基于具身人工智能的多模态智慧学伴应用开发与教学实践研究

□ 严一格

摘　要：随着教育数字化战略的深入推进，传统基于文本的智能问答系统已难以满足课堂教学与学生个性化学习支持的需求。为此，本文构建了一个基于具身人工智能的多模态智慧学伴系统“浙小安”，融合语音、视频、视觉感知、校本知识库、物联网控制及 MCP 工具链等关键技术，实现“可看、可说、可行动”的智能交互能力。系统采用本地部署的大模型 Qwen3-30B-A3B，并通过 CPT、SFT 与 LoRA 等技术完成校园语境适配；同时基于 RAG 框架实现教务、课程与准警知识的高准确度检索。依托实体端、网页端与移动端三端一体的架构，智慧学伴可支持课堂答疑、随堂回放、项目指导、课后个性化学习与情绪关怀等功能。在两学院开展的教学实践中，系统表现出良好的应用效果：课堂互动率提升 25%，任务完成率提升 30%，学生满意度达到 4.8/5。上述结果表明，“浙小安”在响应速度、任务完成率与教学满意度方面均取得显著成效，为多模态 AI 深度融入课堂教学提供了可行路径与实践范式。

关键字：具身人工智能；多模态交互；智慧学伴；大模型本地部署；RAG 检索增强；MCP 工具链

随着教育数字化战略行动的全面推进，人工智能技术在高校教学场景中的应用正不断深化。大语言模型在自然语言理解、知识组织与任务决策方面展现出显著优势，为智慧校园建设、个性化学习支持及课堂辅助教学带来了新的可能性。然而，目前高校普遍部署的智能问答系统仍以文本交互为主，交互方式单一，缺乏对多模态信息的理解能力，对复杂任务的执行能力有限。此外，系统通常依赖云端计算，存在延迟高、场景适配性弱、隐私风险突出等问题，难以满足课堂即时反馈、跨场景陪伴式学习以及高敏感度校园业务的需求。另一方面，学生在课堂学习、课后复习以及项目实训中，对“即时、可信、可互动”的学习支持需求日益增长。传统教学辅助手段难以实现对学生学习行为的实时感知，也无法在课堂、宿舍与校园场景之间保持连续陪伴。随着具身智能与多模态感知技术的发展，能够“看见环境、理解语境、执行任务”的学习型智能体成为可能，为构建智慧学伴提供了新的技术路径。

基于以上需求，本文提出并构建了一个基于具身人工智能的多模态智慧学伴系统“浙小安”。系统融合语音识别、视频理解、情绪识别、物联网控制、本地化大模型推理与校本知识库检索等关键技术，形成“可看、可说、可行动”的综合能力。通过在端侧部署 Qwen3-30B-A3B 并结合 CPT、SFT 与 LoRA/QLoRA 轻量化模型优化策略，系统能够在保证隐私性与低延时的前提下，高质量完成课程问答、学习指导、任务分解及心理关怀等功能。此外，通过 MCP（Model Context Protocol）与 Dify 的结合，智慧学伴可完成包括成绩查询、请假申请、教室灯光控制等校园业务操作，实现从“能回答”向“能执行”的跃迁。在教学应用方面，“浙小安”被嵌入课堂、实训与课后多场景，实现了随堂答疑、要点回放、个性化学习资源推送以及学习画像生成等功能。通过对两学院的教学实验表明，智慧学伴能够提高课堂参与度，改善学生学习习惯，降低教师重复性工作负担，并在心理支持方面展现潜在价值。本文将对系统的架构设计、关键技术路径、功能实现及教学实践效果进行系统性阐述，以期为高校构建多模态智能教学助手提供可复制、可推广的实践样例。

一、关键技术

智慧学伴系统“浙小安”在设计与开发过程中融合了人工智能、Web 前后端开发、物联网控制、多模态交互与嵌入式硬件等多类关键技术，以实现“可看、

可说、可行动”的具身智能体验。本节将从后端技术、前端技术、AI 技术栈、RAG 检索、MCP 工具链以及硬件端实现等方面系统阐述项目的核心技术基础。

（一）基于 Spring Boot 的微服务治理与接口集成

系统后端采用 Spring Boot 构建，具备稳定性高、生态完善、扩展性强等优势，能够满足智慧校园类项目对高并发稳定访问、权限管理与模块化服务的需求。核心功能包括：

