

基于知识增强大语言模型的历史学科试题生成系统

纪天昀¹, 张征¹, 赵宇泽¹, 黄振亚¹, 黄威², 佟威², 刘淇¹, 陈恩红¹

1. 中国科学技术大学认知智能全国重点实验室, 安徽 合肥 230026;
2. 教育部教育考试院, 北京 100084

摘要

随着大语言模型的问世, 其凭借强大的语言能力和推理能力能够模仿教师的题目设计方法, 对出题材料进行分析并生成相应题目, 并通过自我检查保证生成题目的质量。受此启发, 设计了一个基于知识图谱增强的大语言模型历史学科试题生成系统 KG-QGH, 旨在生成符合教师需求、内容准确、匹配考纲且丰富多样的题目。该系统以历史学科的简答题为例, 将历史知识图谱中的关系信息与课本知识点相结合, 自动生成简答题目, 并引入反馈机制对生成的题目进行验证和修正。对比实验评估结果表明, KG-QGH 系统在题目生成的正确性、相关性和多样性等关键指标上均显著优于单一的大语言模型方法。本研究不仅验证了知识图谱增强方法在题目生成领域的有效性, 也为教育领域的智能化应用提供了新的研究思路和实践参考。

关键词

大语言模型; 知识图谱; 题目生成; 历史教育; 智能教育

中图分类号: G434

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025059

Knowledge-enhanced large language model based history subject exam question generation system

Ji Tianyun¹, Zhang Zheng¹, Zhao Yuze¹, Huang Zhenya¹, Huang Wei², Tong Wei², Liu Qi¹, Chen Enhong¹

1. State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei 230026, China
2. National Education Examinations Authority, Beijing 100084, China

Abstract

With the advent of large language models, their powerful language and reasoning abilities can mimic the question design methods of teachers, analyze the materials for questions, generate corresponding questions, and ensure the quality of the generated questions through self-checking. Inspired by this, a knowledge graph-enhanced large language model historical question generation system (KG-QGH) was designed, aiming to generate questions that meet teachers' needs, are content in accurate, aligned with the syllabus, and diverse in form. Taking short answer questions in the history subject as an example, the system automatically generated questions by integrating relational information from the historical knowledge graph with textbook knowledge points, a feedback mechanism was introduced to verify and

correct the generated questions. The comparative experimental evaluations was demonstrated that the KG-QGH system significantly outperformed the single large language model approach in key metrics such as correctness, relevance, and diversity of question generation. This study not only validated the effectiveness of the knowledge graph enhancement method in the field of question generation but also offered new research insights and practical references for intelligent applications in the educational domain.

Key words

large language model, knowledge graph, question generation, history education, intelligence education

0 引言

近年来，随着深度学习技术的突破性发展，自然语言处理领域取得了显著进展，作为其中代表性成果之一的大语言模型已在多个领域得到广泛应用^[1]。教育领域作为社会发展的关键环节，已多角度、多层地面地与大语言模型相结合^[2]。相较于传统语言模型，大语言模型在文本生成质量、语义理解等方面具有显著优势，这使其在题目生成、作业批改等教育场景中展现出独特价值。有了大语言模型的辅助，教师可以在习题设计上投入更少时间，同时获得更多灵感，进而创作出更高质量的教学材料^[3]。

在传统的历史学科出题过程中，教师首先需要对出题所用的史料进行分析，并依托自身的知识储备以及参考书籍进行联想，以确定合适的考点。**图1**为历史简答题示意图，教师在设计该试题时，首先根据材料确定了“孔子”这一考点。随后，教师结合考点与史料的关联，并参考学生的知识水平，设计出能够考查学生史料分析问题能力的问题。例如，教师结合君臣关系和儒学地位的变化，设计问题考查学生分析材料并解析儒学地位变化的原因。然而，这一联想考点的过程往往需要消耗大量的精力。教师需要手动查询大量的史料，并进行筛选，才能确定适合出题的考点。即便教师具备丰富的历史知识，其联想能力仍可能受到限制，难以兼顾所有可能的考点。因此，教师在命题过程中迫切需要借

儒家学说的道德准则和文学、思想方面的遗产，主要由后人修改和润色**孔子**的教诲组成。阅读材料，回答问题。

材料一

君君、臣臣、父父、子子。

——《论语》

材料二

君有过则谏，反复之而不听。则去……君之视臣如土芥，则臣视君如寇仇。

——《孟子》

材料三

唯天子受命于天，天下全受命于天子……君为臣纲。

——《春秋繁露》

问题：

(1)据材料一、二，指出孔子和孟子关于**君臣关系**的主张。

(2)结合所学知识，指出西汉以后**儒学地位**发生的变化。综合上述材料并结合所学知识，概括其地位变化的主要原因。

图1 历史简答题示意图

助大语言模型等人工智能技术，自动生成题目，为教师设计题目提供辅助^[4]，从而提升出题效率和质量。

为了保证生成的题目能够正确地起到辅助作用，试题生成系统的输出需要满足如下3个标准^[5]。

- 题目内容正确。生成的题目不应存在知识性错误，避免为教师提供错误信息。同时，生成的题目应符合历史学科的题目考查要求，能够为教师提供参考。
- 题目与考点相关。生成的题目应符合高中历史的考纲。在题目设计过程中，教师需确保题目知识点在高中学纲之内，从而保证学生能够顺利地作答。
- 题目丰富多样。系统生成的题目应足够多样，以达到为教师提供灵感的目的。

虽然当前的大语言模型有强大的题目生成能力，但是直接应用大语言模型命题存在诸多不足，主要体现在以下3个方面。

- 由于缺乏专业的知识储备，大语言模型在生成专业内容时可能出现错误。并且，在面对不了解的历史事件时，大语言模型可能会凭空编造、产生“幻觉”^[6]，导致题目包含虚假信息，无法保证生成结果的正确性。

- 大语言模型不了解高中历史考纲，无法保证生成题目涉及的知识在考纲要求范围内。

- 大语言模型回复的多样性不足。如图2（a）所示，当多次用相同的提示词生成题目时，其回复都只围绕“君臣关系”和“道德”展开，重复度较高。即针对某一特定主题，大语言模型难以生成足够丰富的考点，无法有效地为教师提供足够广泛的参考。

为了解决上述问题，本文结合知识图谱^[7]与大语言模型设计出题算法，并搭建出题系统进行历史学科简答题生成。通过引入知识图谱中丰富、专业的历史关系信息，增强大语言模型的联想能力，在弥补大模型知识储备不足的同时增加生成结果的多样性，从而解决以上不足，自动生成内容正确、符合考纲且丰富多样的题目。如图2（b）所示，实体“孔子”与“儒家”“《论语》”“董仲舒”“《春秋》”“李贽”等相关联。通过向大语言模型补充上述的关联实体，可增强大语言模型自身的联想能力，从而实现为教师联想知识点提供灵感的目的。具体来说，首先，本文利用维基百科的页面跳转数据库构建知识



