Computer Science and Application 计算机科学与应用, 2025, 15(4), 64-69
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154078

Hans汉斯

大语言模型在高中数学解题中的效能提升研究

曹建伟

华南师范大学数学科学学院，广东 广州

收稿日期：2025年3月1日；录用日期：2025年3月31日；发布日期：2025年4月8日

摘　要

本文探讨了如何通过构建基于LangChain的知识库和利用LMDeploy推理加速技术，提升大语言模型在解答高中数学题目中的正确性和响应速度。通过OCR技术将解析卷转化为LaTeX格式，结合BGE-M3模型进行文本向量化并存储于Faiss数据库，模型在解答时通过动态检索知识库内容来增强准确性；同时，通过LMDeploy量化加速推理技术，显著提升了模型的推理效率。实验结果表明，多个大模型在构建知识库前后的得分有显著差异，总体回答正确率提升了71.84%；在回答速度上，自部署模型总体第一题回答速度提高了121.10%，后续题目回答速度提高了259.94%。这些改进显著提升了大语言模型在解答高考数学题时的正确率和速度。

关键词

大语言模型，LangChain知识库，推理加速，LMDeploy

Research on Enhancing the Effectiveness of Large Language Models in Solving High School Mathematics Problems

Jianwei Cao

School of Mathematical Sciences, South China Normal University, Guangzhou Guangdong

Received: Mar. 1st, 2025; accepted: Mar. 31st, 2025; published: Apr. 8th, 2025

Abstract

This paper explores how to enhance the accuracy and response speed of large language models in solving high school mathematics problems by constructing a knowledge base based on LangChain and utilizing the LMDeploy inference acceleration technology. OCR technology is used to convert scanned papers into LaTeX format, and the BGE-M3 model is employed for text vectorization, which

文章引用: 曹建伟. 大语言模型在高中数学解题中的效能提升研究[J]. 计算机科学与应用, 2025, 15(4): 64-69.
DOI: 10.12677/csa.2025.154078

is then stored in a Faiss database. The model dynamically retrieves knowledge from the database to improve accuracy during problem-solving. Meanwhile, LMDeploy's quantization and inference acceleration technology significantly boosts the model's inference efficiency. Experimental results show that there is a significant difference in the scores of various large models before and after constructing the knowledge base, with an overall improvement of 71.84% in answer accuracy. Regarding response speed, the model's answer speed for the first question improved by 121.10%, and the speed for subsequent questions improved by 259.94%. These improvements substantially enhanced the correctness and speed of large language models in solving high school math problems during the college entrance examination.

Keywords

Large Language Models, LangChain Knowledge Base, Inference Acceleration, LMDeploy

1. 研究背景

近年来，大语言模型在自然语言处理领域取得了显著进展，在医疗、法律、金融、教育等多个科学领域发挥作用。2024 年 6 月，上海人工智能实验室公布了首个 AI 高考全卷评测结果。该实验室的司南评测体系 OpenCompass 选取了 6 个开源模型及 OpenAI 的 GPT-4o 进行大模型高考“语数外”全卷能力测试。结果显示，阿里通义千问 Qwen2-72B、OpenAI 的 GPT-4o 及书生浦语 2.0 文曲星(InternLM2-20B-WQX)成为本次大模型高考的前三甲，对应得分率分别为 72.1%、70.5%和 70.4%。大部分模型在“语言”本质上的表现良好，语文平均得分率为 67%，英语更是达到了 81%，而数学则是所有大模型的短板，平均得分率仅为 36%；书生浦语 2.0 文曲星取得了 75 分的最高分，超过所有受测模型，然而仍未达到及格水平[1]。

由此可见，大语言模型在高中理科题目解答方面的能力还有待提高，分析大语言模型在解答高中理科题目方面的能力，有利于为大语言模型在教育领域的使用和发展提供更多探索思路和可能性[2]。

本研究通过构建基于 LangChain 的数学知识库和利用 LMDeploy 加速推理来提升数学解答的效果和推理速度。LangChain 作为一款集成了多个大型语言模型的框架，能够将海量的数学知识和解题策略进行有效存储和组织，从而为模型的解答过程提供丰富的支持[3]。LMDeploy 则通过量化加速推理技术，进一步提高大模型的推理速度和响应效率，使其在实际应用中更加高效。

