

基于人工智能指令语生成高三英语阅读理解试卷

——以ChatGPT 5.6为例

喻梦旗

浙江师范大学外国语学院, 浙江 金华

收稿日期: 2026年9月8日; 录用日期: 2026年10月9日; 发布日期: 2026年10月15日

摘 要

随着生成式人工智能的快速发展, 大语言模型逐渐被应用于试卷开发, 为教师开展自动化命题提供了新的可能。本研究旨在设计AI指令语, 运用ChatGPT生成一份高三英语阅读理解试卷并探讨可行性及其存在的问题。研究依据《普通高中英语课程标准(2017年版2025年修订)》以及《中国英语能力等级量表》的相关阅读技能, 确定本试卷拟考查的阅读能力构念, 包括提取关键细节信息、概括主旨要义、推断词义、推断上下文隐含的语义逻辑关系等。在此基础上, 研究者编制命题细目表, 对语篇难度、体裁、题材、题型、阅读技能和认知层级作出规定, 并将上述要求转化为AI指令语, 运用ChatGPT (GPT-5.6)生成一套高三英语阅读理解试卷。随后, 从iTEST英语文本指难针分析试题难度以及从构念效度视角和布鲁姆认知目标层次考察阅读技能可行性与一致性。结果显示, ChatGPT虽能快速生成结构完整、题型多样的试卷, 但部分题目存在目标构念与实际考查不一致和题干表述晦涩等问题, 因此, AI可以提高命题效率, 但最终生成的试题仍需教师进行专业审核与修订。

关键词

高三英语阅读理解, 试卷开发, ChatGPT

Prompt-Driven Generation of Senior Tree English Reading Comprehension Tests

—A Case Study of ChatGPT 5.6

Mengqi Yu

College of Foreign Languages, Zhejiang Normal University, Jinhua Zhejiang

Received: September 8, 2026; accepted: October 9, 2026; published: October 15, 2026

Abstract

As generative artificial intelligence has developed rapidly, large language models have been applied to generate examination papers, which offers a promising prospect for teachers to implement automated item-generation. This study aimed to use ChatGPT to generate English reading comprehension test for Chinese Grade 12 students by designing AI prompts and to explore its feasibility and existing problems. Based on the relevant reading skills outlined in the Curriculum Standards for General Senior High School English (2017 Edition, Revised in 2025) and the Chinese English Proficiency Scale, this study identified the reading ability constructs to be assessed in this test, including extracting key details, summarizing main ideas, inferring word meanings, and deducing implicit semantic and logical relationships in context. Based on this, the researcher developed a test-item specification table by specifying requirements for text difficulty, genre, topic, question types, reading skills, and cognitive levels, and then converted these requirements into AI prompts. ChatGPT (GPT-5.6) was then used to generate a set of senior high school English reading comprehension test papers. Subsequently, the difficulty of the test items was analyzed by using the iTEST English Text Difficulty Index, and the feasibility and consistency of the reading skills were examined from the perspectives of construct validity and Bloom's taxonomy. The results showed that while ChatGPT can quickly generate test papers with complete structures and diverse question types, some problems exhibited inconsistencies between the intended constructs and what was actually assessed, as well as obscure question stems. Therefore, although AI can improve test-development efficiency, the final test items still require professional review and revision by teachers.

Keywords

Grade 12 English Reading Comprehension, Test Development, ChatGPT

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

自 2022 年 11 月 30 日 ChatGPT 正式发布后, Fütterer 等统计超 1600 万条推特, 梳理全网围绕 ChatGPT 产生的热点话题, 发现公众极其关注 ChatGPT 赋能教育带来的发展机遇、现实局限以及多重深远影响[1]。学生可借助 ChatGPT 获取个性化学习辅导与即时反馈; 写作时, 学生也可以运用 ChatGPT 来头脑风暴获得写作思路[2]。教师也可借助 ChatGPT 生成对话文本, 或是围绕同一主题生成多种文体的学习素材; 同时教师能够指令 ChatGPT 灵活调整语言难度, 以此设计分层差异化教学材料[3]。从实践应用的角度看, ChatGPT 在语言教学、语言学习与语言测评等领域具有较高的应用价值[4]。对教师而言, ChatGPT 有效缩短教师命题时间以及缓解教师在试题编制方面的工作压力; 对语言学习者而言, 人工智能语言模型能够根据学习者兴趣、能力水平和题型需求提供更具针对性的学习内容; 对语言测评人员而言, ChatGPT 不仅能够降低试题开发所需的时间与人力成本, 提高多项选择题的编制效率, 还可为后续自适应测评与个性化学习系统的建设提供技术支持。

