

人机协同视域下生成式人工智能辅助高中英语词汇选择题编写研究

王欣雅

浙江师范大学外国语学院, 浙江 金华

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

摘要

在生成式人工智能技术快速发展的背景下, AI为语言测试开发提供了新的可能, 但现有研究对教师如何运用AI进行人机协同命题的实践过程关注不足。本研究以DeepSeek为辅助工具, 探索编写高中英语词汇选择题的可行路径与协作模式。研究发现, DeepSeek在试题格式规范、教材主题契合和语言表达方面具有较好的辅助价值, 但AI生成试题仍存在考查内容覆盖不均衡、语境真实性与支持程度不足、正确答案唯一性和干扰项有效性不稳定等问题, 需经人工审查与修改方可达到使用标准。研究建议, 教师在使用AI辅助命题时应坚持教师主导、人机协同的原则, 以提高英语词汇测试开发的效率与质量。

关键词

生成式人工智能, 人机协同, 词汇测试, 选择题编写, 高中英语

A Study on Generative AI-Assisted Senior High School English Vocabulary Multiple-Choice Item Generation from the Perspective of Human-AI Collaboration

Xinya Wang

College of Foreign Languages, Zhejiang Normal University, Jinhua Zhejiang

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

Abstract

Against the backdrop of rapid advances in generative artificial intelligence, AI has opened up new

文章引用: 王欣雅. 人机协同视域下生成式人工智能辅助高中英语词汇选择题编写研究[J]. 教育进展, 2026, 16(9): 1085-1097. DOI: 10.12677/ae.2026.1692000

possibilities for language test development. However, existing research has paid insufficient attention to how teachers can engage in human-AI collaborative item generation in practice. This study employs DeepSeek as an assisting tool to explore feasible pathways and collaborative models to develop multiple-choice vocabulary test items for senior high school English. The findings reveal that DeepSeek demonstrates considerable value in terms of item formatting consistency, thematic alignment with textbook content, and naturalness of language expression. Nevertheless, AI-generated items still exhibit limitations, including uneven coverage of target vocabulary, inadequate contextual authenticity and support, and instability in answer uniqueness and distractor effectiveness, all of which necessitate human review and revision before the items can meet quality standards. The study recommends that teachers adhere to the principle of teacher leadership and human-AI collaboration when utilizing AI for item development, so as to enhance both the efficiency and quality of English vocabulary test development.

Keywords

Generative Artificial Intelligence, Human-AI Collaboration, Vocabulary Testing, Multiple-Choice Item Generation, Senior High School English

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

词汇是语言能力发展的基础,也是英语教学和评价的重要内容。《普通高中英语课程标准(2017年版2025年修订)》(以下简称《课标》)强调,高中英语教学应关注学生语言能力的发展,引导学生在真实语境中理解和运用语言知识[1]。词汇学习不应停留于单词形式和中文释义的机械记忆,而应关注学生在语境中理解词义、辨析词义以及灵活运用词汇表达意义的能力。因此,词汇测试不仅承担检测学生词汇掌握情况的功能,也应服务于后续教学调整,以实现“教-学-评”一体化。

然而,在实际教学中,词汇测试仍存在一定问题。部分教师设计词汇测试时主要通过听默写、英汉互译或简单选择题检测学生记忆情况,难以全面反映学生词汇理解和应用能力。宋德龙(2008)指出,传统词汇教学与测试容易造成学生记住单词但不会使用的现象,学生虽然能够完成词汇默写,却可能无法在句子或文本中准确理解和运用词汇[2]。再者,高质量词汇选择题的编写需要命题者同时考虑测试目标、题目难度、语境真实性以及干扰项有效性,对教师专业能力提出了较高要求。

近年来,生成式人工智能(Generative Artificial Intelligence,以下简称 AI)技术迅速发展,以 ChatGPT 为代表的大语言模型开始应用于语言测试领域[3]-[5]。在此背景下,探索人与 AI 协同命题的有效模式,既能丰富 AI 辅助语言测试开发的实践案例,也能为一线教师提供可操作的命题流程参考。因此,本研究尝试将 DeepSeek 引入高中英语词汇选择题编写过程,探索如何通过指令设计引导 DeepSeek 产出符合构念和命题规范的词汇选择题。