1.RESTful API 服务：对接前端、实体端、移动端，实现大模型调用、知识检索、用户会话管理、工具指令执行等功能。

2. 校园业务 API 聚合：后端负责与校内统一身份认证系统、成绩管理系统、课表系统、准警系统等进行安全通信，为 MCP 工具子系统提供统一数据入口。

3.IoT 控制接口：后端通过 MQTT/HTTP/WebSocket 等协议连接智慧校园物联网网关，支持教室灯光、空调、投影仪等远程控制，实现“能执行”的具身智能行为。

（二）大模型本地部署、多模态理解与情绪识别

在智慧学伴的智能核心构建中，本地化大模型部署与多模态理解技术发挥了关键作用。本研究采用 Qwen3-30B-A3B 作为底层语言模型，并在校内 GPU 服务器上实现全本地化部署，所使用 GPU 的详细参数如图 1-2 所示。通过 vLLM 推理加速框架显著降低推理延迟，具备适应教学场景的高并发访问能力。为进一步提升模型在校园语境中的理解力与表达准确性，系统结合 CPT 连续预训练、SFT 监督微调以及 LoRA/QLoRA 轻量化适配等技术，使模型可精准响应课程知识点讲解、教务制度问答、实验步骤指导及心理关怀类对话等需求。此外，智慧学伴还引入语音识别、语音合成、视频感知、OCR 文本识别、基础物体检测与情绪识别技术，实现听觉与视觉的融合感知，使系统能够在真实环境中理解用户语气、表情及学习情境，从而提供更拟人化、情境化的学习支持。

（三）MCP 工具链与 Dify 编排：从“会说”到“会做”

为使智慧学伴具备执行真实校园任务的能力，而不仅仅停留在语言问答层面，本研究引入 Model Context Protocol（MCP）作为跨系统操作的标准化接口机制，通过工具化抽象实现成绩查询、请假申请、课程检索、宿舍到寝核验及 IoT 设备控制等多类型任务的自动执行。系统同时采用 Dify 作为 MCP 客户端，负责自然语言指令解析、参数抽取、对话式表单填充、执行摘要确认以及多工具链工作流编排，确保操作过程的准确性与可控性。相应地，MCP 服务端承担权限验证、事务一致性控制及审计记录写入等安全治理任务，保障所有校园事务均在合规边界内执行。借助 MCP+Dify 的结合，智慧学伴从传统的“语言助手”拓展为能够代学生办理实际事务的“行动型智能体”，实现从“会说”向“会做”的关键能力跃迁。

二、技术在系统中的实现

智慧学伴“浙小安”的系统实现过程是在总体架构和关键技术方案的基础上，将大模型、本地推理、多模态感知、工具执行链路以及具身化硬件终端进行高度集成，以实现跨场景、跨介质、跨模态的连续学习支持体验。该系统不仅在技术上满足了智慧校园对实时响应、精确答复和设备联动的要求，同时在教学应用中实现了与课堂、课后及实训环节的深度融合。本节将从系统架构设计、大模型推理服务以及实体硬件端等三个方面阐述技术的具体落地方式。

（一）系统架构设计

本智慧学伴“浙小安”的系统架构采用分层解耦、端云协同与多模态融合的设计理念，通过对大模型推理、本地服务治理、多模态交互能力、工具调用链路以及具身化硬件的系统集成，构建了一个可在课堂、课后、自习室以及宿舍等多场景持续运行的智能学习支持体系。从整体上看，系统由表现层、交互层、服务层、智能层、数据层与硬件层构成，各层通过标准化接口通信，使系统在保证高可扩展性的同时，具备良好的稳定性与可维护性，整体系统架构如图 2-1 所示。

图 2-1 智慧学伴整体架构图

在架构顶部，表现层面向不同使用场景提供三种

交互形态：网页端采用文本与数字人语音界面，用于课堂教学与在线学习；移动端以 App 与小程序形式实现随身学习支持；实体端基于 ESP32 与 3D 打印外壳形成具身智能载体，用于宿舍、自习室等场景的持续陪伴。表现层通过 WebSocket、HTTP 及 MQTT 等协议与后端通信，为用户提供多终端一致的智能交互体验。

交互层负责统一处理语音、文本与视觉输入，通过 VAD、ASR、TTS 及 OCR 等技术模块完成对语音流、图像流的解析与特征抽取，并将处理结果封装为结构化提示输入至智能层，实现多模态信息的高效融合。对课堂场景而言，交互层可实时捕获学生提问、识别板书信息；在课后场景，则可通过摄像头识别作业纸内容或学习资料，从而为大模型提供更丰富的语境支持。

