

AI 大模型本地化部署驱动高校智能应用生态构建

康宁

山东中医药大学, 山东 济南 250355

DOI: 10.61369/ETR.2026200031

摘 要：在教育数字化转型与人工智能深度融合背景下，高校亟需突破传统管理模式，依托 AI 大模型自然语言处理、多模态交互、智能分析等核心能力，搭建覆盖智能审批、数据研判、师生服务的一体化智能新生态。本文以山东中医药大学“智慧山中医”融合信息门户为载体，聚焦 AI 大模型本地化部署技术路径与应用创新，构建包含智能消息中心、AI 附件摘要、智能公文助手、史记胶囊及 AI 对话管理平台的多元应用生态。研究阐释“数据层-模型层-应用层-用户层”四级技术架构设计原理与实现细节，介绍五大核心模块应用逻辑，为高校数字化转型提供可复制、可推广的技术范式与实践经验。

关 键 词： AI 大模型；本地化部署；智能应用；数字化转型

AI Large Model Local Deployment Driven Construction of University Intelligent Application Ecosystem

Kang Ning

Shandong University of Traditional Chinese Medicine, Jinan, Shandong 250355

Abstract： Against the background of the deep integration of educational digital transformation and artificial intelligence, universities are in urgent need of breaking through traditional management models. Relying on the core capabilities of AI large models such as natural language processing, multimodal interaction, and intelligent analysis, it is necessary to build an integrated intelligent new ecosystem covering intelligent approval, data research and judgment, and teacher-student services. Taking the integrated information portal "Smart SDUTCM" of Shandong University of Traditional Chinese Medicine as the carrier, this paper focuses on the technical path and application innovation of AI large model local deployment, and constructs a diversified application ecosystem including an intelligent message center, AI attachment summary, intelligent document assistant, historical record capsule, and AI dialogue management platform. The research explains the design principle and implementation details of the four-level technical architecture of "data layer - model layer - application layer - user layer", and introduces the application logic of the five core modules, providing replicable and promotable technical paradigms and practical experience for the digital transformation of universities.

Keywords： AI large model; local deployment; intelligent application; digital transformation

引言

AI 大模型在自然语言理解与业务流程自动化处理中表现突出，凭借内容生成与语义检索能力，为高校管理服务创新提供新路径，可支撑行政事务智能管理、资源配置优化、教学模式创新、校园安全防控等目标落地。但通用大模型云端部署存在明显短板：高校数据、师生隐私信息面临泄露风险；校园公文规范、业务流程、专业术语的特殊性，导致通用模型难以精准适配。当前多数高校 AI 应用仅停留在单一功能或场景的浅层探索，从生态视角构建全链路一体化教学体系仍处于起步阶段。

山东中医药大学作为中医药特色高校，在数字化转型中完成 AI 大模型本地化部署，依托“智慧山中医”信息门户，打造适配中医药院校管理特色与师生服务需求的智能体系。本研究以该校为实践案例，探索大模型本地化部署技术路径，开发多场景核心智能模块，为高校数字化转型提供实践参考^[1]。

一、AI 大模型技术原理

AI 大模型是基于深度学习的大型语言模型，是 AIGC 核心支

撑技术，具备强学习能力、高效计算推理、广知识覆盖与深度数据挖掘能力，核心技术包含 Transformer 架构、预训练与微调机制、注意力机制。Transformer 架构以自注意力机制实现文本序

基金信息：山东中医药大学2025年教育教学研究基金资助。

作者简介：康宁（1988-），女，山东菏泽人，硕士，山东中医药大学讲师。研究方向：软件工程、数字媒体。

列并行处理，精准捕捉语义关联与上下文信息；预训练阶段依托海量文本学习语法、语义与知识图谱；微调阶段基于领域语料二次训练，适配具体场景。

大模型核心能力覆盖自然语言理解、内容生成、语义检索、逻辑推理，可准确识别文本意图与关键信息、生成规范连贯内容、实现语义关联检索、处理复杂逻辑关系，为高校管理、师生服务、科研支持等场景提供技术支撑^[2]。

