心理科学进展 2025, Vol. 33, No. 10, 1766–1782 https://doi.org/10.3724/SP.J.1042.2025.1766
Advances in Psychological Science https://cstr.cn/32111.14.2025.1766

·研究方法(Research Method)·

题目自动生成的技术革新与现实挑战*

韩雨婷[1,2,3] 王文轩[4] 刘红云[5,6] 游晓锋[7]

([1]北京语言大学心理与认知科学学院; [2]北京语言大学生命健康研究院;
[3]语言认知科学教育部重点实验室, 北京 100083) ([4]中国人民大学信息学院, 北京 100086)
([5]应用实验心理北京市重点实验室; [6]北京师范大学心理学部, 北京 100875)
([7]南昌师范学院数学与信息科学学院, 南昌 360111)

摘 要 题目自动生成(Automatic Item Generation, AIG)技术通过自动化生成测验题目, 旨在解决心理与教育测验中题目开发成本高、效率低、维护困难和安全风险等问题。该技术经历了从规则驱动方法到大语言模型(Large Language Model, LLM)的演进历程, 虽显著提升了生成效率与内容多样性, 但在应用过程中面临专业知识表达准确性、文化公平性与构念效度、多模态内容生成、开放性题目发展、智能化质量控制、资源环境适应及技术可访问性等现实挑战。针对这些挑战, 有效应对策略包括检索增强生成技术和多模态生成模型应用、多阶段心理测量学验证、云算力资源整合及用户友好型系统开发等。这些方法为提升自动生成题目的科学性与实用性提供了可行路径。

关键词 题目自动生成, 心理与教育测量, 大语言模型, 知识增强技术, 测验质量
分类号 B841

1 引言

心理与教育测验作为评估个体心理特征、能力水平和行为表现的重要工具, 在基础研究、临床诊断、教育教学、人才选拔等领域发挥着关键作用(Hambleton, 2004)。随着心理测验在研究和实践中的广泛应用, 对测验题目的质量和数量需求急剧增长。特别是计算机化自适应测验(Computerized Adaptive Testing, CAT)的推广, 要求建立包含大量等值题目的题库以保证测验的信度和效度(Embretson & Yang, 2006)。然而, 传统的人工编制题目方式存在显著局限。首先, 命题专家需要投入大量时间和精力进行题目编写、评审和修订(Falcão et al., 2023; Kurdi et al., 2020), 用于高风险测验的单个选择题开发费用可能高达1500至2000美元(Kosh et al., 2019; Lai et al., 2009; Rudner, 2010)。其次, 题目资源有限, 在高利害测验中容易出现题目被记忆和传播的问题, 严重影响测验的安全性(Bejar et al., 2003)。此外, 人工编制难免受到命题者主观经验和认知偏差的影响, 难以保证题目质量的一致性(王蕾, 2023)。因此, 如何高效地生成质量高、数量充足的测验题目, 已成为心理与教育测量领域的重要挑战。

题目自动生成(Automatic Item Generation, AIG)是指计算机根据特定要求, 基于认知理论和心理测量模型, 自动生成符合指定参数的测验题目(Embretson & Yang, 2006; 李中权, 张厚粲, 2008)。AIG系统能够快速批量生成大量题目, 即使是半自动生成系统, 其题目生成效率也会显著提升(Mitkov et al., 2006)。在AIG技术实现层面, 早期主要形成了两种方法路径(Bejar et al., 2003; Embretson & Yang, 2006): 一是基于强理论的认知设计系统法(Cognitive Design System), 该方法基于分析问题解决过程中隐含的心理学原理, 通过建模学生认知过程来设计和生成题目; 二是基于弱理论的题目模型法(Item Modeling Approach),

收稿日期: 2025-02-22
* 国家自然科学基金面上项目(32471145)和青年科学基金项目(32400934)资助。
通信作者: 刘红云, E-mail: hyliu@bnu.edu.cn

该方法以一组具有内容和难度代表性的校准题目为基础, 是一种侧重经验数据和实际应用的题目自动生成方法。近年来, 人工智能生成内容(Artificial Intelligence Generated Content, AIGC)技术的突破性进展, 特别是大语言模型(Large Language Model, LLM)的快速发展, 为 AIG 的发展开辟了新的技术路径, 推动 AIG 向更智能、更灵活的方向迈进。基于人工智能技术的 AIG 不仅显著提高了生成效率(Gierl et al., 2021; Kosh et al., 2019), 还在一定程度上克服了传统方法中模板化、缺乏创新性等局限(Götz et al., 2024)。

LLM 的发展为 AIG 技术带来了新的发展机遇, 但这一技术进步也引发了对生成内容质量的深层思考。本研究将系统梳理 AIG 的技术演进轨迹, 揭示从传统规则方法到新兴技术方法的发展脉络, 探索技术革新如何推动 AIG 向更加智能化的方向发展。在此基础上, 深入分析当前 AIG 技术在质量保障、功能拓展和应用普及三个层面面临的核心挑战及其应对策略。本研究旨在为心理与教育测量领域的 AIG 技术应用提供前沿视角与实践参考, 促进测验开发的科学性与实用性提升。

2　题目自动生成的技术演进

AIG 技术经历了从规则驱动方法到基于 LLM 的智能生成的演进, 这一发展不仅显著提升了生成效率, 推动题目生成向智能化与多样化方向发展, 也凸显了各阶段生成题目质量与适用性的差异性挑战。

2.1　传统方法概述

2.1.1　基于规则的方法

基于规则的方法主要依赖专家知识和认知理论, 通过预定义规则或模板控制题目生成。此类方法主要包括认知设计系统法(Embretson, 1998)、题目建模法(Bejar et al., 2003)和基于本体的方法(Papasalouros et al., 2008)。如图 1 所示, 认知设计系统法通过识别影响题目难度的基本成分与随机成分, 构建基于认知理论的生成算法; 题目建模法以优质题目为模板, 通过替换非关键特征生成平行题目; 基于本体的方法则依赖领域知识的形式化表示, 通过概念关系生成题目。基于规则的方法尽管能产生结构严谨的题目, 但这些方法需要大量专业知识投入, 维护成本高且灵活性受限, 难以处理复杂语言现象和适应创新性需求。

2.1.2　基于语料库和浅层文本分析的方法

基于语料库的方法通过分析大规模语言数据生成题目, 包括语料选择、预处理、题目生成三个主要阶段(Smith et al., 2010)。以选择题生成为例, 流程如图 2 所示。

语料选择需考虑规模与专业性, 如 Biber 等(2004)和 Barker (2006)证实了语料专业性与测试目标匹配度的重要性。

预处理阶段将原始语料转化为结构化数据, 构建文本的多层语言学表示, 为后续题目生成提供质量保证。基础处理构建文本的语法框架和基础特征; 专业处理则侧重于提取学科专业内容并建立语义层面的表征。Mitkov 和 Ha (2003)使用 FDG 浅层解析器进行句法分析, 通过频率统计筛选关键术语, 并结合 WordNet 进行词义消歧, 同时应用特定领域启发式规则排除干扰内容。Karamanis 等(2006)在预处理阶段使用 Charniak 的解析器进行句法分析, 利用 UMLS 术语库而非 WordNet 识别医学领域的关键术语, 并开发了专门的过滤模块来选择合适的源句子, 该模块比 Mitkov 等人(2006)的方法提高了约 30%的源句子选择效率。

选择题的题目生成主要涉及题干生成和干扰项生成。题干生成技术经历了从简单规则到复杂算法的演进。早期研究如 Mitkov 和 Ha (2003)采用浅层句法分析和术语抽取技术从电子书生成题干; Sumita 等(2005)利用词形分析和考点相关性从商务旅游语料库中确定空白项位置; 随后 Karamanis 等(2006)结合句法分析器与关键术语识别技术从医学文本中提取专业术语并自动生成相应测验题干。至 2010 年后, Goto 等(2010)应用条件随机场(CRF)估计最佳空白项位置, Smith 等(2010)利用 Sketch Engine 语料库管理软件中的“优质词典实例提取器” (Good Dictionary Example Extractor, GDEX)功能从大型网络语料库中筛选符合特定语法结构和搭配的句子作为题干。

干扰项生成经历了从简单匹配到语义理解的技术演进。早期如 Coniam (1997)主要依靠词频匹配, 随后 Sumita 等(2005)引入同义词典并用网络搜索检查题干与候选项的共现性以验证可靠性, Mitkov 等(2006)利用 WordNet 计算语义距离选择概念相近词语作为干扰项。技术成熟期, Agarwal 和 Mannem (2011)整合词频、句法和语义特征进

