

基于重排序和后检索反思的教育大模型问答增强方法

孙浩然, 王志豪, 吴一帆, 高晓影, 向阳

同济大学计算机科学与技术学院 (国家特色化示范性软件学院), 上海 200000

摘要

计算机教育是现代信息社会教育的重要组成部分, 随着大语言模型的发展, 将大语言模型应用于计算机教育过程受到越来越多的关注, 但大语言模型存在的“幻觉”问题为其应用带来了挑战。为了应对这一挑战, 检索增强生成技术通过引入外部知识库, 有效提升了大语言模型的回答质量。然而, 传统检索增强生成技术缺乏对检索信息的精细筛选机制, 导致大量低相关知识被保留, 无关信息的干扰使模型幻觉问题并未得到有效解决。收集计算机相关教材和知识文档, 根据内容结构将其划分为知识文档块, 以构建外置知识数据库; 在此基础上, 提出基于重排序和后检索反思的教育大模型问答增强方法。该方法利用基于交叉编码器的高性能多语言重排序模型捕捉深层语义信息, 过滤信息检索, 筛去无关信息以提高检索质量, 并将检索增强生成技术应用于模型反思, 令模型通过自我检查进一步增强模型回答质量, 有效提升了大语言模型在计算机问答时的准确性。对所提方法在多个主流生成模型上进行测试, 其在 CS-Bench 上取得了良好的效果, 并在计算机知识问答时使模型回答正确率提升约 5%。

关键词

大语言模型; 检索增强生成; 计算机教育

中图分类号: TP391

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025062

Question-answering enhancement method for large educational models based on re-ranking and post-retrieval reflection

SUN Haoran, WANG Zhihao, WU Yifan, GAO Xiaoying, XIANG Yang

School of Computer Science and Technology National Pilot Software Engineering School with Chinese Characteristics, Tongji University, Shanghai 200000, China

Abstract

Computer education is one of the requirements of modern information society education. With the development of large language models, there has been increasing attention on applying of large language models to the computer education process. However, the hallucination problem associated with large language models poses significant challenges to its application. To solve the challenges, RAG techniques by incorporating external knowledge bases can effectively enhance the quality of responses generated by large language models. However, the traditional RAG techniques lack a fine screening mechanism for the retrieved information, which leads to the retention of a large amount of low-correlation knowledge, and the interference of irrelevant information makes the model hallucination problem not effectively solved.

We collected computer-related textbooks and knowledge documents, dividing them into knowledge document blocks according to the content structure to construct an external knowledge database. On this base, we introduced the large educational models question-answering enhancement method based on re-ranking and post-retrieval reflection, which utilized a high-performance multilingual re-ranking model based on a cross-encoder to capture deep semantic information, filter the retrieval information, filter out irrelevant information to improve the retrieval quality. The proposed method applied RAG techniques for model reflection so that the model can further enhance the quality of the model's answers through self-examination, and effectively improve the accuracy of the large language model in computer question-answering tasks. This approach significantly improves the accuracy of large language models in computer question-answering tasks. The proposed method has been tested on several popular current generative models, achieving promising results on CS-Bench, with an approximate 5% increase in accuracy for computer question-answering tasks.

Key words

large language model, retrieval-augmented generation, computer education

0 引言

随着人工智能的快速发展,大语言模型(简称大模型),如GPT^[1]、LLaMa^[2]在自然语言处理任务中发挥着重要作用,并在其他多个领域得到广泛应用。通过在大量通用语料上的训练,大模型拥有了推理能力、长文本建模能力以及文本生成能力^[3]。在教育过程中,大模型可以用于习题解答、辅助评分、个性化教学、提问生成等任务^[4]。在计算机教育领域,基于大模型构建的知识问答系统可以为学生解答疑问,提高学生学习效率。

然而,教育领域要求大模型高度可靠,以确保传播知识的正确性。对此,大模型在计算机教育领域中的应用仍存在两点不足。一是大模型基于通用语料库进行预训练,缺少各领域内细节的知识材料,导致其在特定领域掌握的知识不够具体,在计算机知识问答时回答不够准确;二是大模型的训练机制使其存在“幻觉”问题,可能会生成不真实的信息,导致错误的知识传播^[5-6]。

为了解决上述问题,Dan等^[3]提出使用

有监督微调的方式,利用少量数据对预训练大模型进行微调训练,以增强其在特定领域的知识和任务处理能力。但使用有监督微调的方法可能导致大模型原有的能力下降,并且在知识更新时需要对模型重新微调训练,会消耗大量时间和计算资源^[5]。Dhuliawala等^[7]提出CoVe框架,首先让模型生成初稿,随后将初稿划分为多个子问题逐个检查和修正,但其框架高度依赖模型预训练,当模型缺少领域细节知识时其无法解决幻觉问题。Wang等^[8]、Liu等^[9]、Neupane等^[10]提出使用检索增强生成(retrieval-augmented generation, RAG)技术,在外置的知识库检索适合的知识内容,将其作为大模型回答时的参考材料^[11]。使用RAG技术不会使模型原有能力丢失,且外置知识库更新成本低。但传统RAG技术仅使用余弦相似度进行搜索,未考虑检索知识与查询文档的上下文关联和交互语义,检索能力不足,且召回的无关参考材料可能对模型生成造成干扰,降低模型回答质量。

