

大模型远程 API 赋能高校就业指导的实践与实证分析

龚翔, 袁小刚, 高旭昊, 常有康, 马亚燕
甘肃政法大学网络空间安全学院, 甘肃 兰州 730070
DOI: 10.61369/VDE.2026040037

摘 要 : 针对当前高校毕业生就业指导服务存在的覆盖不足、精准度低及西部高校算力有限等问题, 本文将 DeepSeek 大模型本地化部署重构为远程 API 调用的轻量化智能体平台, 以解决本地算力不足、并发性能弱、模型精度受损、运维成本高等痛点。平台采用五层架构设计, 全程无本地模型加载, 通过 API 接入、Prompt 工程、RAG 知识库与容器化部署实现快速落地。以甘肃政法大学 800 名毕业生为对象开展对照实验, 结果表明, 远程 API 方案在响应速度、并发稳定性、内容准确率及岗位转化率上均显著优于本地部署, 且成本更低、运维更简便。

研究验证了 API 化方案在区域高校的适用性, 可为西部高校提供轻量化、低成本的智能就业指导落地路径。同时, 研究存在 API 持续成本、小众专业数据覆盖不足、断网容灾薄弱等问题, 未来可通过混合架构、动态数据库及政校企协同进一步优化完善。

关 键 词 : 智能体; 远程 API 调用; 就业指导

Practice and Empirical Analysis of Large Model Remote API Empowering Employment Guidance in Universities

Gong Xiang, Yuan Xiaogang, Gao Xu hao, Chang You kang, Ma Yayan
School of Cyberspace Security, Gansu Political Science and Law University, Lanzhou, Gansu 730070

Abstract : Aiming at the problems of insufficient coverage, low accuracy in college graduates' employment guidance services, and limited computing power in universities in western China, this paper reconstructs the localized deployment of the DeepSeek large model into a lightweight intelligent agent platform based on remote API calling, so as to solve the pain points such as insufficient local computing power, weak concurrent performance, impaired model accuracy, and high operation and maintenance costs. The platform adopts a five-layer architecture design without local model loading in the whole process, and realizes rapid deployment through API access, Prompt engineering, RAG knowledge base and containerized deployment. A controlled experiment was conducted on 800 graduates of Gansu University of Political Science and Law. The results show that the remote API scheme is significantly better than the local deployment in response speed, concurrent stability, content accuracy and post conversion rate, with lower cost and simpler operation and maintenance.

The research verifies the applicability of the API-based scheme in regional universities, and can provide a lightweight and low-cost path for the implementation of intelligent employment guidance in universities in western China. Nevertheless, there are still limitations including continuous API costs, insufficient data coverage for less-popular majors, and weak network disconnection disaster recovery. In the future, the platform can be further optimized through hybrid architecture, dynamic database and government-university-enterprise collaboration.

Keywords : intelligent agent; remote API calling; career guidance

引言

(一) 研究背景及意义

近年来我国高校毕业生规模持续增长, 2025 年预计突破 1200 万人, 就业工作面临总量压力与人岗错配双重挑战^[1]。传统就业指导存在覆盖有限、信息滞后、精准度不足等问题, 甘肃等西部地方高校受区域资源限制, 服务短板更为突出^[2]。

前期研究基于 DeepSeek V1.0 实现本地化部署, 在就业指导中取得一定成效, 但受本地算力不足制约, 高峰期并发能力弱、响应延

项目信息: 本项目得到甘肃省教育科学“十四五”规划 2025 年度“大学生职业规划与就业指导”专项课题《AI 大模型驱动的可信大学生职业规划与就业指导智能体平台构建研究》(项目号: GS[2025]GHBZX0059) 的支持。

迟高，模型量化压缩导致输出精度下降，运维与迭代成本偏高。为此，本研究将方案改进为 DeepSeek 远程 API 调用模式，无需本地高性能算力，可显著提升并发、响应速度与服务精度。研究实践意义在于构建轻量化、低成本智能就业指导平台，缓解西部高校就业服务资源不足问题；理论意义在于完善大模型 API 化在就业指导中的应用范式，为区域高校智能化转型提供可借鉴的技术方案与实践参考。

（二）研究目标与内容

本研究针对前期 DeepSeek 本地化部署存在的算力不足、并发受限、模型精度因压缩下降、维护成本偏高等问题，以轻量化、高可用、低成本、区域化落地为导向，将技术架构重构为 DeepSeek 远程 API 调用模式，旨在搭建不依赖本地高性能算力、可支撑高校规模化使用的大学生就业指导智能体平台^[3]。