2. 研究方法

2.1. LangChain 知识库

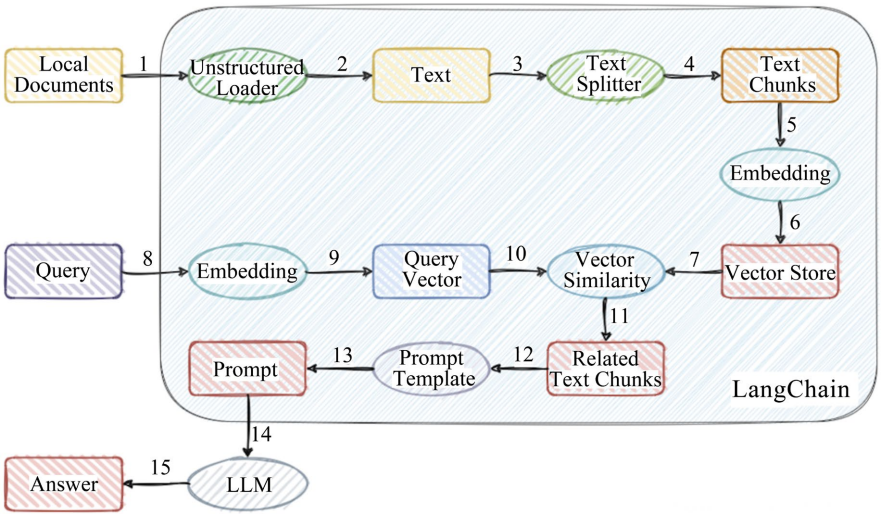
为了提高大模型在解答理科题目中的正确性，本文采用了 LangChain 框架来构建基于本地知识库的问答应用。LangChain 通过扩展大模型的知识内容，有效提升了其对复杂数学题目，特别是高考理科题目的回答能力。

首先，本文通过 OCR 技术处理解析卷，将公式转化为 LaTeX 格式的文本，并将解析卷的内容按题目进行分割，随后使用 BGE-M3 模型进行向量化，并将这些向量存入 Faiss 向量数据库。该过程使得数学问题的解析内容得到高效存储和管理，增强了模型的知识基础。

在实际使用中，当用户提问时，首先将题目向量化，然后在知识库中查找与题目最相似的多个文本(top

k 个)。这些匹配到的文本片段会作为上下文与问题一同添加到大模型的 prompt 中，提交给模型生成答案。这一过程中，模型不仅仅依赖自身的训练知识，还能通过动态检索知识库中的相关内容来提升解答的准确性。

具体来说，LangChain 通过提供一个结构化的知识库，帮助大模型填补其知识盲区。在面对需要综合多种知识和解题步骤的数学题时，模型能够根据检索到的相关信息补充必要的背景知识，从而更好地推理和生成正确的解答[4]。这种知识库辅助机制，显著提升了大模型对复杂题目，特别是涉及多步骤推理和公式计算的题目的解答能力。具体实现流程如图 1。



图片来源：https://github.com/chatchat-space/Langchain-Chatchat。

Figure 1. LangChain implementation flowchart
图 1. LangChain 实现流程图

图示的实现流程清晰展示了 LangChain 如何通过知识库与大模型相结合，提升解答质量。在此过程中，LangChain 不仅提高了模型对问题的理解深度，还增强了其推理和计算的准确性，极大地提升了模型的正确率。这种方法尤其适用于解答包含多种解法或较为复杂推理步骤的题目，如高考中的理科题目。通过这一方式，大模型的表现可以得到显著提升。

2.2. LMdeploy 加速推理

LMDeploy 的核心功能主要集中在推理加速和量化优化两个方面，具体技术实现原理如下：

1) 高效的推理技术

LMDeploy 通过多种技术手段提升大语言模型的推理效率，该技术的持续批处理机制允许在推理过程中动态调整批处理的大小。这一机制能够根据硬件资源的实时变化，优化批处理的执行，从而最大化计算资源的利用率。这种调整能力对于大型模型来说至关重要，因为它有助于避免硬件资源浪费，同时提升整体推理速度。

LMDeploy 的块级键值缓存技术对注意力机制中的键值对进行了优化，将其分块存储和计算，从而有效减少了内存占用和计算开销。这种技术极大地提高了处理大量数据时的内存效率，并减轻了计算瓶颈。

2) 量化优化技术

量化技术是提升推理速度和降低计算成本的核心手段之一。LMDeploy 通过对模型进行量化优化，减少计算中的浮点精度，从而减小了模型的内存需求，提升了推理的计算效率。

