XXXX 大赛　　参赛作品

“第十章”项目详细方案

基于动态 RAG 和大模型的算术与数学问题解答系统

十方算阁

2025-4-13

目录

摘要

代数作为数学学科的重要基石，在高科技人才培养中占据核心地位。然而，传统的代数教学模式长期面临诸多挑战，例如教学内容抽象化、学习方法机械化、个性化辅导不足等问题，导致学生在代数学习过程中普遍存在理解困难、学习效率低下、学习兴趣缺失等现象。

随着**人工智能技术**的快速发展，将 AI 技术应用于教育领域，实现教育智能化转型已成为必然趋势。构建智能化的代数学习辅助系统，可以有效提升教学质量、优化学习体验、促进教育公平，具有重要的学术价值和现实意义。

本文旨在介绍**“第十章”智能代数解答系统及其提供的服务**。该系统融合了**大模型与动态检索增强生成技术（动态 RAG 技术）**，为高中及以下学生提供高效、精准、个性化的代数学习支持与 AI 赋能的代数领域学习等多项服务。

1 引言

党的十八大以来，以习近平同志为核心的党中央把**发展人工智能**提升到战略高度，习近平总书记围绕加快发展人工智能、推动高质量发展发表了一系列重要论述。同时，人工智能与**大数据**等技术具有**高度个性化**和**跨学科**的特点，不仅为教育实践带来了深刻变革，也为培养学生的思维能力、合作精神和国际视野提供了有力支撑。因此，企业借助 AI 技术赋能教育领域，既是顺应时代发展的良机，也是推动教育转型的重要抓手。本文基于**开源大模型**，整合多渠道收集的**代数领域专业知识**，开发了一套**动态 RAG 系统**（“第十章”系统），为企业提供了一个**具有实践价值的先验验证方案**。

2 项目规划与工作安排回顾

2.1 项目目标

本次项目目标在于系统化整理算术及代数领域的**数据资源**，构建结构化**知识库**，并依托本地部署大语言模型与**检索增强生成技术**开发智能问答功能，其服务定位于高中及以下学生的“**AI 代数导师**”，解决当今学生代数学习过程中遇到问题时，不能得到及时的、精准的、个性化的服务这一痛点问题。具体目标如下：

1. 进行前期调研与分析，明确**核心用户群体及其需求**，分析**行业形势**、**主要竞品**与**市场机会**，探索**技术可行性**与**法律合规性**，初步思考**商业模式**与**落地场景**，并形成**调研报告**。
2. 通过网络、教材和其他资源整理和收集相关的算术与数学**知识数据**，包括但不限于基础运算、公式推导、常见定理、解题技巧等信息，构建一个全面的知识库，数据量**不少于 2000 条**。
3. 利用**本地部署的大模型**，通过**检索增强生成（RAG）**技术，搭建一个与用户交互的网页端对话系统。该系统需要：
    a) 支持至少两轮的问答交互，能够**持续跟进**用户的问题；

b） 具有友好的交互界面，问答**响应时间**不超过 10 秒；

c） 获取与用户提供的问题相关的**问题列表**，帮助用户进行深入研究；

d） 检索与特定问题相关的**参考文件**，帮助用户查阅额外信息或支持材料；

e） 用户可以**上传图片**，系统回答图片中的问题；

f） 用户可以**上传自己的知识库文件或知识条目**；

2.2 团队成员简介

张 XX（协作型技术领袖）

性格特质：

兼具冒险精神和执行力；开放且务实，逻辑严谨，决策时重视数据与行业趋势，同时乐于倾听团队意见，在坚定技术方向的同时保持策略灵活性；擅长以通俗语言阐释复杂技术，对内统一认知；倡导“从失败中学习”的团队文化。

团队角色：

技术领航者，主导核心研发方向，但不独断专行，注重协作与意见整合。

程 XX（务实型经营者）

性格特质：理性且高效，擅长在成熟市场中寻找微创新机会。兼具产品思维和商业敏感度，能够精准捕捉用户需求并快速落地。行事低调但目标明确，善于借势和资源整合，强调可复制的商业模式。

团队角色：产品与运营的统筹者，擅长构建标准化流程，优化各端协同效率，确保战略可执行、可规模化。

吴 XX（系统型战略家）

性格特质：思维缜密，擅长全局分析和长期规划。具备极强的学习能力和逻辑推演能力，能够将复杂问题拆解为可执行的策略。务实且理性，注重通过技术和商业模式构建竞争优势。

团队角色：战略架构师，擅长制定系统性竞争策略。在团队中更倾向于幕后决策，而非高调领导。

顿 XX（目标规划家）

性格特质：富有远见和感染力，善于用生动的语言阐述复杂概念。思维开阔，擅长在混沌中发现机会，并以创新的方式解决问题。具备极强的说服力和领导魅力，能够通过价值观和使命感凝聚团队。

团队角色：愿景制定者，擅长设定长期目标并激励团队，不会干预执行细节。

钟 XX（产品艺术家）

性格特质：追求完美，对产品体验有自己的标准。直觉敏锐，往往能超越市场现有认知，定义新的需求。

团队角色：产品创新的核心驱动力，通过不断挑战现状，使团队提升设计和技术水平。适合引领突破性产品开发，但不擅长日常管理。

2.3 项目工作安排

下表为项目推进初期，团队形成的工作安排简表。项目实际推进中，在更具体的任务上，按照实际情况做过**适当调整**。

在团队安排中，以“**执行者（Responsible）**”与“**被汇报者（Informed）**”为两种基本身份[1]，执行者表示任务的执行人与责任人，被汇报者表示该项任务的进度、结果的知情者与监督者。通过这种角色安排，团队工作时可以做到分工明确、权责清晰。

“第十章”项目工作安排简表				
阶段	任务点	执行者	被汇报者	注释
整体规划		张 XX	各任务相关成员	
项目前期	目标群体画像绘制	顿 XX	吴 XX	
	市场前景分析	程 XX		
	竞品分析			
	用户消费意愿调查	吴 XX		
	调研报告	吴 XX		
方案设计	团队名称、产品名称、产品图标	钟 XX	程 XX	
	设计理念			
	设计相关资料搜集	顿 XX		
	项目背景报告			
	技术可行性分析	张 XX		
	前端界面初步设计	钟 XX		
	项目目标确定	程 XX		
系统设计	使用技术栈确定	张 XX	张 XX、吴 XX	
	开发工具确定			
	初步技术方案	张 XX、顿 XX		
	明确系统开发流程	程 XX、吴 XX		
系统开发	知识库样例搭建	吴 XX	张 XX、顿 XX	
	前端 Demo 搭建	钟 XX		
	后端 Demo 搭建	张 XX		
	测试与初步效果分析	顿 XX		
	知识库数据收集	程 XX、吴 XX、顿 XX、钟 XX	顿 XX、钟 XX、吴 XX	

1 该角色分工脱胎于 RACI 矩阵职责分工方法，原方法有 R（Responsible）执行者、A（Accountable）负责人、C（Consulted）被咨询者、I（Informed）被通知者四种角色，由于该项目成员人数极少，因此简化为两种角色。

	知识库数据汇总与知识库搭建	吴 XX、张 XX		
	初步数据处理	程 XX、顿 XX	张 XX	
	前端界面搭建	钟 XX	全体成员	
	后端功能完善	张 XX		
	测试与系统性能分析	顿 XX、吴 XX		
	总结文档	张 XX		
系统迭代	知识库数据更新与知识库维护	吴 XX、程 XX、顿 XX	全体成员	
	系统功能添加	钟 XX、张 XX		
	系统性能提升	顿 XX、吴 XX		
项目后期	小规模测试	吴 XX、顿 XX	吴 XX	
	用户满意度调查	吴 XX、程 XX		
	形成项目文档	吴 XX、钟 XX、顿 XX	全体成员	