二、高校智能应用生态总体架构设计

本研究构建“数据层－模型层－应用层－用户层”四级技术架构，实现大模型本地化部署与多元智能应用的深度融合。各层级既相互独立又紧密关联，形成完整的技术体系，具体架构如表1所示：

层级	核心功能	关键技术 / 组件	技术指标
用户层	提供统一访问入口	电脑端、移动端、融合门户、响应式设计	界面适配率≥98%，单次登录覆盖率100%
应用层	五大智能应用模块，实现业务逻辑	五大智能应用模块、SpringBoot、Golang、Vue.js	并发处理能力≥1000用户/秒，服务可用率≥99.9%
模型层	提供大模型核心能力，支撑应用开发	Qwen32B 模型、Qwen2.5-72B 模型、RAG 引擎、规则引擎	文本生成准确率≥85%，检索响应时间≤1.5s
数据层	数据存储、同步与预处理	MySQL、Redis、MinIO、ETL 工具（Kettle）、Tika 工具	数据同步延迟≤30s，数据清洗准确率≥99%

（表1）

（一）数据层：混合存储与高质量预处理

数据层为整个应用生态提供高质量数据源，采用“结构化＋非结构化＋向量”的混合存储方案。存储架构：MySQL 数据库存储结构化数据（公文元数据、审批流程节点、师生基本信息等）；Redis 作为缓存系统，存储高频访问数据（如近期公文、常用业务规则），提升访问速度；MinIO 分布式文件系统存储非结构化数据（规章制度文档、中医药专业手册、附件文件等）与模型文件，支持 PB 级存储扩展；数据同步：采用 ETL 工具 Kettle 实现各系统数据的抽取、转换与加载，通过 RESTfulAPI 接口实现实时同步，基于 OAuth2.0 协议与 HTTPS 加密传输保障数据安全，同步延迟控制在30s 以内；数据预处理：开发专用数据清洗工具，通过正则表达式过滤冗余信息、标准化数据格式，采用自然语言处理技术对非结构化文本进行分词、词性标注与实体识别，经 Tika 工具提取多格式文件文本内容，确保输入模型层的数据质量^[3]。

（二）模型层：大模型适配与协同运作

模型层为应用层提供核心算法支撑，构建大模型、RAG 引擎与规则引擎的协同架构。大模型适配：选用 Qwen 系列大模型进

行本地化部署，基于校园特色语料（包括公文规范、中医药专业术语、业务流程说明等，语料规模达50GB，涵盖10万余文档）采用 LoRA（Low-RankAdaptation）微调方法进行领域适配训练，微调后模型对校园场景语义理解准确率提升35%；引擎集成：RAG 引擎通过 API 接口调用大模型的向量转换与文本生成能力，实现检索与生成的无缝衔接；规则引擎预设120+ 条高校业务规则，包括公文格式规范、审批流程约束、敏感信息过滤标准等；性能优化：引入 Redis 缓存机制，将高频检索的知识片段与向量数据存储于缓存，缓存命中率维持在80% 以上，使 RAG 引擎响应速度提升40%^[4]。

（三）应用层：插件化集成与功能联动

应用层采用前后端分离架构，通过插件化方式实现智能模块与原有系统的无缝集成。技术栈选型：后端基于 SpringBoot 与 Golang 开发高并发服务，支持负载均衡与水平扩展；前端采用 Vue.js、HTML+CSS+JS 构建可视化界面，支持响应式设计；权限管理：对接信息门户系统，实现用户权限的精细化管理，设置管理员、部门负责人、普通教职工、学生四类角色，明确各角色的检索范围与操作权限；模块联动：五大智能应用模块通过 RESTfulAPI 实现数据互通与流程联动，例如智能消息中心的待办事项可直接跳转至 AI 附件摘要模块查看相关文件摘要，智能公文助手的查询结果可同步至史记胶囊应用进行历史数据关联^[5]。