研究目标为突破本地硬件约束，在普通服务器上快速部署大模型就业服务，提升系统并发、响应速度与输出准确率，降低高校落地与运维成本，形成可在甘肃等西部区域高校推广的 AI 就业指导方案。本研究不涉及大模型底层算法创新，聚焦工程化落地与场景适配，完成 DeepSeek 远程 API 接入、接口封装与稳定性优化，设计就业场景专用 Prompt 模板与调用策略，确保输出合规、精准、可控；同时完成平台整体设计，明确设计理念、技术路线与分层架构，搭建职业画像、岗位推荐、智能问答、政策解读、AI+ 人工协同等核心功能模块，实现轻量化部署，并构建区域化就业知识库，借助 RAG 技术增强输出专业性，配套数据脱敏与加密等安全机制。

研究以地方高校毕业生为对象开展实证实验，设置本地部署组与远程 API 组对照，从性能、服务效果、使用成本等维度进行统计对比，并依据实验结果与用户反馈迭代优化接口、功能流程及推荐逻辑，提出面向区域高校的长效运营与推广策略。

一、相关理论与研究现状

（一）DeepSeek 远程 API 智能体

DeepSeek 远程 API 智能体是依托云端推理接口构建、面向大学生就业指导的轻量化智能服务载体，通过标准化接口调用模型能力，无需在本地加载完整模型参数，可在低算力、低成本条件下完成职业画像、人岗匹配、政策解读、智能问答等就业指导任务。相比本地部署，该模式可摆脱硬件算力限制，直接使用云端持续优化的模型能力，在服务稳定性、并发支撑、输出精度与部署效率上优势显著，更适配高校常态化、规模化使用需求^[4]。

DeepSeek 模型具备良好的文本理解、信息生成与多轮对话能力，可精准识别学生求职特征与岗位需求，结合领域知识库与规则约束后，能够输出贴合区域就业实际、可信可解释的指导内容，与高校就业指导业务高度适配^[5]。

（二）国内研究现状

国内人工智能在大学生就业指导领域的研究已形成学术与实践并行的格局，成果集中在 AI 就业工具研发、大模型本地化部署及轻量化技术应用。多所高校将大模型应用于职业画像、岗位匹配等场景，通过模型量化等轻量化手段试点部署，提升了服务效率，形成“AI 推荐 + 人工核验”模式，为智能化转型提供参考，轻量化技术也不断降低本地运行门槛^[6]。但现有应用仍有突出问题：本地化方案依赖高性能算力，高峰期并发不足、响应延迟；就业数据更新滞后、区域信息覆盖不全，影响匹配精准度；小众专业适配、个性化帮扶效果不佳，区域适配性不足；且研究多聚焦功能实现，对隐私保护、长效运营及区域推广的探讨欠缺，难以满足西部高校低成本、高可用的需求^[7]。

（三）国外研究现状

国外在生成式人工智能与就业服务融合领域起步较早，已形成成熟的研究与实践体系。其核心围绕 AI 职业匹配、智能规划及

API 化模型服务展开，依托完善的云计算生态，通过 API 化接口集成，将大模型应用于岗位匹配、职业评估、求职路径规划等场景，大幅提升就业服务效率^[8]。同时，国外高度重视 AI 应用规范，以相关法规明确要求算法透明性、数据隐私保护，保障 AI 就业服务合规有序^[9]。但国外方案普遍投入成本高，且其服务逻辑、岗位体系、政策适配均基于发达经济体设计，与我国西部地方高校的办学条件、区域产业特征及就业市场环境差异较大，适配性不足，无法直接迁移应用，也难以满足低成本、轻量化的落地需求^[10]。

二、智能体平台设计与部署

（一）平台设计理念

本平台以可信赋能、精准适配、协同高效为核心设计理念，立足地方高校就业指导实际需求与区域产业发展特征，面向前期本地部署存在的算力瓶颈、服务稳定性不足等问题，构建轻量化、低成本、高可用的 AI 就业指导服务体系。平台严格遵循数据安全与隐私保护规范，对学生信息进行全流程脱敏与加密处理，保障数据使用合规安全；通过推荐依据可视化与输出逻辑可校验，强化算法可解释性，破解算法黑箱问题，提升师生使用信任感。

依托 DeepSeek 远程 API 实现无本地算力依赖的轻量化部署，显著降低硬件投入与运维成本，提升系统并发能力与响应效率。在服务层面紧密结合区域就业政策、地方产业岗位需求与高校专业特色，实现就业信息、岗位资源与学生能力倾向的精准对接，同时采用 AI 智能服务与人工指导协同联动机制，兼顾标准化服务效率与个性化指导质量，形成安全可信、轻量高效、贴合区域实际的大学生就业指导智能服务模式。