为了解决RAG技术可能召回无关参考材料,导致无法缓解模型“幻觉”的问题,本文从CS-Wiki、CS-Notes和CS-Basic-SelfLearning中收集计算机中文知

识文档，从 Programming_Books_llama 中收集计算机英文知识文档，根据内容结构进行合理大小的文档块切分后，构建中文和英文外置知识数据库；使用 RAG 技术改善大模型在计算机教育领域的知识性缺陷。本文提出一种基于重排序和后检索反思的教育大模型问答增强方法，通过在 RAG 过程中利用基于交叉编码器的重排序模型，将查询文本与检索结果拼接；通过大模型计算检索结果与用户查询间的相关性得分，过滤无关检索材料，使得到的辅助知识更准确，并用于大模型增强生成更准确的答案；引入后检索反思环节，在初轮回答后再次检索和重排序知识文档，与初轮回答结果一并传入生成模型进行纠错，该过程利用辅助知识增强模型反思和自我纠错的能力，并改善模型最终回答质量。

本文主要贡献如下。

- 本文分别收集来自 CS-Wiki、CS-Notes、CS-Basic-SelfLearning 的中文材料以及来自 Programming_Books_llama 的英文材料，按照文档结构切分后，分别构建不同语言的外置计算机知识库，用高质量知识库结合 RAG 技术提升模型回答质量。

- 本文提出基于重排序和后检索反思的教育大模型问答增强方法。本文首先在传统 RAG 过程中引入基于交叉编码器的多语言重排序模型，对向量检索后的知识文档计算用户查询与每个检索结果之间的相关性；随后根据相关性得分排序，筛选检索结果用于增强生成。此外，本文将模型反思与 RAG 技术相结合，利用初轮生成结果进行检索和重排序查找辅助知识，用于生成模型自我纠错，进一步提升大模型回答准确度。

- 所提方法在 CS-Bench 上利用多个不同生成模型进行测试，实验结果表明，

所提方法适用于不同的生成模型，且均使大模型在计算机知识问答中提升了约 5% 的正确率，缓解了“幻觉”现象。

1 相关工作

1.1 大模型

大模型是具有庞大参数规模以及卓越学习能力的高级语言模型。目前，众多大模型，如 GPT-4^[1]、Qwen2^[12]，采用了 Transformer 架构进行语言建模。有研究表明，模型规模的扩展可以大幅度提高模型的能力。也有研究提出缩放定律，其能有效指导大模型的训练。随着大模型的发展，大模型在多个垂直领域中的使用得到重视^[13-18]。

1.2 M3-Embedding

M3-Embedding 是北京智源人工智能研究院提出的嵌入模型，能够对长度为 8 192 个符号的输入序列进行向量化嵌入，以进行密集检索。密集检索使用查询和文档表示同一语义空间中的嵌入，使用嵌入相似度反映语义关系。通过学习不同语言的共同语义空间，M3-Embedding 能够实现多语言嵌入和检索，在多个跨语言基准测试中取得良好得分表现^[19-20]。

1.3 重排序

重排序主要应用于信息检索任务，旨在过滤无关信息以提升检索结果的准确性^[21]。重排序通常使用交叉编码器架构来构建，使用计算要求较高的模型同时处理查询和文档信息，并根据相关性得分重新排序检索候选项^[22]。重排序模型将用户查

询和文档拼接为单个输入，通过Transformer架构进行联合编码并计算相关性分数。与余弦相似度相比，重排序模型建模词级交互，通过自注意力机制捕捉深层语义关联，在相关性计算时更准确，但计算复杂度更高。

1.4 大模型在教育中的应用

随着大模型的发展，其教育场景也得到广泛应用。大模型可以辅助学生进行阅读和写作，如Prihar等^[23]利用大模型学习现有解释文本，使大模型可以解释数学问题的答案；GrammarGPT^[24]通过收集数据并对大模型进行指令微调，使其可以纠正写作过程中的语法错误。大模型也可以辅助教师设计课程，如Koraishi^[25]探索利用大模型优化英语课程设计和教学材料的编写。Jury等^[26]提出了WorkedGen，利用大模型为编程学习生成示例代码，并设计良好的可视化界面，有效辅助用户学习。大模型还可以用于互动式教学、设计聊天机器人、通过问答方式辅助学生进行学习并提供即时反馈^[4]。大模型在教育中的应用对大模型回答的可靠性提出了更高的要求，这使大模型“幻觉”现象成为一个关键问题^[27]。

1.5 检索增强生成

由于大模型的能力依赖预训练数据，RAG技术通过建立外置知识库，为大模型检索文本段落提供补充信息，从而弥补大模型预训练数据上的知识不足缺陷，大幅提升了大模型在知识密集型任务上的性能。例如，Neupane等^[10]将课程教学材料转化为知识数据库，并通过RAG技术增强模型；Pasquarelli等^[28]通过对学生进行访谈，人工获取外部知识构建数据库，随

