消耗；当负载超阈值时，触发pause\_task\_scheduling()函数暂停任务分配，同时在前端弹出 “资源紧张” 提示；负载恢复后自动调用resume\_task\_scheduling()函数重启调度，整个监控逻辑 CPU 占用率≤5%。

任务控制接口：开发 RESTful 风格的任务控制接口，POST /api/support/task/pause实现任务暂停（记录当前任务进度，如 “ONNX 转换 50%”）、POST /api/support/task/cancel实现任务取消（调用release\_resource()函数释放 CPU / 内存资源，删除临时文件），接口响应时间≤100ms；任务状态更新采用 “前端定时轮询”（轮询间隔 5 秒），避免 WebSocket 长连接带来的内存占用，轮询请求仅返回任务状态与进度，数据量≤1KB，适配调试机内网低带宽环境。

数据存储组件实现：

MySQL 数据库优化：采用 MySQL 8.0 轻量版，仅创建task\_info（任务信息）、user\_config（用户配置）两张核心表，表结构剔除冗余字段（如task\_info仅保留 10 个字段，无任务日志、历史记录等关联字段）；为task\_info表的 “任务 ID” 字段建立主键索引、“状态 + 提交时间” 字段建立联合索引，将任务列表查询耗时控制在 50ms 以内；数据库连接采用 “连接池 + 按需创建” 策略，连接池最大连接数设为 10（满足调试机单用户并发需求），空闲连接超过 30 秒自动关闭，减少数据库资源占用。

MinIO 存储配置：采用单节点 MinIO 部署（无需分布式扩展），通过minio.Client初始化客户端，仅创建raw-model（原始模型）、converted-model（转换后模型）两个存储桶，桶内按 “任务 ID / 模型名称” 路径结构化存储文件（如 “task-123/yolov8.pt”）；启用 MinIO 的 “生命周期管理” 功能，为log文件夹配置 “7 天自动删除” 规则，通过set\_bucket\_lifecycle()函数实现；文件上传采用 “分片上传 + 断点续传”（基于 MinIO 的put\_object()方法，分片大小设为 5MB），适配调试机内网文件传输需求，1GB 模型上传耗时≤5 分钟。

Redis 缓存设计：采用 Redis 6.0 精简版，仅使用 Hash 数据类型存储任务状态（Key：task:{taskId}，Field：status/progress）与模型元数据（Key：model:{modelId}，Field：input\_shape/output\_shape/op\_count）；缓存有效期通过expire()函数设为 6 小时，避免长期占用内存；为减少 Redis 访问压力，核心引擎与插件更新任务状态时，先更新本地内存变量，再批量同步至 Redis（每 30 秒同步一次），同步逻辑封装为sync\_task\_status\_to\_redis()函数，Redis 整体内存占用控制在 256MB 以内。

用户交互组件实现：

前端技术选型：基于 Vue3+Vite 开发单页应用（SPA），Vite 构建工具相比 Webpack 可减少 50% 构建时间，前端资源打包后总大小≤2MB，首次加载时间≤3 秒（适配调试机内网低带宽）；UI 组件库选用 Element Plus 精简版，仅引入Upload（上传）、Select（下拉选择）、Progress（进度条）、Table（列表）4 类核心组件，剔除复杂组件（如Tree、Chart），减少资源加载；界面样式采用 “扁平化 + 少色彩” 设计，CSS 代码量控制在 500 行以内，避免复杂样式渲染消耗 GPU 资源。

核心功能开发：模型上传功能基于el-upload组件开发，仅支持单文件拖拽上传，上传前通过before-upload钩子函数校验文件大小（≤3GB）与格式（.pth/.pb/.onnx 等），校验不通过实时提示；转换配置功能采用 “模板下拉选择 + 简化选项”，预设 “华为昇腾 - FP16”“瑞芯微 - INT8” 等常用模板，参数配置项控制在 5 项以内；可视化查看功能通过iframe嵌入核心引擎生成的静态拓扑图，点击层节点时发起GET /api/support/visual/layer请求获取参数，弹窗展示核心信息；任务监控功能基于el-table组件开发，仅显示 4 列关键信息，表格数据通过setInterval每 5 秒刷新一次，避免频繁请求。

适配调试机环境：针对调试机可能的低分辨率（如 1920×1080），采用响应式布局，界面元素宽度使用%或vw单位，避免内容截断；支持键盘快捷键操作（如 Ctrl+U 快速上传、Ctrl+M 打开模型列表），提升操作效率；前端代码通过terser工具压缩，去除注释与无用代码，最终打包文件大小≤1.5MB，确保调试机快速加载。

## （三）工作内容

结合 “试验场调试机内嵌软件” 的场景约束与 “核心引擎 + 场景化插件 + 轻量化支撑”的技术架构，围绕 “20 类模型适配、4 类平台兼容、2 个月短周期交付” 核心目标，项目工作内容按 “需求-开发-验证-交付” 全流程拆解为 6 大阶段，各阶段任务聚焦技术落地与调试机适配，具体如下：

### **1.需求深度调研与方案细化**

核心目标是明确调试机环境约束与功能边界，输出可落地的技术方案，避免后期开发偏差：

联合试验场甲方开展 “调试机硬件摸底”，确认调试机操作系统（Windows Server 2019/Linux CentOS 8）、算力配置（CPU 型号、内存容量、存储上限）、内网带宽等关键参数，形成《调试机环境适配报告》，作为后续轻量化开发的依据；

细化 20 类模型清单（明确目标检测类含 YOLOv5/v7/v8 等 6 类、目标分类类含 ResNet50 等 5 类、目标跟踪类含 SORT 等 4 类、NLP 大模型类含 BERT-base 等 5 类）与 4 类平台技术参数（如华为昇腾 Atlas 200I 的 NPU 显存、瑞芯微 RK3588 的算力特性），输出《模型与平台适配清单》；

确认可视化功能细节（如是否需支持注意力热力图、拓扑图交互复杂度）与推理测试简化需求（如仅保留核心精度指标），修订《需求规格说明书》并经甲方签字确认，同步输出《技术方案细化文档》，明确核心引擎与插件的接口定义。

### **2.资源筹备与轻量化环境搭建**

聚焦 “低资源占用” 目标，完成硬件租赁、软件部署与测试库搭建，为开发阶段铺路：

硬件资源筹备：租赁 4 类平台核心测试节点（1 台华为昇腾 Atlas 200I、1 台瑞芯微 RK3588、1 台英伟达 Jetson AGX Orin、1 台寒武纪思元 370），接入试验场内网并完成 IP 绑定，通过 Conda 虚拟环境配置最小化运行环境（仅安装驱动、核心 SDK 与依赖库），输出《硬件环境部署报告》；