当前学界关于 ChatGPT 辅助英语命题的研究, 大多集中于大学四级阅读理解题型的生成与分析, 然而, 面向高三高考备考场景的试卷开发以及设计相关 AI 指令语仍存在明显空白, 高中英语教师长期面临学生升学、命题任务繁重等压力, 因此 AI 命题可侧重于高中场景。除此之外, 以往大多数研究采用 ChatGPT 3.5 (或更早版), 也有研究使用了基于 GPT-4 等更新技术来生成试题[5], 但 AI 更新迭代的速度较快, 已具

有更新的 ChatGPT 5.6 模型。基于以上空白，本研究将围绕阅读材料类型、语言难度等提示语，引导更先进的 ChatGPT 5.6 生成一套面向高三学生的英语阅读理解试题。随后，本研究运用了 iTEST 英语文本指难针工具[6]来分析语篇的词汇和语法难度，以及运用布鲁姆认知目标分类和构念对生成试题的认知目标层级进行分析考察是否有效考察阅读技能，从而评价 ChatGPT 5.6 在高中英语阅读试题可行性。

2. AI 辅助高三英语阅读理解命题的生成工具与指令语设计

2.1. 研究工具

本研究采用 GPT-5.6 Sol 模型。GPT-5.6 Sol 是 GPT-5.6 系列中的旗舰推理模型，主要面向研究、知识工作、编程、科学分析及复杂多步骤任务，能够根据用户提供的文本、任务要求和输出规范生成相应内容。在本研究中，研究者通过向模型输入高三英语阅读理解试卷的构念、项目细目表、语篇选材等内容得到一份完整的阅读试卷。

2.2. 基于 AI 指令语的高三英语阅读理解试卷设计框架

2.2.1. 构念

阅读理解测试应主要考查学生的阅读理解能力、技能或策略阅读，试题的编写首先需要明确测试内容，从而保证信度[7]。因此，根据《中国英语能力等级量表》的阅读理解能力总表(第四级)[8]以及《普通高中英语课程标准》(2017 年版 2025 年修订)[9]中的必修和选修理解性技能以及学业质量水平二，系统梳理文件中的有关学生阅读技能(如表 1)并确定本文最终考察的阅读技能，进一步依据修订版布鲁姆认知目标分类确定各构念维度所对应的认知层次：

Table 1. Reading related skills in English Curriculum Standards for Senior High Schools and China’s Standards of English Language Ability

表 1. 课标与中国等级量表的相关阅读技能

必修 阅读技能	辨识不同类型的语篇，把握其语篇结构及语言特征；梳理语篇中主要事件的发生、发展和结局等；提取语篇中的主旨要义和关键信息；分析比较书面语篇中信息、知识或观点之间的关联；概括书面语篇中的主要信息、观点及其支撑论据；辨识语篇中关键词和关键概念以快速获取相关信息；根据上下文线索或非文字信息推断词语的意义；识别书面语篇中常见的指代和衔接关系；分析书面语篇段落间和上下文间的逻辑关系；评价作者的写作策略和技巧；整合不同类型语篇的主要信息，建立信息间的关联；
选修 阅读技能	提取语篇中的主要信息、观点和文化意涵；根据语篇中的定义线索获取概念性词语或术语的意义；梳理语篇中的时间顺序、空间顺序及过程顺序；根据书面语篇标题推测语篇的体裁和结构；根据语境线索或图表信息预测语篇的主要内容；分析书面语篇中的内容要点和相应的支撑细节；根据连接词推断语篇中上下文的语义逻辑关系；根据语篇的语境和语言表达，推断隐含意义；根据书面语篇中的事实进行归因分析和逻辑推理；区分书面语篇中的事实与观点，分析两者间的逻辑关系；概括书面语篇中的核心观点、支撑论据及论证方法；批判、评价语篇中所涉及的观点、文化现象及所反映的情感、态度和价值观；
中国能力 等级量表	提取细节信息，概括主旨要义；分析句子和篇章结构，理解意义之间的关系；区分事实和观点，进行简单推断；
学业水平 质量二	理解、概括、分析和评价学习活动中书面语篇的主要内容；在读的过程中，根据语篇的类型、结构、语言等方面的特点，从错综复杂的信息中推断各种知识之间的内在关系；根据不同学习内容，分析各种知识之间的内在关联和差异，发现产生差异的各种原因，从中梳理出它们之间的逻辑关系；理解语篇中特定语言的使用意图以及语言在反映情感态度和价值观中所起的作用；在语境中理解词语表达的意义以及使用者表达的意图和态度；理解和欣赏优秀英语文学作品的内容及语言表达形式；从作品的意蕴中获得积极的人生启示。

