

大语言模型辅助 HSK 四级试题的自动出题研究*

莫 愁

(北京语言大学 语言科学与资源学院, 北京 100083)

【摘 要】随着国际中文教育的快速发展, HSK 试题资源需求激增。传统人工命题模式在语料获取、考点覆盖和专家资源配置等方面面临系统性挑战。人工智能大语言模型的出现为语言测试资源自动生成提供了新的技术路径。文章以 HSK 四级的选词填空题、排列顺序题和完成句子题研究对象, 构建了一套基于 ChatGPT-4o 的人机协同自动出题框架。该框架以“考点先行”为核心理念, 通过提示语工程结合超纲词检测和相似度过滤机制对 HSK 四级考纲要求的话题、语言点、题型等要素进行精准控制, 能自动生成全面覆盖 HSK 四级考试大纲考点的试题。文章系统考察了大语言模型在试题生成过程中的表现, 并采用问卷调查和统计分析方法对生成试题进行了效度验证。结果表明, 大语言模型能够较好地理解题型要求和考点设置, 在人机协同框架下所命题的试题在语言流畅度、语法准确度和题目难度等维度上与人工命题试题只存在轻微差异。文章表明, 大语言模型支持下的人机协同出题模式能够有效提高命题效率, 实现考纲考点的精准覆盖, 降低命题成本和专家资源依赖, 为国际中文教育领域的智能命题提供了新的实践路径。但由于大语言模型在生成过程中仍会出现语义搭配错误, 且对复杂语篇理解不足等问题, 人工审核与修订仍是保证试题质量的重要环节。

【关键词】大语言模型; HSK 四级; 自动出题; 人机协同; 提示语工程

【中图分类号】H195 【文献标志码】A 【文章编号】1674-5639 (2026) 05-0042-11

DOI: 10.14091/j.cnki.kmxyxb.2026.05.005

一、引 言

随着国际中文教育在全球范围内快速发展, 教师和学习者亟需大量符合 HSK 考纲要求的优质试题资源。传统上, HSK 试题命题依靠人工, 这一过程面临三重系统性问题。一是语料搜集人力成本较高, 需要命题员通过出版物、互联网或者语料库搜集不带有超纲词的真实语料, 语料不允许自编或修改。^①二是“语料先行”模式对考点的覆盖性较弱。人工命题是“语料先行”, 命题员依据语料内容结合考纲中的考点编写题干, 语料本身制约了考点的选择。比如 HSK 四级真题就缺乏对科技和体验感悟等话题的覆盖, 话题间比例也不均衡,^②难以实现从考点出发的、全面覆盖考纲的针对性命题 (即“考点先行”)。这种“语料先行”模式从根本上违背了测试工作的“构念效度” (construct validity) 原则——测试内容应系统反映理论构念的全部维度, 而非便利样本的碎片化呈现。三是命题专家资源稀缺且分布不均。国内许多院校缺少命题专家资源, 缺乏自主命题能力, 教学资源鸿沟持续扩大。基于以上问题, 也有专家讨论通过建设 HSK 题库, 利用计算机自动生成试卷“粗坯”, 在此基础上进行人工干预, 最终形成正式用卷的出题模式。^③这种通过题库拼接组卷的方式可以缓解问题一和三, 但易造成试题内容重复, 也无法解决问题二。

* 【作者简介】莫愁, 女, 湖南长沙人, 北京语言大学助理研究员, 博士, 研究方向为句法语义学、国际中文教育。

【基金项目】北京语言大学校级科研项目 (中央高校基本科研业务专项资金) “汉语体标记语言知识库资源建设与技术研发” (23YBB18); 北京语言大学 2026 年“大思政课”建设项目 (DSZK202604)。

① 黄理兵, 郭树军. HSK 阅读理解试题的语料和命题 [J]. 世界汉语教学, 2008, (2): 135-144.

② 申琳. HSK 四级真题完成句子题型考点分析 [J]. 考试周刊, 2019, (10): 31.

③ 张晋军. 新汉语水平考试 (HSK) 题库建设之我见 [J]. 中国考试, 2013, (4): 21-26.

人工智能技术的发展,尤其是基于提示语工程(prompt engineering)^①的大语言模型(large language model,下文简称大模型),可以有效改善上述“命题难”的问题。大模型具有涌现能力^②,可执行多种自然语言处理任务,如生成符合语法且流畅自然的语句^③。利用大模型进行自动出题的技术在国内外教育领域已成为新兴热点话题。^④有鉴于此,本研究旨在构建并验证一套基于大语言模型的 HSK 四级试题人机协同生成框架。具体研究问题包括:(1)如何通过提示语工程实现“考点先行”的精准命题,即根据考纲规定的话题、语言点、词汇等主动生成试题,而非被动依赖语料;(2)该框架生成的试题在语言流畅度、语法准确度和题目难度 3 个核心质量维度上是否与人工真题具有统计可比性;(3)评分员能否有效区分机编试题与人工试题,即图灵测试意义上的“(不)可区分性”。本研究不仅回应了传统命题模式的三重困境,也为人工智能赋能语言测试提供了技术实践与实证证据。

二、研究综述

自动试题生成(Automatic Question Generation, AQG)的研究可追溯至 20 世纪 90 年代的智能辅导系统,早期方法依赖手工编写的生成模板和语言学规则。^⑤这类方法提升了试题生成效率,但其灵活性和内容覆盖面极为有限。大语言模型的突破性进展从根本上改变了 AQG 的技术范式,基于 Transformer 架构的预训练语言模型(如 GPT 系列)通过海量文本的自监督学习,获得了强大的语言理解和生成能力。大模型生成文本类学习资源主要通过提示语工程实现,即通过向大模型输入经设计而适配用户特定需求的提示语,引导其生成受控的高质量、有针对性的文本。^⑥提示语的设计直接影响文本输出效果,一般需做到表述明确且具体、相关且无歧义,^⑦细节性的优化建议包括提供背景知识、让大模型进行角色扮演、逐步增加提示难度等^⑧。许家金等^⑨系统性提供了提示语在外语教学和研究中的编写案例,覆盖词汇语法教学、听说读写译训练及文献阅读与评述、数据采集与分析等多场景,并将编写要领总结为“厘清需求、模仿借用、追问逼近”三步。