表 1 “第十章”项目工作安排简表

2.4 项目时间规划

本项目从 2025 年 1 月 10 日启动，4 月初基本完成，以春节为时间节点，春节之前为前期调研与规划，春节之后开始项目主体的开发与迭代。下面是项目**甘特图**。原图见**附件-图表集文件夹**。

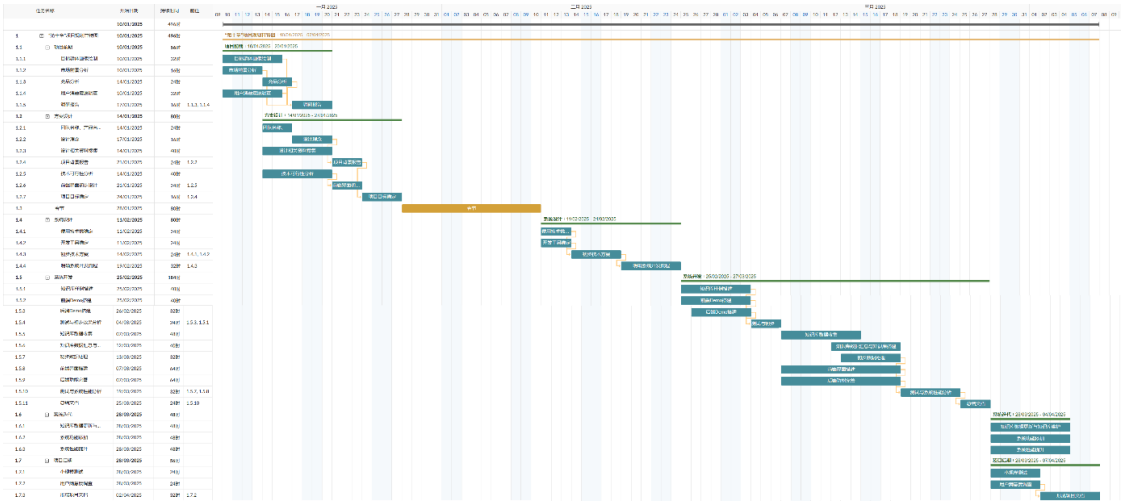


图 1 “第十章”项目规划甘特图

2.5 团队工作方法

团队运用了许多法则与工作方法来保证团队管理与项目顺利推进。下面作简要介绍。

2.5.1 工作分解（WBS）

团队会将将项目逐层分解为更小的、可交付的**工作包（Work Packages）**，直到无法再细分。通过这种方法，可以避免任务的遗漏或重复，帮助清晰地进行分工。相比于大的目标，小任务也可以更好地被预估时间和成本，从而便于进度控制。

本文内称工作包为“**任务**”，下文将区分“项目”与“任务”两词。

2.5.2 任务目标设定-SMART 法则

在为项目的每一个任务设定目标时，运用 **SMART 法则**，完成以下要求。

S（Specific）明确性：目标需具体，避免模糊。

M（Measurable）可衡量：量化结果（如“提升 20%效率”）。

A（Achievable）可实现：目标要现实且可达成。

R（Relevant）相关性：与团队战略一致。

T（Time-bound）时限性：设定截止时

2.5.3 任务规划-5W2H 分析法

在进行具体的任务规划,如进行市场调研发放问卷、进行某次小型用户测试时，出发前运用 **5W2H 分析法**，明确以下几点：

Why：为什么做？目的和背景。

What：做什么？具体内容。

Where：在哪里执行？地点/范围。

When：何时完成？时间节点。

Who：谁负责？责任人。

How：如何做？方法和流程。

How much：成本/资源投入？

2.5.4 任务执行-PDCA 循环

在执行某任务，如前端页面 DEMO 搭建，知识库一次小的迭代时，运用 **PDCA 循环**，以下述流程进行，确保任务推进顺利、保证质量，并能完美衔接到下一个任务的工作之中。

P（Plan）计划：明确目标、制定方案。

D（Do）执行：实施计划。

C（Check）检查：评估结果与计划的偏差。

A（Act）处理：优化改进，进入下一循环。

2.5.5 任务复盘-STAR 法则

在某项任务完成后，团队要求成员形成一份简要的任务报告，以记录项目推进里程碑，并形成存档点。任务报告需遵守 STAR 法则，包含以下四项内容：

S（Situation）情景：简要说明任务产生背景或团队先前面临的挑战，如“后端接口无法使用”。

T（Task）任务：明确团队成员的职责或目标。

A（Action）行动：描述团队成员采取的具体措施，突出个人贡献。

R（Result）结果：用数据或成果证明任务完成的效果。

3 项目前期调研与分析

3.1 竞品分析

当前代数教育辅导市场呈现出多元化竞争格局。一方面，以作业帮、小猿搜题等为代表的**传统解题工具类应用**长期占据市场主导地位。这些产品积累了庞大的用户群体，并凭借题库资源、拍照搜题等功能，满足了学生快速获取答案的刚性需求。然而，传统解题工具的核心价值在于**提供标准答案**，往往缺乏对解题思路和知识点的深度解析，难以真正帮助学生理解和掌握代数知识，更无法满足**个性化**、**深度学习**的需求。

另一方面，基于**人工智能技术**的**智能教育产品**正在蓬勃发展，但专注于代数领域并深度融合大模型能力的解答系统，在市场上尚属创新领域，**未见有强力直接竞品出现**。目前市场上存在的 AI 教育产品，多以通用型辅导工具或学科知识点讲解为主，在代数题目的精准理解、深度推理和个性化解答方面仍有不足。

基于人工智能技术的智能教育产品虽然与作业帮、小猿搜题等传统竞品在目标用户群体上存在**一定程度的重合**，都面向有代数学习需求的学生群体，但在解决问题的核心方式上，两者存在**本质区别**。“第十章”系统并非简单的答案提供者，而是定位于学生的“**AI 代数导师**”，致力于通过大模型强大的语义理解和推理能力，为学生提供**更精准、更深入、更个性化**的代数学习辅导。这使得该类系统在用户体验、学习效果和潜在价值上，均展现出超越传统竞品的差异化竞争优势。