布鲁姆将认知领域的学习目标由低阶到高阶划分为六个层级(如表 2)，依次为知识、领会、应用、分析、综合与评价[10]。在 2001，安德森将其修订版为记忆、理解、应用、分析、评价、创造[11]。Moore

与 Stanley 将布鲁姆认知目标划分为低阶思维与高阶思维，其中知识、理解和应用属于低阶思维，分析、综合和评价属于高阶思维[12]。

Table 2. Bloom’s taxonomy of cognitive objectives
表 2. 布鲁姆认知目标分类表

认知层级	定义	小分类
记忆	从长时记忆中提取相关知识。	识别、回忆
理解	从口头、书面或图像信息中建构意义。	解释、举例、分类、概括、推断、比较、说明
应用	在特定情境中执行或使用某种程序。	执行、实施
分析	将材料分解为组成部分，并判断各部分之间以及部分与整体之间的关系。	区分、组织、归因
评价	根据一定标准作出判断。	检查、评论
创造	将要素重新组合，形成新的整体或新的产出。	生成、计划、制作

低层次思维主要指对文本表层信息的理解，包括理解文中具体信息和推断生词词义；高层次思维则强调对文本深层意义的加工，包括理解主旨要义、进行判断和推理、把握文章基本结构，以及理解作者意图、观点和态度[13]。Li 等对大学英语四级阅读理解所涉及的九类技能进行了认知层次划分，其中将“识别主旨要义”归入理解层次，将“概括主旨要义”归入分析层次，并将“根据上下文推断词语和短语的含义”归入应用层次[14]。然而，本文认为上述分类仍有商榷空间，根据修订版布鲁姆分类，“概括”(summarizing)为理解层次下的具体认知过程，其核心是强调学生能够从已呈现的信息中提炼总体主题或主要观点[10]；本文考察的概括主旨即为对已呈现的显性信息中归纳主旨要义，因此，将其归入理解层次。其次，根据上下文推断词义的核心活动是借助语义、句法和篇章线索理解目标词在特定语境中的意义，其主导认知过程更接近理解层次中的“推断”(inferring)或“解释”(interpreting)，而非应用层次。布鲁姆分类中的“应用”主要强调在具体或新的情境中执行、选择和运用某种已经掌握的程序[10]；若仅因学习者使用了语境线索或词汇知识便将词义推断归为应用，则可能造成使用知识与运用程序之间的概念混淆。因此，本文将根据上下文推断词义题均归入理解层次(如表 3)。

Table 3. Constructs of reading comprehension
表 3. 阅读理解构念

阅读理解构念	布鲁姆层次
1. 提取关键细节信息	理解
2. 根据上下文推断词语的意义	理解
3. 概括主旨要义	理解
4. 识别常见的书面语篇中指代关系	理解
5. 识别常见书面语篇中衔接关系	理解
6. 根据语篇的语境和语言表达，推断隐含意义	分析
7. 根据上下文推断隐含的语义逻辑关系	分析
8. 理解语篇中特定语言的使用意图以及语言在反映情感态度和价值观中所起的作用	分析
9. 根据书面语篇中的事实进行归因分析和逻辑推理	分析
10. 区分书面语篇中的事实与观点，判断事实对观点的支撑关系	分析
11. 评价作者的写作策略和技巧	评价