（四）用户层：统一入口与个性化体验

用户层聚焦使用体验优化，实现多终端统一访问与个性化服务。统一访问入口：通过山东中医药大学融合信息门户实现所有智能应用的一站式访问，用户通过统一身份认证后即可使用全部服务，无需重复登录；多终端适配：开发电脑端（Windows/MacOS）与移动端（iOS/Android）应用界面，采用响应式设计确保不同设备上的使用体验一致性，移动端支持离线查看已缓存的摘要信息与待办事项；个性化服务：基于用户角色、访问历史与业务偏好，生成个性化首页，优先展示高频使用模块与相关信息，支持用户自定义首页布局与消息推送频率^[6]。

三、高校智能应用模块实现

基于“数据层－模型层－应用层－用户层”四级技术架构与相关理论支撑，开发覆盖高校管理服务核心场景的五大智能应用模块，各模块功能互补、流程联动，构建协同高效的智能应用生态。

（一）智能消息中心

对接 OA、教务、人事等 12 类系统，自动汇总个人待办与学校通知，生成含今日待办、未读通知、审批提醒的智能简报，支持一键跳转详情，实现信息一站式获取。

基于 RESTful 与 OAuth2.0 保障通信安全，设置标准化触发规则实现数据自动同步，接入 Qwen2.5-72B 模型提取关键信息生成摘要，以加权排序算法形成个性化简报，推送延迟≤5s。

（二）AI 附件摘要

嵌入审批流程，支持 Word、PDF、Excel 等 10 余种格式

附件核心内容自动提取与结构化总结，生成含核心结论、关键数据、行动要求的摘要，支持在线查看、下载、分享。

用 Tika、POI 解析多格式文件并转为结构化数据，以 Qwen32B 为核心引擎，通过提示词适配校园场景与中医药术语，50 页内文档生成时间 $\leq 2s$ ，同步展示于审批页面^[7]。

（三）智能公文助手

构建校园公文知识库，收录 8 万 + 篇公文，月更新 ≥ 500 篇，提供多模式检索、智能问答、公文生成、格式校验四大功能，支撑公文全流程处理。

以 Qwen32B 为引擎，Kettle 同步公文数据并完成清洗标准化，定制 OCR 识别与文本修复，扫描件文本提取准确率 $\geq 95\%$ ；基于 RAG 实现自然语言问答，规则引擎完成格式校验，权限模块精准控制访问范围。

（四）史记胶囊应用

核心功能：基于高校历年信息发布系统数据，如通知公告、新闻动态、重要事件记录等，通过 AI 摘要技术编纂学校历年工作纪实，支持按年度、月度、日度三个维度查询事记摘要，提供原文跳转、关键词检索、分类筛选功能，支持电脑端与移动端快捷查阅，便于历史信息追溯与工作复盘。

实现原理：通过标准化 API 接口对接学校通知公告系统、新闻网等数据源，抓取新增数据并批量导入 2018 年以来的历史数据（累计数据量 ≥ 10 万条），经数据清洗（去重、错误修正）、分类标注（按业务领域、重要程度）后存入系统；基于 Qwen2.5-72B 模型开发多维度摘要生成模块，按年度（侧重重大事件、成果总结）、月度（侧重重点工作、阶段进展）、日度（侧重具体事项、时间节点）分别设计摘要逻辑；开发双端适配的检索查询界面，支持按时间范围、关键词、业务类型组合查询，通过数据关联机制实现摘要与原文的精准联动，原文跳转成功率 $\geq 99\%$ ^[8]。