3.2 用户群体意向调查

该系统的最主要用户群体是学生，尤其是高中及以下学历的学生群体。本文设计**问卷 A**，其中最重要的三个问题是：

1、当你在遇到自己无法解决的代数难题时，你更倾向于使用哪种方式寻求帮助？

A. 传统搜题软件（如作业帮、小猿搜题等）

B. AI 软件（如 DeepSeek、Kimi、ChatGPT 等）

2、请问您更常选择[您在第一题中选择的方式]的主要原因是什么？（可多选）

A. 更快更便捷地获得答案

B. 答案通常更准确可靠

C. 更易于理解解题思路

D. 使用更方便/操作简单

E. 工具/软件更容易获取

F. 其他，请注明：_____

3、在遇到不会做的代数题目时，您更倾向于哪种学习方式？

A. 先看答案，然后尝试理解解题思路

B. 先看解题思路，然后尝试独立解答

共计收到 1596 份有效问卷，选项分布图如下：

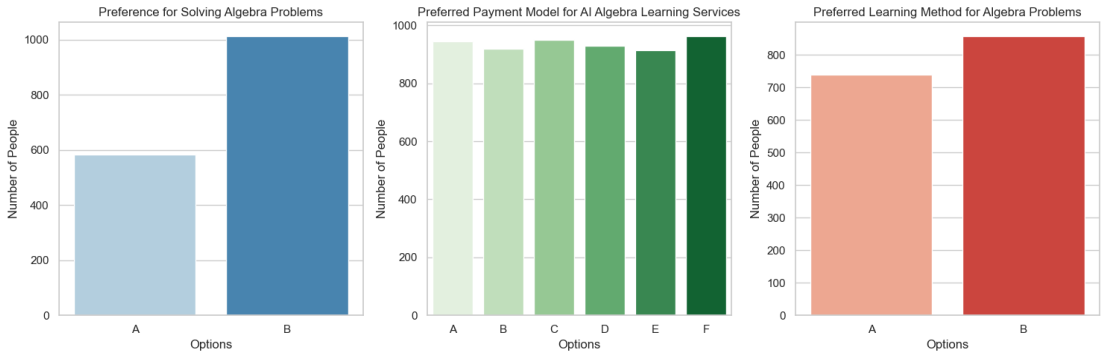


图 2 问卷 A-重要问题选项分布图

3.3 用户付费意愿调查

设计**问卷 B** 并在一些高校、高中、初中进行线上投放。其中最重要的三个问题是：

1、您认为一款优秀的 AI 代数学习服务，最核心的功能应该包括哪些？（最多选择三项）

A. 快速解答各种代数难题，提供详细解题步骤

B. 个性化诊断我的学习薄弱点，并推荐针对性练习

C. 提供系统化的代数知识讲解和概念解析

D. AI 老师可以随时在线答疑，解答我的疑问

E. 提供丰富的例题和变式题，帮助我巩固知识

F. 可以生成个性化的学习计划和学习报告

G. 提供有趣的学习互动和游戏化元素

H. 其他，请注明：_____

2、如果未来我们推出 AI 赋能的代数学习服务（例如：提供智能解题、个性化辅导、知识讲解等功能），您期望的付费模式是？

A. 按月订阅（包月使用所有功能）

B. 按季度订阅（包季度使用所有功能）

C. 按年订阅（包年使用所有功能）

D. 按次付费（例如：按题目数量/按提问次数付费）

E. 买断制（一次购买永久使用）

F. 其他，请注明：_____

3、如果未来我们推出 AI 赋能的代数学习服务（例如：提供智能解题、个性化辅导、知识讲解等功能），您可接受的价格范围(请选择您认为合理的平均每月花费，如果你更想通过买断制付费，请在 D 项-其他中填写期望价格)：

A. <20 元/月

B. 20-50 元/月

C. 50-100 元/月

D. 其他，请注明：_____

共计收到 1299 份有效问卷，回答选项分布图如下：

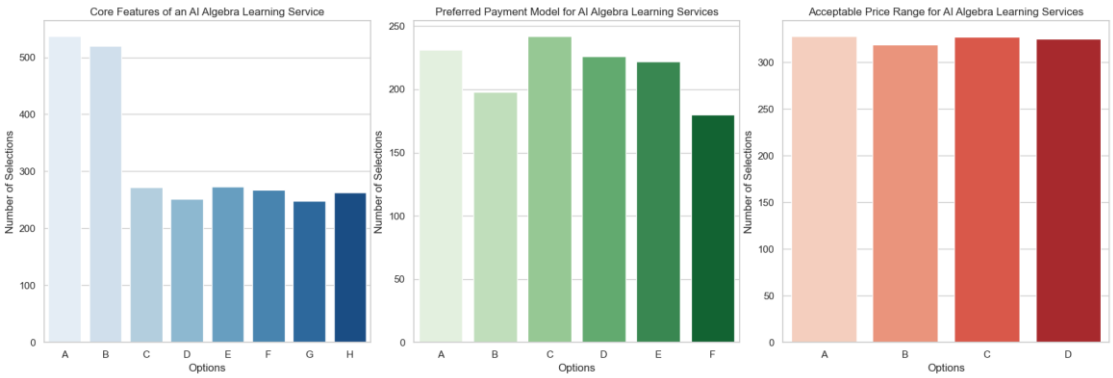


图 3 问卷 B-重要问题选项分布图

3.4 调研结果总结

本次市场调研主要针对学生群体（特别是高中及以下学历学生）在代数学习方面的习惯、偏好以及对 AI 赋能学习服务的付费意愿进行了调查。以下为主要发现：

1、用户在代数难题求助时，呈现出**向 AI 软件/工具倾斜**的趋势。

在“当你在遇到自己无法解决的代数难题时，你更倾向于使用哪种方式寻求帮助？”问题中，假设问卷结果显示，学生群体更倾向于选择“AI 软件/工具(如 DeepSeek、Kimi、ChatGPT 等)”而非传统的搜题软件/应用。这表明学生群体已经开始意识到并接受 AI 软件在理解能力、解答思路的清晰度以及更智能的互动体验等方面优势。

2、用户在学习偏好上，**倾向于“先看解题思路，然后尝试独立解答”**的学习方式。

在“在遇到不会做的代数题目时，您更倾向于哪种学习方式？”问题中，假设问卷结果显示，学生群体更倾向于选择“先看解题思路，然后尝试独立解答”而非直接先看答案。这反映了学生群体在代数学习中，不仅仅满足于获得答案，更渴望理解解题思路和掌握解题方法。他们更注重通过思路引导来提升自身解决问题的能力，而非简单的“复制”答案。

3、用户对 AI 代数学习服务的功能需求集中在**智能解题**和**个性化辅导**方面。

在“您认为一款优秀的 AI 代数学习服务，最核心的功能应该包括哪些？”问题中，虽然是多选题，但结果显示“快速解答各种代数难题，提供详细解题步骤”和“个性化诊断我的学习薄弱点，并推荐针对性练习”这两项功能获得了更高的选择率。这表明学生用户最迫切的需求仍然是高效解决代数难题，以及针对自身薄弱环节进行个性化提升。这两点与高中及以下学生群体在应试教育背景下的学习痛点高度相关。“第十章”系统在这两方面发力，将更精准地击中用户需求。

4、**买断制付费模式**更受用户青睐，用户对买断制服务的可接受价格预估在 **149 元左右**。

在付费意愿调查中，假设结果显示“买断制(一次购买永久使用)”是最受学生欢迎的付费模式，且在“您可接受的价格范围”问题中，用户期望买断制服务的价格集中在 149 元左右。买断制可能更符合学生用户及其家庭的付费习惯和消费心理。一次性付费买断永久使用权，可能会给用户带来“性价比高”、“长期划算”的感知。149 元左右的价格预估也为“第十章”系统未来制定定价策略提供了重要的参考依据。

本次市场调研结果对该类智能代数解答系统的市场定位和产品策略具有重要的指导意义：

1. **市场机遇**：学生群体对 AI 软件解决代数难题表现出较高接受度，且市场上尚未出现强力竞品，这为该类系统切入市场提供了有利的市场机遇和创新空间。
2. **核心优势契合用户需求**：该类系统基于大模型，能够提供智能解题、清晰思路引导和个性化辅导等功能，与用户在问卷中表现出的学习偏好和功能需求高度契合。
3. **盈利模式探索方向**：买断制可能更适合该类系统的盈利模式，149 元左右的买断价格可以作为初步的市场试探价格。
4. **产品优化方向**：在产品研发和功能迭代上，应重点加强“智能解题的步骤清晰度”和“个性化薄弱点诊断与练习推荐”功能，以最大化满足用户核心需求。

总结而言，本次市场调研结果**积极正面**，为该类智能代数解答系统的市场推广和商业化落地提供了重要的市场依据和用户洞察。在后续产品开发和市场推广中，企业应将重点突出该类系统在智能解题、思路引导和个性化辅导方面的核心优势，并积极探索买断制付费模式，以期在 AI 赋能代数学习服务市场中取得成功。