软件工具链部署：部署开发工具链（ONNX Runtime 1.15、Netron 7.0 精简版、Flask 框架、Vue3+Vite 前端环境），搭建 “MySQL 8.0 轻量版 + 单节点 MinIO+Redis 6.0 精简版” 存储环境，配置日志 7 天自动清理策略，确保调试机存储无溢出风险；

测试资源搭建：收集 20 类模型原生文件（PyTorch/TensorFlow/ONNX 格式），整理标准测试数据集（COCO 2017 验证集、ImageNet 精简集、GLUE 文本集），建立 “模型测试库” 与 “数据测试库”，标注每类模型的输入输出维度、精度基准值，便于后续适配调试。

### **3.核心引擎与场景化插件开发**

按 “引擎先行、插件跟进” 逻辑，完成架构核心模块开发，确保功能与轻量化需求匹配：

核心引擎开发：完成多格式模型解析引擎（支持 PyTorch/TensorFlow/ONNX 三类格式，解析耗时≤90 秒）、跨平台转换核心引擎（实现 “原生格式→ONNX” 统一转换与算子融合优化）、模型可视化渲染引擎（支持静态拓扑图渲染与基础参数查看）开发，输出《核心引擎开发文档》，并通过单元测试（如解析成功率≥99%、渲染响应时间≤1.5 秒）；

场景化插件开发：开发 4 类平台专属转换插件（每类插件≤50MB，加载时间≤10 秒），实现平台工具调用与算子兼容处理；开发 20 类模型适配优化插件（每类模型模板≤100KB），注入专属解析规则与精度参数；开发推理测试简化插件（内存占用≤256MB），实现功能验证与基础性能测试逻辑，输出《插件开发清单》；

接口联调：完成核心引擎与插件的接口对接（通过注册接口、调用接口实现协同），验证 “引擎调用插件” 的流畅性（如 ONNX 模型调用华为插件转换为 OM 格式的响应时间≤500ms），修复接口兼容性问题，输出《接口联调报告》。

### **4.模型与平台适配调试**

针对 20 类模型与 4 类平台开展全量适配，解决算子不兼容、精度损失等问题：

模型适配调试：基于 20 类模型适配插件，逐一调试模型转换流程（如 YOLO 模型确保检测头 Anchor 信息保留、BERT 模型优化 Transformer 层调度），记录每类模型的转换耗时、精度损失，针对精度损失超 8% 的模型（如 MobileNetV3），优化量化参数或算子映射规则，确保适配成功率≥95%，输出《20 类模型适配调试报告》；

平台兼容调试：在 4 类平台测试节点上，部署转换后模型并执行推理验证，解决平台特有的兼容性问题（如华为昇腾算子替换、瑞芯微卷积拆分），验证模型在调试机环境下的运行稳定性（连续推理 100 轮无崩溃），输出《4 类平台兼容调试报告》；

性能优化：针对调试机低算力特性，优化转换与推理流程（如缩短轮询间隔、减少临时文件生成），确保模型转换平均耗时≤5 分钟（1GB 模型）、推理延迟符合甲方要求（如瑞芯微 RK3588 运行 YOLOv8 的延迟≤300ms），输出《性能优化报告》。

### **5.功能性能测试与问题修复**

开展全维度测试，确保软件满足调试机使用需求，避免验收风险：

功能测试：验证核心功能（模型解析、跨平台转换、可视化查看、推理测试）完整性，如测试 20 类模型的解析成功率、4 类平台的转换有效性、可视化界面的交互流畅性（点击节点参数加载≤300ms），输出《功能测试报告》；

性能测试：在调试机上测试软件资源占用（CPU 利用率≤80%、内存占用≤2GB）、任务并发处理能力（5 个任务同时执行无卡顿）、大模型处理效率（3GB NLP 模型解析耗时≤2 分钟），输出《性能测试报告》；

问题修复与优化：针对测试中发现的问题（如某类模型转换失败、可视化界面分辨率适配异常），集中修复并记录《问题修复清单》；根据内部预验收意见，优化软件交互（如简化转换参数配置步骤），确保符合调试机操作习惯。

### **6.部署交付与售后培训**

完成调试机部署、成果交付与技术培训，确保甲方能独立使用：

现场部署：在试验场内场调试机上安装软件（Windows/Linux 版适配），逐台验证全流程功能（模型上传 - 转换 - 可视化 - 推理测试），确保每台调试机的软件运行稳定（连续 72 小时无故障），输出《部署验收清单》；

成果交付：跨平台智能算法模型转化工具（含可执行程序、可二次开发源代码）、全套技术文档（技术方案、设计说明、需求说明书、设计文档、自测试报告、使用说明书、自主可控测试报告）、测试资源（模型测试库、数据测试库），并归档电子版与纸质版各 1 套，签署《成果交付确认单》，提供软件授权书一份；

技术培训与售后：开展 2 次现场培训（操作培训：软件使用与常见问题处理；维护培训：日志分析与版本更新），覆盖甲方所有技术人员，输出《培训纪要》；建立 1 年售后服务机制（远程支持 2 小时响应、2 次现场支持），明确售后联系方式与服务范围，确保甲方后续使用无顾虑。