图2 大模型题目生成示例

图谱，最终形成了一个包含超过千万个词条，其中与历史学科相关词条超过百万的知识图谱。其次，本文提出了知识图谱增强的大语言模型历史学科试题生成系统(knowledge graph based question generation system for history subject, KG-QGH)，系统包括知识图谱检索模块和题目生成模块。知识图谱检索模块根据用户输入的材料段落在知识图谱中检索相关关键词，而题目生成模块则基于课本知识点及知识图谱的检索结果初步生成题目。系统通过引入知识图谱中丰富的历史关系数据保证生成题目的多样性。为确保题目的正确性，系统设有反馈机制(refine, RF)^[8]对生成的题目进行质量检查，确保其正确、符合质量标准。此外，系统引入课本知识点，将课本知识点与输入材料相结合，以保证生成结果与课本知识紧密相关。

为保证系统输出结果符合教师需求，本文从正确性、相关性和多样性3个维度对系统性能进行评估。其中，正确性检验生成题目的内容是否准确，是否存在常识性错误；相关性评估题目与课本知识点的关联程度及其在考纲范围内的适宜性；多样性则关注在使用同一材料进行题目生成时，生成内容的丰富性。

本文的主要贡献如下。

- 本文根据维基百科的PageLink数据构建千万级别的中文历史知识图谱并进行开源，辅助系统进行题目生成。
- 本文设计KG-QGH，综合知识图谱和课本知识点进行题目生成，并通过反馈机制检验生成结果的质量。
- 本文使用真实试题数据进行实验，从正确性、相关性和多样性3个角度检验生成结果，验证模型性能的有效性。
- 本文构建在线出题系统，辅助教师设计题目。系统经教师实际体验，取得

了良好效果。

1 相关工作

1.1 大语言模型

自然语言处理^[9]问题的研究有悠久的历史。早在20世纪，研究者们便开始尝试利用计算技术模拟语言生成与对话系统，这为教育技术领域的智能化发展奠定了基础。近年来，得益于计算能力和算法模型的飞速发展，自然语言处理研究取得了重大突破。其中，Transformer架构^[10]的提出，以及以GPT家族^[11]为代表的大语言模型的快速发展，标志着语言技术从传统规则驱动向深度学习驱动的转变。尤其是ChatGPT^[12]等先进的大语言模型，展现出超越预期的语言理解与生成能力，能够支持多样且复杂的任务，在教育领域的应用前景更是广泛而深远。

在教育场景中，大语言模型正在逐步融入教学实践^[13]，推动个性化学习与智能教育的深入发展^[14]。例如，大语言模型可用于智能求解题目、自动评估学生作业、生成针对性学习材料，甚至作为“智能助教”回答学生问题。利用这些应用程序或模型，教育工作者可以大幅减轻重复性工作负担，同时为学生提供更高效、更个性化的学习支持^[15]。

本文聚焦于大语言模型在试题生成中的具体应用，提出并设计了一个基于大语言模型的框架，旨在实现历史学科简答题的自动化生成，为教育工作者提供高效的辅助工具。这一探索不仅为学科教学创新提供了技术支持，也为智能教育研究注入了新的思路^[16]。

1.2 知识图谱

知识图谱 (knowledge graph, KG)^[7] 是一种用于存储和组织结构化信息的技术, 通常采用图的形式表示, 记作 $G=(V,E)$, 其中 V 表示图中的顶点, 代表不同的实体; E 则表示实体之间的关系, 体现实体间的相互联系。以 ConceptNet^[17] 为例, 标准的知识图谱通常由三元组 (e_h, r, e_t) 组成, 其中 e_h 表示头实体, r 表示关系, e_t 表示尾实体。这些三元组构成了图谱中各个实体及其关系的基础。

通过对知识图谱中的数据进行检索, 教育领域可以获得丰富的关系信息^[18], 从而为学习者提供更精准的知识查询与推理支持^[19]。随着大语言模型的快速发展, 越来越多的研究开始探索将大语言模型与知识图谱结合^[20], 以弥补传统语言模型在知识存储和推理上的局限性。传统的语言模型通常将知识图谱的检索结果与文本信息一起编码输入, 作为增强模型性能的手段。例如, David 等^[21]提出的方法中, 将从知识图谱检索的知识与文档编码拼接后输入语言模型, 帮助语言模型获取外部知识, 提升其理解与推理能力。

随着大语言模型的兴起, 研究者们逐渐采用更加灵活的方式, 将知识图谱的信息通过提示词 (prompt) 的形式输入大语言模型, 从而增强其知识基础, 缓解模型生成时的“幻觉”现象, 并提升模型在特定领域, 尤其是教育领域中的应用效果。例如, 通过输入包含学科知识的三元组信息, 大语言模型能够在自动化试题生成、智能答疑等任务中, 表现得更精准和富有逻辑性。

1.3 试题生成

已有的试题生成研究主要集中在使

语言模型生成数学题或问答题。一方面, Agrawal 等^[22]提出的 CyberQ 模型, 将大语言模型与知识图谱结合, 用于生成网络安全领域的问答题。该模型从知识图谱中检索相关三元组, 并将其以提示词的形式传递给大语言模型, 不仅增强了模型的知识基础, 还显著减少了“幻觉”现象。另一方面, Wei 等^[23]采用 Memory Retrieval 机制辅助语言模型生成数学简答题, 将检索到的逻辑描述和场景描述补充至语言模型的生成过程, 有效提升了模型的生成准确性与相关性。

与数学应用题^[24-25]和问答题^[26-27]不同, 历史简答题侧重于考查学生对历史事件或事物间联系的分析能力, 这对生成模型提出了更高的要求。现有的研究方法大多聚焦于基于逻辑推理或领域知识的补充生成, 无法很好地满足历史学科试题生成的需求。为此, 本文通过深入分析历史简答题的特征, 创新性地从知识图谱中的关系信息与历史学科知识相结合, 设计了一种新的生成方法。该方法通过知识图谱辅助大语言模型提取历史事件间的联系, 为生成质量更高、更符合教学需求的历史简答题提供了技术支撑。

2 历史学科试题生成系统关键技术

为模拟教师真实的出题过程, 系统分别设计了知识图谱检索模块和题目生成模块。知识图谱检索模块根据输入的材料找出相关的历史知识点, 从而得以在知识图谱上进行检索。随后, 题目生成模块根据知识点信息与知识图谱检索结果进行题目生成。系统会使用反馈机制检验初步生成的题目, 最终输出检验合格的生成结果。

本节将首先详细介绍知识图谱的构建过程，随后深入分析系统的整体结构设计，以阐明各个模块之间的相互作用及其对提升历史学习效果的重要性。

2.1 问题定义

系统的问题定义如下：用户向系统输入多段材料 $T = \{T_1, T_2, \dots, T_n\}$ ，其中， T_i 代表第 i 段材料。对于每一个输入的材料段 T_i ，系统会根据 T_i 设计相应的历史简答题 Q_i 。