LMDeploy 的量化策略包括对模型权重进行量化，将浮点数权重转换为低精度的整数表示。此举不仅

减少了内存占用，还加速了推理过程，尤其是在硬件资源有限的环境下，量化带来的效益尤为显著。

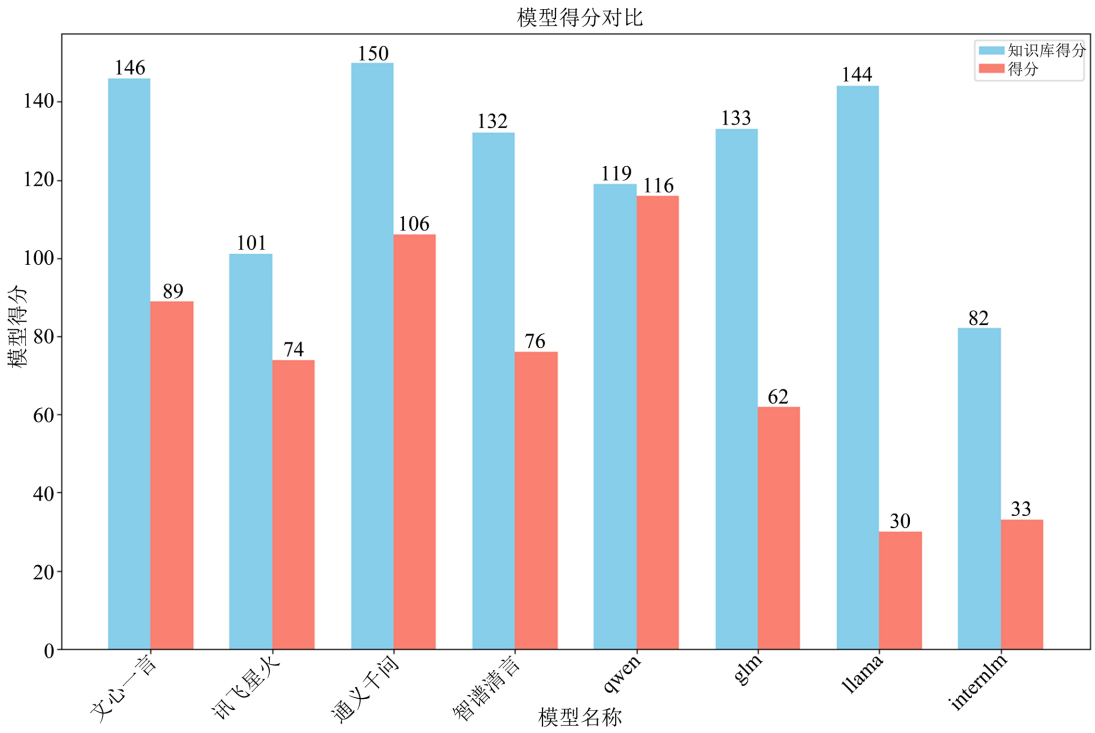
此外，LMDeploy 还引入了 4-bit 量化技术，显著提高了推理效率。4-bit 量化使得模型的推理速度达到原来 FP16 精度模型的 2.4 倍，极大地加速了推理过程，同时保持了较高的计算精度。量化后，推理的时间和计算消耗大幅降低，尤其在处理大规模数据集时表现得尤为明显[5]。

3. 结果分析

本研究选取了 2020 年至 2024 年高考数学新课标Ⅰ卷的题目作为评测数据，利用现有的大语言模型(商用的文心一言、讯飞星火、智谱清言、书生浦语、ChatGPT-4o、通义千问和开源的 qwen2.5-math-7b-instruct、interlm2-math-7b、glm4-9b-chat、Meta-Llama-3-8B-Instruct)对相应的数学题目进行解答，并整理不同大语言模型的测试结果如下。

3.1. 正确性

在构建知识库后，本文对 2023 年高考数学新课标Ⅰ卷进行问答测试，测试结果分数如图 2：



图片来源：作者自绘。

Figure 2. Model score comparison histogram
图 2. 模型得分对比柱状图

从柱状图中可以看出，多个大模型在构建知识库前后的得分有显著差异。总体回答正确率提升了 71.84%。接下来从总体趋势、各模型的具体表现以及可能的原因和启示等方面进行详细分析，以探索知识库构建对大模型回答高考题表现的影响。