后使用RAG技术提升模型性能。但是，上述研究在RAG过程中均沿用了传统RAG技术。传统RAG技术遵循索引、检索和生成流程。其首先对外置知识进行切分和向量嵌入；在接收用户查询时，对查询进行向量嵌入；然后根据向量相似度查找相似知识块；最后将用户查询与相关知识块拼接为一个连贯提示词，发送给大模型进行回答。然而，传统RAG技术使用余弦相似度检索独立计算查询向量和检索片段的相似度，仅从全局向量进行浅层语义建模和匹配，检索结果不够精准，无关段落可能降低模型的回答质量，且无法缓解大模型的“幻觉”现象。为了提高模型检索质量，Multi-Meta-RAG^[29]引入重排序模型，通过Transformer架构实现更细粒度的语义交互和深层建模，并对低相关性语义材料进行过滤。然而，Multi-Meta-RAG使用的重排序模型的支持语种单一，在计算机教育和代码场景下表现不佳。本文基于多语言重排序模型过滤无关信息段落，并在生成过程中引入后检索反思，以缓解大模型的“幻觉”现象，从而有效提高大模型在计算机知识问答中的回答准确性^[5,30]。

2 研究方法

传统的模型问答检索增强生成方法，检索过程仅使用向量余弦相似度查找相关文本段，导致检索结果往往包含与用户提问无关的外部知识文本，影响生成模型的回答质量。此外，传统的模型问答检索增强生成方法仅使用单轮检索生成，缺少对回答文本的检查机制，不能有效地缓解模型“幻觉”现象。为此，本文提出了基于重排序和后检索反思的教育大模型问答增

强方法，引入重排序筛选检索结果中的无关片段，并通过有检索的反思环节赋予模型一定的自我纠错能力，缓解模型“幻觉”现象。

本文提出的方法框架如图1所示，其分为以下3个步骤。

- 本文构建了外置知识库，对文本段落进行分块、向量化嵌入和数据持久化；后根据用户查询的向量化嵌入检索出相关知识段向量。

- 本文使用重排序模型在知识段中进行相关性排序，找出几个最相关段落并舍去其余段落，将用户查询和知识段拼接输入大模型，得到首次大模型输出。

- 本文根据首次大模型输出的向量化嵌入再次检索相关知识段，使用重排序模型对新知识段和用户查询进行相关性排序，将用户查询、大模型首次响应以及新知识段拼接输入大模型，得到最终的模型输出。

2.1 外置知识库构建

外置知识库的信息质量与回答结果的质量直接相关。Markdown是一种轻量级标记语言，Markdown文档内容往往具有明显的层次结构。LangChain是一个致力于简化大模型应用开发环节的框架，其提供轻量的Markdown文件解析切分接口。本文收集了大量计算机知识相关Markdown文档，并使用LangChain Markdown加载器进行文档切分。具体地，加载器根据文档的标题层级进行切分，将正文部分切分为多个子文本块，并将标题和文件名写入元数据部分。

得到多个子文本块后，本文使用预训练语言模型对子文本块进行向量化嵌入，这使每一个子文本块转化为一个多维的知识向量。本文使用Chroma数据库对知识向量进行持久化存储，将知识向量保存在磁盘中用于问答时的检索过程。