（二）技术路线与架构

1. 技术路线

平台遵循需求分析→API 对接→系统开发→测试优化→实证

验证的闭环技术路线推进实施。首先结合高校毕业生求职痛点、就业指导业务流程与区域产业需求，完成功能边界、接口规范、性能指标与安全要求的梳理与界定；其次完成 DeepSeek 远程 API 的申请、接口封装、调用策略与 Prompt 模板设计，实现大模型能力的轻量化接入；在此基础上开展应用端、服务逻辑、数据管理与安全机制的全流程开发，集成 RAG 检索增强、规则引擎与 AI + 人工协同模块；随后开展接口压力测试、功能可用性测试、安全合规测试与并发性能验证，修复问题并优化体验；最后以区域高校毕业生为对象开展实证实验，对比评估远程 API 方案相较于本地部署的综合效能，形成可落地、可推广的技术方案。

2. 技术架构

平台采用五层分层架构设计，以 DeepSeek 远程 API 为核心能力支撑，各层级协同联动，保障系统轻量化、高可用与安全稳定。

应用层面向学生与管理员提供 Web 端与小程序端交互入口，承载职业画像、岗位推荐、智能问答、政策解读、数据可视化等功能呈现，覆盖就业指导全流程服务。服务层作为业务调度核心，负责请求转发、接口封装、并发控制、缓存管理与异常降级，统一封装对外调用逻辑，降低前端接入复杂度。算法层以 DeepSeek 远程 API 为核心，结合 RAG 检索增强、规则引擎与混合推荐算法，完成语义理解、内容生成、匹配计算与合规校验，保障输出精准、可控、可解释。数据层负责脱敏学生信息、区域岗位库、就业政策库、简历语料的清洗、存储与更新，采用混合存储架构保障多源数据高效读写与时效性。安全层贯穿全架构，提供数据脱敏、传输加密、身份认证、权限控制与操作审计能力，确保平台在数据安全、隐私保护与合规使用上达到高校服务标准。

3. 平台部署过程

本平台采用全程无本地大模型加载的远程 API 轻量化部署方案，整体流程按照环境搭建、接口集成、业务配置、数据处理与安全上线有序推进。平台选用通用配置服务器与 Ubuntu 系统，部署 Python、Nginx、MySQL、MongoDB、Redis 等基础组件，完成环境初始化，全过程无需高性能 GPU 支撑。

在此基础上完成 DeepSeek 远程 API 的接入与封装，进行密钥与权限配置，基于 FastAPI 框架统一接口，构建请求分发、超时重试、限流熔断与缓存机制，保障模型调用稳定可靠。同时围绕就业指导场景开展 Prompt 工程建设，针对职业画像、岗位推荐、政策解读、智能问答、简历优化等业务设计结构化模板，规范输出内容与格式，提升结果精准度。

平台同步构建 RAG 检索增强知识库，对区域就业政策、地方岗位信息、高校专业要求及求职知识进行整理、向量化与索引构建，为模型提供专业领域支撑。在数据处理环节，对毕业生基本信息、实习经历、技能证书、求职意向等数据进行脱敏、清洗、归一化与标准化预处理，实现安全入库与合规使用。最后通过 Docker 容器化技术封装前后端及接口服务，借助 Docker Compose 完成服务编排，配置 Nginx 反

向代理与 HTTPS 加密传输，完善身份认证、权限控制与日志审计，搭配接口监控实现系统轻量化、安全化、稳定化上线运行。

三、平台实证实验与结果分析

（一）实验目的

本次实验旨在通过对照研究的方式，全面验证 DeepSeek 远程 API 方案相较于本地部署方案的综合改进效能，重点检验平台在响应速度、并发支撑能力、内容生成精度、岗位申请转化率以及部署与使用成本等关键指标上的提升效果，客观评估远程 API 模式在解决本地算力不足、服务稳定性不足、精准度偏低等问题上的实际作用，为智能体平台的技术选型、优化迭代与区域高校规模化推广提供科学、可靠的实证依据。

（二）实验对象

本次实验选取甘肃政法大学 2026 届全体毕业生共 800 名学生作为研究对象，涵盖网络空间安全、人工智能等相关本科与硕士专业，样本结构与学校实际生源构成保持一致。为保证实验结果的客观性与可比性，将 800 名实验对象随机均分为两组，其中 400 人为本地部署方案使用组，采用前期基于 DeepSeek 本地化部署的就业指导智能体平台获取服务；另外 400 人为远程 API 方案使用组，使用基于 DeepSeek 远程 API 重构优化后的就业指导智能体平台，两组在专业分布、学历层次、就业准备基础等方面保持均衡，确保实验数据具备可比性。