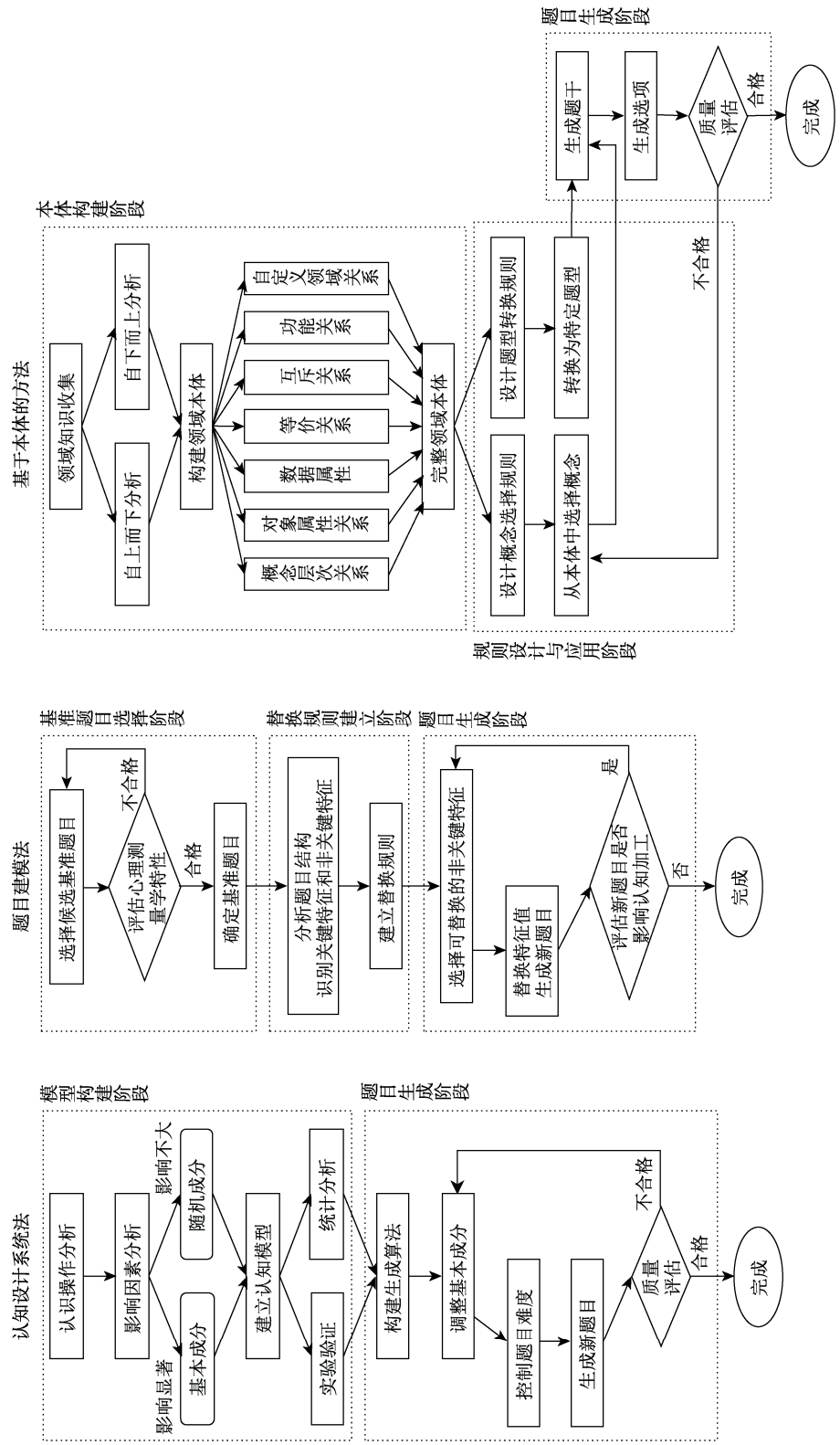

图 1 基于规则的题目自动生成流程图

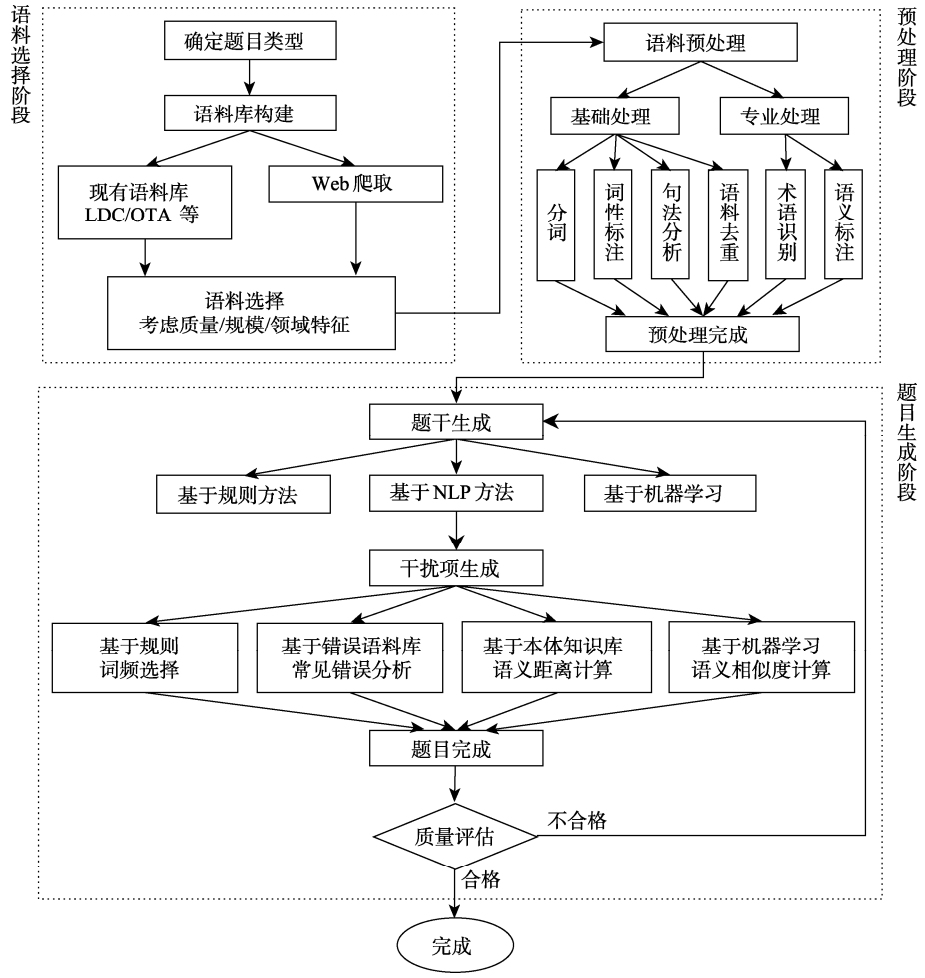

图 2　基于语料库和浅层文本分析的题目自动生成流程图

行多维度选择, Sakaguchi 等(2013)从学习者错误模式中训练支持向量机分类器, 生成具有语义混淆作用的干扰项, Karamanis 等(2006)的快速题目生成(Rapid Item Generation, RIG)系统则结合句法分析、专业词库和分布相似度计算, 实现了领域专业干扰项的精准生成。

基于语料库的方法虽能从真实语言材料中提取自然语言模式, 提升题目的语言真实性和效度, 但其应用效果受限于语料库的领域覆盖范围、内容质量和更新频率, 在处理新兴领域或特定专业知识时灵活性不足。

2.2　新兴技术方法及其应用

2.2.1　经典深度学习技术及应用

随着神经网络技术发展, 深度学习方法显著提升了自然语言处理(Natural Language Processing, NLP)在 AIG 中的应用效果, 推动 AIG 向语义理解与自然表达的方向发展。

词嵌入模型(如 Word2Vec)通过分布式表示计算词汇语义相似度, 优化了干扰项设计。Jiang 和 Lee (2017)将此技术应用于汉语名词多项选择题干扰项自动生成, 取得了较好的效果。

循环神经网络(RNN)和长短期记忆网络(LSTM)通过序列建模增强了文本生成的连贯性和流畅度。von Davier (2018)首次将 LSTM 应用于人格测验题目生成, 虽生成的题目在语法和形式上正确, 但尚未能针对特定人格特质实现定向生成。Liyanage 和 Shingo (2020)则将 LSTM 与形态学和词性标记(POS)相结合, 通过将字符、单词和词性标记以嵌入的方式组合提供给 LSTM, 成功生成符合语法规则的初级数学应用题。

基于编码器−解码器架构的序列到序列(Sequence-to-Sequence, Seq2Seq)模型和注意力机制(Attention Mechanism)增强了对上下文信息的捕捉和利用能力。Du 等(2017)开发的 Seq2Seq 模型利用双向 RNN 编码器提取句子和段落级信息, 再通过 RNN 解码器预测下一个词来生成问题, 实现了阅读理解题目的自动生成。针对模型生成题目包含答案部分的问题, Zhao 等(2018)和 Kim 等(2019)在 Seq2Seq 框架中引入了答案感知机制和位置感知机制, 通过计算上下文词与答案的距离关系, 使生成问题与答案更好地分离。随后发展出融合语义理解和知识表示的神经网络模型, 如 Zhou 和 Huang (2019)提出的数学应用题生成网络(MAGNET)采用方程编码器、主题编码器和解码器的结构设计, 利用双重注意力机制提取方程和主题信息, 并结合方程−主题融合机制与专门的损失函数确保题目与方程的一致性, 显著提高了生成题目的流畅度和相关性。最近, 徐坚(2024)提出了结合门控循环单元(Gated Recurrent Units, GRU)和图注意力网络(Graph Attention Networks, GAT)的增强型 Seq2Seq 模型, 该模型通过答案引导的图注意网络(Answer-Oriented GAT)机制来捕获文章内部的依赖关系, 并在解码过程中整合注意力机制和指针网络技术, 改善了生成阅读理解题目的语义关联性和答案唯一性, 在精确度、召回率和语义相似性方面都有所改善。

2.2.2 大语言模型技术及应用

Transformer 架构(Vaswani et al., 2017)凭借并行计算和自注意力机制有效解决了长距离依赖问题, 为题目生成提供了更好的语义理解和表达能力。基于 Transformer 架构的预训练大语言模型, 如 BERT (Devlin et al., 2019)、T5 (Raffel et al., 2020)和 GPT 系列(Brown et al., 2020)等, 通过在海量文本数据上预训练, 学习了丰富的语言表示和知识, 获得了强大的语言理解和生成能力, 为 AIG 提供了创新性的技术支持。

部分研究者并未直接利用 LLM 的文本生成能力, 而是应用其嵌入表示、相似度计算或掩码预测等功能间接实现题干或干扰项的生成。如 Mitkov 等(2023)分别利用 SBERT (基于 BERT 的句子编码器)和 doc2vec 为新的段落生成嵌入向量, 再通过段落匹配技术从已建构的包含“段落—问题对”的法律领域语料库中检索语义相近段落, 并由专家基于该段落的现有问题模板生成新的多项选择题。结果发现, SBERT 在段落相似度匹配任务中的表现显著优于 doc2vec, 专家评估显示 57%自动生成的题目可直接使用或经少量编辑后使用。此外, Chan 等(2022)应用 BERT 的掩码预测能力开发了自动语法阅读练习(AGREE)系统。该系统通过识别句中目标语法结构, 将其替换为[MASK]标记, 然后利用 BERT 预测可能的替代词, 并结合语言学知识进行筛选, 确保干扰项既合理又具有区分度。结果显示, 85%的题目被认为只有唯一正确答案, 且评估者难以区分系统生成题目与教师创建题目, 证明了该方法在自动生成语法练习方面的良好表现。

当前, 更多的 AIG 研究利用 LLM 生成内容的能力直接创建题目, 网络版附表 1 总结了基于 LLM 生成题目的主要研究。从附表 1 可以看出, 这些研究涵盖了多样化的应用领域: 包括教育测验领域的数学测验(如: 窦若琳, 2023)、语言测验(如 Attali et al., 2022; Dijkstra et al., 2022; Rathod et al., 2022; Zu et al., 2023 等)、医学考试(如 Kıyak, 2023; Kıyak et al., 2024; von Davier, 2019; Zuckerman et al., 2023 等), 以及心理测量领域的人格测验(如 Hernandez & Nie, 2023; Hommel et al., 2022; Götz et al., 2024)等。其中大多数研究表现出良好的题目生成质量——如在专家评审方面, Dijkstra 等(2022)报告生成的题目在流畅性(99%)和相关性(97%)方面表现优异; 在心理测量特性方面, Götz 等(2024)生成的人格测验项目显示出良好至优秀的心理测量学特性, 重测信度超过 0.70; Lee 等(2023)的研究表明生成的人格测验具有良好的信度(Omega = 0.77~0.86)和测量不变性。然而, 这些研究也存在共同挑战, 尤其是在干扰项质量(如 Dijkstra 等, 2022 报告的干扰项干扰能力仅达 60%)和专业知识准确性方面(如 von Davier, 2019 指出生成文本中仍有不准确内容, 需要医学专家修改)。

为了提升模型在特定领域的适用性和生成质量, 研究者通常采用领域微调和提示工程两种优化策略。模型微调通过在特定任务数据集上进行训练, 使模型学习任务相关的知识和模式。如网络版附表 1 所示, 部分基于 LLM 的题目生成研究采用了领域微调技术, 如张津旭(2022)对 UniLM 和 mT5 的优化训练, 以及 Hernandez 和 Nie (2023)

对 GPT-2-XL 的特定任务微调等。当前采用微调策略的 AIG 研究主要使用传统的全参数微调, 需要改变模型的所有参数, 计算资源需求大, 实施门槛高。近期发展的参数高效微调技术, 如 LoRA (Hu et al., 2022)和 QLoRA (Dettmers et al., 2023)等, 通过仅更新少量适配器参数而非全部模型参数, 显著降低了资源需求。然而, 从附表 1 中可以看出, 这些高效微调技术在当前 AIG 研究中几乎未见应用, 表中所列研究要么采用传统全参数微调方法, 要么完全依赖提示工程策略, 为题目生成领域引入参数高效微调技术提供了明显的发展空间。