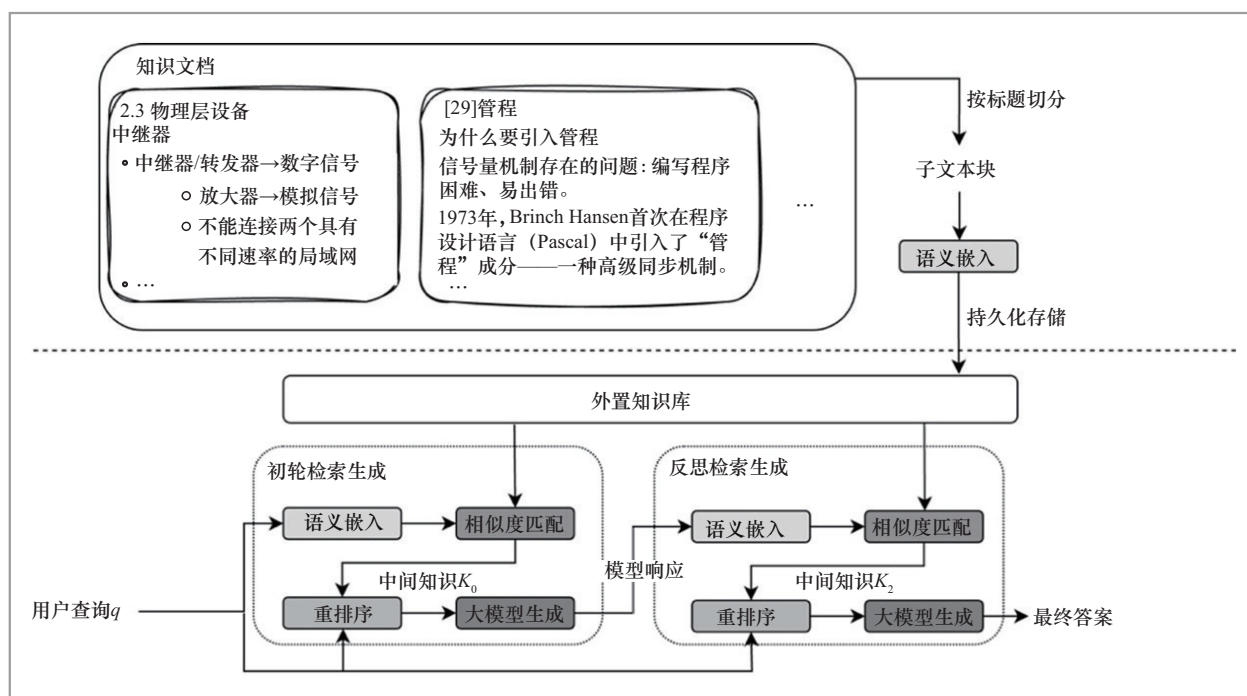


图1 所提方法的整体框架

$$S_K = \{f_e(k_1), f_e(k_2), \dots, f_e(k_i)\} \quad (1)$$

其中, S_K 是编码后得到的知识库, f_e 是使用预训练的语言模型进行嵌入编码, k_i 是切分后的单个子文本块。

2.2 重排序知识段过滤和初次生成

将用户查询记作 q , 首先计算用户查询语义嵌入 s_q 。

$$s_q = f_e(q) \quad (2)$$

余弦相似度计算方法可以计算知识库中每个文本块嵌入 s_k 和查询语义嵌入 s_q 的相似性得分, 并挑选出最相似的前 m 条知识 K_0 , 如式 (3) 所示。

$$\text{sim}(s_q, s_k) = \frac{\vec{s}_q \cdot \vec{s}_k}{\|\vec{s}_q\| \|\vec{s}_k\|} = \frac{\sum_{i=1}^d s_{q_i} s_{k_i}}{\sqrt{\sum_{i=1}^d s_{q_i}^2} \sqrt{\sum_{i=1}^d s_{k_i}^2}} \quad (3)$$

$$K_0 = \left\{ f_d(s_k) \mid \text{sim}(s_q, s_k) \in \text{Top}_m(\text{sim}(s_q, s_k)) \text{ and } s_k \in S_K \right\} \quad (4)$$

其中, d 为语义嵌入向量的维度, f_d 为取得知识嵌入向量对应的知识文本段。

由于知识库中可能包含多种不同的语言以及代码文本段, 得到前 m 条知识后, 本文使用预训练的多语言重排模型 (如 BGE-M3 系列模型) 计算每个知识文本段与用户查询文本的相关性得分 $f_r(q, k)$, 筛选出相关性得分大于预设阈值 cut_{off} 的前 n 条知识嵌入向量 T_1 , 随后获取知识向量对应的知识文本段 K_1 。通过引入重排序过程, 检索器可以通过 Transformer 架构进行深层语义建模, 进一步过滤无关知识, 用精准的检索结果增强生成, 弥补大模型的细节知识不足, 提升大模型的回答质量。

$$T_1 = \left\{ k \mid k \in K_0 \text{ and } f_r(q, k) \geq \text{cut}_{\text{off}} \text{ and } f_r(q, k) \in \text{Top}_n(f_r(q, k)) \right\} \quad (5)$$

$$K_1 = f_d(T_1) \quad (6)$$

最后, 本文使用用户查询 q 和相关知识 K_1 组成提示词, 将其输入大模型, 得到首轮回答, 记作 resp 。

2.3 后检索反思

得到大模型首轮生成结果后, 本文通过余弦相似度检索进行初步召回前 m 条相关知识 K_2 , 然后使用重排序模型进一步筛选获得知识向量 T_2 , 查找与首轮回答最相关的前 n 个知识段落 K_2 。

$$K_2 = \left\{ f_d(s_k) \mid \text{sim}(f_e(\text{resp}), s_k) \in \text{Top}_m(\text{sim}(f_e(\text{resp}), s_k)) \text{ and } s_k \in S_K \right\} \quad (7)$$

$$T_2 = \left\{ k \mid k \in K_2 \text{ and } f_r(q, k) \geq \text{cut}_{\text{off}} \text{ and } f_r(q, k) \in \text{Top}_n(f_r(q, k)) \right\} \quad (8)$$

$$K_3 = f_d(T_2) \quad (9)$$

最后, 本文将用户查询 q 、首轮回答 resp 和二轮知识 K_3 组成提示词, 将其输入大模型得到最终回答 answer 。因其融合了单轮检索的知识信息, 提示词中保留了模型的首轮回答 resp 。在模型生成答案过程中, 大语言模型可以根据二轮检索知识识别出首轮回答中的“幻觉”并进行更正, 模型的自我反思和答案生成同时进行。通过引入有检索的反思环节, 本文所提方法可以缓解大模型“幻觉”现象, 从而提高大模型最终回答的质量。

3 实验

3.1 数据源和生成模型

本文收集了来自 CS-Wiki、CSNotes

和 CS-Basic-SelfLearning 提供的大量 Markdown 文档构建外置知识库，以下称为 CS-Materials。CS-Materials 的数据构成如图 2、图 3 所示。CS-Wiki 是包含数据结构、计算机组成原理、操作系统和计算机网络 4 门课程知识的计算机通用基础知识百科；CSNotes 是包含常用算法知识文档、Java 基础知识、数据库原理、操作系统、计算机网络、设计模式、分布式系统等多个不同课程内容的知识学习笔记；CS-Basic-SelfLearning 是包含数据结构、计算机网络和操作系统知识的学习笔记。本文认为上述材料覆盖了计算机学习过程

中的众多学科，利用上述知识材料构建外置知识库可以弥补大模型的细节知识缺陷。此外，本文使用由 Open Phi 提供的 Programming_Books_llama 公开数据集构建外置知识库。Programming_Books_llama 包含约 11 万条数据，每条数据由几个知识点和相关英文教学材料的 Markdown 格式文本组成，覆盖了编译原理、算法、数据结构、图形学等计算机科学知识。本文选择 Programming_BOOKS_llama 作为英文知识外置知识库进行对比实验。