2. 文献综述

2.1. 英语词汇测试

词汇测试是语言测试的重要组成部分。Read 和 Chapelle (2001)提出,词汇能力具有多维特征,包括词汇知识广度和深度,因此词汇测试应关注学习者词汇知识的不同维度,包括词汇形式、意义以及使用

能力[6]。换言之,有效的词汇测试应避免单纯考查词汇记忆,而应关注词汇在具体语境中的使用。

目前,高中英语词汇测试主要采用选择题、填空题、匹配题等形式。其中,多项选择题因具有评分客观、操作便利、适合大规模测试等特点,被广泛应用于学校考试。然而,多项选择题的质量高度依赖命题设计,尤其是干扰项设计。Heaton(1975)指出,高质量选择题不仅需要保证正确答案唯一,还需要确保干扰项具有合理性和迷惑性,否则容易降低测试效度[7]。

在词汇选择题设计方面,陈晓扣和郑庆珠(2000)提出,词汇测试需要关注被测试词与辅助词之间的难度关系,避免由于选项词汇难度差异导致测试结果受到非目标因素影响[8]。换言之,在设计高中英语词汇测试时,需要控制选项难度、词性以及语境一致性。

此外,词汇测试还应体现情境化特点。近年来英语考试评价逐渐从知识检测转向能力评价,强调学生在真实或模拟情境中理解和使用语言。乔辉(2024)指出,情境化试题能够避免脱离实际语境的机械考查,更有助于评价学生综合语言运用能力[9]。因此,词汇选择题的设计应优先选用能提供充分语境的题目形式。

2.2. AI 辅助语言测试与人机协同命题

AI 的快速发展为教育领域的测试开发带来了新的技术路径。在语言测试方面,已有研究者开始探索利用大语言模型自动生成各类测试项目,以期缓解传统人工命题在时间和人力资源上的压力。Khademi(2023)的研究发现,ChatGPT 能够生成具有一定测试一致性的语言测试项目,但生成结果仍需要人工检查[4]。Lin 和 Chen(2024)基于大学英语四级考试阅读材料,考察了 ChatGPT-3.5 生成多项选择题的能力,结果表明 AI 生成的题目在心理测量特征上与人工命题具有一定可比性,具备教学和测评的应用潜力,但生成题目在测试目标的精准对应方面仍有不足[5]。上述研究从生成可行性和试题质量两个层面初步验证了 AI 辅助命题的潜力,但其结论也共同指向一个核心问题:AI 生成的测试项目尚不能独立达到高质量命题标准,亟需人类专家的介入与校准。

在此背景下,人机协同命题逐渐成为学界的关注焦点。周艳琼(2026)基于项目反应理论(IRT),比较了专家主导、AI 生成和人机协同三种模式下的英语阅读测试效度,结果显示专家主导模式在内容效度、结构效度和效标关联效度上均显著优于 AI 生成模式和人机协同模式,表明当前阶段 AI 仍难以独立完成高质量的命题任务,专家参与仍是保证测试效度的关键因素[10]。该研究建议采用 AI 初筛、专家优化等新型协同机制,以兼顾命题效率与质量。

综上,单纯依赖 AI 命题存在质量隐患,同时也印证了人机协同的必要性。已有研究为 AI 辅助语言测试提供了理论基础和实证参考,但针对高中英语词汇选择题开发的具体实践案例仍较少。因此,本研究以 DeepSeek 为辅助工具,聚焦高中英语词汇选择题的编写实践,探讨人机协作在该领域的具体应用路径与效果。

2.3. 语言测试质量验证

语言测试的质量验证是确保测试能够合理反映被测试者语言能力的关键环节。在测试效度理论的发展历程中,多位学者从不同视角对测试质量的评估标准进行了系统阐述,为本研究的试题质量分析提供了理论基础。

Messick(1989)在其经典的效度理论[11]中提出,效度并非测试本身的属性,而是用以判断现有的数据和理论能在多大程度上证明测试分数所作出的解释和决定是合理且恰当的。他将效度框架划分为两个维度:证据基础与后果基础,以及测试解释与测试使用,并由此形成四个相互关联的方面,即建构效度(证据基础/解释)、相关性与效用(证据基础/使用)、价值含义(后果基础/解释)以及社会后果(后果基础/使用)。