需要注意的是, 尽管这两个构念“根据上下文推断隐含的语义逻辑关系”和“根据书面语篇中的事实进行归因分析和逻辑推理”都涉及英语阅读中的推理活动, 但二者的考查重点并不相同。前者主要考查读者能否依据上下文, 识别句子、段落或信息之间未被显性连接词标示的逻辑关系, 包括因果、转折、递进、并列、条件和解释说明等; 后者则主要考查读者能否整合语篇中明确呈现的事实、行为、数据和结果, 对事件发生的原因、人物动机、可能后果或潜在结论作出合理推断。二者的区别在于, 前者侧重判断语篇内部各部分如何关联, 属于对信息组织方式和逻辑结构的识别; 后者侧重解释事实为何发生以及能够推出什么, 属于在事实基础上的归因与结论推导。

2.2.2. 项目细目表

Heaton 系统梳理七大类阅读测试题型为匹配题、正误判断题、多项选择题、填空题、排序题、完形填空、开放式问答题[15]。英文阅读测试中的匹配题通常可分为词汇匹配题和句子匹配题, 而考虑到这类题型整体难度相对较低, 更偏向基础语言知识与信息对应能力的考查, 与高三学生应重点考查的阅读理解能力匹配度不高。董曼霞建议完形填空题的命制在考点设计方面应以意义因素为主, 同时适当覆盖语法因素和词汇因素, 以发挥完形填空题型的优势, 而本研究测试的主要是阅读技能。因此, 确定本研究考查的阅读题型为正误判断题、多项选择题、填空题、七选五、开放式问答题[16]。

Table 4. Test specification table

表 4. 项目细目表

题型	技能	认知层次	频率
选择题	提取关键细节信息	理解	4
	概括主旨要义	理解	3
	根据上下文推断词语的意义	理解	2
	识别书面语篇中常见的指代关系	理解	1
	根据上下文推断隐含的语义逻辑关系	分析	2
	根据语篇的语境和语言表达, 推断隐含意义	分析	3
	根据书面语篇中的事实进行归因分析和逻辑推理	分析	2
	区分书面语篇中的事实与观点, 判断事实对观点的支撑关系	分析	1
判断题	理解语篇中特定语言的使用意图以及语言在反映情感态度和价值观中所起的作用	分析	1
	提取关键细节信息	理解	1
	识别书面语篇中常见的衔接关系	理解	1
	根据书面语篇中的事实进行归因分析和逻辑推理	分析	1
七选五	区分书面语篇中的事实与观点, 判断事实对观点的支撑关系	分析	1
	概括主旨要义	理解	1
	识别书面语篇中常见的衔接关系	理解	1
开放题	根据上下文推断隐含的语义逻辑关系	分析	3
	提取关键细节信息	理解	1
	概括主旨要义	理解	1
	理解语篇中特定语言的使用意图以及语言在反映情感态度和价值观中所起的作用	分析	1
	评价作者的写作策略和技巧	评价	1
总计			32

在编制项目细目表时，试题难度与能力层级的均衡性是需要重点考虑的因素。为保证题型设置和能力考查分布具有合理依据，本文对研究了近几年对高考英语阅读理解真题分析的文章，重点考察其在低阶能力与高阶能力之间的分布情况。倪晗和罗晓杰对 2004~2016 年阅读理解试题进行考察后发现，试题对高层次思维能力的考查总体少于对低层次思维能力的考查；其中，有五年二者数量相等，仅有一年高层次思维能力考查数量多于低层次思维能力[13]。刘科对 2019 高考卷进行分析，发现阅读理解试题整体以低阶思维能力考查为主。全国卷低阶思维题占 66.67%，自主命题卷占 63.33%，均高于高阶思维题比例。不同试卷之间存在一定差异，其中全国卷 3 低阶思维占比最高，为 73.33% [17]。栗瑞莲发现 2020 年高考阅读理解题整体更侧重高阶思维能力[18]。王石桂梳理 2021~2023 年新高考英语全国 I 卷阅读理解的能力考查较为稳定。三年中，每年低阶思维题均为 9 题，高阶思维题均为 11 题，高阶思维考查略多于低阶思维。整体上呈现出低阶能力与高阶能力相对均衡的特点[19]。因此，本文在设计项目细目表时会保持低阶思维题目与高阶阅读题目的考查比例保持平衡。最终，整套阅读试题共 32 题，考察低阶思维共 16 题，高阶思维 16 题，二者比例等同(如表 4)。本文将运用 iTEST 英语文本指难针工具，对 ChatGPT 生成的阅读语篇的词汇以及语法难度进行检测，以判断其是否符合高三学生的阅读水平。此外，该指南针的难度划分标准依据中国能力等级量表来对语篇难度进行客观测量，有助于提高生成试题的适用性和可靠性。