本文使用的预训练语言模型为 BGE-M3，使用的预训练重排模型为 BGE-Reranker-M3^[19]。BGE-M3 和 BGE-Reranker-M3 均支持符号长度为 8 192 个的输入序列，这使 BGE 模型适用于长文本教学材料的语义嵌入过程。BGE-M3 和 BGE-Reranker-M3 均为多语言模型，在多个语言测试任务中均有良好的性能表现，这使 BGE 模型适用于不同语言的外置知识库 CS-Materials 和 Programming_Books_llama。

本文分别使用闭源模型和开源模型进行实验。对于闭源生成模型，本文选择 GPT-4o-mini^[1]进行实验；对于开源生成模型，本文选择 GLM-4-9B^[31]、LLama-3.1-7B^[2]和 Qwen2.5-7B^[12]进行实验。

3.2 评估指标

CS-Bench 是一个用于评估计算机科学领域大模型性能的基准，包括来自数据结构、计算机组成原理、操作系统和计算机网络 4 门课程的测试题^[32]。本文将 CS-Bench 中提供的中文选择题和判断题作为评估数据，将答题准确率作为模型回答质量的评价指标。测试数据的数据统计如图 4、图 5 所示，测试数据样例见表 1。

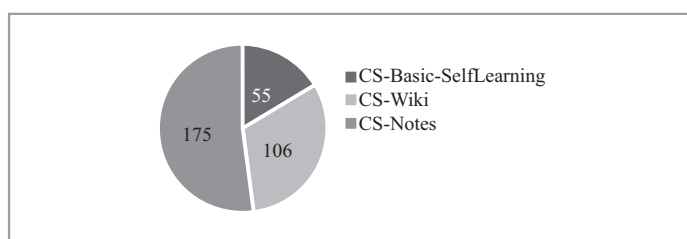


图2 CS-Materials中不同数据源的文档个数

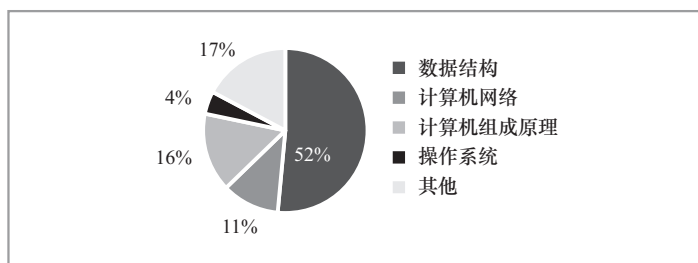


图3 CS-Materials中不同科目知识的占比

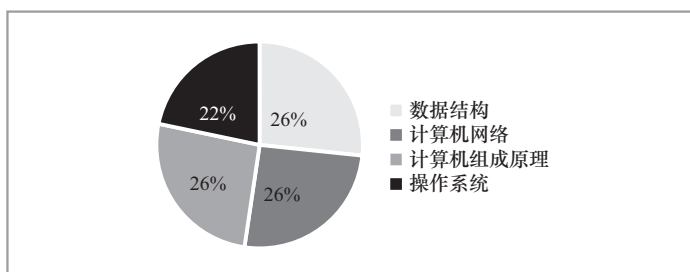


图4 CS-Bench中测试数据包含的不同科目的占比

3.3 实验环境

本文通过 Python 语言中 OpenAI 包调用应用程序编程接口使用 GPT-4o-mini, 并通过 Transformers 包下载其他开源模型进行实验。实验在初轮生成及后检索反思过程中使用的提示词格式见表 2。对于 CS-Materials 知识库, 单条知识文本长度较短, 上下文窗口内可以包含较多的知识文本数, 因此本文设置参数粗粒度召回目标数 $m=200$ 、重排序召回目标数 $n=8$; 对于 Programming_Books_llama 知识库, 单条知识文本较长而大模型的上下文窗口有限, 为了避免过长的外部知识超出上下文窗口, 本文设置参数 $m=100$ 、 $n=5$, 限制单次生成中最多使用的知识文本个数。此外, 本文设置 $\text{cut}_{\text{off}}=0.3$, 在过滤无关材料的同时避免长文本块难以召回。为了减少模型幻觉, 本文将生成模型温度均设定为 0。

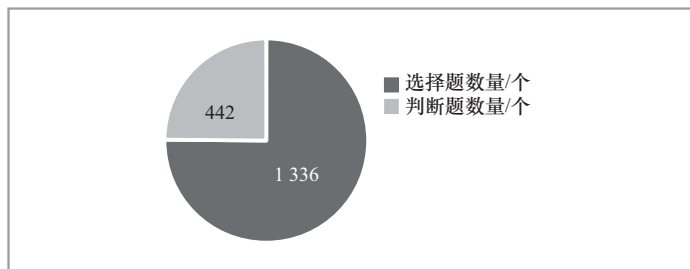


图5 CS-Bench中选择题与判断题的数量

3.4 实验结果

本文使用相同的外置知识库分别用不同生成模型在 CS-Bench 上测试, 测试结果见表 3。其中, 原模型表示直接使用生成模型进行问答, w/o Reflection 表示不使用后检索反思, 仅使用引入重排序检索的初轮生成结果进行问答。为了进一步验证所提方法的有效性, 本文保持生成模型不变, 分别用 CS_Materials 和 Programming_Books_llama 作为外置知识库进行