其中,建构效度被视为统摄其他效度证据的核心概念,内容包括效度证据、内容适切性、实质性、结构性和外部性等多个层面。这一效度观为本研究从多维度审视 AI 生成试题的质量提供了理论依据,即不仅关注试题是否测量了预设的构念,还需考察其内容适切性、任务过程的实质性以及与测试使用相关的潜在后果。

Bachman 和 Palmer (1996)提出了更为可操作化的测试有用性框架[12],将测试质量分解为六个核心品质:信度、建构效度、真实性、互动性、影响性和实用性。他们认为,测试的有用性不是各品质的简单叠加,而是需要在具体测试情境中寻求六者之间的适当平衡。该框架的一个突出贡献在于将测试任务特征和考生特征纳入测试设计的有意识考量之中,强调测试开发应从目标语言使用域出发,通过任务特征描述来确保测试任务与真实语言使用情境之间的可论证对应关系。这一思路为语言测试的开发提供了从构念定义到任务设计的系统化路径,也为评估 AI 生成试题是否贴近真实语言使用情境提供了具体的分析维度。

基于上述理论梳理,本研究在构建试题质量分析框架时,整合了效度理论中建构效度的多维度内容与测试有用性框架的系统化思路,使框架兼具理论上的全面性与实践上的可操作性,能够从多个层面逐题审视 AI 生成试题的质量。

3. 研究方法

3.1. 研究材料

本研究采用 DeepSeek 作为辅助命题工具,所有交互均通过其网页端对话界面完成。词汇测试目的为诊断性评估,测试对象为完成目标单元学习的高一学生。试题开发的输入材料包括教材文本,测试构念与命题细目表,以及参考样题与选择题编写原则。其中,教材文本采用人教版必修一 Unit 3 Sports and Fitness 作为词汇来源和话题依据。该单元主题涵盖体育赛事、运动员、健身活动、健康及相关个人经历,词汇表中目标词汇约 40 项,包括名词(如 athlete, strength, failure)、动词(如 compete, pretend, cheat)、形容词(positive, slim)、副词(rather)及短语动词(如 work out, make it, give up)等。测试构念涵盖不同测试目标层次,命题细目表规定了选择题的两种题型,相应的测试技能与题量。样题中提供两种题型的具体例子,编写原则涵盖题干与选项的约束要求。

3.2. 研究流程

本研究按照四个阶段的流程展开:

第一阶段为准备阶段。根据词汇教学目的,确定测试构念,并在此基础上制定命题细目表,明确题型比例和测试技能。

第二阶段为提示语设计阶段。研究者从不同维度设计初始提示语,包括角色设定、教材范围限定、测试规范输入、题型要求、输出格式要求等。

第三阶段为 AI 生成阶段。将指令语输入 DeepSeek,由 AI 生成初版试题。

第四阶段为人工审查与修改阶段。研究者对 AI 生成的试题进行审查,重点分析所编试题是否符合测试构念和教材主题,正确选项是否具有唯一性,干扰项是否有效等。针对试题问题,研究者进行修改,形成最终版本。

3.3. 研究分析框架

由于本研究旨在考察 DeepSeek 辅助高中英语词汇选择题编写过程中的应用价值与局限,不涉及大规模学生测试,因此本研究主要采用质性分析方法。

本研究综合参考 Read 和 Chapelle (2001)的词汇测试要求[6]，Heaton (1975)的选择题编写原则[7]，Messick (1989)的效度理论[11]，以及 Bachman 和 Palmer (1996)的语言测试有用性框架[12]，构建试题质量分析框架(详见表 1)。