为了保证 Q_i 的生成满足内容正确、考点相关以及题目多样的要求，在生成过程中系统会根据 T_i 在 KG 上进行检索，并结

合课本知识点 K 来进行问题设计。同时， Q_i 还会经过 RF 的检验，并将检验后的结果 Q'_i 进行输出。

2.2 算法设计

在接收到用户输入的提示词后，知识图谱检索模块会根据用户输入的材料进行检索，获得用于生成题目的相关考点。**图 3** 是算法的结构示意图。例如，用户向算法输入了一段关于孔子的材料，知识图谱检索模块根据输入的材料识别出课本中对应的知识点“中国哲学思想”“春秋社会变革”等，并根据知识点生成相应查询词“论语”“仁”等。模块使用检索词在知识

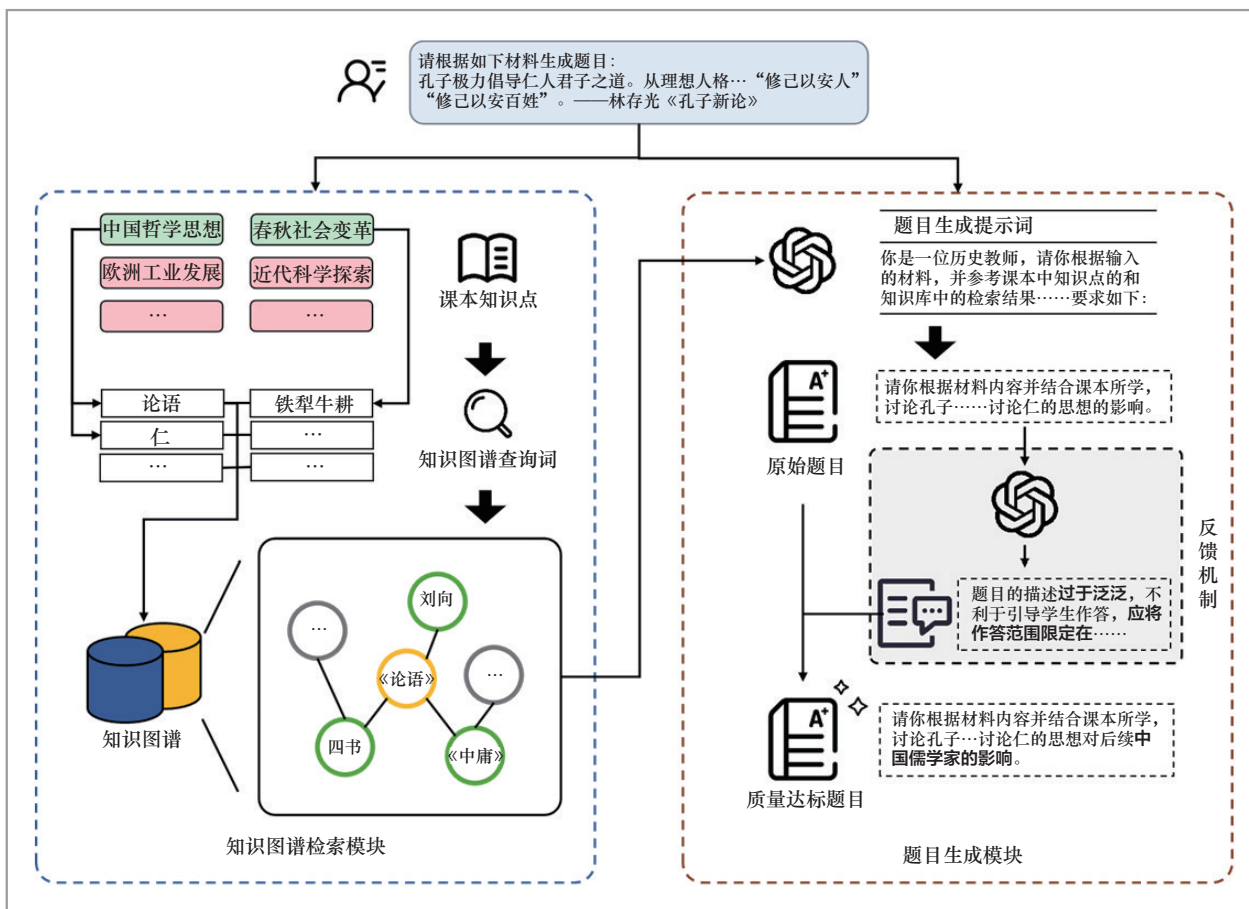


图3 算法结构示意图

图谱中进行检索，得到了“刘向”“四书”等查询结果。随后，题目生成模块会使用检索结果以及课本知识点进行题目的生成。最后，反馈机制会对初步生成的题目进行检验，确保生成的题目正确且与考点相关，如图3题目生成模块部分，反馈机制判定生成题目的作答范围过于宽泛，并对原题目进行修改，限定题目作答范围在“后续中国儒学家”并进行输出。

下面将具体介绍知识图谱的构建过程和知识图谱是如何与提示词、反馈机制进行结合的。

2.2.1 知识图谱的构建

目前常见的开源知识图谱包括 ConceptNet、YAGO^[28] 以及 DBpedia^[29] 等，均为英文知识图谱。全历史网站提供了可供查询的历史知识图谱，但是并未进行开源。由于缺乏历史学科的公开知识图谱，本文利用中文维基百科的 PageLink 数据手动创建知识图谱。PageLink 数据记录了维基百科各页面之间的跳转信息，其数据由实体对构成，实体对中的实体包括跳转链接的源页面和目标页面。这些页面跳转记录反映了源实体与目标实体之间的某种内在联系。因此，利用 PageLink

数据集构建知识图谱，可以为系统提供历史学领域重要的关系数据，从而增强题目生成。

知识图谱的构建流程如图4所示。首先，下载并读取维基百科公开的数据库文件，并依据数据库中的“源实体ID-目标实体ID”和“实体ID-实体名称”2张表将实体的ID与名称进行对应，形成“源实体-目标实体表”。随后，对该表进行整理和筛选得到“关键词-值关系对”，并将该键值对转为 csv 格式的表格文件进行储存。基于维基百科的定期数据更新，通过自动化上述构建流程知识图谱可以不断地根据维基百科中的数据进行更新，保证了知识图谱数据能够持续得到改进。

在构建的知识图谱中，数据结构由“关键词”和“值”2列组成。其中，“关键词”列代表源页面实体名称，“值”列代表跳转页面的实体名称。例如，关键词“孔子”，其对应的值包括“儒家”“庄子”等。鉴于维基百科庞大的知识体量，最终构建的知识图谱包含千万条涵盖各个领域的信息记录，见表1。在实际应用中，系统对图谱中的关键词进行了筛选，保留了与历史相关度较高的 300 万余条数据，表2中给出了部分数据示例。