# 资质文件

## 事业单位法人证书复印件

我校中南大学为国家高等学校教育事业单位，在中华人民共和国境内注册，具有独立承担民事责任的能力和经营许可，非外资独资或外资控股企业。

附：事业单位法人证书复印件



法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

供应商单位公章：中南大学

2025 年 09 月 18 日

## 开户行出具的银行资信证明

我校中南大学为国家高等学校教育事业单位。我单位声明，提供的2024年9月至2025年9月内银行资信证明信息真实可信，如有虚假情况，愿承担一切法律责任。

附：开户行出具的银行资信证明



法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

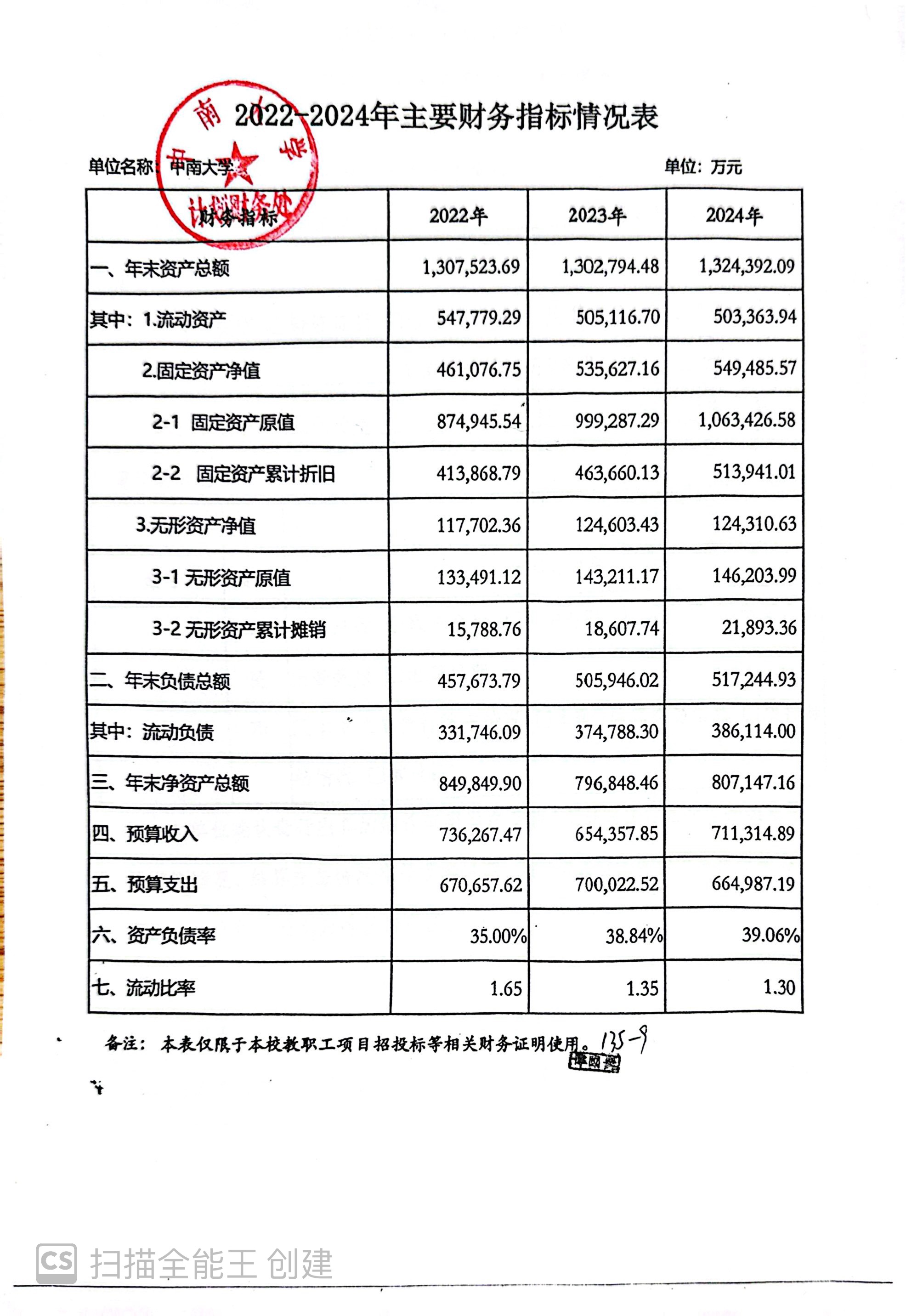
供应商单位公章：中南大学

2025 年 09 月 18 日

## 2022年至2024年资产负债表

我校中南大学为国家高等学校教育事业单位。我单位声明，提供的2022-2024年主要财务指标情况表信息真实可信，如有虚假情况，愿承担一切法律责任。

附：2022-2024年主要财务指标情况表复印件



法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

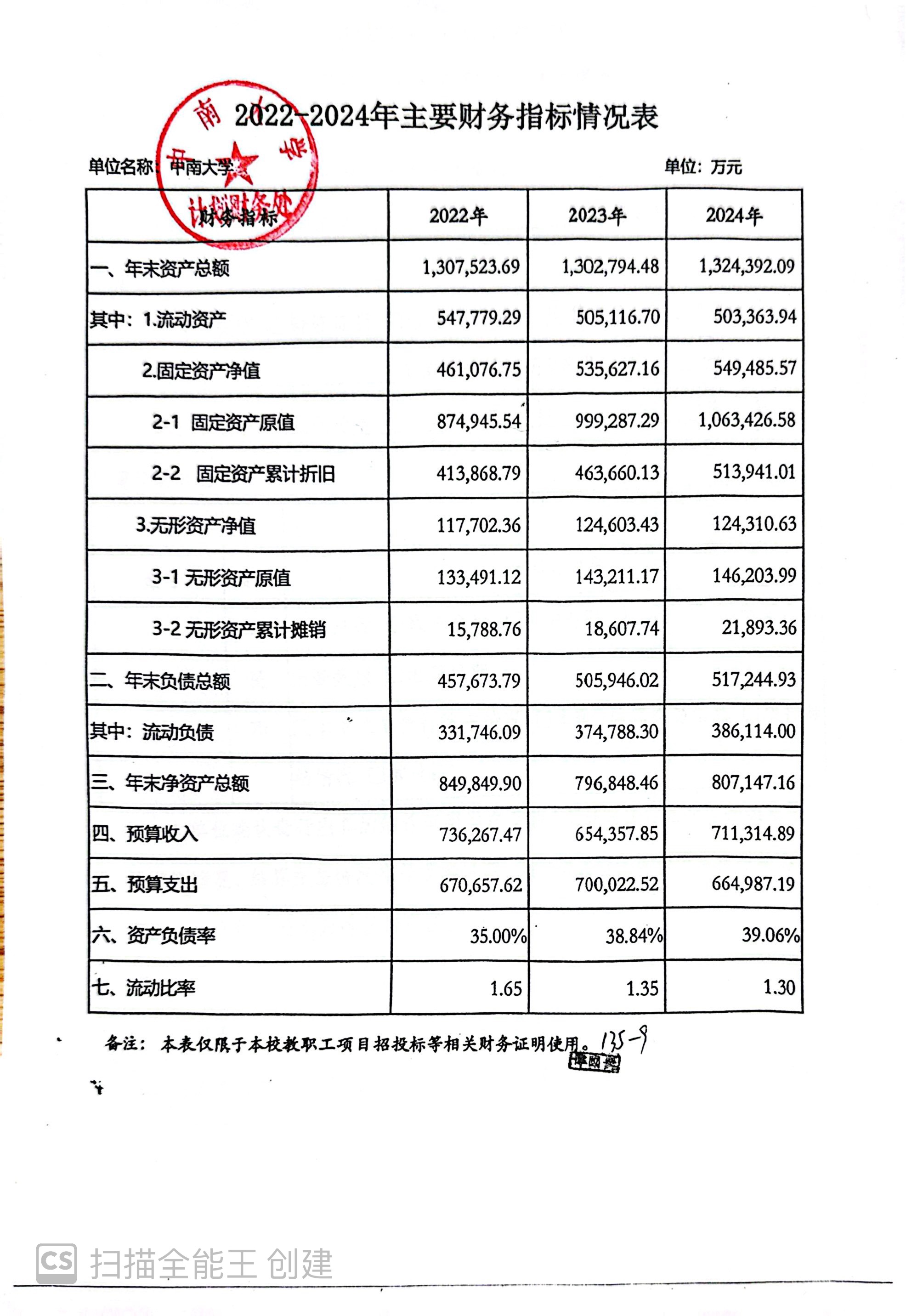
供应商单位公章：中南大学

2025 年 09 月 18 日

## 2022年至2024年收入支出表

我校中南大学为国家高等学校教育事业单位。我单位声明，提供的2022-2024年主要财务指标情况表信息真实可信，如有虚假情况，愿承担一切法律责任。

附：2022-2024年主要财务指标情况表复印件



法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

供应商单位公章：中南大学

2025 年 09 月 18 日

## 履行合同所必须的设备和专业技术能力证明材料

中南大学是中华人民共和国教育部直属的全国重点大学，中央直管高校，是国家“211工程”首批重点建设高校、国家“985工程”部省重点共建高水平大学、国家“2011计划”首批牵头高校，入选世界一流大学建设高校、第二轮“双一流”建设高校。中南大学占地面积317.02万平方米，由原湖南医科大学、长沙铁道学院与中南工业大学于2000年4月合并组建而成。学校现有岳麓山、麓南、潇湘、开福、杏林、天心6个校区，设有32个二级学院。学校师资力量雄厚，具有两院院士16人，教授及相应正高职称人员2213人。资金实力丰厚，注册资金为109542.00万元，截至2025年6月，学校固定资产总值1023343.60万元，2024年营业收入为711314.89万元。业务范围包括：文史类、理学类、工学类学科高等专科学历教育；工学类、理学类、医学类、文学类、法学类、经济学类、管理学类、哲学类、教育学类学科本科、研究生班，硕士研究生学历教育；工学类、理学类、医学类、管理学类学科博士研究生学历教育；博士后培养、护理中专生培养；相关科学研究、技术开发、继续教育、专业培训、学术交流、科技和法律咨询与临床医疗保健。在计算资源方面，中南大学拥有国际领先的高性能计算平台，平台配置1056个计算节点，以及100块NVIDIA Tesla V100系列专业计算GPU卡，采用Infiniband HDR实现计算网络互联，平台聚合计算能力达5.6千万亿次，存储容量2.3PB，完全能满足本项目在数据存储、数据分析、算法测试、机器学习等过程所需要的计算资源。同时，平台部署了Matlab、Ansys、Materials Studio、Deform、Comsol、Abaqus、Pandat™、Gaussian等正版商业软件，为科学计算提供了软硬一体的解决方案。

中南大学拥有完备的交通学科体系，交通运输工程入选为国家“双一流”建设学科。交通运输工程学院设有4个系，2个研究中心，1个实验中心，11个研究所（中心）。现有教职工124人，其中教授及相应职称42人，副教授及相应职称46人，博士生导师49人。拥有中国工程院院士、国家“万人计划”科技创新领军人才等高层次人才10余人，拥有教育部“轨道交通安全关键技术”长江学者创新团队1个，科技部科技创新人才推进计划重点领域创新团队1个，湖南省“轨道交通创新人才培养”教学团队1个，获“全国专业人才先进集体”、“全国党建工作标杆院系”。近年来，学院承担了国家支撑、国家“973”、国家“863”重大专项、“985”科技创新平台建设项目、国家自然科学基金重大项目等一系列国家和省部级重大课题；获国家科技进步特等奖4项，国家科技进步一等奖1项、国家科技进步二等奖2项、国家科学技术进步创新团队奖1项；国家技术发明二等奖1项、中国专利金奖2项；省部级科研奖58项。

综上所述，我方中南大学在参加跨平台智能算法模型转化工具项目中，具有履行合同所必须的设备和专业技术能力。

附：近年的类似项目情况表及证明材料

近3年承担过的同类项目情况表

（2022年9月至2025年9月）

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 合同名称  （产品名称） | 合同编号 | 甲方名称 | 采购数量 | 合同  金额  （万元） | 合同签订时间 | 合同约定  交付（约定节点）时间 | 实际交付  （节点完成）时间 | 备注 |
| 1 | 应急物资运送全链路大规模运力协同动态配置多目标鲁棒优化 | 72474231 | 国家自然科学基金委员会 | 1 | 41 | 2024.08 | 2028.12 | 2024.10报送计划书 | 论证类 |

