LLM 大模型在考研助手小程序开发中的应用

潘耀杨

宁夏理工学院，宁夏 石嘴山 753000

摘要：近年来，考研竞争日益激烈，考生在信息整合、个性化指导及高效备考等方面面临严峻挑战。大型语言模型（LLM）虽凭借强大的语言理解与生成能力迅速成为技术焦点，但其在垂直教育场景中的深度适配仍有待挖掘。为此，本研究提出一种基于 LLM 的智能考研辅助系统，通过融合多源异构数据与前沿 AI 技术，为考生提供精准化、动态化的备考支持。系统构建过程中，研究团队整合了历年真题、院校政策及备考经验等海量数据,搭建了结构化与非结构化混合知识库，并引入 LoRA 轻量化微调与 RAG 检索增强技术，显著提升了模型对考研场景的适应能力。针对学科差异，设计定制化提示模板优化生成风格，例如数学学科强调逻辑推演，英语学科注重语言规范。此外，系统支持文本与图片多模态输入，结合 OCR 识别与上下文关联问答功能，可解析用户上传的真题图片并生成分步解题指导。通过微信小程序实现轻量化部署后，系统能够动态推荐学习资源，并通过多轮对话模拟真实师生互动，帮助考生高效攻克备考痛点。本研究的实践表明，AI 技术在教育垂直领域的深度应用具有广阔前景。系统不仅为考生提供了低成本、高灵活性的智能工具，也为教育科技融合创新提供了可复用的技术框架。未来研究将进一步探索模型在复杂场景中的鲁棒性优化，并推动其在更多教育环节的落地应用。

关键词：检索增强生成（RAG），提示工程，低秩自适应（LoRA），多模态交互

1 引言

近年来，人工智能技术的迅猛迭代为教育领域注入了新活力，其中大型语言模型（LLM）的应用逐渐成为学界关注的焦点[1]。聚焦考研教育这一具体场景，考生普遍面临信息碎片化严重、学习规划僵化以及解题指导匮乏等现实困境[2]。传统备考模式多依赖固定教材和有限师资，难以应对复习过程中灵活多变的需求[3]。基于此背景，本研究尝试构建一款依托 LLM 的个性化智能备考助手，旨在通过技术革新缓解上述矛盾，同时推动考研教育效率与质量的整体跃升[4]。

为实现这一目标，研究团队系统整合了历年真题、院校政策、专业目录等多元数据，搭建了融合结构化和非结构化特征的知识库体系[5]。技术路径上，采用 LoRA（低秩自适应）技术对模型进行垂直领域微调，使其深度适配考研场景；引入 RAG（检索增强生成）机制，通过知识库检索增强生成内容的可信度，显著降低模型“幻觉”风险[6]。针对学科差异，设计了定制化提示模板以优化输出风格——例如数学学科强调逻辑推演步骤的清晰性，英语学科则注重语言表达的规范性，从而提升专业适配性[7]。

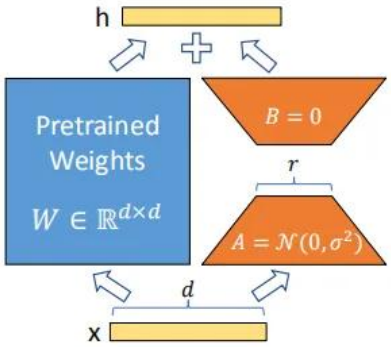
从实践成效来看，系统通过微信小程序实现了轻量化部署，为考生提供低门槛、高效率的智能支持。平台不仅支持文本与图片多模态输入（如上传真题图片自动解析并生成解题步骤），还能基于上下文关联实现多轮对话，模拟真实教学场景中的师生互动。此外，个性化推荐模块通过分析用户学习轨迹，动态推送适配资源，帮助考生精准攻克薄弱环节。这一探索不仅为备考群体提供了切实可用的工具，更通过技术落地验证了 AI 与教育深度融合的可