提示工程作为一种更轻量级的方法, 无需改变模型参数, 而是通过设计合适的提示引导已有模型生成目标内容, 有助于提升 AIG 质量。学术界已发展出多种提示策略, 如少样本学习(Few-shot Learning, Brown et al., 2020)、角色定义(Shanahan et al., 2023)、思维链(Chain-of-Thought; Wei et al., 2022)、思维树(Tree of Thoughts; Yao et al., 2023)、自洽性(Self-Consistency; Wang et al., 2022)、自我反思(Self-Reflection; Madaan et al., 2023)、自我验证(Self-verification; Weng et al., 2022)、自动提示工程(Automatic Prompt Engineer; Zhou et al., 2022)和代理协作(Multi-Agent Collaboration; Du et al., 2023)提示等。但如附表 1 所示, 当前 AIG 研究仍主要采用较为基础的提示方法。结构化指令(如 Lee et al., 2023; Sayin & Gierl, 2024; Wang et al., 2025)、角色定义(如 Kıyak, 2023 中“扮演医学考试题库开发者”的角色定义)和提供示例(如 Attali et al., 2022; Götz et al., 2024; Maertens et al., 2024)是最常用的提示策略, 而高级推理提示策略如思维链和自我反思在题目生成中的应用尚待深入探索。

2.2.3　知识增强技术及应用

尽管 LLM 在题目生成中展现出丰富的应用价值, 但在内容真实性和专业知识表达准确性等方面仍面临重要挑战(详见 3.1 节)。知识增强技术通过引入外部知识库、构建专业领域表示和设计高效检索策略等方法整合外部结构化知识, 能够克服 AIG 系统在专业知识表达方面的固有限制, 显著提升自动生成题目的质量。

(1)检索增强生成(RAG)技术

检索增强生成(Retrieval-Augmented Generation, RAG)是近年来最具影响力的知识增强技术之一, 其核心原理是在生成过程中实时检索外部知识库, 为模型提供可靠信息支持, 减少“幻觉”问题(Lewis et al., 2020)。如图 3 所示, RAG 框架通过三个关键阶段运作: 首先, 构建专业知识库, 将专业教材、文献、专家知识、测验标准和实践案例等转化为向量表示, 为生成过程提供事实依据和专业信息; 其次, 检索引擎利用基于关键词的稀疏检索(如 BM25 算法)和基于语义的密集检索(如向量相似度匹配)等方法, 定位知识库中与当前查询或任务最相关的信息片段; 最后, LLM 接收包含检索信息的增强提示, 产出融合外部知识与模型能力的内容, 大幅提升输出的准确性和可靠性。如网络版附表 2 所示, 当前已有研究对 RAG 技术在 AIG 系统中的应用进行了初步探索。

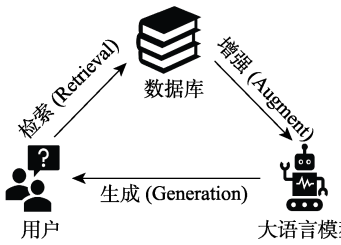

图 3　检索增强生成(RAG)技术的基本原理

(2)知识图谱技术

知识图谱作为一种结构化知识表示方法, 通过节点(实体)、边(关系)和属性的图形结构组织信息(Ji et al., 2022; Wang et al., 2017), 其核心优势在于能够表示复杂的概念网络和语义关系, 支持基于图结构的知识推理和关联分析(Nickel et al., 2016), 这为 AIG 提供了另一种重要的知识增强路径。不同于传统数据库, 知识图谱能够明确捕捉领域知识的层次结构(Xie et al., 2016)、概念依赖和逻辑关系, 使机器能够处理知识间的内在联系。基于知识图谱的 AI 系统首先将专业知识转化为结构化的图谱表示, 然后利用其结构化特性对生成过程进行约束和引导, 这种架构特别适合需要严格逻辑推理和概念关系的学科测验题目生成, 如数学、物理和医学诊断等领域。

在 AIG 实践中, 知识图谱已在保障学科专业知识完整性和逻辑关系方面展现出独特价值。高凯(2024)构建的初等数学题目检索式生成系统结合了知识图谱与 LLM 能力: 首先构建数学领域概

念知识图谱作为顶层规范，然后引导 LLM 从题目文本中抽取结构化三元组，如<椭圆，离心率，1/2>，进而通过 TransE 模型将知识图谱嵌入为特征向量，实现高精度的相似题目检索。Qin 等(2023)提出的数学应用题生成模型(MaPKG)则采用“先规划后生成”策略：利用 ConceptNet 和 HowNet 构建关键词子图，捕捉数学概念间的语义关联；基于方程结构树设计动态规划模块，使系统按照数学逻辑顺序组织句子；最后通过双重注意力机制同时关注方程信息和主题知识，实现词级别的精确生成。

综上，在 AIG 领域，深度学习技术推动了多轮技术创新。词嵌入模型、RNN 和 Seq2Seq 等经典方法优化了语义表示和文本连贯性，Transformer 架构和 LLM 则进一步提升了语言生成能力。领域微调通过调整模型参数提升专业适应性，提示工程则利用巧妙设计的提示引导模型行为，共同扩展了 AIG 在教育测验、认知评估和人格测量等多领域的应用。知识增强技术如 RAG 和知识图谱则通过结构化表示和精准检索提供专业知识支持，推动 AIG 向更加专业化的方向发展。图 4 展示了基于大语言模型的 AIG 技术路径。

3 AI 时代题目自动生成的挑战与对策

随着 LLM 与新兴技术的融合应用，AIG 技术向智能测验系统快速演进，在这一转型过程中也催生了多层面的技术与应用挑战：在质量层面，生成内容的真实性与准确性、文化公平性与构念效度成为关键挑战；在功能层面，多模态内容生成、高阶开放性题目与智能化质量控制构成主要发展需求；在普及层面，资源受限环境下的应用策略与技术可访问性是亟待解决的现实问题。针对这些挑战，创新应对策略不仅能提升 AIG 系统的科学性与实用性，也为心理与教育测量领域的

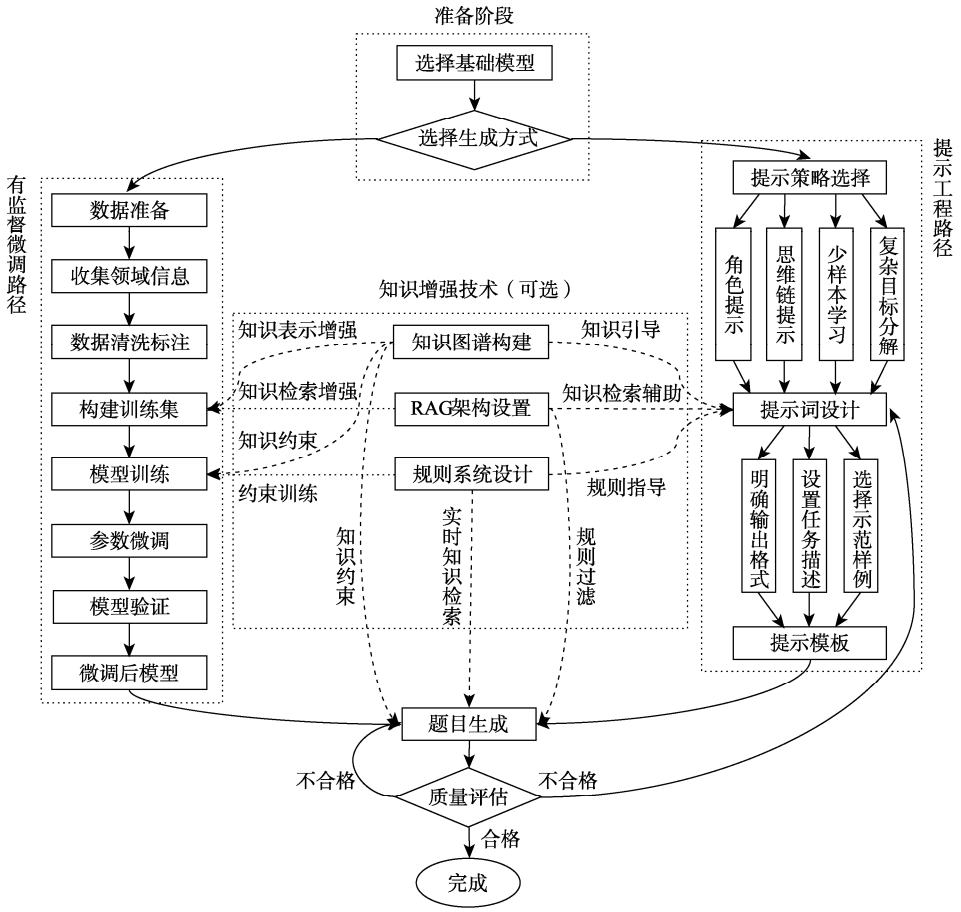

图 4 基于大语言模型的题目自动生成流程图

智能化发展提供新的技术路径。

3.1 保障生成内容的真实性与准确性

LLM 在题目生成中面临的首要挑战是确保内容真实性与专业知识表达的准确性。LLM 易于产生“幻觉”现象, 生成表面上合理但实际不准确的内容(Alkaissi & McFarlane, 2023; De Angelis et al., 2023; Gao et al., 2023)。在医学、法律等专业领域, 这一问题尤为突出, 因专业测验题目需同时满足术语准确性、理论一致性, 并真实反映专业实践场景。研究显示 LLM 在专业领域主要面临知识错误和细节不足等问题(Indran et al., 2024): Han 等(2024)在评估 ChatGPT 生成医学教育内容时, 发现其在描述专业医学概念和机制时存在错误, 如错误地声称低密度脂蛋白(LDL)是在肝脏合成的。Ngo 等(2024)指出当前模型在处理医疗文档中的错误信息和噪声时表现不佳, 且 LLM 生成的临床场景缺乏完整性, 如在代谢综合征案例中忽略了体重指数、腰围等关键诊断参数。Zuckerman 等人(2023)发现 ChatGPT 生成的医学题目错误主要与概念理解不准确有关, 而非明显事实错误。虽然能生成高质量题目, 但仍需内容专家编辑, 包括移除未教授的干扰项、调整措辞及补充题干细节。这些问题一方面源于训练数据存在的质量问题, 如未经筛选的资源、无法更新的知识库, 导致模型在生成过程中可能复制或放大错误信息和过时知识(Han et al., 2024; Indran et al., 2024); 另一方面则源于 LLM 的运作方式是基于概率预测而非明确知识推理(Zuckerman, et al., 2023), 这使模型难以区分正确与错误信息, 缺乏对专业知识的深层理解。针对这一影响测验质量的关键问题, 可以尝试从以下几个方面进行改进:

(1)模型规模与专业化训练

使用更大规模、更先进的模型是提升生成质量的直接路径。Brown 等(2020)研究表明, 从 GPT-2 (1.5B 参数)到 GPT-3 (175B 参数)的发展过程中, 模型参数规模与专业任务表现呈正相关。Indran 等(2024)指出 GPT-4.0 比 GPT-3.5 能生成更复杂的临床情景和详细解释。此外, 针对特定领域的微调训练能显著提升专业表达能力。在提示工程方面, Indran 等(2024)建议提供精确指令, 明确题目类型、数量、目标学生群体等具体参数, 并指出大规模生成需全面的提示设计, 小规模场景则可采用迭代方法。