（三）实验结果统计

实验以甘肃政法大学 2026 届 800 名毕业生为有效样本，将其划分为本地部署组与远程 API 组开展对照测试，围绕系统性能、服务精度与应用效果三大维度进行数据统计与对比分析。

在系统性能层面，本地部署组职业画像与岗位推荐的平均响应时间为 2.8 秒，但在求职高峰期 200 人以上并发访问时，响应延迟上升至 5.7 秒，请求超时率达 4.1%；远程 API 组依托云端弹性算力支撑，常规服务平均响应时间稳定在 2.2 秒以内，高并发场景下无明显延迟波动，请求超时率控制在 0.3% 以下，并发承载能力与服务稳定性显著优于本地部署模式。在内容准确率方面，经就业指导教师统一核验，本地部署组受模型量化压缩与算力限制，输出内容平均准确率为 71%，远程 API 组使用官方完整版模型并结合 RAG 检索增强，内容平均准确率提升至 82%，信息有效性与专业贴合度明显改善。

在核心应用效果上，本地部署组有效岗位匹配率为 50%，基于平台推荐获得面试邀请的岗位申请转化率为 42%；远程 API 组依托更精准的语义理解与区域知识匹配，有效岗位匹配率提升至 61%，岗位申请转化率提升至 53%，在人岗匹配精度与求职转化效果上实现明显提升。同时，从全生命周期投入来看，远程 API 方案无需高性能硬件投入，运维开销大幅降低，整体成本显著低于本地部署方案，综合表明采用 DeepSeek 远程 API 的智能体平台在各项关键指标上均优于本地化部署方案，能够更高效、更稳定、更精准地支撑大学生就业指导服务工作。

四、总结与展望

本研究针对 DeepSeek 大模型本地化部署存在的算力不足、并发性能有限、模型精度受损、运维成本较高等问题，重构为基于 DeepSeek 远程 API 的轻量化就业指导智能体方案。实证结果表明，远程 API 模式有效突破本地硬件约束，显著提升系统响应速度、并发能力、内容准确率与岗位转化效率，在降低部署门槛、简化运维、增强稳定性等方面优势明显，可为区域高校提供轻量化、低成本、易落地的 AI 就业指导方案，为生成式 AI 在高

校就业服务中的规模化应用提供可行路径。同时，研究仍存在不足：长期 API 调用产生持续成本，大规模推广需优化调用策略；小众专业与跨专业就业场景数据覆盖不足，精准度有待提升；平台依赖网络，断网容灾与降级机制尚不健全。未来可从三方面优化：一是探索远程 API 与轻量本地模型结合的混合架构，兼顾成本、速度与离线能力；二是联合地方政府及园区搭建区域动态就业数据库，实现岗位与政策信息实时更新；三是完善政校企协同机制，整合多方资源，推动平台向区域化、体系化的智能就业服务生态发展，为缓解毕业生结构性就业矛盾提供技术支撑。

参考文献

[1] 曾小龙. 生成式人工智能时代大学生就业面临的机遇与挑战 [J]. 经济师, 2024, (12): 51–52+57.

[2] 张青松, 代伟. 青年高质量充分就业的现实差距与提升路径 [J]. 青年探索, 2024, (04): 82–92.

[3] 杨静. 生成式人工智能赋能应用技术型高校大学生就业能力提升的路径研究 [N]. 河北经济日报, 2025-05-09(009).

[4] 周文吉米. 基于 AI 与大数据挖掘的大学生就业岗位自动推荐算法 [J]. 无线互联科技, 2025, 22(15): 47–50.

[5] 孙庆英, 周含, 刘思妍, 等. 基于图卷积网络的岗位匹配模型研究 [J]. 河南师范大学学报 (自然科学版), 2025, 53(06): 93–101.

[6] 任皓. 人工智能视角下大学生核心就业能力提升路径探究 [J]. 现代职业教育, 2025, (17): 51–54.

[7] 张格. 高校学生生成式人工智能技术使用意愿影响因素研究 [D]. 安徽大学, 2024.

[8] Zheng Z, Qiu Z, Hu X, et al. Generative job recommendations with large language model[J]. arXiv preprint arXiv:2307.02157, 2023.

[9] Nandi A, Das A K, Sur H, et al. AI-Driven Intelligent System for Personalized Job Recommendations Using Real-Time Skill Identification and Industry Trend Analysis[C]//2025 International Conference on Computing, Intelligence, and Application (CIACON). IEEE, 2025: 1–6.

[10] Dewasurendra S V, Udunuwara U K C, Wanniarachchi T T, et al. Cutting-edge AI-driven higher educational and career advisory platform[C]//2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE, 2024: 1–7.