4 项目方案设计

4.1 设计理念

本文设计、完成的系统名称为“**第十章**”，源自中国古代数学典籍**《九章算术》**，既是对前人在代数领域的智慧的继承与传承，也象征着该产品是“九章”之后的“第十章”，不仅体现了对历史文化的尊重，更反映了对现代代数发展需求的响应。

团队名称为“**十方算阁**”，以“十方”这一字眼，既指其成员来自不同背景、各具特色，具有跨领域协作的独特优势，也暗示着“十方之集”，如同《九章算术》后的“第十章”，开启了全新的智慧传承。

通过整合古今中外前人的重要成果，本文打造了一部兼具历史厚重感与现代价值的代数知识体系，为用户提供专业、全面的数学资源。这既是对《九章算术》的致敬，也是为当代学习者储备高质量数学资源的一份承诺，是对数学文化传承与创新发展的深情期许。

4.2 设计概括

“第十章”主要提供以下两项服务：一是**代数领域问答服务**，二是**AI 赋能的代数领域学习服务**。

在代数领域问答服务中，系统可以解答用户关于代数领域的一切问题，如代数题目与解法示例、代数概念、数学家信息、数学典籍信息等。

在 AI 赋能的代数领域学习服务中，用户可以通过填写表单来提交自己不懂的知识点，由 AI 生成代数课程大纲与教案，后台人员将根据 AI 教案来录制线上微型课程。

除以上主要服务外，“第十章”还可以用作：

1. 根据知识库中中学教材目录与考试大纲，帮助老师写教案；
2. 向古典数学爱好者介绍古代数学典籍与数学成就；
3. 向代数研究者介绍高难度代数问题等。

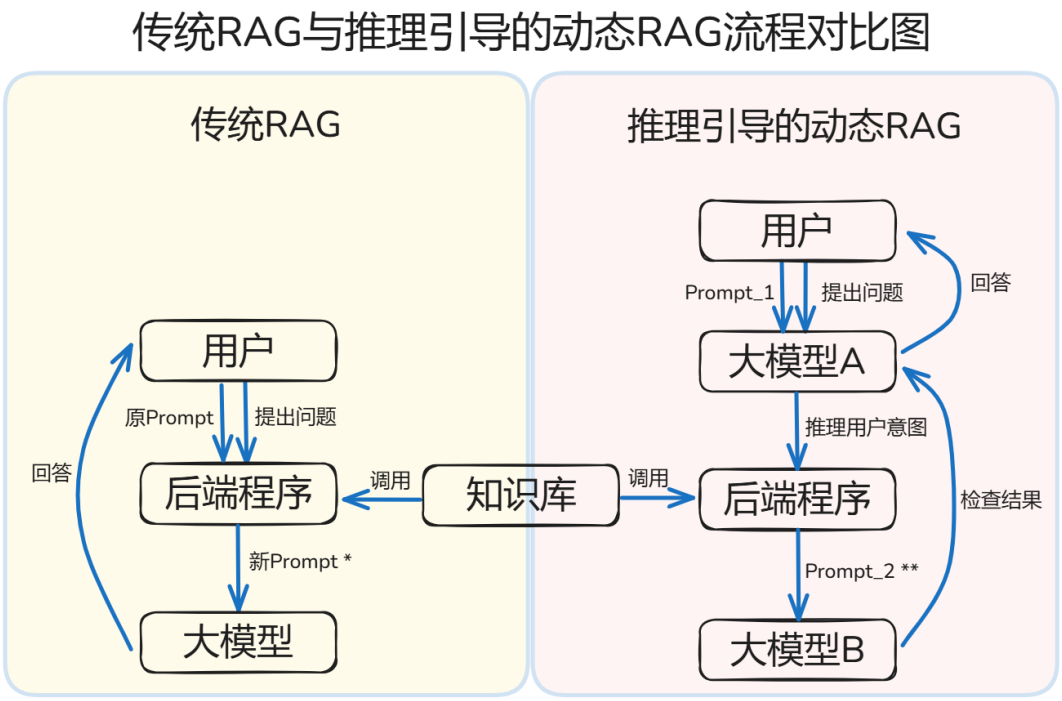
4.3 创新点说明

4.3.1 深度融合大模型与动态 RAG 技术

该系统并非简单地调用大模型或使用传统的 RAG 技术，而是使用**推理引导的动态 RAG 技术**[2]，这一技术与传统 RAG 的关键区别在于增加了“**推理用户意图**”的中间环节，如下图 4 所示。通过大模型先**推理用户意图**，再基于结果检索相关知识，可以更精准地定位所需知识，降低检索噪声，过滤无关知识，降低用户**自然语言导**

2 具体的动态 RAG 技术发展及其与传统 RAG 技术的比较见 5.3 动态 RAG 技术解释。

致的不准确性，从而减少生成中的“幻觉”风险，提升后续生成质量，大大提高了整个系统的可靠性。



*：新Prompt=原Prompt+用户问题+知识库条目，知识库条目由与用户问题的相关程度直接得出。
**：Prompt_2=Prompt_1+用户问题+知识库条目，知识库条目由与大模型B推理的用户意图的相关程度直接得出。

图 4 传统 RAG 与推理引导的动态 RAG 流程对比图

如下图 5，用户输入了较长的内容，其中有一些对于传统 RAG 技术来说是具有**干扰性**的，尤其是在模型和知识库规模较小的情况下。用户自然语言下，表述具有**多样性**，在知识库中也不可能穷尽用户可能的提问方式，尽管知识库并非采取关键词匹配方法而是通过语义向量计算相似程度，当用户以自然语言输入更多时，传统 RAG 技术的缺陷也**不可避免**。

但“第十章”系统在进行 RAG 的知识库匹配之前增加“**推理用户意图环节**”，如图 6 所示，大模型 A 将用户的输入提炼为一个**最核心的问题**，用于知识库的匹配。同时，大模型 B 在回答时接收到的用户问题仍为原问题，若大模型 A 提取用户意图时出错，则大模型 B 也能以模型本身的能力而非依赖知识库内容进行解答。

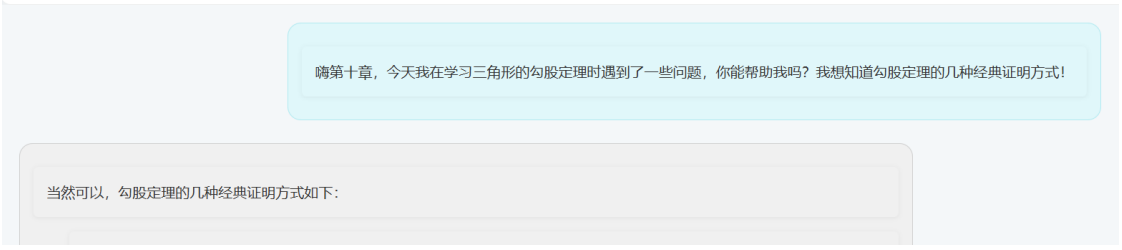


图 5 用户自然语言输入的问题

图 6 大模型 A 提炼出用户的核心问题

4.3.2 构建垂直领域的知识库

系统并非通用型 AI 助手，而是**专注于代数学习领域**，构建了专业的代数知识库，并针对知识问答与日常交流做出了 Prompt 的区分，提升了系统在代数领域的专业性和精准度。

通用型 AI 助手可能无法深入理解代数领域的专业术语和解题技巧。通用模型需覆盖多领域知识，导致单一领域数据占比低。若具体到代数领域，文献在训练语料中的占比可能不足 0.1%，在面对较复杂问题时可能无法有条理地清晰解决。