表1 使用的测试数据样例

选择题	判断题
问题: 数据对象在数据结构中指的是什么? A: 具有相同性质的数据元素的集合, 是数据的一个子集 B: 计算机程序中处理的主要数据源 C: 数据存储的主体 D: 数据库管理系统中的基本组成部分 回答: A	问题: 存储字长指存放在一个存储单元中的二进制代码组合 回答: 错误

表2 初轮生成及后检索反思过程中使用的提示词格式

过程	初轮生成	后检索反思
提示词	你是一个计算机领域的专家, 擅长回答计算机相关问题。接下来你会得到一些与问题相关的辅助知识, 请帮助回答问题 ---问题:< q >; ---参考资料:< K ₁ >	你是一个计算机专家, 擅长根据问题和提供的辅助材料, 评估答案是否正确。接下来你会得到问题、相关回答以及辅助材料, 请根据辅助材料对答案进行修改 ---问题:< q >; ---对应答案< resp >; ---评估资料:< K ₃ >

表3 使用相同的外置知识库分别用不同生成模型在CS-Bench上测试的结果

生成模型	原模型答题正确率	w/o Reflection	所提方法答题正确率
GPT-4o-mini	68.90%	72.44%	76.89%
GLM-4-9B	68.34%	71.61%	75.09%
Qwen2.5 7B	73.50%	74.91%	76.94%
LLama-3.1-7B	54.19%	54.19%	61.21%

表4 使用不同外置知识库的测试结果

外置知识库	CS_Materials		Programming_Books_llama	
生成模型	GPT-4o-mini	GLM-4-9B	GPT-4o-mini	GLM-4-9B
原模型答题正确率	68.90%	68.34%	67.88%	66.93%
所提方法答题正确率	76.88%	75.09%	72.61%	70.81%

测试，实验结果见表4。为了评估所提方法与现有工作的差异，本文选取基于传统RAG技术的Chated^[8]作为对比基线，且为保证实验的公平性，在Chated的实现中使用与所提方法相同的CS-Materials作为外置知识库，并使用GPT-4o-mini作为生成模型。两种方法在CS-Bench上的测试

结果见表5。此外，本文对比了所提方法与传统RAG技术的生成内容，对比结果见表6。

由表3、表4可知，所提方法与不同的生成模型相结合时均可以令大模型回答准确率提升5%左右。由表5、表6可知，所提方法相比基于传统RAG技术的问答增强方法有更好的性能提升，在GPT-4o-mini生成模型上回答准确率提升了6.02%。其原因为外置知识库补充了大模型的知识细节，重排序对知识库信息进行了精细过滤，减少了无关材料对模型生成的干扰，后检索反思根据知识库提供的辅助材料，修正了生成过程中出现的部分“幻觉”，从而提

表5 所提方法与Chated的答题情况对比结果

使用方法	答题正确条数/条	答题错误条数/条	正确率
Chated	1 260	518	70.87%
所提方法	1 367	411	76.89%

表6 所提方法与传统RAG技术生成内容的对比

问题描述	传统RAG技术	所提方法
当两个顺序栈共享同一个一维数组空间时，它们的栈底和栈顶分别如何设置？ A: 两个栈底在同一端，栈顶向相反方向延伸 B: 两个栈顶在同一端，栈底向相反方向延伸 C: 栈底和栈顶都在同一端 D: 两个栈底在两端，栈顶向中间延伸	当两个顺序栈共享同一个一维数组空间时，正确的设置是：两个栈底在同一端，栈顶向相反方向延伸。因此，答案为A	根据辅助材料中的信息，两个顺序栈共享同一个一维数组空间时，其栈底和栈顶应该是：两个栈底在数组的两端，栈顶指针向中间延伸 选项A是不正确的，虽然它描述了一种栈的生长方式，但并不符合共享栈的配置方式。根据共享栈的描述，栈底应当分布在两端，而栈顶则向中间生长。正确答案应为D

升了大模型回答的准确性。本文还发现，在一些生成模型上仅使用单轮 RAG 技术并不能明显提升大模型回答准确度，这可能是由于检索知识结果不够精确，仍存在无关知识材料。因此，为了进一步提升大模型准确率，研究者可以考虑更换性能更强的语义嵌入和重排序模型。此外，使用不同外置知识库所带来的模型回答质量提升幅度存在明显差异，这表明外置知识库的质量直接影响大模型的回答效果，为了使模型回答准确率继续提高，研究者可以进一步补充和筛选外置知识库的内容，提高模型回答质量。

为了检验单轮后检索反思是否充分，本文进行了多轮后检索反思的对比实验。本文使用 Programming_Books_llama 作为外置知识库，使用 GLM-4-9B 作为生成模型进行实验，统计了从用户提问至大模型给出答案的时间，实验结果见表 7。

由表 7 可知，在多轮后检索反思的情况下，大模型回答准确率没有明显提升，但反思次数的增加会显著影响推理时间，降低大模型答案生成的效率。本文认为，在不同轮次后检索中使用了相同的知识库、检索模型和重排序模型，导致得到的知识重合度较高，难以辅助大模型自我纠错和增强生成，因此大模型准确率无明显提升。