为进一步呈现不同阅读技能在各类题型中的考查情况，并比较各技能的出现频率及题型分布，现将相关数据整理为频率分布图(如图 1)。该图能够较为直观地反映各项阅读技能的考查重点及其在选择题、判断题、七选五和开放题中的具体分布。

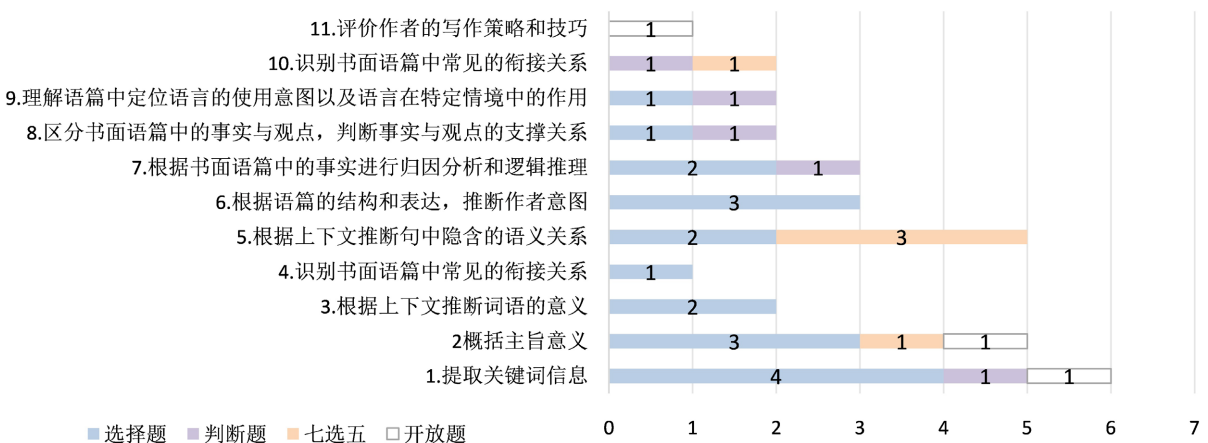


Figure 1. Frequency distribution of four reading skills across four types of English reading comprehension items
图 1. 各阅读技能在四类英语阅读理解题型中的频率分布图

2.2.3. 阅读语篇选材原则

语篇是阅读理解试题设计的基础，语篇质量直接影响试题的效度和学生作答的公平性。语篇形式具有多样性，既包括对话、访谈、记叙文、说明文、应用文、议论文、歌曲和诗歌等连续性语篇，也包括图表、图示、网页和广告等非连续性语篇。但是，本次阅读试题所采用的材料总体上仍以连续性语篇为主。

《普通高中英语课程标准》(2017 年版 2025 年修订)梳理了人与自我、人与社会和人与自然三大主题语篇主题，本文将此作为 AI 指令语。课标提出应设置日常生活、学习活动和跨文化沟通等情境，考查学生在情境中分析和解决问题的能力。情境设置需体现教育性、真实性和适用性[9]。