(2)构建人机协同框架

von Davier (2019)提出结合专家输入和智能系统的人机协作模式: “一旦预训练模型可用, 可开发工具让医学专家输入草稿或关键词, 应用程序将其转化为案例描述草稿, 再由人类专家完成并提交”。与此相似, Indran 等(2024)提出的人机协同框架则通过精心设计的提示词引导 AI 生成初步内容, 随后由专家进行质量评估和必要修正, 从而形成“专家提供框架→AI 生成初稿→专家审核反馈→AI 调整→专家确认”的互补互促循环, 使题目生成过程在效率与专业性之间达到平衡。在这些协作模式中, AI 与人类专家均扮演着明确而互补的角色: 一方面, AI 作为辅助工具, 担任内容生成引擎和初稿提供者, 主要负责快速批量生成候选题目、提供多样化表达以及处理重复性工作; 另一方面, 人类专家则作为专业知识把关者、质量控制者和最终决策者, 主要负责确保内容的专业准确性、设定测验目标和边界条件, 并对 AI 输出进行筛选和修正。这种“AI 辅助、人类主导”的协作方式不仅保留了 AI 的效率优势, 还有效融合了专业人员的知识判断, 因此特别适合对内容质量要求严格的高利害测验情境。

(3)检索增强生成(RAG)的应用

Lewis 等(2020)提出的 RAG 框架通过整合外部知识资源、实时检索相关知识, 能减少模型在事实表达上的错误, 提升生成内容的专业准确性。在 AIG 中应用这一技术需综合考虑多个方面: 首先明确测验构念和测量目标, 作为知识库设计的基础; 其次科学设计检索策略, 平衡覆盖度与精确度; 最后优化提示设计, 通过模板融合、多阶段生成和动态权重调整, 整合专业知识与生成需求。此外, RAG 与知识图谱的协同应用(Sarmah et al., 2024)可提供概念间的结构化表示, 增强内容逻辑准确性, 而检索机制则提供更灵活多样的文本知识支持。

(4)必要的审核机制

无论采用何种先进技术, 自动生成的测验题目均需经过严格审核。Indran 等(2024)强调确保最终质量的责任仍在专家身上, 而 Götz 等(2024)提出的双盲专家评审方法则要求每道题目经多位专家一致认可才能使用。Han 等(2024)建议将 LLM 视为医学教育者的辅助工具, 但所有生成的评估题目必须经过领域专家审查和修改。Attali 等

(2022)报告，在其交互式阅读任务研究中，生成的文章经内容和公平性多轮专家审查后仅 58%被保留，证实人工专业评审在确保测评内容的质量和适用性方面不可或缺。

表 1 系统梳理了现有基于 LLM 的 AIG 研究中的专家审核维度，包括已实施验证的维度和研究者建议的关键指标。与传统规则驱动 AIG 系统相比，LLM 生成内容的专业准确性、概念深度表达和伦理公平性是需要专家审核，并特别关注的关键质量控制点。这主要是因为 LLM 可能产生

表 1 基于大语言模型的题目自动生成系统专家审核维度体系

审核维度	审核子维度	具体内容
内容有效性与准确性	1. 概念相关性与内容有效性	题目与测验目标的一致性；题目与测量目标构念相关；题目与源领域一致；题目准确测量其预期测量内容；题目明确表现构念特征
	2. 内容(专业)准确性	学科概念的准确性；专业术语的正确使用；专业测验中事实性内容的验证；不包含误导性信息、防止错误信息传播
	3. 信息(概念)深度与完整性	解释性内容的详细程度与质量；对测验中复杂概念的完整解释；超出表面层次的专业原理阐述；专业测验与实际专业实践的关联性；医学测验中临床情境的合理设计与描述完整性(如实验室检测值的临床意义、诊断标准的合理性、临床场景的真实性、与实际医疗实践的关联、医学概念的整合应用)
	4. 题目内容多样性	测验题目中不同情境与场景的多样体现；医学测验中临床情境的代表性多样化(医学案例中患者人口统计学的多样化、不同疾病阶段与共病情况的体现)
	5. 难度符合预期	符合布鲁姆分类法的不同认知水平；题目难度与目标群体水平的匹配度；问题的认知挑战程度；答案不应过于明显；区分事实型问题(可从单一句子推断)与推理型问题(需要深入理解或更长上下文)；阅读测验中语料的适当水平
题目结构与语言表达	1. 语言表达质量	表述的语法正确性、拼写正确性、表述清晰度、阅读流畅性、文本可读性连贯性、语言无歧义性
	2. 题目结构与格式	题目整体结构和格式的适当性；符合标准题目模板与规范；避免题干在选项中重复
	3. 选项设计	选择题中选项格式(如长度、数字格式、百分比符号等)、词性类别、语法形式、详细程度的一致性；正确答案的唯一性；正确答案的准确性与有效性；干扰项的迷惑性、连贯性；干扰项语法正确、符合人类的逻辑和常识；干扰项在语境中合理但不正确；避免选项之间的重复或过度重叠
	4. 阅读语料	阅读语料不存在令人困惑或分散注意力的元素；阅读语料的趣味性(是否能够吸引受试者回答)
	5. 关联性与逻辑	题目与阅读语料主题相关、语义相关、逻辑一致；题目涵盖阅读语料中重要、有趣的方面；选项与题目和语料的相关性
	6. 完善性与可回答性	包含题目、答案和必要解析等完整组成部分；不需人工填补缺失内容；题目提供了足够的信息使问题可回答；答案明确
	7. 内容冗余与重复	题目集合中无相似(冗余)题目；自动生成的题目避免与源题目过于相似
伦理与公平性	1. 公平性与偏见	测验内容以尊重的方式对待各群体；最小化与构念无关的知识或技能的影响；避免不必要的有争议、煽动性、冒犯或令人不安的材料；使用适当的术语指代人群；避免刻板印象；在描述人物时体现多样性；不包含过于文化特定的内容、技术或特定领域的专业术语、对测试者可能敏感的材料
	2. 道德性	题目内容不挑战任何既有法律与法规；尊重作答者的认知和道德水平；不作任何价值观上的误导；不包含可能有害的内容
教育适用性	1. 与教学内容的匹配度	匹配课程中教授的内容(不超纲)、考查所学课程涵盖的知识点；反映课程中强调的重点；
	2. 教学实用性	具有教学价值和意义；对特定教学背景/培训情境的适合度
元评估与生成过程	1. 评估确定性	评估者对自己评分的确定程度
	2. 提示有效性	用来生成题目的提示框架和模板的有效性、可解释性、实用性和全面性

“幻觉”导致专业内容不准确, 在复杂概念解释方面深度不足, 以及可能隐含训练数据中的偏见。这种系统化的质量把关机制是 AIG 技术应用于高利害测验的必要保障。

3.2　解决伦理责任、文化公平性与构念效度问题

基于 LLM 的 AIG 技术在人格、社会态度等心理测验领域可能面临伦理责任、文化公平性、与构念效度的问题。首先, LLM 可能从训练数据中习得并“放大和延续有害的社会偏见”, 导致生成题目存在潜在伦理风险和误导性结果(Gallegos et al., 2024; Indran et al., 2024)。确保生成题目符合伦理标准是最基础的要求和前提条件。这涉及多个方面: (1)生成内容需尊重不同群体, 避免刻板印象和偏见性表述(LaFlair et al., 2023); (2)题目设计应避免包含有害或误导性内容, 并充分尊重作答者的认知能力与道德观念(陈欣 等, 2024); (3)当系统处理敏感信息时, 需确保数据处理符合相关隐私法规和伦理标准(Indran et al., 2024)。其次, 在确保伦理责任的基础上, 文化公平性是测验能否应用于多元文化背景下的关键。训练数据在不同语言和文化间的分布不平衡问题在跨文化应用场景中尤为突出——高资源语言(如英语)与低资源语言在数据量、质量和多样性方面存在明显差距, 导致模型在跨文化应用中表现差异显著(Navigli et al., 2023)。最后, 构念效度反映了测验内容能否准确测量其目标构念, 是测验专业质量的核心指标。人格特质等复杂且内隐的心理构念本身就具有测量困难, 而在自动生成中这一挑战更为突出。Hommel 等(2022)研究发现, 微调 GPT-2 生成的人格测验题目中, 约三分之二达到良好心理测量标准, 但剩余题目在内部一致性和区分度方面表现不足, 尤其在尽责性等构念上较弱, 表明 AI 生成内容在复杂构念表达上仍存在局限。

针对伦理责任、文化公平性与构念效度的多重挑战, 可尝试以下应对策略:

(1)数据增强与模型优化

通过数据增强技术和多样化训练集策略, 平衡不同文化和语言背景的训练数据, 从源头改善模型的文化公平性。Hommel 等(2022)建议使用数据增强技术解决训练数据不平衡问题, 而 Götz 等(2024)通过整合不同来源的测验样本, 减轻了生成内容的文化偏向。在模型参数层面, Götz 等(2024)发现调整 GPT-2 的温度(temperature)和核采样(nucleus sampling)参数, 能有效平衡文本创造性与构念相关性, 更精确地控制生成内容与目标心理构念的匹配度。

(2)多阶段心理测量学验证

建立从专家评审到实证验证的系统化流程, 并将明确的伦理与公平性要求整合到系统设计中, 是确保自动生成题目达到心理测量学标准的关键。Götz 等(2024)提出了从双盲专家筛选到实证验证的五步流程, 全面评估生成题目的心理测量特性。在专家评审环节, LaFlair 等(2023)实施了多元文化背景的专家团队机制, 应用 Zieky (2006)的六项公平性指南进行系统评估, 有效识别和消除题目中的文化偏见和伦理问题。表 1 所示的伦理与公平性维度为 AIG 专家审核提供了具体操作标准。专家初筛后, 通过预测试计算难度、区分度等心理测量指标, 可为修改和筛选题目提供信息。正式施测阶段, 基于大样本数据可进行更全面的心理测量学评估。在公平性验证方面, 项目功能差异(DIF)分析是确保测量等价性的关键技术。Hommel 等(2022)证明, 对模型生成题目应用 DIF 分析可有效筛除存在性别或年龄差异的题目, 确保跨群体的测量公平性。在结构效度验证方面, Götz 等(2024)发现传统验证性因素分析(CFA)不足以评估 AI 生成的人格测验的结构效度, 而探索性结构方程模型(ESEM)能更准确捕捉生成项目的因素结构。完整的构念效度验证还需考察测验分数与相关构念测量之间的关系, 可通过多特质多方法矩阵(MTMM)等技术, 分析 AIG 测验的聚合效度和区分效度。

3.3　突破测验内容单一模态限制

瑞文推理、韦氏智力测验中的图形推理测验涉及对称性、规则性等图像特征, 需整合空间关系和逻辑规则。STEM 学科评估中也常包含图表、图像和公式等复杂表现形式。但现有 AIG 技术主要局限于文本模态, 难以处理视觉和空间推理内容, 影响测验的多样性和构念效度。