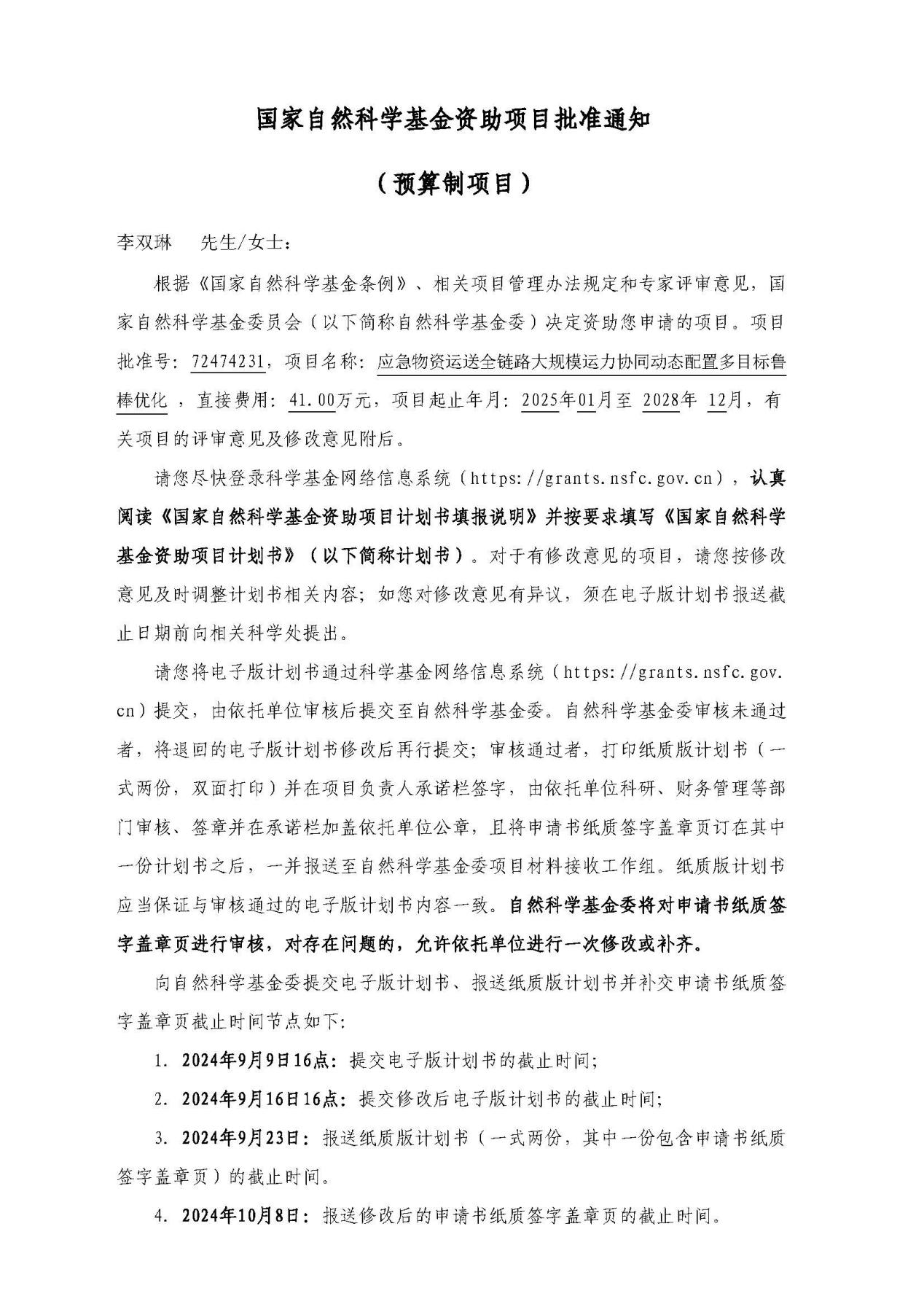
法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