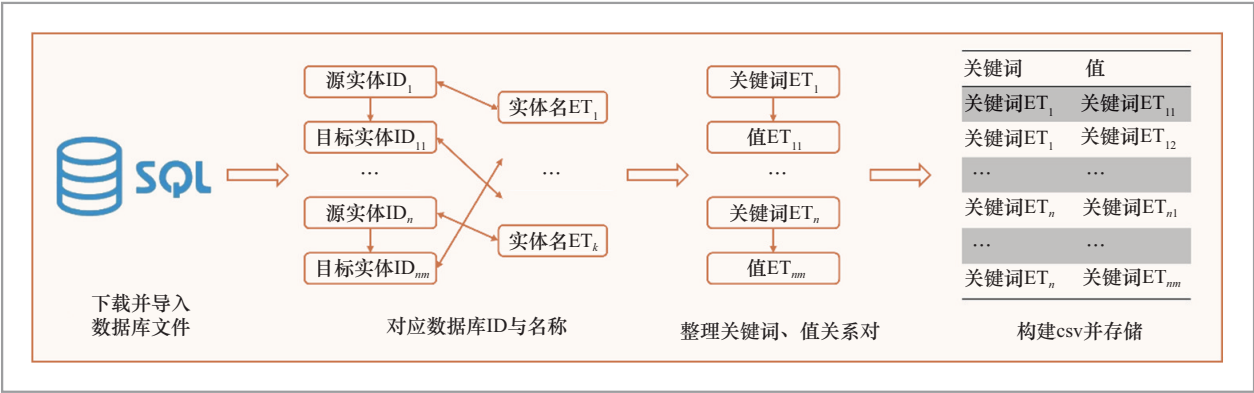


图4 知识图谱的构建流程

2.2.2 知识图谱检索模块

知识图谱检索模块利用知识图谱中的丰富历史关系信息对大语言模型进行增强，补齐大语言模型的专业知识并使生成的题目更加多样，满足教师的使用需求。试题生成系统以历史材料段为输入，而为了利用知识图谱中的信息，系统首先要根据输入的材料获得检索关键词。因此，知识图谱检索模块首先根据用户输入的材料生成相应的查询关键词。对于输入的材料 T_i 和教材知识点 $K_1, \dots, K_n$ ，具体的做法为首先令大语言模型根据输入的材料 T_i 从 $K_1, \dots, K_n$ 中找到对应的学科知识点 K_i 。例如，针对一段讲述孔子经历的输入，知识图谱检索模块可能会从课本知识点中找到“中国古代思想”“儒家思想”等知识点。在获得了学科知识点后，为了结合材料与课本知识点获得知识图谱的检索关键词，知识图谱检索模块将根据知识点 K_i 联想出与知识点相关的关键词 k_i 。例如，对于知识点“儒家思想”，相关的关键词便有“董仲舒”“李贽”等。相关关键词会作为检索关键词在KG中进行检索，检索结果 $kw_{i1}, \dots, kw_{in}$ 会与材料共同输入题目生成模块，用于向大语言模型提供可供参考的考查内容。知识图谱检索模块结合课本知识与输入材料生成检索关键词，从而加强知识图谱检索结果与教材间的联系。这样便能够避免系

表1 知识图谱概览

名称	总数/个
总边数	52 036 074
总节点数	4 217 251
关键词节点数	1 651 843
值节点数	2 565 408

统从知识图谱检索结果中获取超纲内容，导致考生无法顺利地解答生成的题目。

2.2.3 题目生成模块

题目生成模块负责根据知识图谱中的关系信息生成参考题目，为教师命题提供灵感。在获取知识图谱检索信息后，题目生成模块同时利用输入的材料、课本知识点与知识图谱检索结果进行题目生成，即：

$$Q_i = \text{LLM}(T_i, K_i, kw_{i1}, \dots, kw_{in})$$

(1)

由于构建的知识图谱仅包含“关键词”和“值”两项数据，并不包含两者之间的联系，因此系统使用大语言模型分析材料与关键词的联系，并据此进行题目设计，初步生成的题目会传至RF中进行质量检验后输出。试题生成的提示词如下。

你是一位历史教师，请你根据输入的材料，并参考课本中的知识点和知识库中

表2 知识图谱数据展示

关键词	值	总数/个
孔子	墨子、东周列国春秋篇、中国历史、东汉、儒法之辨、六经、六艺、四书、国子监、三纲五常、儒家、哲学	382
唐朝	两税制、中书令、中书侍郎、佛道之争、儒学、中亚、古文运动、吐蕃、兵部、兵部司、唐人宫乐图、唐代三夷教、唐代宗、唐代经济、唐令、唐会要、唐传奇	621
第一次工业革命	商人、商务印书馆、商品经济、商业、启蒙运动、乔治·史蒂芬森、蒸汽时代、圈地运动、国立国会图书馆、土耳其	3 680
亚里士多德	伙友骑兵、大伦理学、大卫·休谟、大卫·希尔伯特、大学文件系统、大希腊、天主教百科全书、天文学、哲学逻辑、哲学分析	551

的检索结果生成一道历史简答题，要求如下：

1. 试题应该考查学生能否分析历史事物人物之间的联系，而不是单纯的考查背诵能力
2. 请从知识库图谱的检索结果中选取与输入材料相关的关键词作为参考来设计问题

输入的材料为：{material}

待筛选的关键词有：{keywords}

请按如下格式输出

[EXPLAIN]选取的关键词，关键词与材料的联系[EXPLAIN]

[START]试题[END]

为了避免大语言模型生成的题目中包含错误，同时确保题目的正确性，系统引入了RF。RF会根据提示词中提供的标准检验初步生成题目的质量，如有需要则进行修改。RF使用的提示词如下。

你是一名历史老师，请你对输入的历史试题质量进行检查，如果质量不达标请根据参考资料对题目进行修改，如果质量达标则直接进行输出

参考资料

教材的知识点包括：{knowledges}

知识图谱中的关键词包括：{KGResult}

输入的试题

{question}

请按如下格式输出

[EXPLAIN]对试题质量的分析，如质量不达标则说明质量不达标的原因[EXPLAIN]

[START]输出的试题[END]

具体来说，大语言模型首先对题目的质量进行分析，若合格则直接进行输出；若题目质量不合格，则要求大语言模型分析题目不合格的原因，最后根据知识图谱检索结果及课本考纲，在原有题目的基础

上进行修改并进行输出，即：

$$Q'_i = \text{Rf}(Q_i, T_i, K_i, kw_{i1}, \cdots, kw_{im})$$

(2)