多模态大语言模型(MLLM)通过整合文本、图像、音频和视频等多种数据类型, 为多样化题目自动生成提供了技术基础。从 CLIP (Radford et al., 2021) 建立视觉与语言关联的开创性工作到 Flamingo (Alayrac et al., 2022)实现少样本学习的突破, 多模态技术持续发展。在跨模态能力方面, GPT-4o 整合了文本、视觉和音频处理于单一模型,

实现更自然的多模态交互(OpenAI, 2024); Phi-3-Vision 则针对图像理解进行了专门优化(Microsoft, 2024)。在高效与轻量化方面, MiniCPM-V 成功将 GPT-4V 级别的能力压缩至可在终端设备运行的规模(Yao et al., 2024)。此外, MLLM 正从实验原型向实用系统转变, 如 Google 的 Robotic Transformer 2 (RT-2)实现了视觉-语言知识向机器人控制动作的转化(Brohan et al., 2023), DALL·E 3 则展现了多模态技术在创意图像生成中的应用潜力(OpenAI, 2023)。

在 AIG 应用中, MLLM 可用于创建包含图形、表格和文本的综合题目, 关键在于为不同测验类型设计专业提示工程框架。这些框架需通过理解学科特点、认知要求与测量目标, 建立专业知识与多模态表达间的精确映射, 确保生成内容符合学科标准并精确控制测量难度。例如, 认知能力测验中的图形推理题目需精确控制对称性、规则数量和复杂度等参数, 要求构建能将抽象认知成分转化为可视化表示的专业框架。在学科测验方面, 数学测验需整合图形生成与公式表示能力, 还需在几何题中创建符合精确度量关系的图形; 物理学测验要创建准确表达力学关系的图示; 生物学测验则需呈现复杂的结构与过程图解。此外, 多模态题目的质量审核也需要在传统文本型题目审核框架基础上扩展新的维度, 如图形视觉元素质量(如清晰度、分辨率和视觉吸引力)、多模态内容间的一致性(图表数据与文本描述是否匹配)、视觉内容的专业准确性(如医学图像或物理示意图是否符合学科标准), 以及视觉表达形式的认知负荷评估和对不同视力条件用户的适用性等。

3.4 拓展开放性题目自动生成能力

尽管已有研究探索了简答题(张津旭, 2022)和数学应用题(窦若琳, 2023)等非选择题的自动生成, 但 AIG 研究仍主要集中在结构化题型上(Circi et al., 2023)。且现有简答题生成多聚焦于事实性知识检索, 对需要深度思考和主观评价的高阶开放性题目研究相对匮乏。如张津旭(2022)在中文阅读理解简答题生成方面虽取得进展, 但仍指出生成的问题多为基于既定事实的浅层提问, 缺乏激发批判性思维和多元阐释的有深度的问题。这种局限性难以满足培养高阶思维能力和复杂问题解决能力的现代心理与教育评估需求。

解决高阶开放性题目自动生成挑战需要多方面策略整合。首先, 可以将布鲁姆认知分类法等教育理论融入生成框架, 通过设计特定的提示模板和知识库, 引导模型生成针对分析、评价和创造等高阶认知能力的问题。其次, 开放性题目生成系统需同步产出详细的评分标准和典型答案示例, 支持评分者准确理解期望的答案质量标准, 确保评分客观性和一致性。最后, 开放型题目的评审标准应包含问题的认知层次、语言表达清晰度和评分标准适用性等多维指标, 确保生成题目的信度与效度。

3.5 建立智能化质量控制机制

AIG 技术虽能批量生成题目, 但人工逐一审核的需求与追求高效的初衷相悖。随着技术发展, 质量控制正朝智能化方向演进。Gorgun 和 Bulut (2024)初探了基于 LLM 的题目质量自动评估系统。该研究使用了 Becker 等人(2012)创建的 Mind the Gap 数据集, 该数据集包含 2 252 个基于维基百科文章生成的完形填空题。在原始研究中, Becker 等人招募了 85 位众包工作者对题目进行评分。Gorgun 和 Bulut (2024)保留了评分一致的 1, 825 个题目用于实验。在自动评估方面, 研究者通过小规模测试(10 个问题)确定了最优提示模板, 评判标准包括内容重要性、可回答性、明确性和具体性。随后利用 QLoRA 技术对 Llama 3-8B 模型进行指令调优。结果显示, 该系统在测试样本上整体准确率达 78%, 对劣质题目的精确率、召回率和 F1 分数分别为 0.83、0.82 和 0.82, AUC-ROC 值为 0.77, 表明模型具有中等区分能力, 可作为题目质量初筛工具。尽管该研究存在局限性, 如使用众包评分而非专家评判作为基准、仅进行二分分类评估, 但仍展示了 LLM 评估方法作为高效替代方案的潜力。当前阶段, 自动化评估可以作为题目生成和实地测试之间的中间环节, 但并不能完全替代人工评估。Gorgun 和 Bulut (2024)建议将 LLM 评估视为减少专家修订轮次和评估成本的辅助工具, 而非独立的质量保证机制。

未来研究可朝着人机协同、多元化、智能化方向发展: 首先, 使用与生成模型不同的训练数据和优化目标, 为质量评估任务专门优化模型, 使评估系统能够独立于生成系统运行; 其次, 构建领域特定的评估框架, 针对不同学科提供定制化评估标准; 第三, 探索多模态评估技术, 以评估更复杂的题目类型; 第四, 设计多 AI 评审机制,

通过不同架构、不同训练策略的模型组成评审团, 分析它们评判的一致性与差异, 潜在地提高评估的可靠性。此外, 开发评估解释系统也是关键发展方向, 使 AI 不仅能评估题目质量, 还能具体指出问题所在、解释评分原因并提供针对性修改建议, 形成题目质量改进的闭环。最终, 这些技术应当整合到人机混合评审体系中, 将 AI 初筛与人类专家终审相结合, 保持人类专家在质量控制链条中的主导地位, 同时大幅提升评审效率。

3.6　优化资源受限环境下的应用策略

构建和训练专用模型需要大量计算资源, 成为 LLM 在 AIG 领域实践应用的主要障碍。根 von Davier (2019)的研究, 中等规模 GPT-2 (345M 参数)重训练需耗费 6 天双 GPU 时间, 而 XLNet 和 Grover-Mega 训练成本分别高达$30K ~ $245K 和 $25K, 对多数研究机构形成实质性限制。

解决策略围绕技术优化与资源整合两个方向。在技术层面, 参数高效微调(Parameter-Efficient Fine-Tuning)能有效降低资源需求。LoRA (Hu et al., 2022)和 QLoRA (Dettmers et al., 2023)通过调整少量参数或添加小型适配层, 显著减少 GPU 内存占用, 同时保持性能水平。实证表明, QLoRA 能减少约 6GB 内存需求, 训练时间仅增加 30%, 使研究者能在单个 48GB GPU 上微调 65B 参数模型(Lightning AI, 2023)。学术界近期提出的 QALoRA 和 LoftQ 等方法在超低精度量化条件下仍保持良好性能(WordPress, 2024)。资源整合方面, 研究机构可通过云计算服务按需获取算力, 如 Vast.AI 提供的 GPU 租赁服务和 Google Colab 提供的免费计算资源。需要长期计算能力的机构可构建经济型计算集群, 整合多台设备建立适合特定应用的高性价比基础设施。

对资源极度受限的研究者, 提示工程优化提供了可行替代方案。这类方法无需训练微调 LLM, 而通过设计提示模板与外部知识增强, 提升通用模型在专业领域的表现。研究表明, 链式思维(Chain-of-Thought)提示和概念深度探索策略能显著提高专业任务性能(Sahoo et al., 2024)。这种低资源方法虽在专有性上有所妥协, 但降低了应用门槛, 使更多研究者能探索基于 LLM 的 AIG 技术在测验开发领域的应用。

3.7　提升 AIG 技术的可访问性与易用性

当前基于 LLM 的 AIG 系统雏形普遍要求使用者具备编程背景和深度学习专业知识, 这一技术门槛限制了教育工作者、心理测量专家等领域专业人员的实际应用能力, 导致先进 AIG 技术与心理与教育测量实践需求之间形成实施鸿沟。

为弥合技术与应用之间的差距, 开发用户友好型 AIG 系统是重要发展方向。通过构建允许领域专家通过简单提示或关键词输入的交互式系统, 可有效屏蔽底层技术复杂性, 为非技术背景用户提供直观的端到端解决方案, 使专业人员能够将注意力集中于内容设计、评估或教学目标实现。理想的系统设计应将题目生成、编辑与质量评估功能有机整合, 并通过嵌入专业指导框架和预设模板, 引导用户基于明确的教育目标, 生成符合特定学科标准和难度要求的测验题目, 同时提供实时质量反馈机制。这种面向用户体验的方法适用于各类资源环境条件, 尤其能够使资源有限的研究与实践机构也能有效应用基于 LLM 的 AIG 技术, 从而提升测验开发效率与质量标准, 促进先进技术在心理与教育测量领域的广泛普及与应用深化。

参考文献

陈欣, 李蜜如, 周悦琦, 周同, 张峰. (2024). 基于大语言模型的试题自动生成路径研究. *中国考试*, (12), 39−48.

窦若琳. (2023). *基于数学语义理解的自动解题与出题系统实现* [硕士学位论文]. 北京邮电大学, 北京.

高凯. (2024). *基于预训练模型的初等数学题目自动生成* [硕士学位论文]. 电子科技大学, 成都.

李中权, 张厚粲. (2008). 计算机自动化项目生成概述. *心理科学进展, 16*(2), 348−352.

王蕾. (2023). 人工智能生成内容技术在教育考试中应用探析. *中国考试*, (8), 19−27.

王鹏, 封迅, 康艳俊, 康春花. (2024). 青少年心理危机量表的项目自动生成: 基于生成式人工智能和 RAG 技术. *ChinaXiv*. https://chinaxiv.org/abs/202406.00361V1

徐坚. (2024). 语义图支持的阅读理解型问题的自动生成. *智能系统学报, 19*(2), 420−428.

杨生文. (2021). *基于深度学习的阅读理解题目生成研究* [硕士学位论文]. 华中师范大学, 武汉.

张津旭. (2022). *基于预训练语言模型的智能提问系统的设计与实现* [硕士学位论文]. 西南大学, 重庆.

Agarwal, M., & Mannem, P. (2011). Automatic gap-fill question generation from text books. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the sixth workshop on innovative use of NLP for building educational applications* (pp. 56−64). Association for Computational Linguistics.

Alayrac, J. B., Donahue, J., Luccioni, P., Aghajanyan, A.,

Clark, S., Sriram, A., … Vinyals, O. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems, 35*, 23716−23736.

Alkaissi, H., & McFarlane, S I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 903077.

Barker, F. (2006). Corpora and language assessment: Trends and prospects. *Research Notes*, *26*, 2−4.

Becker, L., Basu, S., & Vanderwende, L. (2012). Mind the gap: Learning to choose gaps for question generation. In *Proceedings of the 2012 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 742−751). Association for Computational Linguistics. https://aclanthology.org/N12-1092

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, *2*(3), 1−29.