供应商单位公章：中南大学

2025 年 09 月 18 日

承担相关项目扫描件





## GJB质量管理体系证书

法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

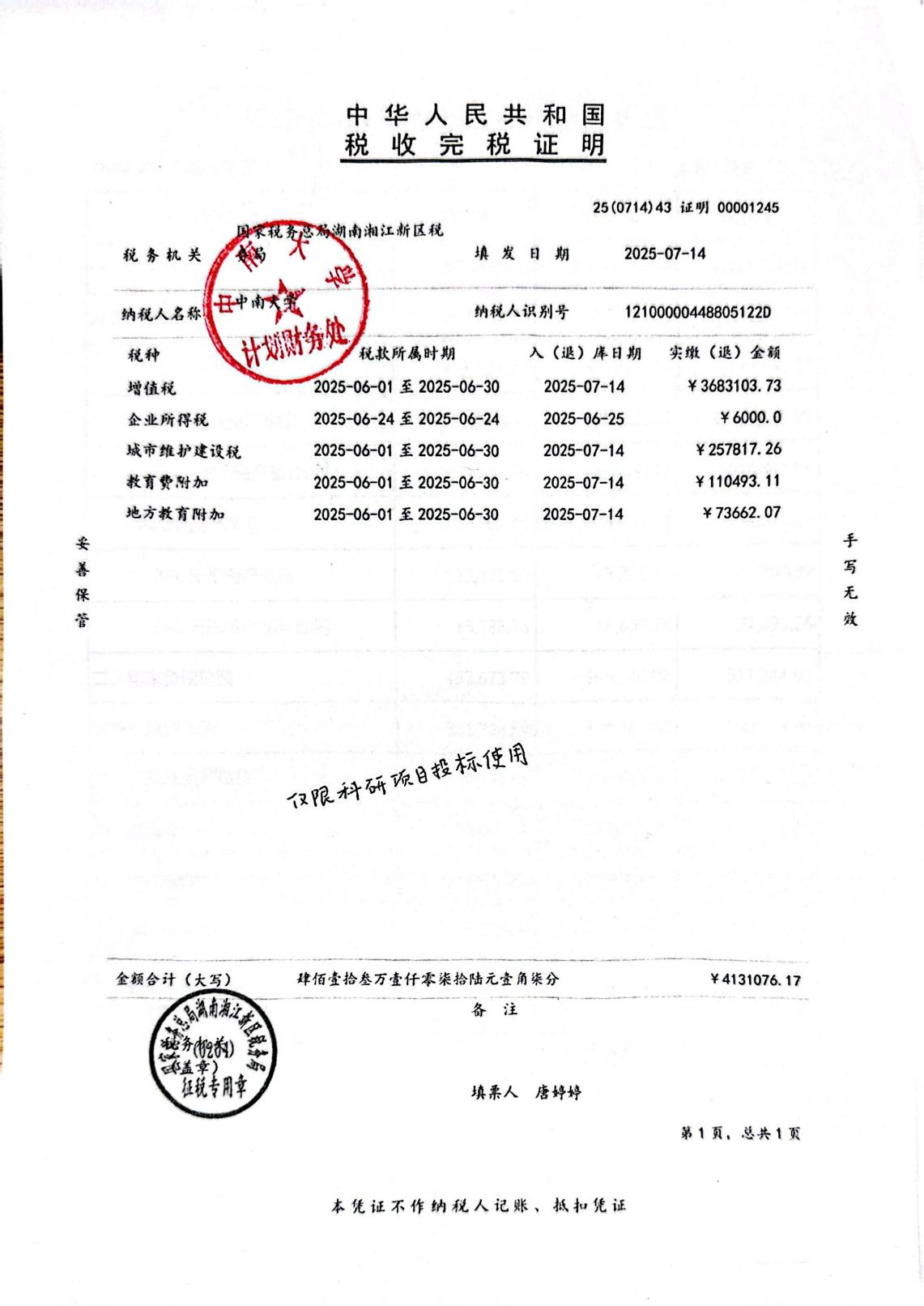
供应商单位公章：中南大学