国际中文领域利用大模型生成教学资源的研究已初具规模。通过思维链形式的提示语指令,大模型可用于构建教学所需例句库^⑩和图片库^⑪。通过零样本形式的提示语指令,大模型可以辅助学生进行学术

① Prompt 常被译成“提示词”,但这一翻译并不准确,因为 prompt 往往不是一个词而是一串词语乃至一连串问话,故本文采纳李宇明的译法将其译作“提示语”。参见:李宇明. 智能时代的语言生活 [J]. 昆明学院学报, 2025, (2): 1-6.

② Wei, Jason, Tay, Yi, Bommasani, Rishi et al. Emergent Abilities of Large Language Models [Z]. arXiv: 2206.07682 [cs. CL]. 2022.

③ Ouyang, Long, Wu, Jeff, Jiang, Xu et al. Training language models to follow instructions with human feedback [Z]. arXiv: 2203.02155 [cs. CL]. 2022.

④ 刘月嫦,陈紫茹,杨敏,等. 国内外大语言模型在临床检验题库中的表现 [J]. 临床检验杂志, 2023, (12): 941-944; 江哲涵,奉世聪,王维民. 生成式大语言模型在医学考试题库建设中的实践探索 [J]. 中华医学教育杂志, 2024, (8): 561-569; 陈哲,郎为民,安海燕,等. 一种基于大语言模型的命题智能辅助系统 [J]. 电信快报, 2024, (4): 1-6; Mulla, Nikahat & Gharpure, Prachi. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications [J]. Progress in Artificial Intelligence, 2023, (12): 1-32; Hasan, A. S. M Mehedi, Ehsan, Md. Alvee, Shahnoor, Kefaya Benta & Tasneem, Syeda Sumaiya. Automatic Question & Answer Generation Using Generative Large Language Model (LLM) [D]. Bachelor thesis, Brac University. 2024.

⑤ Heilman, Michael, & Smith, Noah, A. Extracting simplified statements for factual question generation [C]. The 3rd Workshop on Question Generation. 2010.

⑥ Floridi, Luciano & Chiriatti, Massimo. GPT-3: Its nature, scope limits and consequences [J]. Minds & Machines, 2020, (30): 681-694; 陈峥. 与 AI 对话: ChatGPT 提示工程揭秘 [M]. 北京: 电子工业出版社, 2023.

⑦ Schulhoff, Sander, Ilie, Michael, Balepur, Nishant et al. The prompt report: A systematic survey of prompting techniques [Z]. arXiv: 2406.06608 [cs. CL].

⑧ 陈峥. 与 AI 对话: ChatGPT 提示工程揭秘 [M]. 北京: 电子工业出版社, 2023.

⑨ 许家金,赵冲,孙铭辰. 大语言模型的外语教学与研究应用 [M]. 北京: 外语教学与研究出版社, 2024.

⑩ 朱奕瑾,饶高琦. 基于 ChatGPT 的生成式共同价值标准例句库建设 [J]. 云南师范大学学报(对外汉语教学与研究版). 2023, (3): 71-80.

⑪ 何晴,饶高琦. 国际中文教育生成式图片库的构建与应用 [J]. 国际中文教育(中英文), 2025, (4): 59-73.

写作,比如生成选题和论点,修订论文等。^① 吴琼和唐盖盖^②采用“提示语 + 零样本学习”“提示语 + 少样本学习”“提示语 + 少样本学习 + 迭代优化”3 种方法,成功引导大模型自动生成符合《国际中文教育中文水平等级标准》的一级对话和独白文本。

具体在自动出题方面,已有研究表明大模型能够进行题型层面的模仿,具有一定的测试资源自动生成能力。徐娟和马瑞祯指出 ChatGPT 可以协助教师自动生成选择、填空、连线等多样化的语言测试习题,但是习题的质量、信度和效度还需检验;王鸿滨和吕海辉利用思维链提示技术让大模型在零样本场景下自动生成 HSK 五级随堂阅读测试题,生成试题经统计检测能达到近似人工出题的水平;刘玉屏等人通过少样本提示语尝试生成 HSK 三级的阅读理解题,发现大模型生成的试题质量虽能基本达标,但在稳定性、内容准确性方面存在不足,仅靠少样本提示语在文字层面进行约束无法确保大模型一定生成有效的题目形式和内容,大模型有时会给出错误答案、生成没有正确答案的题目、生成不符合题目要求的选项数量或答案数量。^③ 这意味着大模型在自动出题方面的能力尚未达到“能够按照考试要求规范、稳定地生成题目”的水平。尤其值得注意的是,王鸿滨和吕海辉虽然对生成题目的质量进行了统计检验,但其研究并未严格控制生成文本与 HSK 五级真题在难度方面的一致性,因此,其生成机制是否能够满足标准化考试中严格的考点控制要求,仍有进一步讨论的空间。