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, 100161.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., ... Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFFL 2000 spoken and written academic language corpus*. Educational Testing Service. https://starryvalve.com/wp-content/uploads/2024/09/using-mustang-fuse-when-recording-with-audacity-aaf779.pdf

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... Zitkovich, B. (2023). *RT-2: Vision-language-action models transfer web knowledge to robotic control*. arXiv. https://doi.org/10.48550/arXiv.2307.15818

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th conference on neural information processing systems* (pp. 1877−1901). Curran Associates, Inc.

Chan, S., Somasundaran, S., Ghosh, D., & Zhao, M. (2022). Agree: A system for generating automated grammar reading exercises. *arXiv*. https://doi.org/10.48550/arXiv.2210.16302

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*, Article 858273.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, *14*(2−4), 15−33.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Front Public Health, 11*, 1166120.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems, 36*, 10088−10115.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 4171−4186). Association for Computational Linguistics.

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading comprehension quiz generation using generative pre-trained transformers. In *Proceedings of the Fourth International Workshop on Intelligent Textbooks 2022 co-located with 23rd International Conference on Artificial Intelligence in Education* (pp. 4−17). CEUR Workshop Proceedings. https://www.e.humanities.uva.nl/publications/2022/dijk_read22.pdf

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. In R. Barzilay, & M. Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 1342−1352). Association for Computational Linguistics.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning* (pp. 11733−11763). Proceedings of Machine Learning Research.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380−396.

Embretson, S. E., & Yang, X. (2006). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 747−768). Elsevier.

Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., & Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of automatic item generation. *Advances in Health Sciences Education*, *28*(5), 1441−1465.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097−1179.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital*

Medicine, 6(1), 75.

Gierl, M. J., Lai, H., & Tanygin, V. (Eds.). (2021). *Advanced methods in automatic item generation*. England: Routledge.

Gorgun, G., & Bulut, O. (2024). Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educational Measurement: Issues and Practice*, *44*(1), 96−107.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning*, *2*(3), 210−224.

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2024). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, *29*(3), 494−518.

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema*, *16*(4), 696−701.

Han, Z., Battaglia, F., Udaiyar, A., Fooks, A., & Terlecky, S. R. (2024). An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. *Medical Teacher*, *46*(5), 657−664.

Hernandez, I., & Nie, W. (2023). The AI‐IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011−1035.

Hommel, B. E., Wollang, F. -J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, *87*(2), 749−772.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR, 1*(2), 3.

Indran, I. R., Paranthaman, P., Gupta, N., & Mustafa, N. (2024). Twelve tips to leverage AI for efficient and effective medical question generation: A guide for educators using Chat GPT. *Medical Teacher*, *46*(8), 1021−1026.

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems, 33*(2), 494−514.

Jiang, S., & Lee, J. S. (2017). Distractor generation for Chinese fill-in-the-blank items. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 143−148). Association for Computational Linguistics.

Karamanis, N., Ha, L. A., & Mitkov, R. (2006). Generating multiple-choice test items from medical text: A pilot study. In N. Colineau, C. Paris, S. Wan, & R. Dale (Eds.), *Proceedings of the fourth international natural language generation conference* (pp. 111−113). Association for Computational Linguistics.

Kim, Y., Lee, H., Shin, J., & Jung, K. (2019). Improving neural question generation using answer separation. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 6602−6609.

Kıyak, Y. S. (2023). A ChatGPT prompt for writing case-based multiple-choice questions. *Revista Española de Educación Médica*, *4*(3), 98−103.

Kıyak, Y. S., Coşkun, Ö., Budakoğlu, I. İ., & Uluoğlu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal of Clinical Pharmacology*, *80*(5), 729−735.

Kıyak, Y. S., & Kononowicz, A. A. (2024). Case-based MCQ generator: A custom ChatGPT based on published prompts in the literature for automatic item generation. *Medical Teacher*, *46*(8), 1018−1020.

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost−benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice, 38*(1), 48−53.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121−204.

LaFlair, G., Yancey, K., Settles, B., & von Davier, A. A. (2023). Computational psychometrics for digital-first assessments: A blend of ML and psychometrics for item generation and scoring. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 107−123). Routledge.

Lai, H., Alves, C., & Gierl, M. J. (2009). Using automatic item generation to address item demands for CAT. In 2009 *GMAC Conference on Computerized Adaptive Testing* (pp.1−16). Graduate Management Admission Council.

Laverghetta, A., & Licato, J. (2023). Generating better items for cognitive assessments using large language models. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, ... T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications* (pp. 414−428). Association for Computational Linguistics.

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from “human writing” to “machine generation” in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38*(1), 163−190.

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, *29*(9), 11483−11515.

Lelkes, A. D., Tran, V. Q., & Yu, C. (2021). Quiz-style question generation for news stories. In *Proceedings of the web conference 2021* (pp. 2501−2511). Association for

Computing Machinery.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459−9474.

Lightning AI. (2023, October 12). Finetuning LLMs with LoRA and QLoRA: Insights from hundreds of experiments. *Lightning AI*. https://lightning.ai/pages/community/lora-insights/

Liyanage, V., & Ranathunga, S. (2020). Multi-lingual mathematical word problem generation using long short term memory networks with enhanced input features. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4709−4716). European Language Resources Association.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., ... Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, *36*, 46534−46594.

Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., ... van der Linden, S. (2024). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, *56*(3), 1863−1899.

Maity, S., Deroy, A., & Sarkar, S. (2024). A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), *Advances in information retrieval: 46th European conference on information retrieval* (pp. 268−277). Cham: Springer Nature Switzerland AG.

Microsoft. (2024, May 21). *Azure AI Vision at Microsoft Build 2024: Multimodal AI for Everyone. Microsoft Community Hub.* https://techcommunity.microsoft.com/blog/azure-ai-services-blog/azure-ai-vision-at-microsoft-build-2024-multimodal-ai-for-everyone/4146911

Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing* (pp. 17−22). Association for Computational Linguistics.

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, *12*(2), 177−194.

Mitkov, R., Maslak, H., Ranasinghe, T., & Sosoni, V. (2023). Automatic generation of multiple-choice test items from paragraphs using deep neural networks. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment* (pp. 77−89). Routledge.

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, *15*(2), 1−21.

Ngo, N. T., Van Nguyen, C., Dernoncourt, F., & Nguyen, T. H. (2024). Comprehensive and Practical Evaluation of Retrieval-Augmented Generation Systems for Medical Question Answering. *arXiv*. https://doi.org/10.48550/arXiv.2411.09213

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE, 104*(1), 11−33.

Oldensand, V. M. (2024). *Developing a RAG system for automatic question generation: A case study in the Tanzanian education sector* [Unpublished master's thesis]. KTH Royal Institute of Technology. https://www.diva-portal.org/smash/get/diva2:1903662/FULLTEXT01.pdf

OpenAI. (2023). *DALL·E 3 system card.* OpenAI. https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf

OpenAI. (2024). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies. In M. B. Nunes & M. McPherson (Eds.), *Proceedings of the IADIS International conference on e-learning 2008* (pp. 427−434). International Association for Development of the Information Society.

Qin, L., Liu, J., Huang, Z., Zhang, K., Liu, Q., Jin, B., & Chen, E. (2023). A Mathematical Word Problem Generator with Structure Planning and Knowledge Enhancement. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval* (pp. 1750−1754). Association for Computing Machinery.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the international conference on machine learning* (pp. 8748−8763). Proceedings of Machine Learning Research.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1−67.

Rathod, M., Tu, T., & Stasaski, K. (2022). Educational multi-question generation for reading comprehension. In *Proceedings of the 17th workshop on innovative use of NLP for building educational applications (BEA 2022)* (pp. 216−223). Association for Computational Linguistics.

Rudner, L. (2010). Implementing the graduate management admission test computerized adaptive test. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 151−165). Springer.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*. https://doi.org/10.48550/arXiv.2402.07927

Sakaguchi, K., Arase, Y., & Komachi, M. (2013). Discriminative approach to fill-in-the-blank quiz generation for language learners. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 238−242). Association for Computational Linguistics.

Sarmah, B., Hall, B., Rao, R., Patel, S., Pasquali, S., & Mehta, D. (2024). HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. *arXiv*. https://doi.org/10.48550/arXiv.2408.04948

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice*, *43*(1), 5−18.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493−498.

Smith, S., Avinesh, P. V. S., & Kilgarriff, A. (2010). Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of the ICON-2010: 8th international conference on natural language processing* (pp. 1−6). Macmillan Publishers.

Sumita, E., Sugaya, F., & Yamamoto, S. (2005). Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In J. Burstein & C. Leacock (Eds.), *Proceedings of the second workshop on building educational applications using NLP* (pp. 61−68). Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 31st international conference on neural information processing systems* (pp. 5999−6009). Curran Associates Inc.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847−857.

von Davier, M. (2019). Training optimus prime, M.D.: Generating medical certification items by fine-tuning OpenAI's GPT2 transformer model. *arXiv*. https://doi.org/10.48550/arXiv.1908.08594

Wang, L., Song, R., Guo, W., & Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, *33*(3), 2559−2584.

Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering, 29*(12), 2724−2743.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv*. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th conference on neural information processing systems* (pp. 24824−24837). Curran Associates, Inc.

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., ... Zhao, J. (2022). Large language models are better reasoners with self-verification. *arXiv*. https://doi.org/10.48550/arXiv.2212.09561

Xie, R., Liu, Z., & Sun, M. (2016). Representation learning of knowledge graphs with hierarchical types. In *Proceedings of the 25th international joint conference on artificial intelligence* (IJCAI) (pp. 2965−2971). AAAI Press.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*, 11809−11822.

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., ... Sun, M. (2024). Minicpm-v: A gpt-4v level mllm on your phone. *arXiv*. https://doi.org/10.48550/arXiv.2408.01800

Zhao, Y., Ni, X., Ding, Y., & Ke, Q. (2018). Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3901−3910). Association for Computational Linguistics.

Zhou, Q., & Huang, D. (2019). Towards generating math word problems from equations and topics. In *Proceedings of the 12th international conference on natural language generation* (pp. 494−503). Association for Computational Linguistics.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. In *Proceedings of the eleventh international conference on learning representations*. OpenReview.

Zieky, M. J. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359−376). Routledge.

Zou, B., Li, P., Pan, L., & Aw, A. (2022). Automatic true/false question generation for educational purpose. In *Proceedings of the 17th workshop on innovative use of NLP for building educational applications (BEA 2022)* (pp. 61−70). Association for Computational Linguistics.

Zu, J., Choi, I., & Hao, J. (2023). Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychological Testing and Assessment Modeling*, *65*(1), 55−75.

Zuckerman, M., Flood, R., Tan, R. J., Kelp, N., Ecker, D. J., Menke, J., & Lockspeiser, T. (2023). ChatGPT for assessment writing. *Medical Teacher*, *45*(11), 1224−1227.