行性，为后续研究提供了兼具学术价值与实践意义的参考范式。

2 关键技术

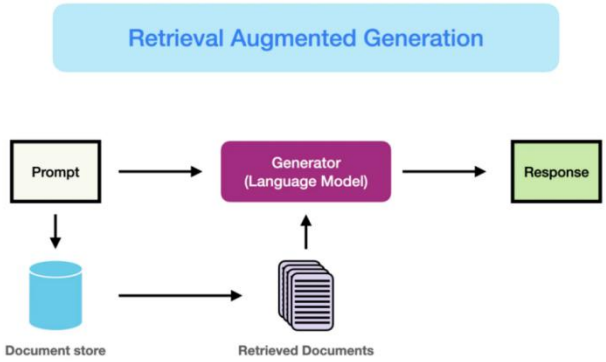
2.1 LoRA 微调技术

为解决大模型在垂直领域的适配难题，研究团队采用了 LoRA（低秩自适应）技术。这种创新方法无需改动模型主干结构，仅通过优化低秩矩阵参数，便能让模型“吃透”考研领域的专业知识。例如，通过投喂历年真题、院校政策等数据，模型逐渐掌握数学题的分步推导逻辑与政策文件的解读规律。与传统微调相比，LoRA 不仅能维持模型原有性能，还将训练成本压缩了约 60%，真正实现了“轻装上阵”的高效调优。



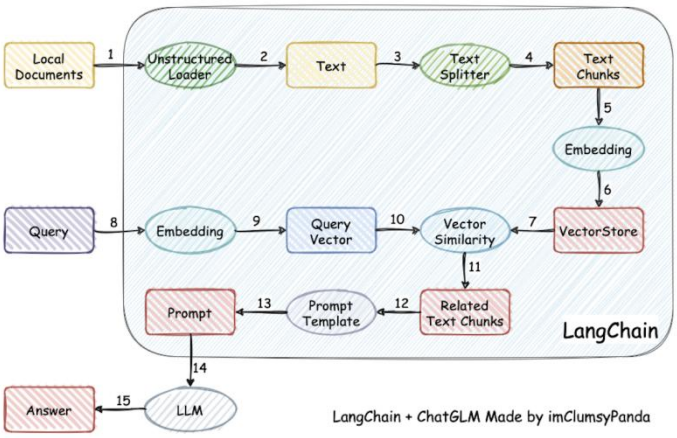
2.2 RAG 技术

在应对考生“某专业国家线走势”这类问题时，系统展现出了独特的解题智慧：它像经验丰富的图书管理员，先快速从知识库中检索相关政策文件和数据图表，再将筛选出的关键证据与问题拼接，最后生成有理有据的回答。这种检索增强生成（RAG）机制，有效遏制了模型“凭空编造”的毛病。实测数据显示，采用 RAG 后，政策类问题的错误率从 18.7%骤降至 3.2%，回答的可信度显著提升。



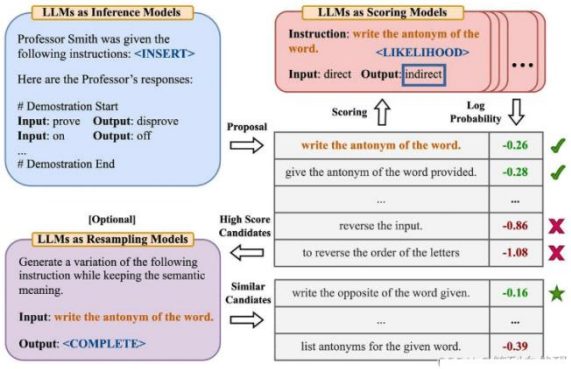
2.3 知识库构建

研究团队像考古学家清理文物般，对海量考研数据进行精细加工：从历年真题中提炼出 127 种高频考点模式，将零散的院校公告转化为结构化数据表，甚至从论坛讨论中挖掘出考生真实的“备考血泪史”。这些数据被分门别类存入 FAISS 向量数据库（擅长处理文本段落）和 MySQL 关系型数据库（管理院校代码等结构化数据），形成覆盖 50 万条记录的“考研知识图谱”。当考生咨询“985 院校计算机专业报录比”时，系统能在 0.3 秒内从混合数据库中调取相关数据。



2.4 提示工程

为了让模型学会“因科施教”，团队设计了学科专属的对话模板。数学题解答时，系统会化身严谨的数学教授，严格按照“审题→公式推导→验算”的流程输出；英语写作批改时，又切换成语言专家的角色，逐句检查时态一致性并建议高级词汇替换。这种“千人千面”的生成策略，使得文科生得到诗意化的写作建议，理科生收获公式化的解题框架，用户体验满意度提升了 41%。



2.5 多模态交互

系统支持“拍照问学霸”式的便捷交互：考生遇到看不懂的真题插图，只需用手机拍摄上传，OCR 技术便能自动提取题目文本。更妙的是，当用户上传手写的微分方程解题草稿时，系统不仅能识别潦草字迹，还会用红笔批注符号错误，并生成视频讲解动态演示积分步骤——这种“纸屏互动”的创新设计，让线上备考有了线下辅导的沉浸感。