综上,现有研究多采用提示语指定题目的形式和内容,缺乏对精准设置话题、语言点等考点的研究,即仅告诉大模型“生成一道什么题”,而非“针对哪个考点生成”,更多解决的是题目的“形式生成”问题,缺乏对“考点控制”话题的探讨,传统出题模式中语料先行的问题(问题二)仍未被解决。此外,现有自动出题研究也缺乏针对 HSK 四级——这一报考人数最多、应用最广的级别的命题研究。有鉴于此,本研究将以上述两项研究缺口为切入点,在已有自动出题研究的基础上,将研究问题由“生成能力”进一步推进到“考点可控性”,构建“考点先行”的人机协同式 HSK 四级模拟试题出题框架,并实施效度验证。本研究的开展在解决问题二的同时,也能助力解决问题一和三:基于人机协同的提示语自动出题研究一方面能快速生成 HSK 试题,降低命题的时间和人力成本,提升试题更替速率;另一方面也能减少对命题员资源的依赖,解决教学资源不均衡问题,提升教育公平性。

三、人机协同式 HSK 四级阅读试题出题框架设计

国家汉办编制的新 HSK 四级考试大纲(2015 年版,即 HSK2.0 版本)要求考生掌握 1200 个常用词语以及 12 类语言点(包括多种词类、特殊句型、复句、固定格式)。考试可能涉及的话题包括日常生活、各类信息、教育、职业工作、文化、体验感悟、文学艺术、科技、自然、经济,涵盖出行准备、存款取款、谈论健康等 11 个语言任务。四级试卷分为听力、阅读、书写三部分,每部分包含多种题型。囿于篇幅限制,本研究聚焦于题干为纯文字的三类题型,即阅读及书写部分的选词填空题、排列顺序题、完成句子题,^④ 不涉及包含图片、音频或长语段的阅读理解题,以控制变量并保证生成任务的明确性。

选词填空题每题提供一个句子或由两个句子组成的对话,题干中会空缺一个核心词语,考生从含义或用法相近的备选项中选词填空。此题型考查词性辨析、词语搭配运用、结合语境推断词义等微技能。排列顺序题每题提供 3 个句子,考生需将这 3 个句子按顺序排列。此题型考查学生对语篇知识的掌握、考查关联词运用、按时间或逻辑顺序排列句子等微技能。完成句子题每题提供几个语块(一般不超过 5

① 李姝姝,王佳. ChatGPT 在学术汉语写作教学中的应用探讨 [J]. 海南师范大学学报(社会科学版), 2024, (2): 71-78.

② 吴琼,唐盖盖. 基于提示工程的汉语二语文本自动生成研究 [J]. 语言文字应用, 2024, (4): 44-57.

③ 徐娟,马瑞祯. ChatGPT 浪潮下国际中文教育的技术变革 [J]. 国际汉语教学研究, 2023, (2): 41-52; 王鸿滨,吕海辉. 基于大语言模型的中文阅读测试题自动生成研究 [J]. 国际汉语教学研究, 2025, (1): 41-54; 刘玉屏,欧志刚,武晓琴. 生成式人工智能赋能国际中文教学的效果测评——以教学设计、HSK 模拟试题编写及作文评分为例 [J]. 民族教育研究, 2025, (1): 156-166.

④ 本文在写作后期正值 HSK 考试由 2.0 版本更新至 3.0 版本过渡期,最新版 HSK 四级考试大纲于 2026 年 7 月正式实施,新版四级题型不含排列顺序题和完成句子题。本研究虽使用国家汉办 2014 和 2018 年出版的 2.0 版本真题集,但本研究的出题思路和流程仍可应用于 3.0 版考试出题。

个), 考生需用这几个语块造一个语义通顺、语法正确的句子。此题型考查汉语基本句式运用、语块语义理解等微技能。

通过对历年真题进行分析, 前人已发现上述题型的人工出题模式存在三类问题。^① 一是语言点设置不够精准, 排列顺序题和完成句子题并非每题都考查四级语言点, 常出现一至三级语言点, 完成句子题中四级语言点的频率远低于三级语言点^②, 甚至考查非考纲规定的语言点^③。具体语言点设置也不够均衡, 选词填空题中多考查名词和动词, 形容词和副词次之, 较少考查量词; 排列顺序题常考关联词^④和逻辑顺序类题 (占比分别为 38% 和 26%)。代词 (12%)、时间顺序 (8%)、描述现象 (4%) 等较少考查。二是话题内容覆盖不均衡, 三类试题中日常生活类占比均超过 50%, 涉及职业工作和各类信息的题目平均占比不到 20%, 其他话题各自占比平均不到 10%, 科技和体验感悟两类话题未曾在完成句子题中出现。^⑤ 四级考纲对考生能力的要求为“能用汉语就比较复杂的话题进行交流, 表达较为规范、得体”, 真题对复杂话题如经济、科技、文学艺术等的覆盖力度明显不足。此外, 话题和语言点还呈现出一定固化关联, 如排列顺序题中各类信息话题多考查代词提示。三是相同考点反复考查, 选词填空题中部分备选词在不同年份的真题中被反复考查。

这三类问题的产生很大程度上是受制于人工出题“语料先行”的模式, 由于人工可搜集的不含或仅含少量超纲词的自然语料有限, 导致命题员无法依据内容受限的语料针对考点 (话题、语言点等) 作出满足考纲规定的理想化设计。因此, 本研究利用大模型能自动生成内容多样的语料的优势, 通过人机协同的方式作出如下实验设计, 试图解决上述人工出题的问题 (试题的生成流程见图 1)。