2025 年 09 月 18 日

## 近一年中任意一个月依法纳税的凭证

我校中南大学具有依法缴纳税收记录及社会保障资金的良好记录。

附：2025年6月依法纳税凭证复印件



法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

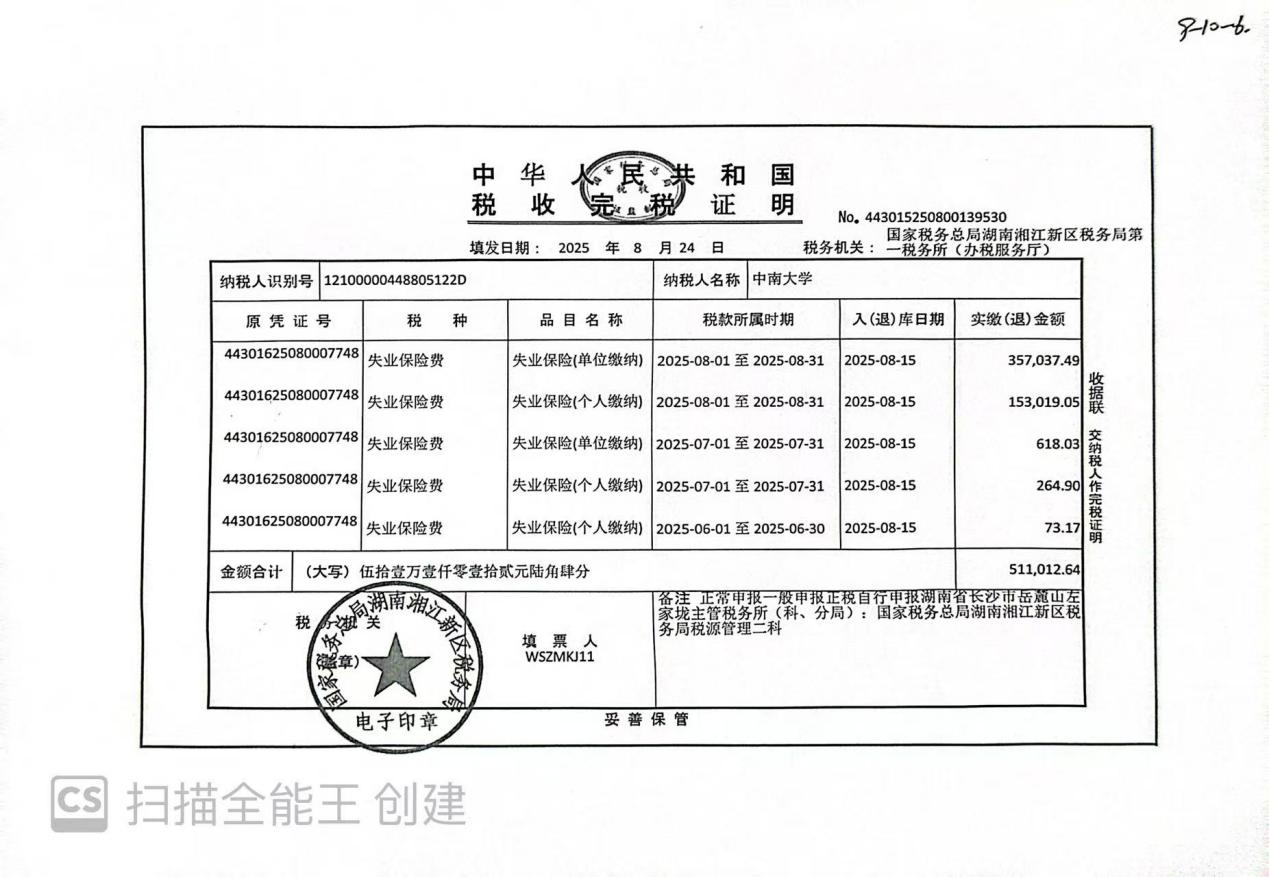