服务层是系统运行的枢纽，实现包括 API 网关、用户会话管理、权限控制以及 MCP 工具链编排在内的核心功能。服务层不仅承担将用户自然语言指令映射为可执行任务的职责，还负责调用教务系统、准警系统、校园新闻系统以及物联网平台的接口，实现成绩查询、请假申请、到寝核验、灯光空调控制等具体操作。通过 Dify 作为 MCP 客户端，系统将工具调用过程抽象为标准化流水线，使智慧学伴具备“从理解到执行”的完整任务能力。

智能层作为系统核心，主要包括本地部署的 Qwen3-30B-A3B 服务、CPT/SFT/LoRA 的模型适配模块以及 RAG 检索引擎。推理服务通过 vLLM 加速机制实现低延迟生成，保证课堂和实训场景中的实时响应；模型适配模块根据课程资料、教务制度等校本语料对大模型进行领域增强；RAG 引擎与校本知识库耦合，确保智慧学伴在处理课程知识点、准警条例及常见问答时具备高准确度和可溯源性。

数据层为智慧学伴提供知识支撑与动态反馈机制，包括课程知识库、校园 API 数据、学生学习画像与互动日志。通过该层的结构化管理，系统能够持续更新知识内容、记录学生行为数据，并在教学中提供个性化推送与学习诊断。

硬件层由 ESP32 控制模块、语音采集模块、扬声器及可扩展传感器组成，通过 MQTT 与后端通信，实现本地唤醒、环境感知、设备联动等功能。结合 3D 打印外壳设计，实体端学伴能够作为物理载体，以高度拟人的交互方式融入学生学习生活，实现持续陪伴与即时反馈。

（二）大模型推理服务的实现

“浙小安”的智能核心通过本地部署的大规模语言模型 Qwen3-30B-A3B 实现。在校内 GPU 服务器上构建推理服务，使系统在处理课堂问答、课后辅导、心理关怀等场景时能够保持低延迟响应，同时确保学生隐私数据不离开校园网络环境。推理过程基于 vLLM 加速框架，通过连续批处理、KV 缓存复用等机制有效提升并发能力，使系统能够在课堂高负载场景下稳定运行。

为进一步适配教育领域，本研究采用 CPT、SFT 与 LoRA/QLoRA 轻量化方法对模型进行多阶段优化。CPT 用于将模型语言风格迁移至教学语境；SFT 基于课程问答、实验步骤说明、教务制度解析等指令数据提升模型专业性；LoRA 使不同学院或课程能够以较低成本获得定制化模型能力。在知识对齐方面，推理服务与 RAG 模块协同运作，通过向量检索获取课程讲义、制度规范等校本知识，并在推理时注入上下文，使回答有据可依、减少幻觉，满足教学场景的准确性要求。

在多模态支持方面，模型可融合 ASR 语音识别、OCR 图像文本提取、基础物体检测等外部输入，通过统一的提示模板进行语义整合，使其可应用于课堂演示理解、作业识别、学习材料解析等场景。例如，当学生拍摄作业纸时，系统可自动识别题干文本并指导下一步学习行为。

总体来看，本地模型推理服务实现了低延迟、高稳定性与教育场景定制化的统一，使智慧学伴能够在课堂、课后与实训环境中稳定提供准确、可信的智能支持。

三、多终端协同与系统集成实现

本智慧学伴“浙小安”的系统实现基于端云协同架构构建，核心理念是在保持大模型本地推理能力的前提下，实现网页端、移动端与实体端的统一接入及一致交互体验。后端采用 Spring Boot 3 构建服务框架与 API 网关，负责用户会话管理、意图解析、工具调用编排与多模态数据转发；前端基于 Vue 实现聊天界面、数字人交互界面与教学管理端界面，支持实时文本流展示、数字人口型同步与语音收发。系统数据由 MySQL 8.0 统一存储，涵盖用户画像、课程知识条目、会话记录、行为日志及 MCP 工具调用回执，并通过 Redis 提供热点知识缓存、快速会话态存储及消息队列式异步调度，显著提升系统整体响应速度。

平台的知识增强链路与任务执行链路通过 Dify 统

一管理：其承担 MCP 客户端职责，实现自然语言意图识别、参数抽取、任务表单生成、工具链编排及执行摘要确认，确保智慧学伴不仅能够回答问题，更能够通过标准化工具链执行成绩查询、到寝核验、课程检索与 IoT 设备控制等实际操作。在静态资源管理方面，Nginx 负责承载前端项目部署与反向代理，同时作为 WebSocket 连接的网关节点，保障实体端与网页端的实时通信；整个系统采用 Docker Compose 容器化交付，实现快速部署、迁移及版本回滚。