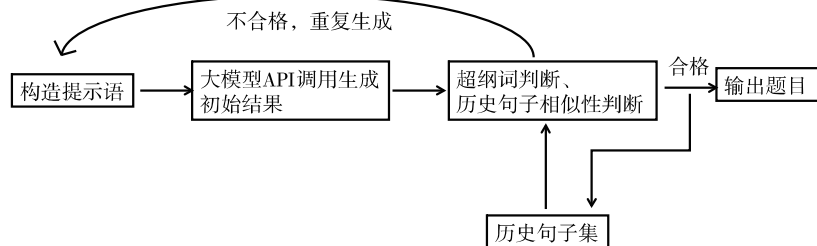


图1 试题生成流程

第一步：针对不同题型的考点设计规定考点的提示语，生成覆盖考纲要求且话题^⑥内容多样的批量题目。

第二步：判断生成题目的超纲率，保证除专有名词外无超纲词，当出现专有名词时超纲率小于 5% (传统 HSK 四级命题的接受度)。

第三步：为避免生成内容重复的题目，对生成的每个题目中的句子与之前生成的历史句子进行相似性检测。先采用 tf-idf 从词形角度检测句子，过滤掉词语组成高度相似的句子。鉴于 tf-idf 判断语义相似性能力不足，再通过少样本提示语对剩余句子进行语义相似性判断，判断结果为“强不相似、弱不相似、不确定、弱相似、强相似”其中一种，过滤掉“强相似”结果。

① 汪洁玉. 新 HSK 四级阅读真题分析及备考策略研究 [D]. 湖北工业大学硕士学位论文, 2021; 申琳. HSK 四级真题完成句子题型考点分析 [J]. 考试周刊, 2019, (10): 31.

② 申琳共研究 8 套真题, 即 2018 年人民教育出版社出版的《汉语水平考试真题集四级》中给出的五套真题、2015 年人民教育出版社出版的《HSK 考试大纲 (四级)》中的一套样卷, 2016 年高等教育出版社出版的《HSK 真题解析 (四级)》中列出的两套真题。参见: 申琳. HSK 四级真题完成句子题型考点分析 [J]. 考试周刊, 2019, (10): 31.

③ 汪洁玉. 新 HSK 四级阅读真题分析及备考策略研究 [D]. 湖北工业大学硕士学位论文, 2021.

④ 关联词内部也存在不均衡现象, 常考查转折、因果、条件、递进关系, 较少考查并列、假设和目的关系。

⑤ 申琳. HSK 四级真题完成句子题型考点分析 [J]. 考试周刊, 2019, (10): 31.

⑥ 话题和任务均起到限定题目内容的效果, 二者的限定效果有重叠, 鉴于话题的外延更抽象、广泛, 本研究在提示语中将使用话题而不是任务来限定题目内容。

四、提示语在自动出题过程中的运用实例

本部分根据第三部分提出的流程进行出题,重点展示前文提及的三种题型所涉及的提示语内容。提示语需要限定专业领域、提供具体的任务及操作要求。^① 本研究将从“考点先行”角度通过提示语限制出题形式和考点要求。超纲词检查在生成试题后完成,故不在提示语中加以限制。本研究使用 Python 3.11 作为编程环境,通过 API 请求的方式发送实验参数和提示语给 ChatGPT-4o 模型,其中参数设置参照官方推荐的配置。

(一) 选词填空题

本题主要考察语义搭配,据此设计提示语如下,其中(1)、(2)、(5)限制题目形式,(3)、(4)为考点要求,包括话题与词类:

提示语:假定你是一个国际中文教师,正在出一套 HSK 四级模拟题,请仿照下面的选词填空例子,出一道新的选词填空题。要求:(1)给出六个备选词,用 A、B、C、D、E、F 给每个备选词标号,标明词性;(2)给出六个句子(每个句子 25 个字以内),对每个句子挖空,被挖空的词在句子中用括号加以标识,针对每个被挖空的句子从六个备选词中选出一个合适的词填入;(3)句子 1 与“日常生活”话题相关、句子 2 与“教育”话题相关、句子 3 与“文化”话题相关、句子 4 与“科技”话题相关、句子 5 与“经济”话题相关、句子 6 与“职业工作”话题相关,标明话题;(4)要求备选词中包括名词、动词、其他词类;(5)最后给出答案,备选词间不能重复。例子如下:

备选词: A 躺 B 内 C 通过 D 坚持 E 因此 F 基础

句子 1: 她每天都 () 走路上下班, 所以身体一直很不错。

句子 2: 白天迷路时, 我们可以 () 太阳来判断方向。

句子 3: 很多城市都不允许在地铁 () 吃东西。

句子 4: 他辛苦了一天, 现在只想 () 在床上休息。

句子 5: 语言是人们交流的 (), 其他任何方法都不能与它相比。

句子 6: 我从小就想成为一名演员, () 考大学时我报了表演专业。

答案: D C B A F E

ChatGPT-4o 模型输出结果如下:

备选词: A 讨论 (动词) B 交通 (名词) C 理由 (名词) D 显然 (副词) E 创新 (动词) F 成熟 (形容词)

句子 1 (日常生活): 每天选择绿色出行, 不仅能减轻城市的 (), 还能为环境保护做出贡献。

句子 2 (教育): 课堂上, 老师鼓励学生大胆 (), 表达自己的观点, 提升沟通能力。

句子 3 (文化): 春节是中国的传统节日, 家人团聚的 () 之一是共同庆祝这一重要时刻。

句子 4 (科技): 在人工智能领域, 技术 () 的速度直接影响到新产品的研发效率。

句子 5 (经济): 经济发展需要 () 的政策支持, 以及市场条件的逐步改善。

句子 6 (职业工作): 他在这个行业工作了十年, 经验非常 (), 得到了同事和领导的一致认可。

答案: B A C E D F

可见, 大模型能生成形式正确的题目, 但生成的部分句子存在语义或语法问题: 如句子 1 中“减轻城市的(交通)”, “减轻”与“交通”语义搭配不当; 句子 5 中“经济发展需要(显然)的政策支持”, “显然”作为副词不能修饰政策支持。另外, 大模型将句子 4 中“技术(创新)的速度”中的名词“创新”误判为动词。

上述结果显示直接通过少样本提示语让大模型生成选词填空题的效果不佳, 可能原因是初始提示语内容过于复杂, 大模型难以理解和操作。因此本研究采取“加强提示语限定+分话题生成”的方法优化出题思路。一方面在初始提示语中增加“每个句子需具有合理的语义并符合汉语语法, 备选项词性需根

^① 饶高琦, 朱奕瑾. 大语言模型的语言特征计量及智能语伴应用探究 [J]. 语言治理学刊第 1 辑, 2024: 139-153.