从总体上看，大多数大模型在构建知识库后得分都有了显著提升，表明知识库在提升模型回答复杂问题的准确性和深度方面发挥了重要作用。高考题目覆盖了广泛的知识领域，对大模型的知识储备和推理能力提出了较高要求，而知识库显然在填补这些需求方面提供了支持。这一趋势从各模型得分的普遍

上升可以明显看出，尤其是一些基础较强的模型，知识库构建后得分提升尤为显著，这些模型能够较好地利用知识库补充自身的知识缺口。

具体而言，“文心一言”在知识库构建前得分为 89 分，构建后提升至 146 分，增幅高达 57 分；“通义千问”得分从 106 分提升至 150 分，增幅为 44 分，表现为所有模型中的最高分。这表明，部分模型能够充分利用知识库来补充自身的知识空白，并显著提升解答的准确性。例如，“通义千问”可能在知识检索和信息整合方面有较强优势，能够较好地适应和应用知识库中的信息。

“Llama”模型在构建知识库前的得分为 30 分，知识库构建后提升至 144 分，增幅达到 114 分，这是所有模型中得分提升幅度最大的一例。这表明，知识库对于基础知识储备较少的模型尤其重要，它能够迅速弥补模型的知识不足，显著提升其表现。

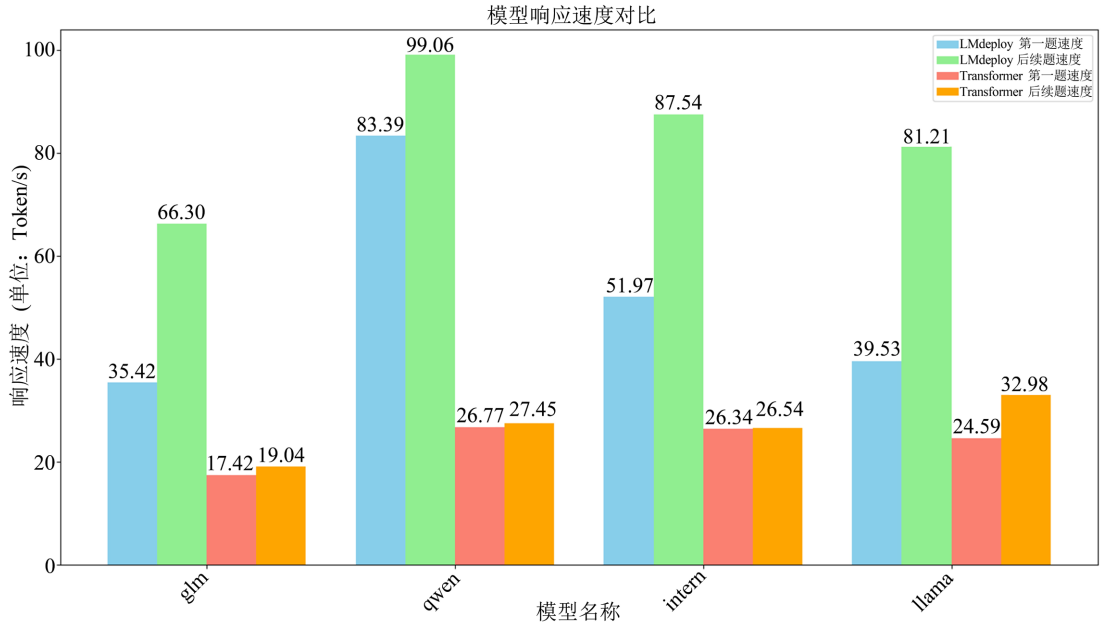
然而，并非所有模型都对知识库的构建产生相同的效果。例如，“讯飞星火”和“智慧清言”的得分增幅虽然显著，但相对其他模型略低。可能原因在于这些模型的架构、训练数据和对知识库内容的适应能力有所不同。某些模型可能无法充分吸收知识库内容，导致知识库对其提升的作用有限。

“GLM”和“InternLM”模型的得分提升幅度较小，这可能与模型的结构设计和知识库的适配性有关。尽管它们在某些领域表现出色，但对于高考题目所涉及的广泛知识领域，知识库的构建对这些模型的提升效果较为有限。

综上所述，知识库的构建在提高大模型回答高考数学题目时具有重要作用，尤其在提升模型的知识覆盖面、信息检索能力和推理深度方面效果显著。不同模型对知识库的依赖程度和适应能力不同，因此，模型架构、知识获取方式和参数规模等因素均影响知识库对模型表现的提升效果。