本文还针对超参数 cut_{off} 的设置进行了探究实验。该参数的作用是限制检索材料

表 7 GLM-4-9B 经过多轮后检索反思的知识问答正确率及平均推理时间

后检索轮数	正确率	平均推理时间/s
$t=1$ (所提方法)	71.34%	31.7
$t=2$	71.00%	48.7
$t=3$	71.17%	71.4

的最低相关度，相关性得分低于 cut_{off} 的检索材料会被直接丢弃，避免不相关材料对模型生成造成干扰。本文从 CS-Bench 中采样了 300 条测试数据，使用 CS-Materials 和 Programming_Books_llama 作为外置知识库，并在此基础上分别测试了 cut_{off} 为 0.1、0.3 和 0.5 时模型生成时的平均召回数，生成模型均采用 GPT-4o-mini，实验结果见表 8。

表 8 中在 CS-Materials 的实验结果显示，当 cut_{off} 为 0.1 时，平均召回文本数量与检索返回数量接近，而当 cut_{off} 达到 0.3 和 0.5 时，召回文本数得到了控制。这表明，如果相关性阈值设定过低，会使过滤无关文本材料的能力下降。此外，在 Programming_Book_llama 上的实验中， cut_{off} 达到 0.5 时，召回文本数过少，这是因为知识文本块较长时，即使语义相关，检索材料与用户查询的相关度仍会有一定程度的下降，过高的相关性阈值会使长知识块难以召回。因此，设置合适的 cut_{off} 值可以提升大模型检索增强生成的效果，本文中令 $\text{cut}_{\text{off}}=0.3$ 。

表 8 不同 cut_{off} 下大模型生成答案的平均召回数

外置知识库	Programming_Books_llama($n=5$)	CS-Materials($n=8$)
$\text{cut}_{\text{off}}=0.1$ 时召回条目数/条	3.92	7.22
$\text{cut}_{\text{off}}=0.3$ 时召回条目数/条	2.88	5.75
$\text{cut}_{\text{off}}=0.5$ 时召回条目数/条	2.04	4.64

4 结束语

为了更好地将大模型应用于计算机教育问答过程，本文首先构建了计算机相关外部知识库，通过RAG技术缓解大模型在特定领域细节知识的不足以及“幻觉”现象；其次，为了缓解传统RAG技术中无关检索结果对生成模型问答的负面影响，使用了重排序过滤无关检索材料，使检索结果更加精准；最后，为了进一步提升大模型生成答案的质量，引入了后检索反思，增强大模型自我修正答案的能力。本文提出的基于重排序和后检索反思的教育大模型问答增强方法在CS-Bench数据集的实验中取得了良好的结果，在使用多个不同生成模型时均使大模型回答正确率提升了5%左右，缓解了“幻觉”问题，并使大模型在计算机问答中更准确。此外，使用不同数据源构建不同外置知识库的对比实验表明，本文所提方法能有效提升大模型在计算机问答中的正确率。后检索反思次数对模型回答准确率和推理时间的影响的实验结果表明，多轮后检索反思并未显著提升回答准确率，但明显延长了推理时间。本文的下一步工作是尝试在更小规模的生成模型上应用RAG技术，并进行改进。

参考文献：

[1] OpenAI, ACHIAM J, ADLER S, et al. GPT-4 technical report[EB]. arXiv preprint, 2024, arXiv: 2303.08774.

[2] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The llama 3 herd of models[EB]. arXiv preprint, 2024, arXiv: 2407.21783.

[3] DAN Y H, LEI Z K, GU Y Y, et al. Edu -

Chat: a large-scale language model-based chatbot system for intelligent education[EB]. arXiv preprint, 2023, arXiv: 2308.02773.

[4] WANG S, XU T L, LI H, et al. Large language models for education: a survey and outlook[EB]. arXiv preprint, 2024, arXiv: 2403.18105.

[5] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey[EB]. arXiv preprint, 2023, arXiv: 2312.10997.

[6] JI Z W, YU T Z, XU Y, et al. Towards mitigating LLM hallucination via self reflection[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL, 2023: 1827-1843.

[7] DHULIAWALA S, KOMEILI M, XU J, et al. Chain-of-verification reduces hallucination in large language models[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 3563-3578.

[8] WANG K, RAMOS J, LAWRENCE R. Chated: a chatbot leveraging ChatGPT for an enhanced learning experience in higher education[C]//Proceedings of the INTED Proceedings. Valencia: IATED, 2024: 6580-6589.

[9] LIU S Q, YU Z Z, HUANG F R, et al. Can small language models with retrieval-augmented generation replace large language models when learning computer science? [C]//Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1. New York: ACM, 2024: 388-393.