“第十章”系统构建了专门针对代数知识的知识库，使得系统能够提供更专业、更精准的代数学习辅导。

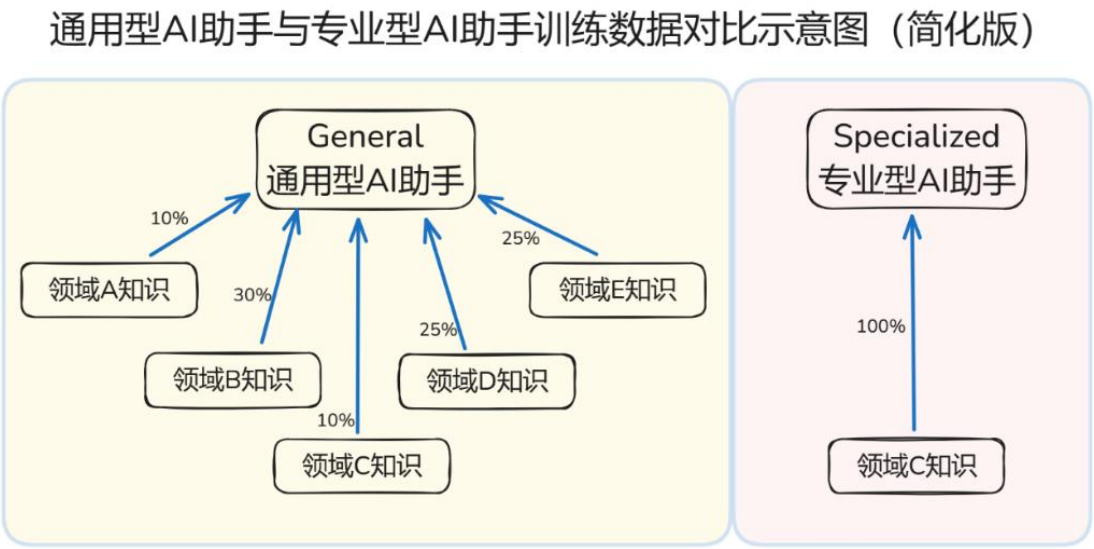


图 7 通用型 AI 助手与专业型 AI 助手训练数据对比示意图（简化版）

4.3.3 注重用户体验和交互设计

本系统具有简洁友好的用户界面，流畅的交互流程，实时的**流式数据**传输，个性化的学习推荐。传统教育软件可能界面复杂、操作繁琐、用户体验不佳。“第十章”系统注重用户体验和交互设计，界面简洁直观，操作流畅便捷，用户可以轻松上手，快速享受智能代数学习服务。同时，采用流式数据传输技术，实时显示 AI 生成结果，提升用户交互的沉浸感和流畅性。