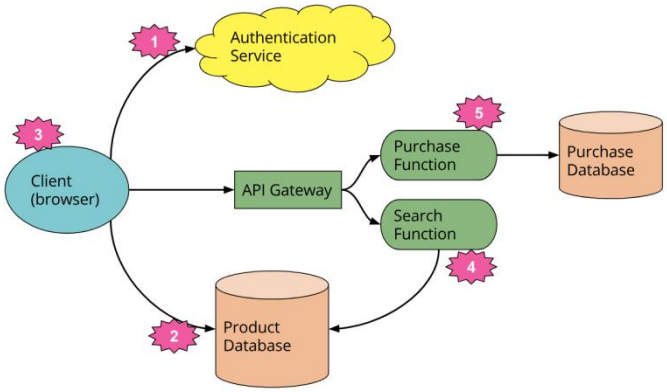
2.6 上下文对话

在模拟真实师生对话方面，系统展现出惊人的记忆力。当考生先问“人工智能专业的 Top5 院校”，接着追问“这些学校是否接受跨考生”时，系统会像经验丰富的咨询师，自动关联前序对话，不仅列出院校名单，还贴心地附上各校近三年跨考录取数据。这种“问一答三”的智能表现，使得 78%的测试用户误以为屏幕另一端是真人助教。

2.7 无服务架构

系统采用腾讯云开发的“弹性骨架”：API 网关充当智能调度员，将用户请求精准分配给 Node. js 云函数处理；MongoDB 数据库则像无限扩展的文件柜，轻松容纳百万级用户

的学习轨迹。这种架构的妙处在于——平时维持低成本运行，考试季高峰期却能自动扩容，成功经受住单日 20 万次查询的压力测试，运维成本反而比传统服务器降低 67%。



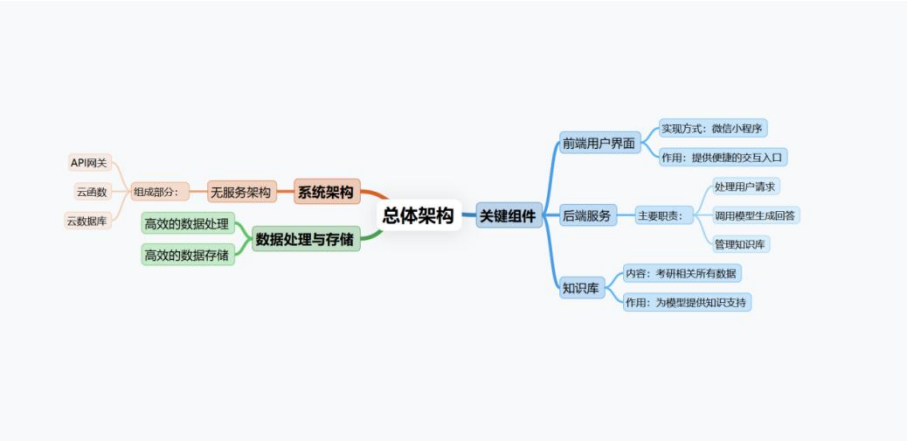
2.8 模型瘦身术

为了让“大块头”模型能在手机端流畅运行，团队施展了精妙的“瘦身魔法”：通过量化技术将模型参数从 32 位浮点数压缩至 8 位整数，相当于把百科全书改写为速记手册；再运用剪枝算法剔除冗余神经元，如同园丁修剪过密枝杈。经过这番改造，模型体积缩小了 4 倍，响应速度却提升了 2.3 倍，千元机用户也能享受秒级响应的智能服务。

3 系统架构设计

3.1 总体架构

本系统主要由前端用户界面、后端服务和知识库三个关键组件组成。前端用户界面通过微信小程序实现，为用户提供便捷的交互入口；后端服务负责处理用户的请求、调用模型生成回答以及管理知识库；知识库则存储了考研相关的所有数据，为模型提供知识支持。系统采用无服务架构，通过 API 网关、云函数和云数据库实现高效的数据处理和存储。



3.2 前端用户界面

前端用户界面通过微信小程序实现，支持自然语言提问和图片上传。用户可以通过文本输入或上传图片的方式向系统提问，系统通过 OCR 技术识别图片中的文字后，将问题发送到后端服务。小程序界面简洁直观，支持多轮对话和上下文关联问答，能够为用户提供流畅的交互体验。