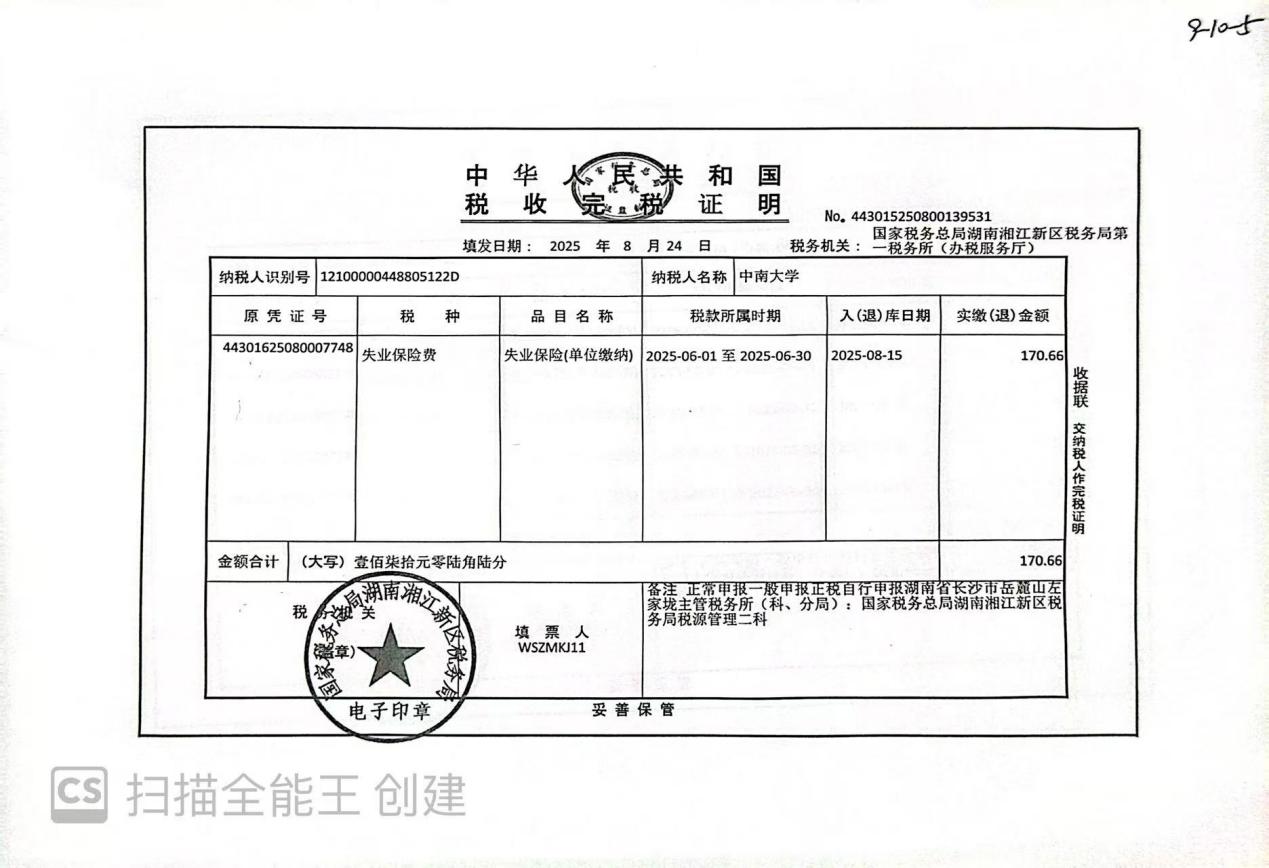
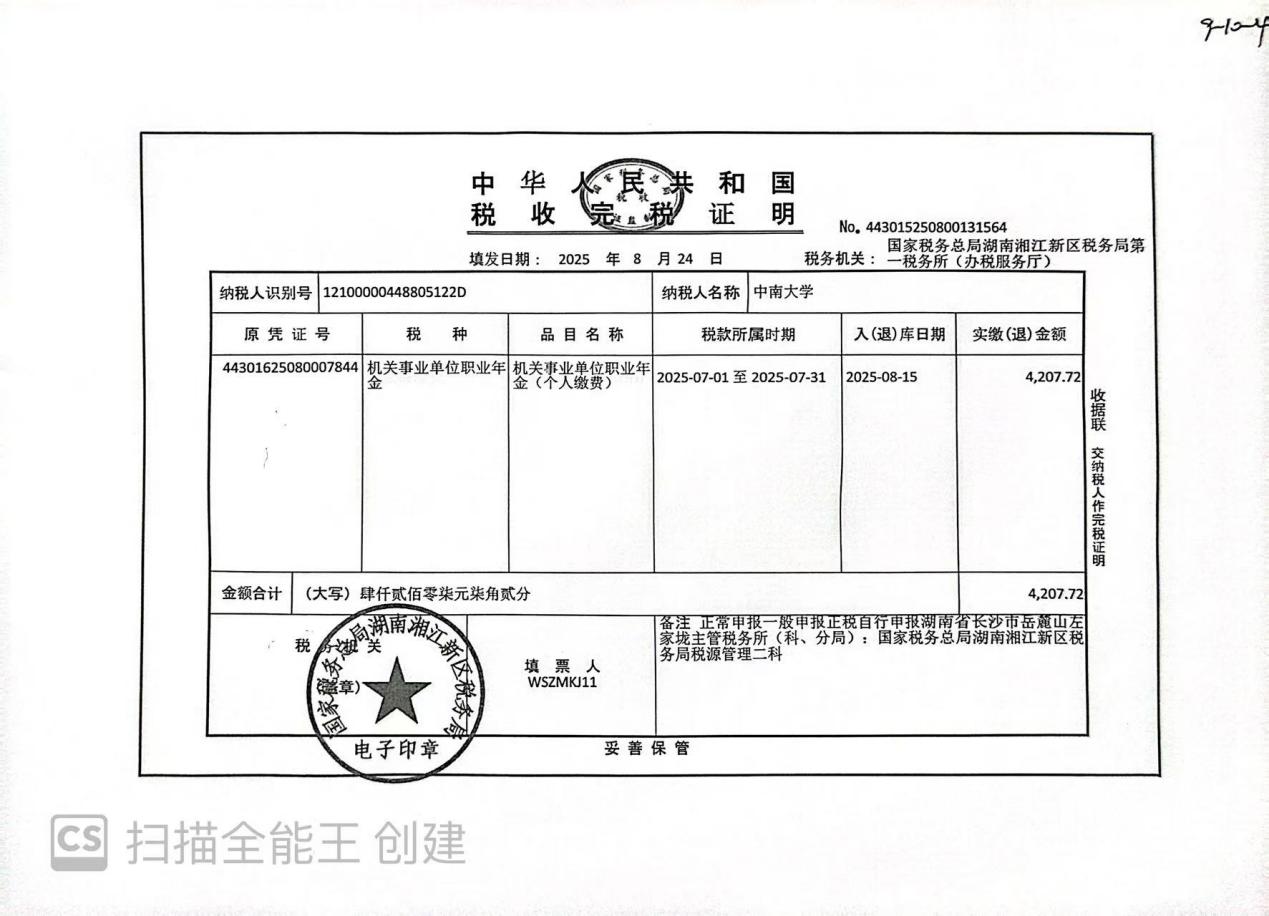
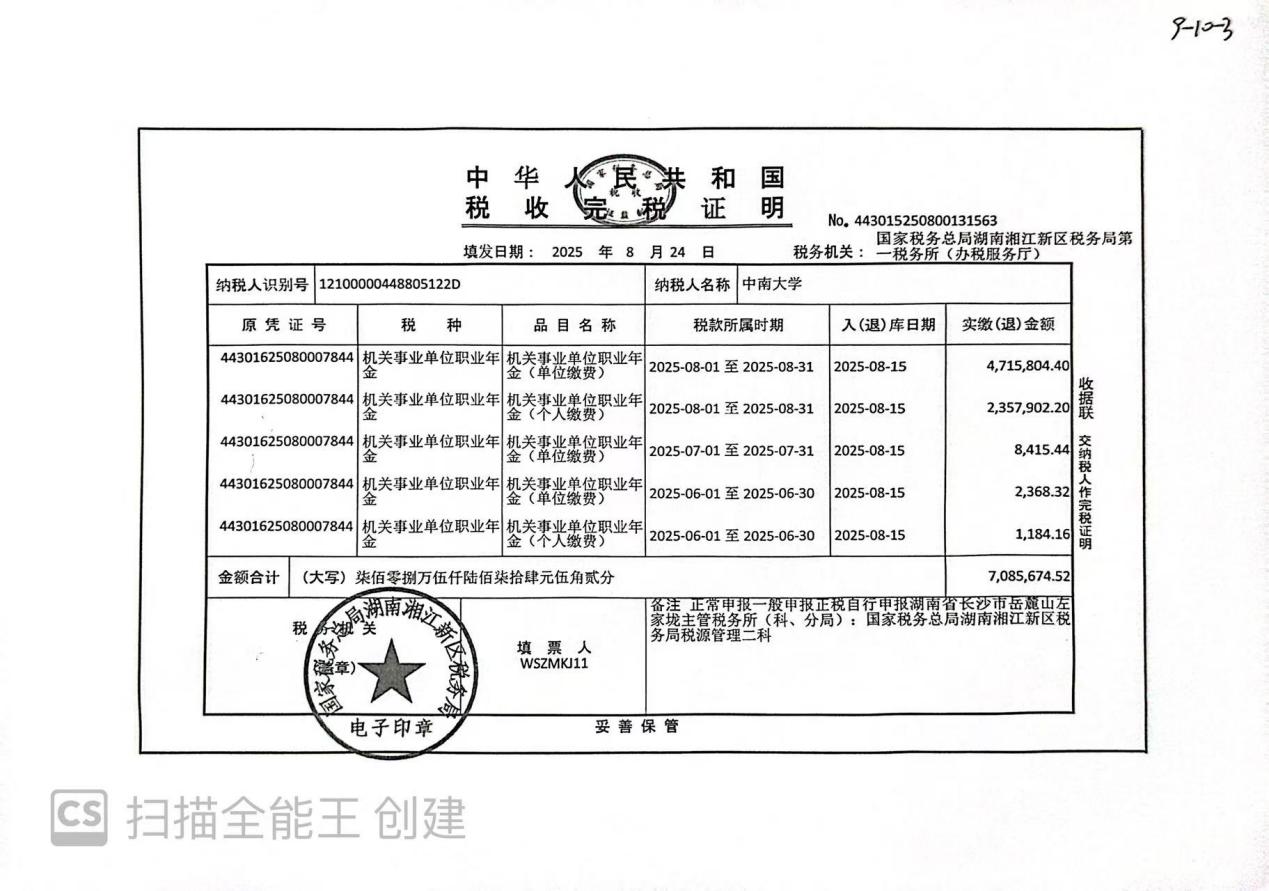
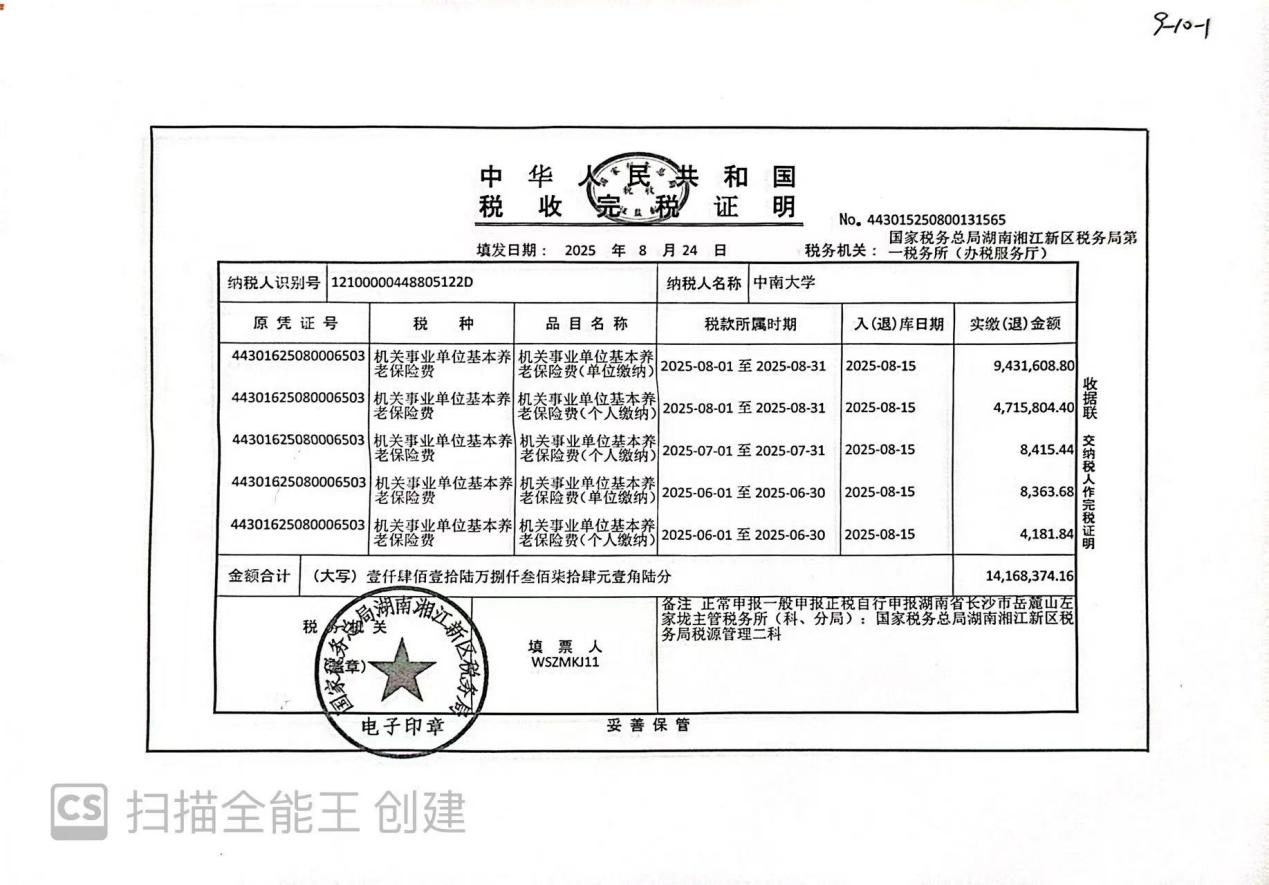
供应商单位公章：中南大学