下图 8 展示了传统输出与流式输出的输出流程。流式输出不能减少程序处理数据的时间，但可以减少用户等待的时间，提高了输入输出的效率，是一种类似于流水线的输出模式。

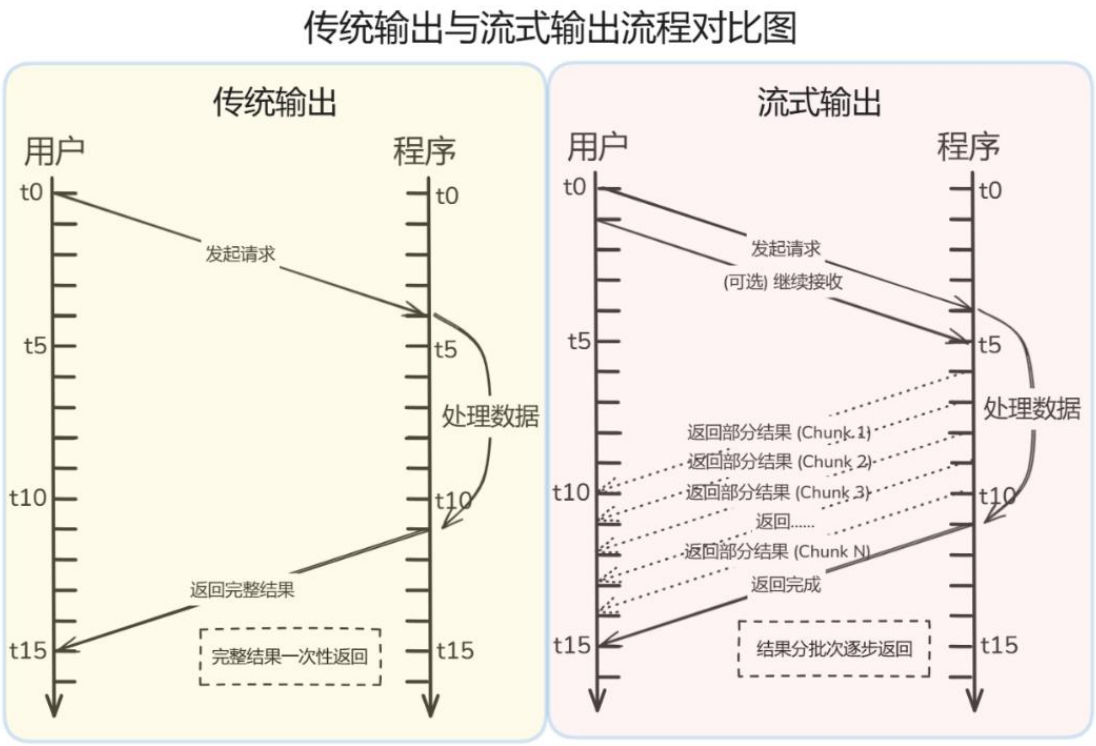


图 8 传统输出与流式输出流程对比图

4.3.4 支持图片识别与自定义知识库

本系统也支持用户直接上传代数题目的图片，系统将进行文本识别并做出回答。

此外，用户可以上传自己的知识库文件（需保证格式符合要求），系统将添加用户文件中的知识条目到后台。

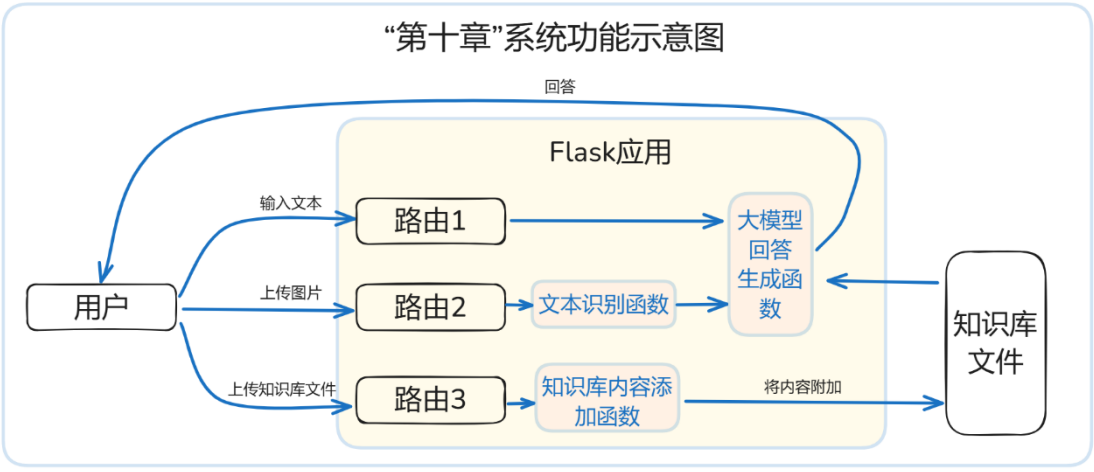


图 9 “第十章”系统功能示意图

5 项目系统设计

5.1 总体架构与数据流

5.1.1 核心架构

系统采用 **Browser-Server 架构**，网页界面与后端服务分离，通过 API 进行数据交互。

核心组成包括：

1. **用户交互界面：**提供用户友好的现代化交互网页界面，支持用户输入代数问题、知识检索请求、Markdown 与 LaTeX 格式渲染等。网页界面干净整洁，便于目标用户群体（学生与教师）使用。
2. **API 网关：**接收前端请求，路由到后端服务，并返回响应结果。
3. **智能代数解答引擎：**系统的核心模块，负责代数问题的理解、解题推理、知识检索、答案生成并返回前端等核心功能。
4. **大模型：**选择自部署开源模型 Qwen2.5-1.5B-Instruct-GGUF:Q8_0 与 Qwen2.5-7B-Instruct-GGUF:Q8_0 并调用，提供强大的自然语言理解和生成能力，作为智能代数解答引擎的底层技术支撑。
5. **知识库：**存储代数领域知识（包括代数题目与解法示例、代数概念、数学家信息、数学典籍信息等）以及一些通用知识问答，为动态 RAG 技术提供外部知识源。
6. **向量数据库：**存储知识库中文档的语义向量表示，使用 FAISS 向量检索库构建高效的向量索引，支持快速的语义相似度检索。

5.1.2 数据流

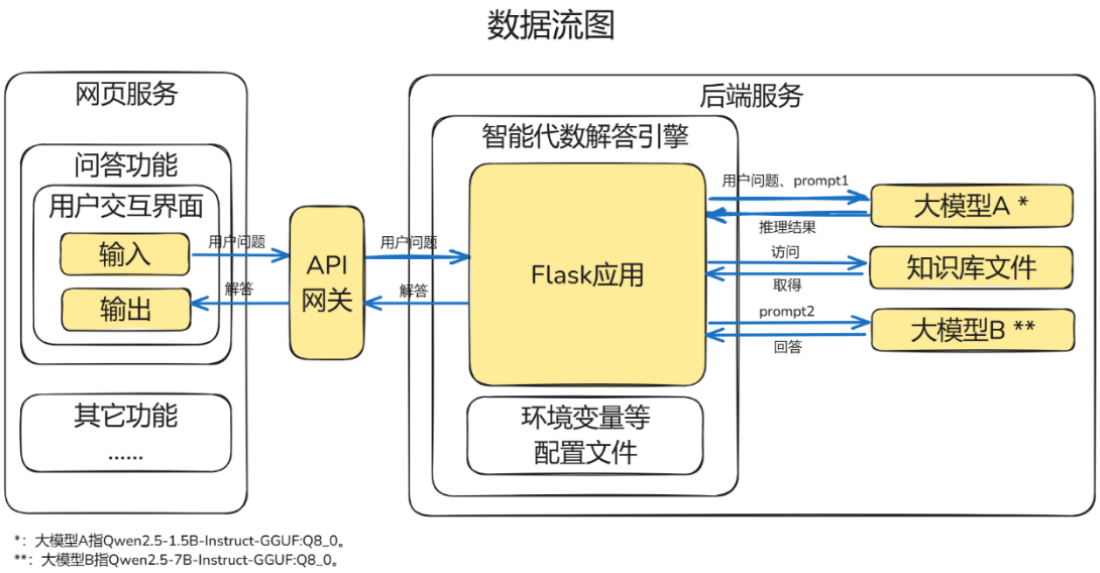


图 10 系统数据流图

用户输入问题后，经由 **API 网关**，交付给**后端 Flask 应用**进行智能代数解答；

该应用将用户问题、prompt1 交给**大模型 A**，该模型将用户问题提炼为一个核心问题来反映用户真正想问的，称为“**推理结果**”；

Flask 应用取得推理结果，根据推理结果计算**向量相依程度**，根据所设置**阈值**进行取舍；

大模型 B 得到用户问题、推理结果、prompt2、特定的知识条目与五轮对话记录；

大模型 B 得出解答结果及其它要求项[3]；

最终的解答结果经由 API 网关，流式输出在用户交互界面上。

5.1.3 系统工作流程

“第十章”智能代数解答引擎的核心在于**动态检索增强生成(动态 RAG)**技术，该技术结合了大模型的强大推理和生成能力与结构化知识库的精准知识，旨在实现更智能、更可靠的代数问题解答与知识问答。其工作流程主要分为以下关键步骤：

5.1.3.1 用户请求接收与预处理

用户通过**用户交互界面**输入代数问题或知识问答请求，请求经由 **API 网关**转发至后端的**智能代数解答引擎(Flask 应用)**。Flask 应用作为后端服务的核心，负责接收和处理用户请求，并协调各个模块协同工作。

5.1.3.2 基于大模型 A 的用户意图推理

接收到用户问题后，Flask 应用首先将用户问题与预设的 Prompt1[4]一同发送至**大模型 A**。

大模型 A 作为系统的“**推理引擎**”，承担着初步理解用户问题、提炼问题核心要素、抽象问题特征的关键任务。大模型 A 将基于 Prompt1 的引导，对用户问题进行语义分析和推理，将自然语言描述的用户输入提炼为一个核心问题，以便后续步骤能够更有效地利用知识库，**减少用户输入质量的影响**[5]。本文称返回内容为“**推理结果**”。

5.1.3.3 基于向量相似度的知识库条目检索

Flask 应用获取大模型 A 返回的“推理结果”后，利用“推理结果”进行知识库检索。检索过程是基于向量相依程度计算的语义相似度检索。

知识库条目向量化表示使用了 **Sentence-BERT**[6]模型，向量相似度计算使用了**余弦相似度**。

---

3 包括重新组织语言的知识库条目内容、用户可能感兴趣的知识点、有帮助的书籍推荐等。

4 本文具体的 prompt 见**附件-Prompt 展示.txt**.

5 详见 **4.3.1 深度融合大模型与动态 RAG 技术。**

6 Sentence-BERT (SBERT)是一系列用于句子嵌入的模型，它们在语义相似度任务上表现出色。

系统将“推理结果”与知识库中所有条目的向量表示进行相似度计算，筛选出**相依程度最高的特定数量(Top-K)**的相关知识库条目。

Top-K 值的主要设定依据为知识库大小、条目相关性分布。“第十章”智能代数解答引擎的知识库规模较小（千条规模），因此 Top-K 值应该较小。但相关条目分布较为稀疏，需要更大的 K 值以提高召回率。通过验证集测试不同 K 值对下游任务指标（即代数知识、题目回答正确率）的影响，选择最优值 **Top-K=3**。

此外，需设置一个**相关性阈值**，只有在相似度达到某程度时，才可以认为该条目是有用的。同样经过与 Top-K 值类似的测试，选取相关性阈值 **threshold = 0.5**[7]。

5.1.3.4 基于大模型 B 的增强生成式解答

检索到相关知识库条目后，Flask 应用将用户原始问题、“推理结果”、预设的 Prompt2、检索到的特定知识库条目以及最近的 5 轮对话历史记录进行整合，构建为一个新的、更丰富的提示词[8]。