据其所在句子来判定”的要求以减少大模型出错概率；另一方面在代码中不再要求大模型一次性生成涉及“日常生活、教育、文化、科技、经济”的完整试题，而是针对每个话题要求大模型批量生成多个句子，过滤超纲词，再从每个话题中抽取一个例子组合成一道完整试题。上述优化思路的机理是通过分步骤给出限定条件，使得大模型单次处理的任务变简单，大模型因此可以对每个限制条件给予充分关注。改进后大模型的生成示例如下：

备选词：A 紧密（形容词）B 动力（名词）C 红包（名词）D 细心（形容词）E 健身房（名词）

F 学习（名词）

句子1（日常生活）：为了保持健康，他每天都会去（ ）游泳。

句子2（教育）：学校的（ ）环境非常好，学生们都很喜欢在这里学习。

句子3（文化）：春节期间，人们会给小孩子发（ ）。

句子4（科技）：科技发展是推动社会进步的重要（ ）。

句子5（经济）：在经济全球化的今天，各国之间的经济联系越来越（ ）。

句子6（职业工作）：他在工作中总是非常（ ），从不出错。

答案：E F C B A D

上述结果无语法语义问题，且备选词词性均标识正确，可见优化后的出题思路能有效激发大模型的出题能力。然而，这一思路也不能彻底避免语义问题，有时仍会生成如“这个问题很复杂，我们需要（仔细）解决”等语义搭配不合理的句子，因此仍需人工核查。

针对对话型选词填空题，本研究使用优化后的思路，在单句型提示语基础上重新规定题目形式（“给出6组对话，每个对话由两个句子组成，每个句子25个字以内”），批量就每个话题单独生成多个对话，过滤超纲词，从符合条件的对话中根据话题要求重新组合成完整试题，示例如下：

备选词：A 爬山（动词）B 病人（名词）C 题目（名词）D 秀丽（形容词）E 性能（名词）F 减少（动词）

对话1（日常生活）

A：你周末有什么计划吗？B：我打算和朋友一起去（ ）。

对话2（教育）

A：这次考试的（ ）有点难。B：是的，我也觉得有点难。

对话3（文化）

A：你去过中国的黄山吗？B：去过，那里的风景真是（ ）。

对话4（科技）

A：你觉得这款新出的平板电脑怎么样？B：我觉得它的（ ）非常高，值得购买。

对话5（经济）

A：最近的出口情况怎么样？B：出口量一直在（ ）。

对话6（职业工作）

A：你为什么选择做医生？B：因为我想帮助（ ）。

答案：A C D E F B

上述示例结果未出现语法语义问题，备选词的词性均标识正确，可见优化后的出题思路也能高效生成对话型选词填空题。与单句型选词填空题类似，这种模式也会生成少量具有语义问题的病句，如“A：这款音箱的语音识别能力怎么样？B：它的语音识别能力非常（准确），几乎没有误差”，该例中“能力”不能被“准确”来修饰，因此还需人工核查结果。

从以上示例可看出，通过少样本提示语限制出题形式和考点要求，并针对不同话题批量生成题目的模式可以快速生成大量选词填空题。虽然个别句子可能存在语义搭配问题，需要人工核查结果；但大模型总体对挖空部分选取合理，通过提示语可精准限定考点，生成覆盖考纲要求的测试词类和话题。

（二）排列顺序题

本题的考点范围较广，包括复句间的语义关系，如递进、假设等，语篇中的代词指代功能、无标记

的多种语义和逻辑关系(如说明关系)等,答案需具有唯一性。鉴于考点的多样性,本文先展示考察复句间语义关系的提示语模板。在该模板中,条件(1)、(2)、(4)限定题目形式和答案的唯一性,(3)限定考点,明确考察话题“日常生活”和复句间的语义关系。

提示语:假定你是一个国际中文教师,正在出一套 HSK 四级模拟题,请仿照下面排列顺序题的例子,出一道新的排列顺序题。要求:(1)给出三个语义连贯的句子,每句字数为7-10个字;(2)随机打乱句子的排列顺序,用A、B、C给打乱顺序后的每个句子标号;(3)这三个句子与话题“日常生活”相关,句子间的语义关系需要具有并列、承接、选择、转折、假设、条件、因果、让步其中一种关系;(4)给出答案,保证答案具有唯一性。例子如下:

例如:A 可是今天起晚了 B 平时我骑自行车上下班 C 所以就打车来公司

答案:B A C

大模型输出结果如下:

A 我们决定去公园散步 B 天气预报说今天有雨 C 如果天气好就去爬山

答案:B C A

大模型给出了符合语法但语义连贯性不足的3句话,其中C句为一个使用“如果……就……”标记词的假设关系复句,可见大模型能理解假设这一术语。由于这一标记词不属于四级考点,因此提示语需明确要求使用四级考纲给出的假设关系标记词(“要是……就……”“……,否则……”),据此更新限制条件(3)中相关部分为“这三个句子间的语义关系需要具有假设关系,并使用标记词为‘要是……就……’或者‘……,否则……’”,同时提供对应例题,大模型生成的试题如下:

A 要是你不早点出发,就会迟到 B 今天的交通情况不太好 C 你最好提前一个小时出门

答案:B A C

上例中,大模型生成了含指定标记词(“要是……就……”)且语义连贯的句子,满足了出题要求。类似地,将复句提示语模板中的考点限定条件(3)改为“这三个句子间的语义关系需要包含专名及代词指代现象”并提供相应例题,即可对代词指代这一考点进行出题,大模型生成题目如下:

A 他每天早上都会去公园跑步 B 小王觉得这样对身体很好 C 这个习惯已经坚持了三年

答案:A C B

上例中,大模型生成的B句使用了专名“小王”,A句和C句中使用代词“他”指代小王。虽然生成结果包含代词指代现象,但使用情况不符合代词指代的语法,正确用法应在A句中先使用专名“小王”,再在B句中使用代词“他”。可见大模型虽能理解术语“代词”,但不能理解“代词指代”这样的语法现象,缺乏必要的语篇知识,其出题结果需要人工调整。

本题型还常考察句子间无标记的逻辑关系。以说明关系为例,此时可将上述提示语模板中条件(3)的考点要求改为“这三个句子间具有语义上的说明关系”并提供相关例题,大模型生成结果如下:

A 我每天早上都会去跑步 B 跑步不仅能锻炼身体,还能放松心情 C 早上空气清新,适合运动

答案:A C B

大模型生成的A句陈述了“我”的一种习惯,B句和C句分别说明了形成这个习惯的不同理由,符合出题要求,可见大模型能理解“说明关系”此类定义明确的术语。然而,如上文所言,大模型缺乏语篇知识,因此不擅长生成无明显标记词、具有一定语义复杂性和深度的逻辑排序题。例如下述真题:

A 仔细想想,原来是手机的短信提醒声 B 以为有人来了,于是忙跑去开门,可外面一个人也没有 C 我刚睡醒,就听到几下敲门的声音

此题是通过实词的语义,即敲门声、开门、手机短信提醒声内隐性地表达事件发生的先后顺序。当要求大模型仿照上述真题出题,并提示“三个句子间具有逻辑顺序,不要在句子中出现明显的标记逻辑顺序的词语”时,会生成如下不合格的结果:

A 我去厨房拿了一杯水 B 刚坐下时,我突然想起来 C 忘记关灯,立刻起身去关

答案:B A C

进一步要求大模型解释上述答案时，大模型回复：“我”首先想到忘记关灯，然后起身去拿水，回忆起忘关灯后，立刻去关。按照此解释，B 句应该与 C 句关系更紧密，但是大模型认为 B 句后应接 A 句，可见其不能理解“逻辑关系”这样定义抽象的考点，无法直接生成合格试题，需要人工辅助。

上述 3 例表明，通过设置不同的考点限定条件，大模型可生成测试具体考点的排列顺序题。然而，大模型的理解局限于定义明确的考点，如“假设、并列”关系、“说明关系”等，无法充分理解定义抽象、涉及语篇知识的考点如“代词指代”、无明显标记词的“逻辑关系”等，因此类似考点出题需要人工辅助。

（三）完成句子题

本题常考察句子中的内部句法关系以及特殊句型，如存现句，且需保证答案的唯一性。据此可以设计如下提示语，其中（1）、（2）、（4）限定题干和答案的形式，（3）限定句子的用词范围：

提示语：假定你是一个国际中文教师，正在出一套 HSK 四级模拟题，请仿照下面的完成句子题的例子，出一道新的完成句子题。要求：（1）给出一个包含 10-15 个字的句子，将这个句子切分成 4 或 5 个语块；（2）随机打乱语块的排列顺序，每个语块用空格隔开；（3）句子与话题“日常生活”相关；（4）给出答案，保证答案具有唯一性。例子如下：

题目：那座桥 800 年的历史 有了

答案：那座桥有 800 年的历史了。

大模型输出结果：

题目：每天早上 去公园 他 跑步 都会

答案：他每天早上都会去公园跑步。

上例中大模型理解了“语块”的意思，输出了理想结果。在此基础上，可进一步要求模型使用特殊句型，比如在条件（3）中增加“使用兼语句句型”的要求，并提供相应兼语句例题，大模型即可生成如下结果：

题目：朋友 我 送去 机场

答案：我送朋友去机场。

上例中大模型正确理解了提示语的要求，输出了兼语句“我送朋友去机场”，并成功将其拆分为 5 个语块。需要注意的是，当批量要求大模型一次性生成 10 道兼语句新题时，可能产生非兼语句的题目，如：一杯热茶 妈妈 给我 准备了（答案：妈妈给我准备了一杯热茶）。此外，上例中的答案也可以是“我去机场送朋友”，不满足“答案唯一”的提示语要求。因此，大模型在批量生成试题时仍需要人工检验。以上示例显示针对完成句子题的考试要点，通过对提示语中的关键词进行调优并不断加入限制条件，可使生成结果不断接近理想效果。

综上，本研究发现，通过精准设计提示语内容，且在明确试题形式和考点要求的情况下，大模型结合人工修订可以高效生成选词填空、排列顺序、完成句子这三类纯文字试题。由于选词填空题题干内容复杂，需要采取“分话题生成”的方法，先就各个话题利用提示语逐一生成批量试题，过滤超纲词后再从各个话题中抽取试题最终组合成一道完整的、含多个话题的题目。排列顺序、完成句子题题干简单，因此可以直接规定话题后通过提示语生成。图 2 显示了在 Python 环境中批量生成 3 种题型的主要代码逻辑。需要注意的是，在批量生成试题时，大模型间或会生成语义、语法不合规的句子，因此人工核查仍是必要的。