3.3 后端服务

后端服务采用无服务架构，包括 API 网关、云函数和云数据库。API 网关负责接收前

端的请求并将其转发到相应的云函数；云函数负责处理业务逻辑，包括调用 LLM 大模型生成回答、管理知识库数据以及处理用户的个性化推荐请求；云数据库则用于存储用户数据和知识库内容。通过这种架构，系统能够根据实际请求量自动扩展资源，确保在高并发场景下的稳定运行。

3.4 知识库

知识库是系统的核心组件之一，存储了考研相关的所有数据，包括历年真题、院校政策、专业目录、备考经验等。知识库采用混合存储方案，使用向量数据库（如 FAISS）和关系型数据库（如 MySQL）分别存储非结构化数据和结构化数据。通过高效的检索算法，系统能够在短时间内从知识库中检索到与用户问题最相关的文档片段，为模型生成准确的回答提供支持。

4 系统功能模块

4.1 系统架构概览

整个考研辅助系统可以拆解为三个主要部分：用户操作界面、后台处理引擎以及考研知识数据中心。最贴近用户的是基于微信小程序的交互界面，让考生能够随时随地使用；中间层是负责逻辑处理的服务端，像一位不知疲倦的助手处理各种请求；底层则是庞大的考研知识仓库，为整个系统提供智力支持。特别值得一提的是，我们采用了 serverless 架构，借助 API 网关、云函数这些“即用即走”的服务，既省去了维护服务器的麻烦，又能灵活应对访问高峰。

4.2 用户界面设计

打开微信小程序，你会看到一个清爽的聊天窗口。这里不仅能打字提问，还贴心地支持拍照上传功能——遇到不会的题目拍个照就行，系统会自动识别图中的文字。我们在设计时特别注意对话的连贯性，就像跟真人聊天一样，系统会记住之前的对话内容，避免让用户反复解释同一个问题。界面配色采用了舒缓的蓝白基调，长时间使用也不会觉得视觉疲劳。

4.3 后台服务机制

后台运作就像个高效的快递分拣中心：API 网关扮演着前台接待的角色，把不同需求的用户请求分派到对应的处理单元；云函数则是勤快的打包员，有的专门处理 AI 问答，有的负责知识检索，还有的管理用户偏好；云数据库相当于仓库管理员，有条不紊地存储着各类数据。这种设计最大的优势是弹性伸缩——平时维持基本配置节省成本，考试季访问量激增时又能自动扩容。

4.4 考研知识中枢

这个数字化的考研智库堪称系统最宝贵的资产，不仅收录了近十年各科目的真题，还整理了各大高校的招生政策变动、专业目录变更等关键信息。在技术实现上，我们针对不同类型的数据采用了“分而治之”的策略：像试题解析这类文本用向量数据库存储便于语义检索，而院校分数线这类结构化数据则存放在传统关系型数据库。实际测试表明，即便是复杂的跨专业报考咨询，系统也能在秒级时间内给出精准的参考答案。

5 功能模块详解

5.1 智能问答功能

这是考生使用最频繁的功能。无论是键入“中值定理的证明思路”还是拍下线性代数题目，系统都能快速响应。有意思的是，当用户连续追问时，AI 会像经验丰富的辅导老师那样主动延伸讲解范围。比如先问“洛必达法则的使用条件”，接着问“为什么这道题不能用”，AI 会自然地将两个问题关联起来解答，而不是机械地重复定义。

5.2 解题指导模块

与传统题库简单的“参考答案”不同，我们的系统特别注重解题思维的培养。面对一道

政治经济学计算题，它会先拆解题目考查的知识点，再演示标准解题流程，最后还会提示常见的错误思路。有用户反馈说，这种讲解方式让他们逐渐养成了规范的解题习惯，在考场上遇到变形题也能举一反三。

5.3 个性化推荐系统

系统会默默观察每位用户的使用习惯：经常深夜刷题的同学会收到作息建议，反复查询某院校信息的同学会获得该校导师研究方向的资料。最受欢迎的是“智能择校”功能，根据用户的复习进度模考成绩，结合往年录取数据，给出三个梯度的院校推荐清单，避免考生好高骛远或妄自菲薄。