（五）AI 对话管理平台

基于对话日志分析，实现情绪意图识别、敏感事件监控与分级预警，重点监测学业焦虑、校园欺凌、心理危机等，为校园管理提供数据支撑。

Golang 后端支持每秒 ≥ 500 条日志处理，Qwen32B 完成智能分析，基于校园语料优化模型，搭建 500 + 敏感词库与 8 类对话倾向分类；规则匹配 + 模型解析双重识别风险，按级别推送预警并关联帮扶资源^[9]。

四、实践结果分析

测试数据：选取山东中医药大学 2025 年 1 月—2026 年 1 月的数据作为测试样本，包括公文数据 8 万余篇、附件文件 3 万余个、对话日志 5 万余条、用户操作记录 10 万余条，覆盖教学、科研、行政、学生管理等核心场景。

功能性能测试：智能公文助手查询准确率达 89%，较传统系统（62%）提升 43.5%，较未集成 RAG 引擎的大模型应用（75%）提升 18.7%；AI 对话管理平台业务问题解决率从 67% 提升至 82%，其中心理危机、学业压力等敏感问题识别准确率达

91%；各模块平均响应时间 $\leq 1.5s$ ，满足实时性需求（表 2）

评价指标	传统系统	本研究方案
查询准确率	62%	89%
页面平均响应时间	1.9s	1.1s
多格式附件支持率	65%	98%
流程审批周期平均时长（h）	9.05	3.12

（表 2）

五、结论与展望

（一）研究结论

AI 大模型本地化部署是高校智能生态构建的优选路径，四级架构有效保障数据安全、提升场景适配力，解决云端部署短板。五大模块形成联动生态，覆盖核心管理服务场景，破解信息孤岛、服务滞后等问题，显著提升管理效率与师生满意度，为高校数字化转型提供可复制经验^[10]。

（二）未来展望

- 拓展应用场景，延伸至教学科研、后勤服务，开发智能教学助手、科研数据分析等模块；
- 深化技术融合，结合区块链、边缘计算、物联网，探索多模态大模型校园应用；
- 构建开放生态，搭建校际开发平台，推动应用共建共享；
- 推动标准制定，联合高校与机构制定数据安全、模型适配、应用开发等行业标准。

参考文献

[1] 孙秋瑞, 杨栋. AI 大模型重塑高校应用新范式 [J]. 中国教育网络, 2025, (Z1): 31-33.

[2] 王春波, 姜晓明. 基于人工智能大模型的高校智慧校园服务体系建设研究 [J]. 华东科技, 2024, (12): 119-121.

[3] 沈丽燕, 李萌, 张紫薇, 等. 基于 AI 技术的高校智慧教学生态体系的构建与应用: 以浙江大学为例 [J]. 现代教育技术, 2022, 32(12): 85-92.

[4] 肖建力, 黄星宇, 姜飞. 智慧教育中的大语言模型综述 [J/OL]. 智能系统学报, 1-17[2025-03-22]. <http://kns.cnki.net/kcms/detail/23.1538.tp.20250205.1354.002.html>.

[5] 李思思. 中文医学文本实体关系抽取方法研究 [D]. 山西大学, 2024.DOI:10.27284/d.cnki.gsxiu.2024.002374.

[6] 赵美展. 人工智能在智慧领域中的应用及风险研究 [C]//“2024 职业教育活动周——钢铁行业大工匠进校园”活动. 哈尔滨信息工程学院, 2024.

[7] 刘宏宇, 武艳丽, 左喆. 零触控智慧系统集成技术应用研究 [J]. 建设科技, 2022(016): 000.

[8] 段磊. 高质量数据是 AI 发挥作用的根基所在 [J]. 中国教育网络, 2025(1): 20-23.

[9] 王慧颖. 基于人工智能技术的高校软件工程专业课程教学 [C]//“AI 赋能, 智能”行政管理综合能力提升研讨会. 南昌应用技术师范学院, 2024.

[10] 沈方园. 人工智能在高校信息化室中的应用前景 [J]. 上海建桥学院学报, 2025(1): 53-55.