(1) 教育性强调情境要坚持正确价值导向，引导学生讲好中国故事，增强文化自信，促进学生全面发展。

(2) 真实性强调情境要贴近学生英语学习和生活实际,体现真实语言使用,具有实际交际意义,便于学生展示综合语言运用能力。

(3) 适用性强调情境要符合学生生活经验和认知水平,能够激发学生思考和探究。

Hughes 提出阅读测试应以测试目标为基本依据,确保材料在文本类型、主题和阅读能力维度上具有充分的代表性。语篇的长度、难度和篇章结构应与题目任务相适配,使细读、略读、查读、扫读等不同阅读能力能够依托相应的文本特征得到有效考查。同时,测试材料应保持一定的多样性,通过设置多篇文本避免单一主题、体裁或背景知识对测试结果的偶然影响,从而提高测试的信度与内容效度。此外,选文还应遵循公平性原则,避免使用学生已接触过的文本、常识依赖性过强的材料、文化负载过重的内容以及可能引发情绪不适的话题,以保证测试结果主要反映学生的真实阅读理解能力[20]。

本文依据课标中词汇范围(参照附录 2 词汇表),语法范围(语法项目表参照附录 3) [9]将其作为 AI 指令语来控制本试卷阅读语篇的语言难度。所选语篇字数控制在 290~325,应以高三学生已经学习或应当掌握的词汇、词组、习惯搭配和语法结构为主,整体语言难度原则上不得超过选择性必修课程规定的水平。为保持语篇真实性,可以保留少量超出课标范围的词汇,但这些词汇不得影响学生理解文章主旨和关键信息,并应能够根据上下文推断词义,必要时提供中文注释。控制长难句的数量和复杂程度,避免连续出现生词、复杂从句或多层嵌套结构,确保语篇难度适中,符合高三学生的实际英语水平。

试题首先应保证考点明确,每道题只能设置一个最佳答案,各选项之间不应相互牵连;同一道题的选项在长度、结构、类属和难度等方面应尽可能保持一致,避免学生依据形式特征而非实际语言能力作答。同时,命题者还应防止题干语境或其他题目无意泄露答案,使干扰项无论单独出现还是放入题干后均具有一定的似真性,并避免通过故意设陷阱或模糊表达形成诡题[21]。Heaton 认为,多项选择题的所有选项都应符合语法规范,不能使学生仅通过排除存在语法错误的选项获得答案。同时,命题者应避免使用双重否定或故意设置语言陷阱[15]。由此可见,阅读理解选择题的选项设计应服务于语篇理解,使学生依据文本内容作出判断,而不是借助形式线索或猜题技巧完成作答。

阅读理解的干扰项编制方式主要包括以下几种[21]:

- (1) 信息组合:把原文中的几个信息组合在一起。
- (2) 混淆关系:混淆原文中的意义关系。
- (3) 局部信息:根据原文中的某一个概念进行扩展发挥,或修改原文中的某一个概念。
- (4) 信息删减:减少原文的信息。
- (5) 添加信息:添加原文的信息。
- (6) 篇外信息:干扰项的信息来源于原文以外的信息。

研究者整合以上信息构建完备的 AI 指令语,通过人工智能生成一份完整试卷,继而针对该试卷展开分析。

3. 结论

3.1. 不同指令输入模式中 AI 命题的局限及优化

本研究在实验阶段采用多种指令方案生成英语试题。研究初期将全部提示指令一次性输入人工智能模型,整套试卷一次性生成。经过五六轮试测后,研究者完成这些试卷后发现该方式产出的试卷质量较差。随后本研究调整为分批次输入指令,进行五轮以上的试测但结果依旧是不尽如人意:尽管人工智能能够按照指令要求覆盖“人与自我、人与社会、人与自然”三大主题,却难以把控语篇情境的丰富度,生成多套试题的场景大多仅围绕校园学习或城市生活展开。针对这一缺陷,本研究采用分步输入指令的方式,并且赋予 AI “专业命题专家”的角色定位,研究者作为命题小组一员与 AI 命题专家开展多轮对话