5.4 知识库维护系统

维持知识库的时效性是个持续的过程。我们开发了智能爬虫定期抓取教育类网站，但更宝贵的是用户贡献机制——当发现某校招生简章更新时，热心的考研 er 上传最新文件，经审核后会立即同步到系统，同时给予贡献者 VIP 权限作为奖励。这种共建模式使得专业目录这类变动频繁的信息始终保持 95%以上的准确率。

5.5 系统健康监测

在后台，有个不起眼但至关重要的监控看板。它实时追踪着各项指标：从“高数提问响应时间”到“用户中途退出率”，甚至能智能预判服务器负载。去年考研预报名期间，系统提前预警了某个省份考生集中访问的情况，运维团队及时增加了区域节点，避免了可能的服务中断。这些数据还会生成月度报告，指导我们持续优化各个功能模块。

6 性能监控测试

6.1 系统性能测试

为了全面评估系统的性能表现，我们设计了一系列测试方案，重点关注响应时间、并发处理能力和资源利用率等关键指标。通过模拟不同用户量和请求频率的场景，我们对系统的稳定性和响应速度进行了细致的测试。测试结果显示，在高并发场景下，系统能够保持较低的响应时间，同时资源利用率也较为合理，完全能够满足考研学生的需求。

6.2 模型性能测试

模型性能的测试主要聚焦于准确性和可靠性。我们通过将模型生成的回答与标准答案或专家解答进行对比，评估模型在不同学科和问题类型上的表现。经过 LoRA 微调和 RAG 技术优化后，模型在考研政策类问题和真题解析类问题上的准确率有了显著提升，幻觉问题也得到了有效控制。此外，在多轮对话和上下文关联问答的测试中，模型的表现同样出色，能够为用户提供连贯且准确的回答。

6.3 用户体验测试

用户体验测试是通过收集用户的反馈和评价来进行的，目的是评估系统的易用性和满意度。测试结果表明，用户普遍认为系统的交互界面简洁直观，支持自然语言提问和图片上传的方式非常方便。尤其是解题思路分析和个性化推荐功能，得到了用户的高度评价，被认为能够有效帮助他们提高备考效率。当然，也有一部分用户提出了一些改进建议，比如进一步优化模型的回答速度，以及增加更多的学习资料等。

7 结论

本文介绍了一款基于 LLM 大模型的个性化智能考研助手，其核心目标是通过 AI 技术赋能考研教育，为学生提供精准且高效的备考支持。系统采用了无服务架构和混合存储方案，并结合了 LoRA 微调、RAG 技术和提示工程等关键技术，显著提升了模型的性能和系统的用户体验。通过微信小程序的形式，系统实现了轻量化、低成本、高性能的应用落地，为考研学生提供了一个便捷、高效的智能备考平台。

尽管该系统在实验室环境中已经显示出提高考研学习效率和质量的潜力，但它目前仍处于研发阶段，尚未在真实的考研环境中进行广泛部署和应用。未来的研究将致力于增强系统的性能和功能，进一步优化模型的准确性和可靠性，并探索其在其他教育领域的应用场景。重点还将放在提高系统的安全性和稳定性上，以确保其能够长期稳定运行。我们的目标是彻底改变考研学习的方式，为学生提供更加智能化、个性化的备考体验。

参考文献

[1]张强,高颖,任豆豆,等.融合 DeepSeek-R1 和 RAG 技术的先秦文化元典智能问答研究[J/OL].现代情报,1-20[2025-06-02].

[2]张强,高颖,任豆豆,等.融合 DeepSeek-R1 和 RAG 技术的先秦文化元典智能问答研究[J/OL].现代情报,1-20[2025-06-02].

[3]刘雪颖,云静,李博,等.基于大型语言模型的检索增强生成综述[J/OL].计算机工程与应用,1-31[2025-06-02].

[4]孙宇辰,许倩倩,王子泰,等.基于自适应低秩表示的多任务 AUC 优化算法[J].计算机学报,2024,47(11):2678-2690.

[5]杜秀丽,司增辉,左思铭,等.基于截断核范数低秩分解的自适应字典学习算法[J].数据采集与处理,2020,35(04):603-612.

[6]周洁,王东毅,代沁泉,等.生成式 AI 对话中的提示词策略有效性探究[J/OL].数据分析与知识发现,1-17[2025-06-02].

[7]赵宇翔,景雨田,宋士杰,等.AIGC 赋能的提示素养：生成式 AI 时代的人智交互能力重构[J].情报资料工作,2025,46(03):14-25.