反馈机制的检查能够有效地修正初步生成试题中的错误，并进一步提高试题生成质量。

3 实验

在试题生成系统设计完毕后，为了验证系统的试题生成效果，将从正确性、相关性以及多样性3个角度对试题生成系统进行实验检验。

3.1 实验数据

实验使用共计500道历史简答题作为实验数据。具体而言，实验使用高考真题中的简答题，通过将简答题的材料部分单独提取，得到共计100段材料段作为系统测试输入。

系统使用的课本知识点数据来自人教版历史高中教材的章节、课时标题，共计23个。这些知识点包含了历史人教版高中必修一、必修二和必修三的主要内容。

3.2 模型设置

实验部分一共设置了3个Baseline模型来检验系统各部分的效果，模型参数设置见表3。其中，Baseline模型为直接使用大语言模型进行题目生成；Baseline + KG模型为使用大语言模型并结合知识图谱检索结果进行题目生成；Baseline + RF模型为在Baseline模型的基础上使用反馈机制对生成结果进行检验。这3个Baseline模型将与KG-QGH系统一并在测试数据上进行生成，随后进行评分。实验

分别选用了 GPT-4o mini 以及 ChatGLM-4-Plus 作为出题算法的基座大模型进行测试, 比较算法在使用不同基座大模型下的表现效果。

对于生成结果的评价, 实验使用大语言模型来对系统和 Baseline 模型的生成结果进行检验, 负责评分的大语言模型选择为 GPT-4o mini。下面将分别介绍检验的内容及方式。

3.3 指标设计

实验从正确性、相关性以及多样性 3 个角度来检验模型的效果, 从而验证系统是否能解决传统大语言模型进行试题生成时出现的缺乏专业知识和出题多样性较差两个问题。

3.3.1 正确性检验

正确性检验主要检查生成的题目是否包含错误、质量是否符合要求。设计的提示词要求大语言模型按如下标准对输入题目进行检查: (1) 输入的内容是否为一个问题; (2) 题目的内容中是否出现历史知识错误; (3) 题目内容是否明确; (4) 题目是否考查了学生对历史事件的分析能力。只有满足上述 4 点要求的问题才会被大语言模型判断为正确, 否则记为错误。对于输入的材料段 $m_1, \dots, m_n$, 检验模型 LLM_{corr} , 正确性得分 C 为:

$$C = \frac{\sum_{i=1}^n (LLM_{corr}(m_i) = \text{True})}{n} \quad (3)$$

正确性得分 C 的取值范围为 0~1, 其值越高代表生成题目更加规范、更加符合历史材料题的设计要求。

表3 Baseline模型参数设置

	大语言模型	知识图谱	反馈机制
Baseline	✓	×	×
Baseline + KG	✓	✓	×
Baseline + RF 反馈系统	✓	×	✓
KG-QGH	✓	✓	✓

3.3.2 相关性检验

相关性检验负责检查题目是否符合课本知识点, 避免生成的试题超纲而使学生无法作答。相关性检验包括知识点匹配度 (Rel) 和知识点相似度 (SBert) 两个指标。

Rel 使用大语言模型判断生成题目与课本知识点是否相关。检验提示词会向大语言模型输入课本的知识点和生成的试题, 要求大语言模型判断生成的试题是否在课本的要求范围内。实验通过比较不同 baseline 模型的生成结果中被判断为相关的题目数量来评判生成题目与课本的相关性, 即对于输入的材料段 $m_1, \dots, m_n$, 课本知识点 $K_1, \dots, K_S$, 模型 LLM_{rel} , Rel 可以表示为:

$$Rel = \frac{\sum_{i=1}^n (LLM_{rel}(m_i, K_1, \dots, K_S) = \text{True})}{n} \quad (4)$$

Rel 的取值范围为 0~1, 其值越高代表生成的试题与课本知识点联系越紧密。需要注意的是相关性检验并不要求生成的试题与课本内容完全相同, 只需要学生能够联系课内所学进行扩展作答即可。

SBert 使用 Sentence Bert^[30] 模型计算生成试题与课本知识点的相似程度, 即对于输入的材料段 $m_1, \dots, m_n$, 课本知识点 $K_1, \dots, K_S$, SBert 可以表示为:

$$SBert = \frac{\sum_{i=1}^n \text{Sentence_Bert}(m_i, K_1, \dots, K_S)}{n} \quad (5)$$

SBert 的取值范围为 0~1，数值越高代表生成的试题与课本知识点的相似度越高。

3.3.3 多样性检验

多样性检验主要检查在使用同一材料进行多次试题生成时，生成试题的多样性是否有所保证。实验指定 KG-QGH 和 3 个 Baseline 模型对 250 篇材料分别进行 3 次独立的试题生成，并对生成的结果进行检验。多样性检验包括大模型多样性打分（Div）以及文本多样性得分（Bleu）指标。

其中，Div 使用大语言模型为使用同一篇历史材料生成的结果多样性进行打分并记录用，即对于输入的材料段 $m_{11}, m_{12}, m_{13}, m_{21}, \cdots, m_{n3}$ ，模型 LLM_{div} ，Div 可以表示为：

$$Div = \frac{\sum_{i=1}^n (LLM_{div}(m_{i1}, m_{i2}, m_{i3}) = True)}{n} \quad (6)$$

表4 GPT-4o mini实验结果

模型	C	Rel	SBert	Bleu
Baseline	0.886	0.892	0.5023	12.7560
Baseline+KG	0.942	0.870	0.5388	6.5538
Baseline+RF	0.980	0.888	0.5220	12.2553
KG-QGH	0.986	0.936	0.5350	6.6747

表5 ChatGLM-4-Plus实验结果

模型	C	Rel	SBert	Bleu
Baseline	0.784	0.902	0.5331	22.5418
Baseline+KG	0.722	0.916	0.5276	20.6642
Baseline+RF	0.966	0.934	0.5504	17.9002
KG-QGH	0.974	0.946	0.5594	20.8456

多样性得分 Div 的取值范围为 0~1，其值越高代表生成的试题更加多样。

Bleu 使用 Self Bleu^[31]模型计算生成文本的丰富程度，即对于输入的材料段 $m_{11}, m_{12}, m_{13}, m_{21}, \cdots, m_{n3}$ ，Belu 可以表示为：

$$Bleu = \frac{\sum_{i=1}^n Self_Bleu(m_i, K_1, \cdots, K_s)}{n} \quad (7)$$

其取数值越低，代表生成试题的文本多样性越高。

3.4 结果分析

（1）对于正确性检验，GPT-4o mini 实验结果见表 4。KG-QGH 与使用了反馈机制的 Baseline + RF 模型的 C 值相较于 Baseline 和 Baseline + KG 模型明显更高，说明反馈机制能够对生成题目中包含的错误进行修正。通过分析被判为错误的试题可以发现，Baseline 模型的错误主要为题干叙述内容不清晰、题目要求较模糊。这些错误大部分能被反馈机制识别出，并在修正中得到调整。