Technical innovations and practical challenges in automatic item generation

HAN Yuting[1,2,3], WANG Wenxuan[4], LIU Hongyun[5,6], YOU Xiaofeng[7]

([1] *Cognitive Science and Allied Health School, Beijing Language and Culture University, Beijing* 100083, *China*)

([2] *Institute of Life and Health Sciences, Beijing Language and Culture University, Beijing* 100083, *China*)

([3] *Key Laboratory of Language and Cognitive Science (Ministry of Education), Beijing Language and Culture University, Beijing* 100083, *China*) ([4] *School of Information, Renmin University of China, Beijing* 100086, *China*)

([5] *Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing* 100875, *China*)

([6] *Faculty of Psychology, Beijing Normal University, Beijing* 100875, *China*)

([7] *School of Mathematics and Information Science, Nanchang Normal University, Nanchang* 360111, *China*)

Abstract: Automatic Item Generation (AIG) technology aims to address challenges in psychological and educational assessment, including high development costs, low efficiency, maintenance difficulties, and security risks through automated test item creation. This technology has evolved from rule-driven methods to Large Language Models (LLMs), significantly improving generation efficiency and content diversity. However, practical implementation reveals several challenges: accuracy of professional knowledge expression, cultural fairness and construct validity, multimodal content generation, development of open-ended items, intelligent quality control, adaptation to resource constraints, and technology accessibility. Effective strategies to address these challenges include Retrieval-Augmented Generation techniques and multimodal generation models, multi-stage psychometric validation processes, consolidation of cloud computing resources, and user-friendly system development. These methods provide viable pathways to enhance the scientific validity and practical utility of automatically generated test items.

Keywords: automatic item generation, psychological and educational measurement, large language models, knowledge enhancement technologies, test quality

附表 1　基于 LLM 生成题目的主要研究

文献	基座模型	应用领域	题型	参数微调策略	微调训练数据集	提示策略	结果
窦若琳 (2020)	BERT	数学应用题自动解题和自动出题	数学应用题(文本数学题)	研究分别训练了三个主要模型：MBERT 预训练模型，通过回归、分类和掩码预测任务对 BERT 进行领域适应；自动解题模型利用预训练模型中间层语义特征构造依赖图，并使用目标驱动的树解码器生成答案；自动出题模型采用图卷积网络编码结构信息，并利用解题模型反馈进行质量评估和参数优化。	Math23K（含 23,161 道中文数学应用题)和 Ape210K (含 210,488 道中文数学应用题)		自动解题模型在 Math23K 上表达式准确率 72.0%，答案数值准确率 84.2%，在 Ape210K 上表达式准确率 65.2%，答案数值准确率 78.1%；生成问题的表达多样性高，与输入表达式的逻辑一致性好，问题描述符合数学应用题的规范
张津旭 (2021)	UniLM 和 mT5	中文阅读理解测验	简答题	两个模型都通过随机替换(构造负样本增强模型泛化能力)缓解训练与推理不一致问题；通过对抗训练(在 Embedding 层添加扰动)增强模型泛化能力。UniLM 额外将问题类型作为特征输入提高问题精确性，利用伪标签再训练扩充数据集，并采用基于词粒度的 WoBERT 预训练参数；而 mT5 则通过自定义输入格式优化任务表示。	CMRC2018（中文阅读理解数据集)和中医药数据集		优化后的 UniLM 和 mT5 模型在问题生成效果(BLEU、ROUGE-L 指标)上均优于基线模型；mT5 模型速度比 UniLM 快 60%，但准确率略低；75%以上学生认为系统有助于提升提问能力
杨生文 (2021)	问题生成使用 UniLM；干扰项生成使用 BERT	英语阅读理解测验	选择题	问题生成部分先通过 BERT+BiLSTM+CRF 结构抽取答案，将答案用特殊标记<a>替换，然后微调 UniLM 模型生成问题，并应用语义相似度计算增加多样性。干扰项生成部分利用 BERT 编码问题和文章，获取上下文表示，使用 LSTM 单独编码正确答案，通过双线性变换提取干扰信息，最后基于干扰注意力机制解码生成具有迷惑性的选项。	问题生成使用 SQuAD 1.1 (一个包含超过 10 万个问题-答案对的阅读理解数据集)；干扰项生成使用 RACE 数据集(来自中国初高中英语考试)		问题生成：BLEU-4=22.15，生成的问题很少包含目标答案；干扰项生成：BLEU-4=12.33，人工评估干扰项的流利度、连贯性和迷惑性均较高
Lelkes 等 (2021)	PEGASUS；T5；T5-Large	新闻知识评估	选择题	先在现有问答数据集上微调，训练了一个问题-答案生成(QAG)模型，再基于自生成问答数据训练干扰项生成(DG)模型；输入前缀标签(如“Style SQuAD:”)区分不同数据集风格。	NewsQuizQA (20K 问答对，5K 新闻摘要) SQuAD、NQ、NewsQA		QAG 模型在 ROUGE-QAG 指标上优于所有基线；DG 模型在多个人工评估指标上表现良好；44%用户认为测验有教育价值，49%希望将其纳入常规新闻阅读体验

续表

文献	基座模型	应用领域	题型	参数微调策略	微调训练数据集	提示策略	结果
Rathod 等(2022)	ProphetNet 用作题目生成、T5 用作题目改写	阅读理解测验	多个语义相似但用词不同的测量相同概念的平行问题	使用 ProphetNet 从 SQuAD 中的问题-答案对生成问题，构建初始训练数据；将生成的问题输入已经在 Quora Question Pairs 上训练过的 T5 改写模型获得改写版本；使用拓展后的数据集对 ProphetNet 进行微调，使其学会直接生成成对的问题。	SQuAD 1.1；训练 T5 改写模型使用 Quora Question Pairs 数据集		问题可回答性与词汇多样性存在权衡；问题与上下文段落的低重叠能提高独特性但保持可回答性；词汇多样性与语义相似性呈反比；随着生成问题数量增加，问题间独特性显著下降
von Davier (2019)	GPT-2 (345M 参数)	医学考试	医学临床病例描述与选择题干扰项	使用 tensorflow-gpu 进行全参数微调；使用内存高效梯度存储技术解决 GPU 内存不足问题。	PubMed 开放获取数据库中约 800,000 篇医学文章(约 8GB 文本)	使用不同类型的提示来引导微调后的模型生成特定类型的医学内容。包括文本续写提示：提供医学情境开头句子；问答格式提示："Q: [问题]? A:"；案例描述提示："[年龄]患者来到急诊室抱怨[症状]"	相比字符级 RNN，微调后的 Transformer 模型(GPT-2)能生成更高质量文本；生成的医学案例描述可作为人类编写临床病例的初稿；能生成合理的选择题干扰项；生成文本中仍有不准确内容，需要医学专家修改
Hommel 等(2022)	GPT-2(355M 参数)	人格测验/心理学构念测量	自陈量表题目	隐式参数化；分段训练模式：使用分隔符将构念标签与题目连接起来(如"#Extraversion@I am the life of the party")；条件生成：输入构念标签(如"#Pessimism@")，模型能生成相应题目(如 "I am not likely to succeed in my goals")。	IPIP (国际人格项目池)中的 1715 个题目		64%的生成题目显示良好心理测量特性(因子负荷>0.40)；其中 32%的题目与人工题目质量相当；未训练构念的生成题目中 76%仍有良好因子负荷，部分量表内部一致性达到 0.70 以上
Zu 等(2023)	GPT-2(1.5B 参数)	词汇测试	选择题	设计 5 种不同具体性的提示模板(含一个完整句子和一个干扰项)，基于训练集构建相应样本，分别微调 5 个模型，其中最具体提示为：[输入句子] The word ["关键词"] in the previous sentence should not be replaced by [干扰项]。	4，572 个来自标准化英语水平测试的词汇题	对于验证集中的每个题目，使用最优提示格式构建输入，提供给微调后的模型；过滤与关键词相同、同义词、高排名填空词和禁用词	更具体的提示产生更好的性能；生成的干扰项在标准化频率指数和语义相似度上与人类专家创建的干扰项高度相似；基于提示的干扰项比原始干扰项更贴近上下文；基于提示学习的方法能够学习并复现人类专家应用的隐含规则
Hernandez 和 Nie (2023)	题目生成用 GPT-2-XL；预测相关性用	人格测验	自陈量表题目	对 GPT-2-XL 进行了 5 个 epoch 的微调；使用 top-k 采样生成策略；采用配对架构微调 DistilBERT 预测项目间相关性。	GPT-2 微调使用 IPIP；相关性预测模型使用 Open Psychometrics 和		成功生成 100 万个人格项目；93%语言可接受；AI 生成的量表与传统量表在心理测量特性上相当；项目间相关性预测模型准确度高

续表

文献	基座模型	应用领域	题型	参数微调策略	微调训练数据集	提示策略	结果
	DistilBERT; 结果验证用多种零样本分类器模型				Eugene-Springfield 的 IPIP 数据集		(r=0.96); 零样本分类可有效进行内容验证
Dijkstra 等 (2022)	GPT-3 (Curie, 6.7B参数)	英语阅读理解测验	选择题	通过 OpenAI API 进行全参数指令微调 (Instruction Fine-tuning); 设计了一个固定的测验结构模板(包含问题-答案-干扰项), 作为微调过程中的目标输出格式; 保持默认微调超参数。	包含英语考试阅读理解题的 EQG-RACE 数据集 (18, 501 个训练实例; 1, 035 个验证实例; 950 个测试实例); 使用完整的文本-测验对作为训练数据		人工评估认为生成的测验在流畅性(99%)、相关性(97%)方面表现优, 总体质量高(90.5%), 主要挑战在于答案的有效性(78%)和干扰项的干扰能力(60%); 端到端生成比步骤式生成效率更高; 自动评估 BLEU-4 得分优异
Zou 等人 (2022)	BART	英语阅读理解测验	判断题			模板基础框架：基于 20 篇英语文章, 采用 7 种基础模板(如移除文本中的否定词来生成错误陈述、交换文本中的数字信息来生成错误陈述等)生成判断陈述句; 生成式框架：使用多种掩码选择协议(如遮蔽从属连词、遮蔽句子中宾语等), 再执行文本填充生成判断陈述句	专家评估生成题目具有良好的流畅性、语义、相关性和可回答性; 基于模板基础框架生成的题目与原文相关度高, 可回答性好; 基于生成式框架生成的题目流畅性更好, 能生成推理性问题
Götz 等 (2024)	GPT-2 (774M 参数)	人格测验/心理学构念测量	自陈量表题目			向模型提供目标构念的示例项目; 分别为正向和反向计分项目创建提示	能快速生成大量表面效度良好的项目; 生成的 N-BFI-20 量表显示良好至优秀的心理测量学特性, 信度、结构效度和预测效度与现有量表相当, 重测信度超过 0.70, 预测效度为 80%
Maertens 等(2024)	GPT-2(1.5B 参数)	错误信息敏感性测试	判断题			使用现有 5 种错误信息量表的项目作为虚假新闻示例	创建了有良好的心理测量特性的三个错误信息敏感性测试工具 (MIST-20/16/8); 测验有良好的聚合效度和跨文化性; 测验得分与