大模型对显性、定义明确的语言学术语理解较好，但对需要整合语篇知识的隐性关系理解较弱。当考点由具有显性形式标记的语法关系转变为依赖上下文推断的代词指代或无标记词的逻辑关系时，大模型的表现明显下降。例如，在代词指代题中，大模型虽然能够识别“代词”这一概念，却无法正确处理专名与代词之间的先行词一回指关系；在无标记的逻辑关系题中，大模型也无法稳定建立句子之间的因果关系。由此可见，大模型能够识别和执行语言学术语所对应的表层指令，但对需要跨句整合语篇信息的关系，其理解仍存在局限。

```

获取选词填空题提示语, 指定各类话题
调用大模型API生成初始结果, 判断超纲词, 得到“话题-选词填空候选集”
for 题型 in 试题:
    for _ in 该题型下生成题目数量:
        if 题型 is 选词填空:
            “话题-选词填空候选集” → 组合生成完整的选词填空题候选
        else:
            获取该题型提示语, 调用大模型API, 得到候选
        if 候选通过 (超纲词判断&历史句子相似性判断):
            题目加入结果集, 更新超纲率和历史句子集
        else:
            重复生成

```

图 2 批量生成 3 种题型的主要 Python 代码逻辑

五、大模型 HSK 四级试题自动出题效果评估

基于前文所述的出题思路, 本研究通过大模型 ChatGPT-4o 针对上述 3 种题型各自生成了 100 道模拟试题。为检验大模型的出题效果, 本研究将大模型生成的题目与 HSK4 级等量真题按照 3 种题型进行了随机混编。将混编试题制成调查问卷交给 6 名评分员进行评分, 并通过 SPSSAU 在线软件对评分结果进行统计分析。为防止问卷内容过长导致评分员疲劳作业, 调查问卷只选择了 16 道大模型所生成的题目与 16 道真题进行检验。包括选词填空题 12 道 (单句类型 6 道、对话类型 6 道)、排列顺序题 10 道、完成句子题 10 道, 每种题型中真题与大模型出题各占一半。由于问卷中包含了 HSK 真题, 评分员未选取熟悉 HSK 真题的考试命题专家, 而是选择具有国际中文教育相关专业背景的人员。评分员中 3 名为具有多年国际中文教学经验的国际中文教师, 3 名为国际中文教育专业的在读博士研究生。评分员知晓调查目的, 但是不清楚每道题目的来源。

由于目前学界尚未形成对大模型自动生成试题的评价体系,^① 本研究参考了王鸿滨和吕海辉^②针对 AQG 类中文测试题所使用的评价体系, 选取了语言流畅度、语法准确度、题目难度 3 个指标。每个指标为等距变量, 问卷采用李克特五级量表形式, 评分员从 1—5 中选取一个数字, 数值越大则表示程度或可能性越高, 1 到 5 分别代表低、较低、中、较高、高这 5 种程度。本研究的 3 个零假设分别是: 大模型出题和人工出题在语言流畅度、语法准确度、题目难度方面不存在差异。相应的备择假设为: 大模型出题和人工出题在语言流畅度、语法准确度、题目难度方面存在差异。

基于回收后的 6 份问卷填写结果, 研究人员计算了各指标的评分平均值, 由于样本数量较小且不满足正态分布, 研究人员选择了曼-惠特尼 U 检验对样本进行了显著性分析。统计分析的具体结果如表 1 所示:

表 1 大模型出题和人工出题问卷调查统计分析结果

	平均数		MannWhitney 检验统计量 U 值	p
	大模型 (n = 96)	人工 (n = 96)		
语言流畅度	4.3	4.6	3616.500	0.003 **
语法准确度	4.3	4.7	3710.500	0.005 **
题目难度	2.9	3.2	3732.000	0.014 *

从平均值角度看, 人工出题的得分在语言流畅度、语法准确度、题目难度方面均略高于大模型 (4.6 > 4.3, 4.7 > 4.3, 3.2 > 2.9), 表明二者的出题水平存在轻微差异, 即人工出题比大模型出题在语言上更流畅、语法上更准确、难度也更高。曼-惠特尼 U 检验结果显示: 在语言流畅度 ($p = 0.003 < 0.05$)、语

① 陈欣, 李蜜如, 周悦琦, 等. 基于大语言模型的试题自动生成路径研究 [J]. 中国考试, 2024, (12): 39-48.

② 王鸿滨, 吕海辉. 基于大语言模型的中文阅读测试题自动生成研究 [J]. 国际汉语教学研究, 2025, (1): 41-54.

法准确度 ($p=0.005<0.05$)、题目难度 ($p=0.014<0.05$) 这 3 个维度上, 零假设均被推翻, 即大模型出题和人工出题在上述 3 个维度存在差异 (接受备择假设)。这一结果与前文生成实验中观察到的现象具有一致性: 大模型生成的题目总体能够满足题型和考点要求, 但仍偶尔出现语义搭配不当、语篇关系处理不足等问题。因此, 大模型与人工命题之间的差异并非体现在“能否生成题目”这一层面, 而更多体现在语言表达的精细程度和复杂语言关系的处理能力上。

此外, 研究员还对每道题进行了图灵测试, 让评分员判断每道题由大模型出题的可能性, 并从 1—5 中选择一个数字进行评分, 1—5 分别代表可能性低、较低、不确定、较高、高。评分结果显示, 大模型生成试题的平均得分为 3.6, 略高于人工所出试题的平均得分 3.4, 可见评分员对大模型和人工出题的差异有一定程度察觉, 但很难明确区分二者 (平均分均趋近于“不确定”选项)。进一步分析评分结果发现, 只有约五分之一 (19.7%) 的大模型出题被评分员判定为 5 分 (即可能性“高”), 有三分之一的大模型出题 (31.2%) 被判定为 4 分 (即可能性“较高”)。可见评分员能正确识别出一半左右 (50.9%) 的大模型出题, 略高于随机概率 (50%)。人工出题中, 约 3% 的试题得分为 5 分, 16.3% 的试题得分为 2 分, 即只有约五分之一 (19.3%) 的人工试题能被明确识别, 约五分之二 (38.6%) 的试题被误判为 5 分或 4 分, 略低于随机概率。由此可见, 评分员较难分辨出大模型和人工出题的差异, 只有少量试题能被明确判定来源。