2025 年 09 月 18 日

## 近一年中任意一个月缴纳社会保障资金的凭证

我校中南大学具有依法缴纳税收记录及社会保障资金的良好记录。

附：2025年8月社保缴纳完税证明及单位参保证明复印件







法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

供应商单位公章：中南大学

2025 年 09 月 18 日

## “信用中国”信息查询截图

我校中南大学未被“信用中国”网站（www.creditchina.gov.cn）中列入失信被执行人和重大税收违法案件当事人名单的响应单位。

附：“信用中国”截图



法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

供应商单位公章：中南大学

2025 年 09 月 18 日

## 主要负责人、股东或出资人信息

不存在控股、管理等关联关系说明

中国船舶集团有限公司第七一六研究所：

我方参加贵单位组织的 跨平台智能算法模型转化工具 采购活动，在此郑重承诺：

我方不存在单位负责人为同一人，不存在直接控股、管理关系的不同供应商，同时参加同一包的采购活动的情况。不存在生产场经营地址或者注册登记地址为同一地址、股东和管理人员（法定代表人、董事、监事）之间不存在近亲属、相互占股等关联。

违反上述承诺，愿承担一切法律责任，接收采购管理部门和采购机构按国家规定作出的相关处罚。

特此承诺！

法定代表人或授权代表（签字或盖章）：

供应商（全称）：中南大学

供应商单位公章：中南大学

2025 年 09 月 18 日