续表

文献	基座模型	应用领域	题型	参数微调策略	微调训练数据集	提示策略	结果
							积极开放思维、认知反思能力呈正相关，与阴谋论倾向和胡说八道接受性呈负相关
Attali 等 (2022)	GPT-3	英语阅读理解测验	完形填空；句子补全题；阅读理解选择题；主旨选择题；标题选择题			少样本学习：向 GPT-3 提供 3-5 个示例；条件生成：提供文本输出的特征；层级式提示：先生成文章，再基于文章生成问题和答案，最后生成干扰项；替代文本生成：生成与原文相似风格但内容不同的文本，用于生成干扰项	58%自动生成内容通过专家审核；题目难度适中(平均 0.70)；良好的区分度(平均 0.27)；题目间依赖性低
Laverghetta & Licato (2023)	GPT-3	自然语言推理能测量	判断前提与假设之间的关系(蕴含、矛盾或中性)			使用高区分度和低区分度题目示例来教导模型目标特性	良好的内容、聚合和区分效度，比人工编写的题目有更好的区分度和信度，更接近最佳难度水平；在命题结构和量词类别的题目上表现更好；初始生成的 400 个题目中 92 个通过了专家内容审查
Lee 等 (2023)	GPT-3(175B 参数)	人格测验	自陈量表题目			每个人格维度提供 5 个示例(来自 IPIP)；包含任务描述和示例的结构化提示；使用“简单冒号”格式的变量参数(如 “Trait: x”和“Items (k):”)来控制生成题目的特质类型和数量	模型拟合优于人类编写测试(CFI=0.93，TLI=0.93)；良好的信度(Omega=0.77~0.86)；与人类编写测试相当的聚合、区分和效标关联效度；高区分度和良好分布的项目参数；88%的题目无 DIF，支持性别群体间的测量不变性
Bezirhan & von Davier (2023)	优化的 GPT-3(text-davinci-002, 175B 参数)	阅读理解测验	阅读理解测验文章(语料)			仅提供指令，不提供示例的零样本(zero-shot)学习；提供一个示例(来自国际阅读素养进展研究 PIRLS 的文章)和指令的单样本(one-shot)学习；在提示中包含目标读者年龄/年级信息；操控 temperature 参数(0.5、0.7、0.9)控制输出多样性；迭代提示生成较长故事；人工后期编辑校正	GPT-3 能生成难度匹配四年级学生的高质量阅读文章；人类评估显示生成文章与 PIRLS 原始文章在可读性、连贯性和适合度方面高度相似；单样本学习加上年级信息的提示在匹配原始文章难度方面效果最佳；信息类文章在识别主题方面稍逊于原始文章；文学类文章在吸引力和减少干扰性方面评分高于原始文章

续表

文献	基座模型	应用领域	题型	参数微调策略	微调训练数据集	提示策略	结果
Lee 等(2024)	最初使用 GPT-3, 后改用 ChatGPT (未指明版本)	英语阅读理解测验	判断题、Wh-问题、完形填空(分别有选择和开放两种形式)			创建了一个包含问题类型(文字理解层次)和题型的提示框架, 为每种问题类型与题型组合设计包含明确指令、问题生成要求和期望输出格式的特定提示模板; 运用多步骤提示策略, 不断根据专家反馈迭代优化问题生成方法	专家评估提示模板有效性高(平均 CVI=0.84)、一致性极高(IRA=1); 专家和教师均指出判断题和完形填空题有效性较低, 而主题句等个别问题类型有效性较高
Sayin & Gierl (2024)	GPT-3.5	阅读理解测验	识别文章中不相关句子			模板化提示: 使用结构化指令模板; 少样本学习: 提供一个示例(父题目); 约束条件指导: 应用语义、组织和文本特征的具体约束; 层级式提示: 先生成相关句子, 再生成不相关句子, 最后组装	生成 12, 500 个题目; 6 个题目样本全部被专家判定为可接受; 题目难度适中(平均 0.62); 良好的区分度(平均 r=0.43, 鉴别指数=0.50); 干扰项表现良好
Wang 等 (2025)	ChatGPT (GPT-3.5)	中文阅读理解测验	选择题、填空题等 11 种问题类型			结构化提示模式: 包含角色定义、输出指示器、类型、定义、特征和示例代码 6 个组件; 集体知识库融合: 从 318 份教学设计中提取问题特征, 建立知识库; 特征引导生成: 为每种问题提供典型特征和表达方式	成功生成多种中文阅读理解题及答案; 90%参与者对系统可用性表示满意; 专家评估认为基于提示模式的系统在可回答性和质量维度上明显优于非提示模式系统, 与人工创建问题相比, 在正确性和流畅性方面没有显著差异
Zuckerman 等(2023)	ChatGPT (未指明 GPT 版本)	医学考试	案例型选择题(包含临床情境、生命体征和检查结果)			通过调整格式和语法描述创建可重用的通用提示模板; 结构化提示: 指定情境长度、包含生命体征等; 目标导向提示: 结合特定学习目标/测试点/疾病状态; 利用再生成功能创建题目的多个变体	题目创建时间从 30-60 分钟减至 5-15 分钟; AI 题目与非 AI 题目表现相似(P 值 0.71 vs 0.72, 区分度略高); 生成的题目格式一致, 遵循最佳实践; 仍需内容专家编辑, 但编辑时间显著减少; 在回忆型题目方面表现优于应用型题目
Kıyak (2023)、Kıyak 等 (2024)	ChatGPT (GPT-3.5)	医学考试	案例型选择题			提供一个包含详细指南的提示词模板; 扮演医学考试题库开发者; 指定题目应包含案例、问题干、选项和解释; 要求遵循医	专家评审认为 10 道生成题目在科学/临床知识方面均正确, 有且仅有一个正确答案; 其中 2 题纳入考试, 有较高区分度(0.41 和 0.39),

续表

文献	基座模型	应用领域	题型	参数微调策略	微调训练数据集	提示策略	结果
						学教育中构建选择题原则；通过填充主题(如“在原发性高血压患者中合理用药”)和难度级别(简单/困难)来定制生成过程	但其中 1 题干扰项选择频率低
Kıyak & Kononowicz (2024)	ChatGPT (未指明 GPT 版本)	医学考试	案例型选择题			整合 Zuckerman 等 (2023) 和 Kıyak (2023)的提示模板；输入相关的学习目标、主题或考点；引导用户遵循特定的使用流程，包括选择提示词、输入详细信息、生成问题三个步骤	提高题目生成效率；生成更符合医学教育语境的题目；解决了标准 ChatGPT 缺乏医学语境的问题
Maity 等 (2024)	GPT-3.5、GPT-4	多语种语言测试	选择题			从 SQuAD (英语)、GermanQuAD (德语)、HiQuAD (印地语)、BanglaRQA (孟加拉语)每个问答数据集中随机抽取 850 个文本作为提示内容一部分；多阶段提示(MSP)：包括释义生成(为每个上下文生成多个释义)、关键词提取(从释义中提取关键词)、问题生成(基于释义和关键词生成问题)和干扰项生成四个阶段；对比不提供任何示例的零样本策略和提供一个示例的单样本策略	MSP 方法在 BLEU、ROUGE-L 和余弦相似度等指标上优于单阶段提示; MSP 生成的问题在语法性、可回答性和难度方面表现更好；英语在单样本设置下表现最佳；单样本比零样本效果好；高资源语言(英语和德语)表现优于低资源语言(印地语和孟加拉语)

注：本表按技术路径和模型规模排序：前部分(1-10 行)为采用微调或微调结合提示工程的研究，按模型参数规模从小到大排列；后部分(11-24 行)为仅采用提示工程的研究，同样按模型参数规模从小到大排列。

附表 2　利用 RAG 技术的题目生成研究

文献	基座模型	题目领域	题型	知识库构建	知识检索方案	生成方案	研究结果
Oldensand (2024)	GPT-3.5 (gpt-3.5-turbo-0125); GPT-4 (gpt-4-turbo-2024-04-09); Llama 3 (llama3-70b-8192)	坦桑尼亚中学二年级地理课程	简答题、长答题、判断题	1.知识来源：坦桑尼亚教育研究所(TIE)出版的《Geography for Secondary Schools, Student's Book Form II》地理教科书; 2.使用 LlamaParse 将 PDF 教科书转换为 markdown 格式; 3.对内容文件和习题文件使用不同的分块策略; 4.使用文本分割器 MarkdownHeaderTextSplitter 按标题、章节、小节分块; 5.对大于 1000 字符的块使用递归字符分割器 RecursiveCharacterTextSplitter 进一步分割; 6.习题文件手动按单个问题分块; 7.使用 all-MiniLM-l6-v2 模型将文本块转化为密集向量表示, 使用 BM25 算法创建稀疏向量表示	1.查询重写：使用(Groq 平台)Llama 3-8B-8192 将用户原始查询重写为更丰富、更具领域相关性的查询; 2.使用相同嵌入模型编码查询; 3.混合检索：使用 Okapi BM25 算法实现 Lucene 稀疏检索、通过余弦相似度计算查询与块的相关性, 实现密集检索; 4.通过排名融合(RRF)合并多种检索结果; 5.使用 cross-encoder/ms-marco-MiniLM-L-6-v2 跨编码器对检索结果进行重排序; 6.从知识库返回 5 个内容文档和 2 个样例习题作为上下文	角色定义：将模型定位为坦桑尼亚教师; 单样本提示：提供示例互动指导模型; 上下文集成：将检索到的文档和样例习题融入提示; 指令细化：明确解释生成任务要求	GPT-4 和 Llama 3 在遵循指令和拒绝不相关查询方面优于 GPT-3.5; 教师在超过 70%的情况下更喜欢 AI 生成的习题而非人类编写的; 查询重写器可能不必要, 但单样本提示很重要; 教师反馈表明系统帮助他们更快完成工作, 适用于生成考题、笔记和评分方案
陈欣 等 (2024)	ERNIE-Bot 4.0	信息检索、数据结构课程	选择题、应用题	1.知识来源：《信息检索》教材(314.62K 字符)、《数据结构》教材(191.48K 字符)、PPT、在线视频等; 2.对教材文本进行解析、分段和清洗; 3.采用千帆企业级大模型服务平台的知识库插件, 使用文心 Embedding_V1 模型将文本转化为向量	1.由教师和专家将课程内容拆分为细粒度知识点; 2.知识点导向检索：以单个知识点作为查询语句检索相关内容; 3.使用相同嵌入模型编码查询; 4.语义检索：采用最大内积搜索(MIPS)进行语义匹配和文档排序	结合结构化知识点和检索内容增强提示效果; 对比实验 5 种提示方法：指令提示、少样本提示、角色提示、正面反馈、任务分解	试题合格率 86.47%; 试题平均难度 0.67; 学生评价积极, 认为试题符合课程要求; 角色提示效果优于其他提示方法; 细粒度知识点比宽泛知识点生成效果更好
王鹏 等 (2024)	文心一言、ChatGPT (FastGPT 配置)	青少年心理危机评估	自陈量表题目	1.知识来源：GitHub 上的心理咨询领域语料库(20, 000 对话) 2.将语料库导入开源架构 FastGPT, 其结构由库、集合和数据三个层次组成; 3.使用 OpenAI 的第二代嵌入模型 text-embedding-ada-002 进行文本向量化	1.使用相同嵌入模型编码查询; 2.使用 PostgresSQL 的 PG Vector 插件作为向量检索器; 3.使用 HNSW 索引优化检索效率; 4.使用双数据库架构：PostgresSQL 负责向量检索, MongoDB 用于其他数据存取	角色定义：明确 AI 在生成内容中的角色和领域背景; 作者信息：增强内容的权威性和可信度; 目标设定：明确生成内容的目的和期望效果; 限制条件：限制 AI 生成内容的范围和方向; 技能描述：强化 AI 在特定领域的知识和能力; 工作流程：指导 AI 如何生成和输出内容; 初始化对话：重申关注重点和引导对话方向	未给出具体的验证结果数据