综上, 统计结果显示遵循本研究所使用的人机协同式出题方法, 其效果和人工出题在语言流畅度、语法准确度、题目难度方面存在轻微差异 (未来有待进一步考察)。这一结果补充了关于大模型自动出题能力的讨论。本研究总体支持王鸿滨和吕海辉^①的判断, 即经过考点约束、超纲词检测和相似度过滤流程后, 大模型确实能够生成接近人工命题水平的测试资源; 但也同时印证了刘玉屏等人^②的结论, 即生成试题的质量仍存在一定波动。由此看来, 依据人机协同式 HSK 四级试题出题框架的工作流程, 大模型虽能够承担大规模试题生成任务, 但尚不足以完全替代人工命题者。

六、结 语

本文探究了如何在 Python 环境中调用大模型 ChatGPT-4o 通过人机协同的方式自动生成三类 HSK 四级模拟试题 (选词填空、排列顺序、完成句子), 在保证试题内容不具有超纲词的情况下, 针对 3 种题型精准设计提示语模板, 高效生成能全面覆盖考纲考点、内容多样且重复率低的模拟题, 且问卷调查结果显示大模型出题效果和人工出题在语言流畅度、语法准确度、题目难度方面只存在轻微差异。本文提出的人机协同式出题流程可以基于 HSK 考纲的话题、语言点、词汇进行精准设置, 实现“考点先行”, 解决开篇提到的问题二。同时, 相比传统命题方式能降低命题的时间与人力成本, 缓解问题一, 同时减少对命题员资源的依赖, 提升教育公平性, 缓解问题三。

然而, 大模型出题也存在一系列问题。一是大模型的中文生成机制不成熟, 间或会生成不符合中文语法或者语义的句子 (常见于语义搭配方面)。二是大模型缺乏复杂的汉语语法知识和语篇知识, 对相关考点 (如代词指代现象、无明显标记词的逻辑关系等) 无法正确理解。在自动出题任务中, 术语层面的指令通过提示语较易实现, 而语言知识层面的考点理解则需要额外的人工审核。三是大模型在批量生成试题时偶尔会生成明显与提示语矛盾的结果 (比如要求批量生成兼语句时会生成少量非兼语句), 即提示语不能彻底约束大模型的输出效果。因此, 借助大模型的自动出题还需人工进一步核查、修订大模型生成的结果, 以此保证试题内容的正确性和可靠性。

本文设计的人机协同式 HSK 四级试题出题框架虽然利用计算机实现了考点约束、批量生成、超纲词检测和相似度过滤等环节的自动化, 但对于涉及语义搭配、复杂语篇关系以及特殊句式的试题, 人工干

① 王鸿滨, 吕海辉. 基于大语言模型的中文阅读测试题自动生成研究 [J]. 国际汉语教学研究, 2025, (1): 41-54.

② 刘玉屏, 欧志刚, 武晓琴. 生成式人工智能赋能国际中文教学的效果测评——以教学设计、HSK 模拟试题编写及作文评分为例 [J]. 民族教育研究, 2025, (1): 156-166.

预仍然是必要的。大模型在上述出题框架中的功能是在明确的考点约束和筛选机制（超纲词检测和相似度过滤）作用下提升命题资源的生成效能。早期自动出题研究主要依赖专家制定规则和生成模板，其优势在于可控性较强，但生成能力和内容覆盖面受到规则库规模限制。大模型虽突破了内容生成能力方面的限制，却将技术瓶颈转移到了输出可靠性和可控性上。本文提出的人机协同式试题出题框架，正是对这一瓶颈的回应。

限于篇幅，本文仅研究了 ChatGPT-4o 大模型，对其他主流大模型如豆包、DeepSeek 等未加以测试。不同大模型在架构、训练数据等方面的差异会影响其理解和生成文字的能力，后续研究可对比不同大模型在命题时的优势与不足，明确哪个（些）大模型最适合 HSK 考试出题。另外，大模型自动出题的结果在试题的实测难度、区分度等方面的问题也值得进一步探讨。

Research on Large Language Model – Assisted Automatic Question Generation for HSK Level 4 Test

MO Chou

(School of Linguistics and Language Resources, Beijing Language and Culture University, Beijing 100083, China)

Abstract: With the rapid development of international Chinese education, the demand for HSK test resources has increased substantially. The traditional expert-based test item generation approach faces systemic challenges in material acquisition, testing point coverage and expert resource allocation. Focusing on the cloze test, sentence-ordering tasks, and sentence-completion questions in the HSK Level 4 test, this paper constructs a human-machine collaborative automatic question generation framework based on ChatGPT-4o. Centered around the core concept of “testing points first”, the framework utilizes prompt engineering combined with out-of-scope vocabulary detection and similarity filtering mechanisms to precisely control elements such as topics, language points, and question types specified in the HSK Level 4 syllabus, to automatically generate test items that comprehensively cover the testing points of the HSK Level 4 test. This paper systematically examines the performance of large language models in the question generation process and examines the validity of the generated items using a questionnaire survey and statistical analysis methods. The results indicate that large language models are capable of generating items that satisfy the question type requirements and testing-point requirements, showing slight differences from expert-generated questions in dimensions such as linguistic fluency, grammatical accuracy and item difficulty. The study demonstrates that the human-machine collaborative question generation framework can effectively improve question generation efficiency, achieve precise coverage of testing points, reduce generation costs and reliance on expert resources, and provide a new practical pathway for intelligent question generation in the field of international Chinese education. However, due to remaining limitations of large language models’ generation such as semantic collocation errors and insufficient understanding of complex discourse-level knowledge, expert review and revision remain crucial steps to ensure the quality of the human-machine collaborative question generation framework.

Key words: large language models; HSK Level 4; automatic question generation; human-machine collaboration; prompt engineering

(责任编辑: 杨 恬)