多模态处理链路由前端与后端协同完成。用户通过网页端或移动端发送语音输入，前端完成 VAD 端点检测后，将音频流转发至后端，由 ASR 引擎实时转写；若同时启用数字人界面，系统将生成的语义内容同步驱动 Live2D 模型，实现嘴型与表情同步。对于图像与视频输入，前端采集内容并传送至后端 OCR 与视觉识别模块，通过结构化提取将文字、物体与场景特征送入大模型提示词，从而实现“看得懂、听得见、能回应”的多模态融合体验。

在大模型推理方面，系统运行本地部署的 Qwen3-30B-A3B，并通过 vLLM 提供高性能推理调度。为适应教学场景需求，系统构建“双通道推理模式”：对于课程概念定义、作业解释等简单提问，直接走快速通道，保障交互流畅性；而涉及课程深度推导、实验报告生成或知识库引用的问题，则进入慢通道进行 RAG 增强推理。慢通道结果将在前端以“无感刷新”方式覆盖初步回答，实现先答复后优化、低延迟与高准确性的平衡。

实体端基于 ESP32 构建具身学伴终端，负责本地唤醒、语音采集、即时响应及 IoT 控制；通过 MQTT 接入后端任务链路后，实体学伴具备执行灯光控制、环境巡检、课程提醒与学习陪伴等功能。实体端外壳采用 3D 打印制作，可结合课程主题、学生喜好及校园文化定制角色形象，使智慧学伴不仅具备技术功能，也具备教育陪伴的情感属性。

整体来看，系统通过本地大模型推理、多模态交互、RAG 增强、MCP 任务链路与具身化实体终端的深度整合，实现了“能理解、会思考、可行动、可陪伴”的一体化智慧学习支持体系，为课堂教学、课后辅导与实训教学提供稳定、精准、持续的智能支撑。

四、结语

本文围绕高校教学与学习支持中“课堂辅助不足、课后陪伴缺失、校园服务割裂”的现实痛点，提出并实现了“基于具身人工智能的多模态智慧学伴”系统。系统在技术层面采用本地化大模型推理、RAG 检索增强、MCP 工具链与 Dify 编排，实现了从智能问答到实际事务执行的能力跃迁；在交互层面融合语音、文本、视觉与情绪识别，并通过网页端、移动端与 ESP32 实体端的多形态终端，为学生构建可随时访问、可持续陪伴的学习支持场景；在数据与知识治理层面，以 MySQL+Redis+ 向量库构建可信、可溯源、可迭代的知识底座，并基于日志、会话记录与学生画像实现过程性评价与智能化诊断。系统已在课程学习、课堂互动、自主实训及校园事务办理等场景中展现出良好的应用前景，为教师提供有效的教学辅助，为学生提供稳定、可信、低门槛的学习伙伴。

未来工作将从三个方向进一步推进：其一，在大模型侧持续优化 CPT、SFT 及 LoRA 等适配策略，提升模型对课程知识、校园制度及学习情境的理解深度；其二，强化具身智能的多模态感知能力，使实体学伴具备更完善的环境识别、任务感知及 IoT 协作能力，进一步拓宽课堂与生活场景的应用范围；其三，推进系统在更多学院、更多课程中的规模化应用，探索基于学生画像的个性化学习路径推荐及面向专业课程的场景化 Agent，构建可复制、可推广、兼具教育价值的智慧学伴体系。随着大模型技术的发展和智慧校园基础设施的完善，本项目有望成为高校信息化、教学创新及学生支持体系的重要组成部分，为未来学习方式变革提供可持续的技术路径与实践依据。

参考文献：

[1] 孙思思 , 贾佳 , 常江 , 等 . 转型时代下的创新实践 :AI 工具在高校中的应用场景设计研究 [J]. 信息与电脑 , 2024, 36(22):189-192.

[2] 张慧芳 .E-learning 时代互动型教材的设计与应用 [D]. 湖南师范大学 .

【项目基金：“浙江安防职业技术学院教学改革研究项目”基于具身人工智能的多模态智慧学伴应用开发与教学实践研究——以“浙小安”为例（编号：JG202502）。】

作者简介：

严一格，男，汉族，浙江温州人，硕士，浙江安防职业技术学院人工智能学院助教，研究方向为大规模预训练模型优化。