（2）对于相关性检验，GPT-4o mini 和 ChatGLM-4-Plus 实验结果见表 4、表 5。在 Rel 指标上，系统在使用 2 种出题大模型时均取得了全部 Baseline 模型中最高的得分。对于 SBert 指标，算法生成的试题分别取得了次优和最优的结果，证明输出试题与课本知识点的相似度较高。由上述得分可以看出，课本知识点的引入可以更好地帮助大语言模型理解历史学科的教学内容，促使大语言模型生成符合课本考纲的试题。

（3）对于多样性检验，GPT-4o mini 和 ChatGLM-4-Plus 的 Div 核密度曲线图如图 5 所示。由 KG-QGH 与 Baseline+KG 生成结果的 Div 数值明显高于 Baseline 与 Baseline + RF 模型的生成结果。这是因为通过知识图谱查询得到的大量结果会被随

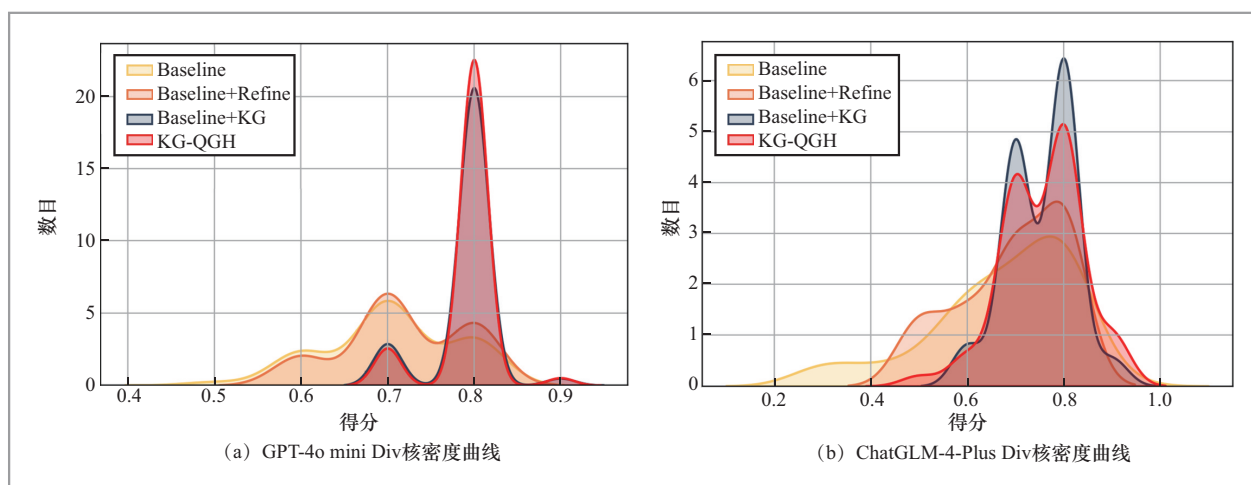


图5 Div核密度曲线

机筛选并输入大语言模型，使在使用相同材料进行题目生成时能够拥有不同的提示信息，从而使系统能够有更加多样的输出。通过计算Bleu指标也证明模型生成结果的多样性较高。在Bleu指标中以GPT-4o mini为基底时算法取得了最佳的成绩，而ChatGLM-4-Plus更倾向于生成附加有原始史料作为题干的题目，导致Bleu值较GPT-4o mini更高。

3.5 案例分析

3.5.1 案例一

输入材料：科技革命使世界发生广泛而深刻的变化，一些国家结合自身条件，抓住机遇，实现了经济的跨越式发展……德国正是依靠一种新的体制和社会机构，推动科技与工业的发展。——摘编自马克圭主编《世界文明史》下

对应课本知识点：['近代科学的发展'，'世界经济的全球化趋势']

知识图谱检索结果：['国家资本主义'，'大萧条']

分析过程：选取的关键词是“国家资本主义”。关键词与材料的联系在于，德国在19世纪晚期的工业化进程中，政府积极支持技术发展和资金投入，体现了国家资本主义的特征。这种通过国家干预市场促进经济和工业发展的模式，使德国能够迅速崛起并超越其他国家。国家对经济的直接参与和对科技的重视是德国能够实现跨越式发展的核心原因。

生成结果：结合19世纪末德国的工业发展背景，分析国家资本主义对德国经济、科技进步的作用，并与其他国家的工业化进程进行比较，指出各自的异同之处。

反馈系统评价：该题目质量较高

知识图谱查询模块首先根据材料获得课本中知识点“近代科学的发展”“世界经济的全球化趋势”，并在知识图谱检索得到“国家资本主义”“大萧条”两个关键词。随后，题目生成模块选择了“国家资本主义”作为考点进行题目设计。题目生成模块随后分析“国家资本主义”与材料主体“德国发展”的联系，并设计习题引导学生进行回答。题目不仅考查了19世纪末德国的发展情况，也同时考查了学生能否联想

课内所学将德国的发展与其他国家进行对比，顺利通过了反馈机制的检验并进行输出。

3.5.2 案例二

输入材料：敦煌位于今天甘肃省的西部，是内地通往西域的门户……吸引了世界各国的学者，在20世纪形成了一门世界性的学问——敦煌学。——摘编自白寿彝总主编《中国通史》等

对应课本知识点：[‘中国传统文化主流思想的演变’，‘古代中国的科学与文艺’]

知识图谱检索结果：[‘佛教传播’，‘儒家思想’，‘道家哲学’，‘科技与文化交流’]

分析过程：……

生成结果：请分析敦煌作为东西方文化交汇的重要门户，其所涉及的儒家思想与其他文化（如佛教、道教等）之间的关系，以及这种多元文化的交融对当时社会及后世的影响。

反馈系统评价：该题目质量不达标。题目过于宽泛，且缺乏明确的结构和重点，尤其是在“儒家思想”与其他文化的关系分析方面，未能提供具体的框架。建议明确要求对“儒家思想”与佛教、道教等文化的具体关系进行分析，并引导探讨其对社会和文化的具体影响。此修改将使题目更具针对性和可操作性。

修改后结果：请分析敦煌作为东西方文化交汇的重要门户，其所涉及的儒家思想如何与佛教、道教等其他文化相互影响，具体阐述这种文化交融对当时社会结构和思想观念的影响，并探讨其对后世文化发展的意义。

本案例中，反馈机制通过分析判断问题的描述过于宽泛，不利于学生进行作答，因此被判断为质量不达标。在对初步生成

结果进行分析后，反馈机制重新修改题目，将考查内容限定为佛教、道教，并更加具体地要求学生根据材料内容，探讨对社会结构和思想观念的影响进行作答，使学生的作答方向更加明确。