[10] NEUPANE S, HOSSAIN E, KEITH J, et al. From questions to insightful answers: building an informed chatbot for university resources[C]//Proceedings of the

- 2024 IEEE Frontiers in Education Conference (FIE). Piscataway: IEEE Press, 2024: 1–9.
- [11] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[EB]. arXiv preprint, 2005, arXiv: 2005.11401.
- [12] YANG A, YANG B S, HUI B Y, et al. Qwen2 technical report[EB]. arXiv preprint, 2024, arXiv: 2407.10671.
- [13] XI Z H, CHEN W X, GUO X, et al. The rise and potential of large language model based agents: a survey[J]. Science China Information Sciences, 2025, 68(2): 121101.
- [14] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1–45.
- [15] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[EB]. arXiv preprint, 2020, arxiv:2001.08361.
- [16] 程大伟, 贾仁军, 李江彤, 等. 知识增强的中文金融大模型研究[J]. 大数据, 2025, 11(2): 5–18.
- CHENG D W, JIA R J, LI J T, et al. Research on knowledge-augmented Chinese financial large language model[J]. Big Data Research, 2025, 11(2): 5–18.
- [17] 王敏, 高晓影, 汪诗蕊, 等. 基于知识图谱和大语言模型增强的推荐系统研究[J]. 大数据, 2025, 11(2): 29–46.
- WANG M, GAO X Y, WANG S R, et al. Research on recommendation system based on knowledge graph and large language model enhancement[J]. Big Data Research, 2025, 11(2): 29–46.
- [18] 乐洁玉, 罗超洋, 丁静姝, 等. 教育大数据隐私保护机制与技术研究[J]. 大数据, 2020, 6(6): 52–63.
- LE J Y, LUO C Y, DING J S, et al. Research on privacy protection mechanism and technology of educational big data [J]. Big Data Research, 2020(6): 52–63.
- [19] CHEN J L, XIAO S T, ZHANG P T, et al. BGE M3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation[EB]. arXiv preprint, 2024, arxiv: 2402.03216.
- [20] LI C F, LIU Z, XIAO S T, et al. Making large language models a better foundation for dense retrieval[EB]. arXiv preprint, 2023, arXiv: 2312.15503.
- [21] NOURIINANLOO B, LAMOTHE M. Re-ranking step by step: investigating pre-filtering for re-ranking with large language models[EB]. arXiv preprint, 2024, arXiv: 2406.18740.
- [22] ZHANG X, ZHANG Y Z, LONG D K, et al. mGTE: generalized long-context text representation and reranking models for multilingual text retrieval[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Stroudsburg: ACL, 2024: 1393–1412.
- [23] PRIHAR E, LEE M, HOPMAN M, et al. Comparing different approaches to generating mathematics explanations using large language models[M]//Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. Cham: Springer Nature Switzerland, 2023: 290–295.
- [24] FAN Y X, JIANG F, LI P F, et al. GrammarGPT: exploring open-source LLMs for Native Chinese grammatical error correction with Supervised fine-tuning [C]//Proceedings of the Natural Lan-

guage Processing and Chinese Computing. Cham: Springer, 2023: 69–80.

[25] KORAISHI O. Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment[J]. Language Education and Technology, 2023, 3(1): 55–57.

[26] JURY B, LORUSSO A, LEINONEN J, et al. Evaluating LLM-generated worked examples in an introductory programming course[C]//Proceedings of the 26th Australasian Computing Education Conference. New York: ACM, 2024: 77–86.

[27] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1–38.

[28] PASQUARELLI L, KOUTCHEME C, HELLAS A. Comparing the utility, preference, and performance of course material search functionality and retrieval-augmented generation large language model (RAG-LLM) AI chatbots in information-seeking tasks[EB]. arXiv preprint, 2024, arXiv: 2410.13326.

[29] POLIAKOV M, SHVAI N. Multi-meta-RAG: improving RAG for multi-hop queries using database filtering with LLM-extracted metadata[M]//Information and Communication Technologies in Education, Research, and Industrial Applications. Cham: Springer Nature Switzerland, 2025: 334–342.

[30] 卢宇, 余京蕾, 陈鹏鹤. 基于大模型的教学智能体构建与应用研究[J]. 中国电化教育, 2024(7): 99–108.

LU Y, YU J L, CHEN P H. Design and application of pedagogical agent with foundation model[J]. China Educational Technology, 2024(7): 99–108.

[31] GLM T, ZENG A H, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[EB]. arXiv preprint, 2024, arXiv: 2406.12793.

[32] SONG X S, DIAO M X, DONG G T, et al. CS-bench: a comprehensive benchmark for large language models towards computer science mastery[EB]. arXiv preprint, 2024, arXiv preprint: 2406.08587.

作者简介



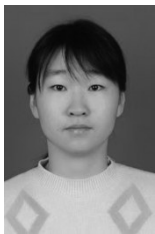
孙浩然（2002–），男，同济大学计算机科学与技术学院（国家特色化示范性软件学院）硕士生，主要研究方向为数据挖掘、智能运维。



王志豪（1998–），男，同济大学计算机科学与技术学院（国家特色化示范性软件学院）博士生，主要研究方向为情感分析、多模态大语言模型。



吴一帆（2002-），男，同济大学计算机科学与技术学院（国家特色化示范性软件学院）硕士生，主要研究方向为事件抽取、自然语言处理。



高晓影（1996-），女，同济大学计算机科学与技术学院（国家特色化示范性软件学院）博士生，主要研究方向为推荐系统、图神经网络。



向阳（1962-），男，博士，同济大学计算机科学与技术学院（国家特色化示范性软件学院）教授，主要研究方向为数据挖掘、自然语言处理、知识图谱。

收稿日期: 2025-01-02
通信作者: 向阳, shxiangyang@tongji.edu.cn
基金项目: 国家自然科学基金项目(No.72071145)
Foundation Item: The National Natural Science Foundation of China (No.72071145)