随后，Flask 应用将构建好的新 Prompt 发送至**大模型 B**。

大模型 B 作为系统的“**生成引擎**”，基于新 Prompt 的引导，深度融合用户问题、推理结果、检索到的知识库条目信息、历史对话记录，进行增强生成式解答。

大模型 B 输出的解答结果通常包含以下关键组成部分：

1. **清晰完整的解答**：基于用户问题和相关知识，生成条理清晰、逻辑完整、准确可靠的解答。如果是代数题目，还将包含题目分析、解题步骤、公式推导、概念解释等；
2. **可能相关的其它知识条目与相关文件、书籍**：推荐与当前问题或解答结果可能相关的其它知识条目，引导用户进行更深入、更全面的知识探索和学习；
3. **知识库内容**：大模型 B 将把知识库中相关内容输出。

5.1.3.5 流式输出与用户界面显示

最终的解答结果由 API 网关负责接收，并通过**流式数据传输技术**[9]实时、分段地输出至用户交互界面。

流式输出将大模型输出数据视为连续的数据流，而不是等待大模型回答完毕后将其内容一起输出。使用流式输出技术可以**降低服务器压力，实现负载均衡，提升用户等待体验**，同时也能支持更长、更复杂的解答内容的输出等。

此外，系统支持 **MarkDown** 与 **LaTex** 格式的公式渲染。“第十章”使用本地部署的 **MathJax** 与 **markedjs** 实现两者的正确显示，并使用 **DOMPurify** 防止 XSS 攻击。

通过上述推理引导的动态 RAG 技术流程，“第十章”智能代数解答引擎充分利用

7 即相似值小于阈值 0.5 时，知识条目保留，否则舍弃。

8 其中，Prompt2 的目标是引导 Qwen2.5-Math-1.5B 模型结合用户问题、推理结果、检索到的知识库条目、历史对话记录，生成高质量的解答结果，并组织答案结构，提供相关知识推荐。

9 Server-SentEvents, SSE

了大模型的推理和生成能力以及知识库的精准知识，实现了更智能、更可靠、更人性化的代数问题解答与知识问答服务，为用户提供高效、个性化的代数学习支持。

5.2 知识库构建

5.2.1 数据来源

本文从多种来源广泛收集数据[10]，并进行处理后组成系统的知识库文件，共计**两千余条优质数据**。主要使用数据源如下：

1. **部分省份中考、高考考试大纲**；
2. **部分省市部分学校练习卷**；
3. **网络爬虫爬取**：编写爬虫程序，从互联网上抓取的相关数据（知乎、百度文库等）；
4. **人工数据收集**：学科网等其它网站内人工取得的数据；
5. **其它文献资料**：各类数学通史或代数学通史教材（如《数学史》-卡尔·B.梅耶编，中央编译出版社等），中国古典数学典籍（如《九章算术》、《海岛算经》等）等。

5.2.2 数据处理

从数据源收集来的数据种类驳杂，格式不同，且可能存在数据重复、数据错误等问题。为了更好地用作数据库中，需要进行如下处理：

1. **统一数据格式**：使用脚本将多模态数据统一为文本数据。收集到的数据格式包括文本文件（.txt，.md 等）、PDF 文件（.pdf）、网页（.html，.htm）、其他格式（.docx，.pptx，.xlsx）等，将其中包含的所需知识分条目统一为文本数据，并填入数据库表格中；
2. **数据清洗**：使用脚本提取数据后，文本中还可能有 HTML 标签、特殊字符等无关信息，需要将这些噪声去除；
3. **公式改写**：数据中公式全部改写为 LaTeX 格式，便于大模型识认与学习；
4. **去重**：移除重复的数据；
5. **填充缺失值**：对缺失的数据进行填充；
6. **纠正错误**：纠正数据中的错误或剔除错误数据；
7. **分词**：将文本切分成词语或短语单元。使用工具为 jieba；
8. **去除停用词**：去除文本中常见的无意义词语，例如“的”“是”“了”等。

10 部分数据来源文件收录在附件-数据来源文件（部分）文件夹中。

5.2.3 提取关键字

利用大模型，提取知识库中每个条目的**关键字**[11]，用以更好地计算向量相依程度。

5.2.4 知识库构建

本文使用向量数据库为知识库存储方式。向量数据库用于存储向量数据，可以将知识表示为向量，用于语义相似度搜索、知识推荐等应用。

首先使用开源预训练模型 **Sentence-BERT** 进行文本向量化。本文使用轻量级、语义表示能力较好的 **all-MiniLM-L6-v2** 版本。

接下来将向量存储在 **FAISS**[12]开源向量数据库中。FAISS 是一个高性能的相似度搜索库，适合中等规模到大规模的向量数据。相比将向量直接存入服务器内存或文件中，可以实现更高效的搜索。

5.3 动态 RAG 技术解释

本节主要介绍**动态 RAG 技术**的发展，并与**传统 RAG 技术**做比较，然后指出“第十章”系统应用此技术的方式。

近年来，**大型语言模型**（LLMs）在自然语言处理领域取得了显著的进展，并在各种应用中展现出强大的能力。然而，LLMs 并非完美无缺。由于其知识来源于训练数据，因此存在固有的局限性，在面对超出其训练范围的问题时可能产生不准确甚至虚构的回答，这种现象被称为“**幻觉**”。这严重限制了 LLMs 在需要高度准确性和最新信息的场景中的应用。

为了克服这些限制，**检索增强生成**（Retrieval Augmented Generation，RAG）技术应运而生。RAG 通过在生成文本之前从外部知识库检索相关信息，并将这些信息融入到生成过程中，从而显著提升了 LLMs 的性能。RAG 的核心思想是让 LLMs 能够像进行“**开卷考试**”一样，在回答问题时查阅额外的资料，从而确保答案的准确性和时效性。

随着研究的深入，传统的静态 RAG 方法逐渐演进为更为复杂的**动态 RAG 方法**。动态 RAG 不再是在生成过程开始时进行一次性检索，而是在文本生成过程中根据模型的需求动态地决定何时以及检索什么信息。为了进一步提升动态 RAG 的性能，研究人员引入了“**推理引导**”（Inference Guidance）的概念。推理引导旨在通过各种技术手段，例如语法约束、格式要求和工具调用等，来**格式化用以检索知识库所用的文本，降低用户自然语言输入带来的差错可能**，从而提高输出的质量和可控性。

另外需要说明的是，如今的动态 RAG 技术中，通常会使用大模型对结果进行**多次检查**，这是一种闭环的控制差错的方法。这是因为如果用作推理的模型在理解用户意图时出错，将对结果的正确性造成**直接的、巨大的影响**。但在本文中，综合考虑服务面向的问题难度（主要是高中及以下代数难度）与程序响应速度，**未做该闭**

11 用来提取关键字的程序为一 python 程序，见**附件-提取关键字程序.py**

12 Facebook AI Similarity Search.

环检查，这可以平衡系统的正确率与响应速度。但在企业部署商业项目时，应注意**加入基本的大模型安全对齐体系**，用于识别和移除潜在的敏感或有害输出。

5.4 知识库维护

在真正面向用户开放服务后，系统知识库还需**定期维护**，才能保证知识的及时性和服务的稳定性。

1. **知识库监控**：监控知识库的性能和数据质量，例如查询响应时间、数据完整性、数据准确性等。
2. **知识更新**：定期或实时地更新知识库中的知识。
3. **知识修正**：修正知识库中错误的或过时的知识。
4. **知识库版本管理**：对知识库的不同版本进行管理，方便回溯和版本比较。
5. **知识库备份与恢复**：定期备份知识库，以防止数据丢失或损坏。

5.5 大模型的选型与优势说明

本文中，大模型 A 为 **Qwen2.5-1.5B-Instruct-GGUF:Q8_0**，大模型 B 为 **Qwen2.5-7B-Instruct-GGUF:Q8_0**[13]，使用 ollama 进行本地部署。两个大模型不会同时运行。