4 出题系统页面展示

本研究构建线上出题系统以展示所涉及算法的功能流程，本节内容将具体介绍并展示试题生成系统的界面与功能。

图6展示了KG-QGH系统的主体功能界面。系统由左侧的“题目设定边栏”和右侧的“对话生成窗口”构成。教师在使用时首先向“题目设定边栏”输入题目的设定信息，例如使用的材料、题目的类型等。系统会根据教师设定的题目难度、类型信息自动生成提示词。在此基础上，教师也可在“其他指令要求”窗口处进一步提出个性化要求。以上的生成指令及要求会被系统自动拼接，并在“最终指令预览”窗口处向教师进行展示。在点击“生成”按钮后，KG-QGH将根据最终指令生成试题，生成结果将以对话的形式出现在右侧窗口。例如，教师向KG-QGH系统输入一段讲述洲际贸易相关的材料，KG-QGH根据材料分析得到对应的考点“世界经济的全球化趋势”，并根据考点设计题目考查学生分析影响经济全球化的因素。

KG-QGH系统还能够获得与教师交互的历史信息，允许教师在多轮对话中完善生成的题目。图7是一个多轮交互示意图。在题目生成后，教师仍可以在右侧窗口以对话的形式与大语言模型进行交互，从而进一步对题目进行修改，直到获得满意的生成结果。

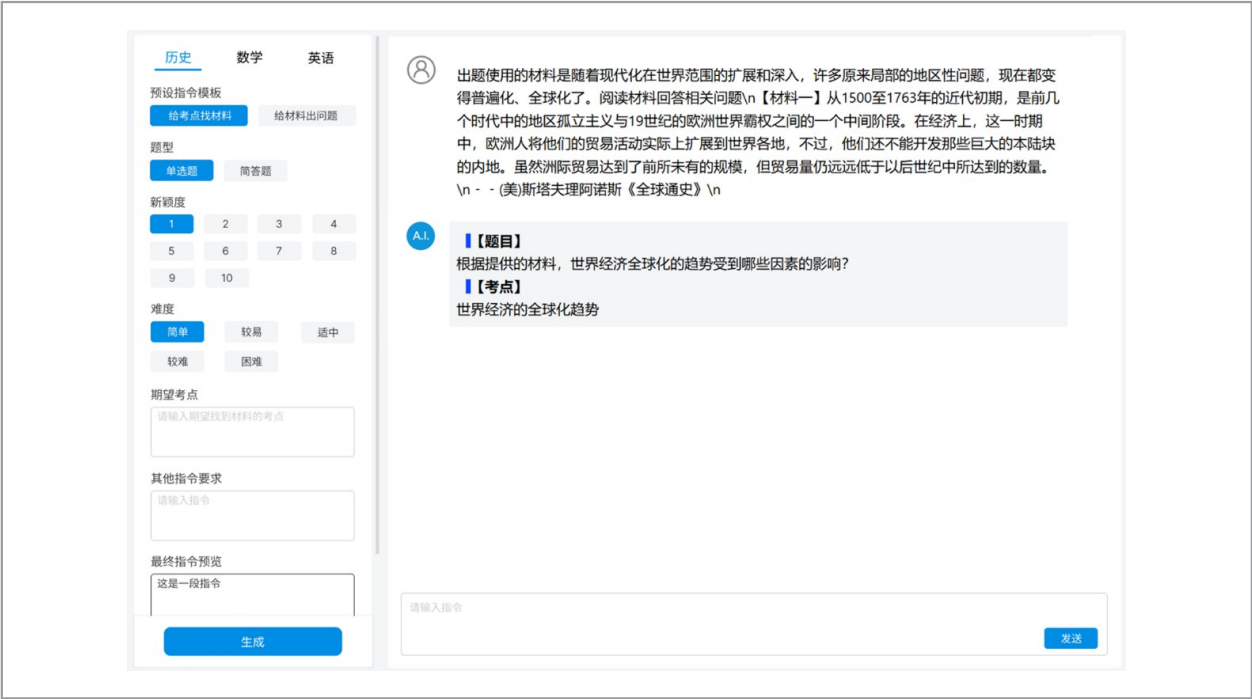


图6 KG-QGH 题目生成界面示意图



图7 多轮交互示意图

5 结束语

为了解决传统大语言模型在进行学科试题生成任务中出现的缺乏专业知识和出题多样性较差问题,本文基于知识图谱设计了一个KG-QGH系统。该系统使用知识图谱对大语言模型进行检索增强并引入了反馈机制,有效改进了题目的生成质量。同时,本文根据高考简答题使用的材料设计实验,从正确性、相关性以及多样性上对系统的性能进行检验,实验结果进一步验证了系统的有效性。

未来,可以从如下角度开展进一步的研究。

- 目前系统重点考虑问题生成,未来可以进一步考虑答案及评分标准的生成。
- 本文的实验设计主要针对生成内容的合格与否进行检验,并未考察题目质量的高低。题目的生成与传统的自然语言处理不同,无法使用Bleu^[32]或ROUGE^[33]等指标来对生成内容的质量进行评估。这是因为上述指标主要考虑的是生成文本与目标文本的相似性,但题目质量的高低并没有统一的标准答案。现阶段针对题目质量的评估仍需专家或教师手动进行,缺少自动评估的指标。在后续的研究中可以进一步关注相关指标的设计,从而更加全面、统一地对生成结果进行评价。
- 本文设计的试题生成系统仅针对历史学科材料题,但在实际历史试卷中除材料题,往往还包括选择题和简答题两种题型。以上3种题型有着相同的核心设计理念,即考查学生对历史事物的分析能力。本文设计的试题生成系统能够利用知识图谱增强的大模型实现核心理念的考查,但由于如选择题涉及混淆选项等特殊设计,

系统仍需对其进行适配,在后续工作中可通过调整提示词或提供参考样例等方法指导大模型完成相应题型设计。

- 除考试试题生成,教师也可结合日常教学利用系统生成课后习题。通过自定义考查知识点及多轮交互修改,教师可利用系统生成高度定制化的题目,从而针对班级学生的学习情况进行适配,提高练习的效率。在未来的研究中,还可以考虑对不同年级的出题进行区分,从而更好地满足日常教学需求。

参考文献:

[1] 王皓,潘昱杉,潘毅.生成式人工智能大模型赋能的元宇宙生命体:前瞻和挑战[J].大数据,2023,9(3):85-96.
WANG H, PAN Y S, PAN Y. Generative AI empowered metaverse organisms: prospects and challenges[J]. Big Data Research, 2023, 9(3): 85-96.

[2] 钱海红,王茂异,熊贇.高等教育数字化转型的现状与发展研究[J].大数据,2023,9(3):56-70.
QIAN H H, WANG M Y, XIONG Y. Digital transformation in higher education: a systematic review[J]. Big Data Research, 2023, 9(3): 56-70.

[3] 郑永和,刘士玉,王一岩.中国教育数字化的现实基础、实然困境与改革方向[J].中国远程教育,2024(6):3-12.
ZHENG Y H, LIU S Y, WANG Y Y. The realistic foundation, practical predicaments, and reform direction of educational digitalization in China[J]. Chinese Journal of Distance Education, 2024(6): 3-12.

[4] 周洋平,李春密,陆建隆,等.评价素养视域下中学物理教师命题能力结构与培养策略[J].中国考试,2024(11):38-46,78.