沟通,逐一确认阅读试题的体裁、题材范围,以此保障 AI 生成试题的合理性与情境多样性。

3.2. 试卷语篇材料的体裁、主题与语言难度分析

本试卷共包含 7 篇语篇,涵盖应用文、记叙文、说明文和议论文。其中,Passage 1 为旅游信息类应用文,Passage 2、4、7 为记叙文,Passage 3、5 为说明文,Passage 6 为议论文,涵盖“人与自我”“人与社会”和“人与自然”三大主题语境,且不同语篇之间题材重复度较低。各篇词数约为 291~325 词,也在范围之内。篇幅较为均衡,体裁分布较为丰富,基本符合高三英语阅读理解试卷的要求。在语言难度方面,采用 iTEST 英语文本指难针对语篇的词汇和语法进行检测,结果显示,试卷中的绝大多数词汇和语法结构均处于高中课程标准规定范围内,仅出现少量超纲词汇和极少量超纲语法结构。

3.3. 目标构念与实际考查不一致

试题审查时,发现部分题目存在目标构念与实际作答过程不一致的问题。首先,Passage 1 第 3 题考察的阅读技能为概括主旨要义而实际考查的是应用文的交际目的 *What is the main purpose of the text?* 其次,Passage 3 第 13 题借助 *Without that information* 以及 Passage 4 第 19 题基础连接词“Instead”,学生仅需识别其替代关系即可作答,无需深入分析上下文。因此,该题实际考查的是“识别语篇中的常见衔接关系”,属于理解层级,而非原本意图考查的“推断上下文隐含的语义逻辑关系”分析这一高阶思维层次。Passage 5 第 24 题,考查的是区分事实和观点题,实际上只是对原文提取相关细节信息。ChatGPT 在生成阅读试题,较难生成有效的高阶题型[22],可能使模型更容易习得训练语料中的表层高频表达,较难完成挖掘文本隐含逻辑的任务[23]。当指令包含多条相互叠加的显式约束时,模型容易出现指令遵从失败的情况[24]。在 AI 生成阅读理解试题任务中,指令往往同时涵盖项目细目表、语篇选材、干扰项设计等多项命题要求,多重约束叠加进一步加大模型执行负担,进而造成生成试题的实际考查目标偏离提示词预先设定的考查构念。

此外,Passage 3 第 15 题,标题 *Which statement is an interpretation rather than a directly reported condition or action?* 晦涩难懂,学生可能对理解命题本身存在问题。因此,后续命题应加强题目任务、文本证据与目标技能之间的一致性,减少术语辨识和表层词语匹配,突出跨句整合与真实推理。

4. 教学应用启示

人工智能辅助教师命题,能够快速、批量生成多样化的阅读文本及配套试题,显著提高命题效率,减轻教师的工作压力。然而,人工智能无法替代教师的专业判断,其生成内容仍需结合课程标准、教学实际和测评目标进行审查与修订。因此,教师的人工审核是确保试题质量与测评效度的必要前提,同时也启示未来需要重视对教师的提示词工程能力与 AI 素养的培养,帮助教师合理运用提示词生成一份高三试卷,甄别与使用 AI 生成的语言教学资源[22][25]。

5. 研究局限性

本研究仍存在一定局限。由于研究条件所限,尚未对全部 32 道试题进行系统编码,也未引入多名专家进行独立评价和评估者间信度检验。除此之外,本研究主要评估了 AI 生成阅读理解题目的质量,并未对 AI 生成阅读文本本身的语言质量展开详细分析,因此后续研究可借鉴[25]的分析,系统评估 AI 产出阅读篇章的质量与教学适配性。本研究主要采取了 ChatGPT,建议后续研究可加入 DeepSeek 或豆包等 AI 工具共同生成一份试卷,进行比较不同的大语言模型生成的质量如何。本研究在实际测试环节仍存在一定局限,未来研究可进一步加入被试,邀请更多高三学生完成 AI 生成的英语阅读试卷,并结合学生的实际作答表现、试题难度、区分度及学生反馈等数据,引入专家评审来对 AI 生成试题的质量进行更全面的