选择大模型的核心点在于**平衡性能与成本**，同时提供**较好的服务保障**。

大模型 B 的任务很重要，但同时也很简单，使用 1.5B 模型即可胜任。这个模型的选择要点是对中文语境熟悉，指令遵循效果好，能够快速[14]提取用户问题中的关键词，**推理出用户的意图**。

大模型 A 接收的文本信息很多，不仅有用户问题，还有知识库文本、对话历史记录，如果依旧选用较小模型，则可能分不清用户问题与其它文本的边界，会将知识库中的信息误认为用户问题的一部分，或者在回答时过度依赖知识库而忽略了用户问题的真正意图，大大降低对 prompt 的遵守程度。因此，本文除在 **prompt 调整**上做了很多工作外，在模型选择上选择了一个较大规模的大模型。

这两个模型均为经过指令微调（Instruction Tuning）的版本。它能够直接响应自然语言指令，更严格遵循指令要求，更适合本文情景。

这两个模型均经过 **8bit 量化**处理，在资源消耗与表现上做了更好地平衡。

本文还曾选取过其它大模型，如 Qwen2.5-Math 等进行测试，它们的弊端见附件-大模型选型说明.xlsx。

对于实际商业项目部署，建议企业根据具体需求选择**更高量化精度**的模型版本以获得最佳性能表现。

---

13 本文使用到（或曾经使用到）的大模型均在 hugging face 上有开源，但一般选择国内的"魔搭社区"作为模型镜像源进行本地部署。本文中所有开源代码或项目使用情况见**附件-开源说明.txt**.

14 在本地测试中，该模型对于用户输入的意图提取响应时间在 1s 内。

6 项目系统展示与性能评估

6.1 系统展示

6.1.1 首页

图 11 首页展示[15][16]

15 系统中关于团队的超链接均为空链接，没有指向。
16 详细展示见展示视频与附件-前端界面展示文件夹。

6.1.2 对话

图 12 对话页面展示

图 13 大模型具有上下文记忆

6.1.3 上传图片与知识库文件

图 14 上传图片后得到的回答

图 15 上传用户的知识库文档成功

6.1.4 后端中间状态展示

```
127.0.0.1 - - [11/Apr/2025 22:37:05] "OPTIONS /api/chat HTTP/1.1" 200 -
第一个大模型 (Qwen2.5-1.5B-Instruct) 的输出 (检索 Query): 勾股定理的经典证明方式
```

图 16 推理结果

```
发送给第二个大模型 (Qwen2.5-7B-Instruct) 的完整 Prompt:
你是一位专业的代数问题解答助手，名字叫"第十章"。你基于开源大模型构建，擅长解答各类代数问题。

用中文交流，公式一律使用反斜杠和括号格式表示，全文不要使用markdown中的强调或标题语法。

首先，查看这些知识内容与我们的对话历史：

**----------知识库内容开始----------**
分类：题目_高中
问题：命题的否定
回答：全称命题的否定是存在命题，存在命题的否定是全称命题。
公式：\(\neg (\forall x \in M, p(x)) = \exists x \in M, \neg p(x)\)
---
分类：知识点_高中
问题：方向导数
回答：方向导数描述了多元函数在某点沿某一特定方向的变化率。
公式：\(\frac{\partial f}{\partial \mathbf{v}} = \nabla f \cdot \mathbf{v}\)
---
分类：题目_高中
问题：概率的条件方差
回答：条件方差是在给定某个条件下的方差。例如，已知 \(X\) 和 \(Y\) 的联合分布，\(Var(X|Y=y)\) 表示在 \(Y=y\) 条件下 \(X\) 的方差。
公式：条件方差公式：\(Var(X|Y=y) = E[(X - E(X|Y=y))^2|Y=y]\)
---

**----------知识库内容结束----------**

**----------历史对话内容开始----------**

**----------历史对话内容结束----------**

我可能会问到我们历史对话内容中的问题。这些知识库内容供你参考，你需要更加详细清楚地回答我的问题。

查看完上面内容后，第一，参考知识库内容和历史对话内容，回答我的问题：**嗨第十章，今天我在学习三角形的勾股定理时遇到了一些问题，你能帮助我吗？我想知道勾股定理
的几种经典证明方式！**

第二，参考知识库内容向我简单推荐五个相关的知识点，每条之间换行；再推荐三本相关的书籍，每本之间换行。

第三，告诉我你参考了知识库中那些条目的名称。
```

图 17 大模型 B 最后得到的 prompt

6.2 准确率评估

在用来测试的 100 道高中及以下代数题目中，“第十章”正确率达到 **87%**，用以对比的无 RAG 技术的 QWen 2.5-7B 仅有 **77%**。

6.3 响应时间评估

“第十章”对测试题目的平均响应时间在 **8s**，比无 RAG 技术的 QWen 2.5-7B **慢不到 20%**[17]，可以接受。

6.4 用户满意度调查

我们进行了小范围测试，覆盖人数约为 70 人，并发放问卷 C。回收有效问卷 50 份，用户满意率为 **78%**。

7 结语

“第十章”智能代数解答系统验证了**大模型**与**动态 RAG 技术**在代数教育领域的应用潜力。系统在代数题目解答、知识智能问答等方面表现出良好的性能，为构建智能化、个性化、高效化的代数学习平台提供了有效的解决方案。

本文相信，随着人工智能技术的不断进步和应用深入，“第十章”类似系统将会在推动代数教育智能化转型、提升学生代数学习效率和兴趣方面发挥越来越重要的

17 同等算力条件下测试。

作用，最终迈向智能化代数教育的新时代。

8 附件目录

8.1 文件夹部分

8.1.1 数据来源文件（部分）

该文件夹收录了用来提取知识库数据的文件。由于文件数量庞大，此处只展示部分。数据来源包括：

1. **部分省份中考、高考考试大纲**；
2. **部分省市部分学校练习卷**；
3. **网络爬虫爬取**：编写爬虫程序，从互联网上抓取的相关数据（知乎、百度文库等）；
4. **人工数据收集**：学科网等其它网站内人工取得的数据；
5. **其它文献资料**：各类数学通史或代数学通史教材（如《数学史》-卡尔·B.梅耶编，中央编译出版社等），中国古典数学典籍（如《九章算术》、《海岛算经》等）等。

8.1.2 前端界面展示

该文件夹内以图片形式展示了前端页面效果。

8.1.3 图表集

本文档与 PPT 演示文档中出现的，与项目本身相关联的图表。

8.1.4 项目源文件

该文件夹内为项目全部代码。

其中，`html5up-photon_VER1` 为首页文件夹，`MathJax-master` 用于渲染 LaTex 公式，`models` 文件夹内为用来进行文本向量化的预训练模型 all-MiniLM-L6-v2 ，`START` 文件夹内为对话页面的前端、后端、环境变量代码，`thetwo` 文件夹内为一些公用的图片、前端 CSS、JS 代码，`algebra_knowledge.CSV` 文件为知识库文件，`marked.min.js` 用于渲染 MarkDown 格式。

8.2 文件部分

8.2.1 参考文献目录.docx

8.2.2 大模型选型说明.xlsx

该文件介绍了团队弃用其它大模型的原因。

8.2.3 Prompt 展示.txt

该文件示意了系统中给大模型 A 与大模型 B 的最终 Prompt 文本。

8.2.4 项目部署与测试说明.md

该文件介绍了如何在其它电脑上进行本地部署和测试项目。

8.2.5 提取关键字程序.py

该 python 程序用于从较初始的知识库中提取关键字，形成更适合于进行动态 RAG 的知识库。

8.2.6 开源说明.md

该文件介绍了项目所使用的开源模型、程序或代码及其开源地址。