3.2. 回答速度

本文对 4 个自部署开源大模型采用 LMdeploy 就 2024 年高考数学新课标 I 卷的回答响应速度进行测试，并与 transformer 架构的回答响应速度进行对比如下图 3。



图片来源：作者自绘。

Figure 3. Histogram of model response rate comparison
图 3. 模型响应速度对比柱状图

由柱状图可得，LMDeploy 优化显著提升了四个大模型(glm、qwen、llama 和 intern)在解答高考题时的响应速度。在 LMDeploy 加速推理后，总体第一题回答速度提高 121.10%，后续题目回答速度提高 259.94%。无论是第一题的速度还是后续题目的处理，使用 LMDeploy 架构的模型相比于原始的 Transformer 架构均表现出显著的改进。

GLM 模型的后续题速度从 Transformer 的 19.04 tokens/s 提升至 66.30 tokens/s，增幅超过三倍。第一题的处理速度也从 17.42 tokens/s 提升至 35.42 tokens/s，说明 LMDeploy 优化显著减少了模型推理中的延迟。在处理连续对话或长对话任务时，优化后的模型能够更加高效地响应，极大改善了用户体验。

Qwen 模型的提升尤为显著，其后续题的速度从 27.45 tokens/s 提升至 99.06 tokens/s，提升幅度接近四倍。第一题的速度也从 26.77 tokens/s 增至 83.39 tokens/s，说明 LMDeploy 在 Qwen 模型的推理过程中有效提高了生成效率，尤其在长文本或多轮对话的场景中表现突出。

Llama 模型同样在 LMDeploy 优化后展示了明显的性能改进，第一题的速度由 24.59 tokens/s 提升至 39.53 tokens/s，后续题的速度也从 32.98 tokens/s 提升至 81.21 tokens/s。这表明 LMDeploy 不仅提高了模型的初始响应速度，还能加快连续对话中的每一轮生成时间，从而提升对话的流畅度和实时性。

Intern 模型的性能提升虽然较为温和，但同样体现了 LMDeploy 架构的优势。后续题速度从原始 Transformer 的 26.54 tokens/s 提升至 51.97 tokens/s，第一题的速度也从 26.34 tokens/s 增至 39.53 tokens/s。这一变化意味着 Intern 模型在面对初次提问和连续对话时，能够提供更迅速的响应，减少了用户的等待时间。

综上所述，LMDeploy 优化对所有四个大模型的响应速度都有积极影响，尤其在减少延迟和提升每秒生成 tokens 的数量方面表现突出。通过 LMDeploy 架构优化，大模型不仅能够在初次响应时更加迅速，在长对话和多轮交互中也能更高效地完成任务，显著改善了用户体验和应用实用性。

4. 总结与展望

本研究通过构建 LangChain 知识库和利用 LMDeploy 推理加速技术，显著提升了大语言模型在解答高中数学题目中的正确性与速度。通过整合丰富的数学知识库，模型能够有效补充其知识空白，提升在解答复杂数学题时的准确性和深度。同时，LMDeploy 架构的优化显著加速了模型的响应速度，使其能够更高效地处理连续对话和长文本生成，极大地改善了用户体验。

未来的研究可以进一步优化知识库的构建与整合，探索更高效的加速推理技术，以应对更大规模、更复杂的数学问题。随着教育需求的多样化，结合个性化学习路径和实时反馈的系统也将成为发展方向。通过对大语言模型的持续优化，将为数学教育和其他学科的智能化教学提供更强大的支持。

参考文献

[1] 刘明, 吴忠明, 廖剑, 任伊灵, 苏逸飞. 大语言模型的教育应用: 原理、现状与挑战——从轻量级 BERT 到对话式 ChatGPT [J]. 现代教育技术, 2023, 33(8): 19-28.

[2] Bai, J.Z., Bai, S., Chu, Y.F., Cui, Z.Y., Dang, K., Deng, X.D., *et al*. (2023) Qwen Technical Report. https://qianwen-res.oss-cn-beijing.aliyuncs.com/QWEN_TECHNICAL_REPORT.pdf

[3] 赵浜, 曹树金. 生成式 AI 大模型结合知识库与 AI Agent 开展知识挖掘的探析[J/OL]. 图书情报知识, 1-14. http://kns.cnki.net/kcms/detail/42.1085.G2.20241103.1003.002.html, 2024-11-10.

[4] Langchain-Chatchat. https://github.com/chatchat-space/Langchain-Chatchat

[5] LMDeploy. https://github.com/InternLM/lmdeploy