验证, 从而进一步检验其在高中英语测试中的适用性与有效性。

参考文献

- [1] Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., *et al.* (2023) ChatGPT in Education: Global Reactions to AI Innovations. *Scientific Reports*, **13**, Article No. 15310. <https://doi.org/10.1038/s41598-023-42227-6>
- [2] Chan, C.K.Y. and Hu, W. (2023) Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. *International Journal of Educational Technology in Higher Education*, **20**, Article No. 43. <https://doi.org/10.1186/s41239-023-00411-8>
- [3] Kohnke, L., Moorhouse, B.L. and Zou, D. (2023) ChatGPT for Language Teaching and Learning. *RELC Journal*, **54**, 537-550. <https://doi.org/10.1177/00336882231162868>
- [4] Lin, Z. and Chen, H. (2024) Investigating the Capability of ChatGPT for Generating Multiple-Choice Reading Comprehension Items. *System*, **123**, Article 103344. <https://doi.org/10.1016/j.system.2024.103344>
- [5] Shin, D., Lee, J.H. and Kim, K. (2025) An Exploratory Study on Two Automated Item Generators for Generating L2 Reading Test Items. *RELC Journal*, 1-16. <https://doi.org/10.1177/00336882251326284>
- [6] 金檀, 陆小飞, 等. 英文分级指难针[EB/OL]. 一针三库智能教研平台. <https://languagedata.net>
- [7] 王笃勤. 英语阅读理解试题的设计[J]. 基础英语教育, 2006, 8(2): 79-83.
- [8] 教育部, 国家语言文字工作委员会. 中国英语能力等级量表[M]. 北京: 高等教育出版社, 2018.
- [9] 中华人民共和国教育部. 普通高中英语课程标准(2017年版 2025年修订) [M]. 北京: 人民教育出版社, 2025.
- [10] Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956) Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. David McKay.
- [11] Anderson, L.W. and Krathwohl, D.R. (2001) A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman.
- [12] Moore, B. and Stanley, T. (2010) Critical Thinking and Formative Assessments: Increasing the Rigor in Your Classroom. Routledge.
- [13] 倪晗, 罗晓杰. 2004-2016年浙江省高考英语阅读理解试题思维能力层次探析[J]. 基础外语教育, 2017, 19(2): 91-99+111.
- [14] Li, B., Qin, Q., Li, W. and Mao, C. (2026) Psychometrics and Linguistics in ChatGPT-Generated Reading Tests Compared with CET-4. *Journal of Psychoeducational Assessment*, **44**, 797-821. <https://doi.org/10.1177/07342829261444828>
- [15] Heaton, J.B. (1988) Writing English Language Tests. Longman.
- [16] 董曼霞. 高考英语全国卷完形填空篇章特点与考点效度研究[J]. 教育测量与评价(理论版), 2014(8): 57-64+47.
- [17] 刘科. 布鲁姆目标分类视角下高考英语阅读理解题思维能力层次分析: 以 2019 年高考试卷为例[J]. 考试研究, 2020(3): 67-73.
- [18] 栗瑞莲. 基于中国英语能力等级量表的高考英语试题分析——以 2020 年高考英语全国卷及新高考 I 卷阅读理解题为例[J]. 基础教育课程, 2020(Z2): 30-35+70.
- [19] 王石柱. 基于高考试题语篇分析优化高中英语阅读教学策略研究——以近四年新高考英语全国 I 卷及新译林版高中英语教材为例[J]. 英语教师, 2024, 24(7): 79-83.
- [20] Hughes, A. and Hughes, J. (2020) Testing for Language Teachers. 3rd Edition, Cambridge University Press, Cambridge.
- [21] 曾用强. 英语语言测试基础[M]. 广州: 华南理工大学出版社, 2009.
- [22] Sihite, M.R., Meisuri, M. and Sibarani, B. (2023) Examining the Validity and Reliability of ChatGPT 3.5-Generated Reading Comprehension Questions for Academic Texts. *Randwick International of Education and Linguistics Science Journal*, **4**, 937-944. <https://doi.org/10.47175/rielsj.v4i4.835>
- [23] OpenAI (2023) GPT-4 Technical Report. OpenAI. <https://openai.com/research/gpt-4>
- [24] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., *et al.* (2022) Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 27730-27744. <https://doi.org/10.52202/068431-2011>
- [25] Allaithy, A. and Zaki, M. (2025) Evaluation of AI-Generated Reading Comprehension Materials for Arabic Language Teaching. *Computer Assisted Language Learning*, **39**, 1296-1328. <https://doi.org/10.1080/09588221.2025.2474037>