- ZHOU Y P, LI C M, LU J L, et al. Structures and training strategies of high school physics teachers' test-item development capacity from the perspective of assessment literacy[J]. Journal of China Examinations, 2024(11): 38-46, 78.
- [5] 王一岩, 吴国政, 郑永和. 生成式人工智能赋能教育信息科学与技术研究: 新机遇、新趋势、新议题[J]. 现代远程教育研究, 2024, 36(6): 46-54.
- WANG Y Y, WU G Z, ZHENG Y H. Generative artificial intelligence empowers educational information science and technology research: new opportunities, new trends and new issues[J]. Modern Distance Education Research, 2024, 36(6): 46-54.
- [6] 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.
- [7] HOGAN A, BLOMQVIST E, COCHEZ M, et al. Knowledge graphs[J]. ACM Computing Surveys, 2021, 54(4): 1-37.
- [8] MADAAN A, TANDON N, GUPTA P, et al. SELF-REFINE: iterative refinement with self-feedback [C]// Proceedings of the 37th International Conference on Neural Information Processing System. New York: Curran Associates Inc., 2019: 46534-46594.
- [9] LAURIOLA I, LAVELLI A, AIOLLI F. An introduction to deep learning in natural language processing: models, techniques, and tools[J]. Neurocomputing, 2022, 470: 443-456.
- [10] VASWANI A, SHAZEER N, VASWANI N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000-6010.
- [11] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[R]. 2018.
- [12] ZHAO X W, ZHOU K, LI J Y, et al. A Survey of Large Language Models[EB]. arXiv preprint, 2023, arXiv: 2303.18223.
- [13] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing System. New York: Curran Associates Inc., 2020: 1877-1901.
- [14] LIU S, FENG J T, YANG Z K, et al. COMET: "cone of experience" enhanced large multimodal model for mathematical problem generation[J]. Science China Information Sciences, 2024, 67(12): 220108.
- [15] 刘三女牙, 郝晓晗. 生成式人工智能助力教育创新的挑战与进路[J]. 清华大学教育研究, 2024, 45(3): 1-12.
- LIU S N Y, HAO X H. Challenges and approaches of generative artificial intelligence in promoting educational innovation[J]. Tsinghua Journal of Education, 2024, 45(3): 1-12.
- [16] LI L, YANG Z K, CHEN M, et al. Automated generation of geometry proof problems based on point geometry identity[J]. Journal of Automated Reasoning, 2024, 68(2): 11.
- [17] SPEER R, CHIN J, HAVASI C. ConceptNet 5.5: an open multilingual graph of general knowledge[C]//Proceedings of the AAAI conference on artificial intelligence. California: AAAI Press, 2017: 4444-4451.
- [18] QIN P G, YU J R, GAO Y, et al. Unified QA-aware knowledge graph generation based on multi-modal modeling[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 7185-7189.

- [19] 佟威, 何理扬, 李锐, 等. 基于无监督语义哈希的高效相似题检索模型[J]. 计算机应用, 2024, 44(1): 206–216.
- TONG W, HE L Y, LI R, et al. Efficient similar exercise retrieval model based on unsupervised semantic hashing[J]. Journal of Computer Applications, 2024, 44(1): 206–216.
- [20] LIU Y, ZHANG K, HUANG Z Y, et al. Enhancing hierarchical text classification through knowledge graph integration[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Toronto: Association for Computational Linguistics, 2023: 5797–5810.
- [21] WILMOT D, KELLER F. Memory and knowledge augmented language models for inferring salience in long-form stories[EB]. arXiv preprint, 2021, arXiv: 2109.03754.
- [22] AGRAWAL G, PAL K, DENG Y L, et al. CyberQ: generating questions and answers for cybersecurity education using knowledge graph-augmented LLMs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2024: 23164–23172.
- [23] QIN W, WANG X W, HU Z Z, et al. Math word problem generation via disentangled memory retrieval[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(5): 1–21.
- [24] XUE S, HUANG Z, LIU J, et al. Decompose, analyze and rethink: solving intricate problems with human-like reasoning cycle[J]. Advances in Neural Information Processing Systems, 2024, 37: 357–385.
- [25] QIN L H, LIU J Y, HUANG Z Y, et al. A mathematical word problem generator with structure planning and knowledge enhancement[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2023: 1750–1754.
- [26] LIU J Y, HUANG Z Y, MA Z Y, et al. Guiding mathematical reasoning via mastering commonsense formula knowledge[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2023: 1477–1488.
- [27] LIU J, HUANG Z, XIAO T, et al. Socrati-cLM: exploring socratic personalized teaching with large language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 85693–85721.
- [28] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: a large ontology from wikipedia and WordNet[J]. Journal of Web Semantics, 2008, 6(3): 203–217.
- [29] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a Web of open data[C]//Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference: ISWC 2007+ASWC 2007. Berlin: Springer, 2007: 722–735.
- [30] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019: 3982–3992.
- [31] ZHU Y M, LU S D, ZHENG L, et al. Texus: a benchmarking platform for text generation models[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York:

ACM, 2018: 1097–1100.

[32] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311–318.

[33] LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004: 74–81.

作者简介



纪天昀（2002–），男，中国科学技术大学认知智能全国重点实验室硕士生，主要研究方向为自然语言处理。



张征（1999–），男，中国科学技术大学认知智能全国重点实验室博士生，主要研究方向为人工智能与公平性。



赵宇泽（1998–），男，中国科学技术大学认知智能全国重点实验室博士生，主要研究方向为自然语言处理。



黄振亚（1992–），男，中国科学技术大学计算机科学与技术学院副教授，主要研究方向为人工智能、认知推理、智能教育。



黄威（1995–），男，教育部教育考试院助理研究员，主要研究方向为人工智能。



佟威（1984–），男，教育部教育考试院副研究员，主要研究方向为计算机教育。



刘淇（1986–），男，中国科学技术大学人工智能与数据科学学院教授，主要研究方向为数据挖掘与知识发现、人工智能技术与应用。



陈恩红（1968–），男，中国科学技术大学计算机科学与技术学院教授，主要研究方向数据挖掘、人工智能、知识工程、智能教育。

收稿日期: 2024-12-06

通信作者: 黄振亚, huangzhy@ustc.edu.cn; 黄威, huangw@mail.neea.edu.cn

基金项目: 新一代人工智能国家科技重大专项(No.2022ZD0117103); 国家自然科学基金项目(No.62477044); 安徽省科技创新攻坚计划(No. 202423k09020039); 中央高校基本科研业务费专项资金(No.WK2150110038)

Foundation Items: The National Science and Technology Major Project (No. 2022ZD0117103), The National Science Foundation of China (No. 62477044), The Key Technologies R&D Program of Anhui Province (No. 202423k09020039), the Fundamental Research Funds for the Central Universities (No.WK2150110038)