具体而言，在测试内容适切性维度上，框架借鉴了 Messick(1989)效度理论对内容适切性、实质性的观点，以及 Read 和 Chapelle (2001)关于词汇能力应包括形式、意义和使用三个层面的要求，从教材词汇覆盖度、测试目标一致性和词汇知识层次三个指标评估试题是否真正触及目标构念的核心要素。在语言情境质量维度上，框架结合了 Bachman 和 Palmer(1996)对真实性的界定，即测试任务特征与目标语言使用域任务特征之间的对应程度，从语境真实性、语言自然度和语境支持程度三个方面判断题干是否为考生提供了充分的语义推断空间。在题目设计质量维度上，框架参照 Heaton (1975)关于高质量选择题必须保证正确答案唯一性及干扰项兼具合理性与迷惑性的论述，从题干设计、正确选项质量和干扰项有效性三个指标逐题审视 AI 生成试题的合理性。上述三个维度的整体构架既能够在宏观层面回应词汇试题效度论证的核心要求，又能够在微观层面为具体试题的分析提供可操作的评判依据。

Table 1. Item analysis framework
表 1. 试题分析框架

一级维度	二级指标	具体分析内容
测试内容适切性	教材词汇覆盖度	分析试题考查词汇是否来源于教材核心词汇范围，是否符合单元主题要求
	测试目标一致性	判断试题是否真正考查词汇知识，而非偏向语法、拼写或机械记忆
	词汇知识层次	分析试题是否涉及词汇形式、意义和使用三个层面
语言情境质量	语境真实性	判断题干是否提供符合真实语言使用特点的情境，是否避免孤立考查词义
	语言自然度	分析 AI 生成句子的表达是否符合英语习惯，是否存在中式表达或不自然搭配
	语境支持程度	判断上下文信息是否能够帮助学生进行合理词义推断
题目设计质量	题干设计	分析题干是否清晰、完整，是否能够有效引导学生完成任务
	正确选项质量	判断答案是否唯一、准确，是否符合目标词汇意义
	干扰项有效性	分析干扰项是否具有合理性、迷惑性和区分作用，而非简单错误选项

4. 试题编写过程与质量分析

4.1. 构念与细目表的编制

词汇测试的有效性首先取决于测试构念的明确。《课标》指出，高中英语课程应围绕核心素养发展，促进学生在真实情境中理解和表达意义，其中语言能力的发展不仅包括语言知识的积累，更强调学生运用语言完成交际任务的能力。因此，高中英语词汇测试应突破传统检测模式，通过情境化任务考查学生理解、辨析和应用词汇的能力。此外，《中国英语能力等级量表》(以下简称《量表》)进一步从语言理解和语言表达两个维度描述英语能力发展水平，强调学习者应能够根据语境理解词汇意义，并选择適切语言形式表达思想[13]。因此，本研究构建了词汇测试构念(详见表 2)。

基于上述测试构念，本研究编制如下命题细目表(详见表 3)，设计了替换题(Substitution)和完形填空题(Gap-filling)两种选择题类型[7]，共 20 题，每题 5 分，总分 100 分。

Table 2. Vocabulary test constructs
表 2. 词汇测试构念

构念维度	构念界定	理论依据
词汇识别能力	识别词汇形式、词性及基本含义	《量表》词汇知识运用能力：能掌握英语核心词语的含义、常见用法以及在特定语境中的意义。
词汇理解能力	根据语境理解词义，把握所传达的态度与意图	《课标》词汇知识内容要求：理解语篇中关键词的词义和功能以及所传递的意图和态度等。
词汇应用能力	在真实情境中选择恰当词汇完成表达	《课标》词汇知识内容要求：在语境中，根据不同主题，运用词汇表述相关事物； 《量表》词汇知识运用能力：能区分常见近义词在词义上的细微差别。

Table 3. Item specification table
表 3. 命题细目表

部分	对应技能	测试重点	题量
第一部分：替换题	a	识别与记忆：识别目标词汇的中英文释义及词性。	10 题
	b & c	理解与匹配：在句子语境中理解词义，并匹配同义词或最接近的含义。	
第二部分：完形填空题	a	识别与记忆：识别目标词汇的中英文释义及词性。	10 题
	b & d	理解与应用：在句子语境中理解词义，并选择恰当的词汇完成表达。	

测试技能：
a) 能识别核心词汇的中英文释义及词性；
b) 能在句子语境中理解核心词汇的意义和功能；
c) 能识别核心词汇的同义词或最接近的含义；
d) 能在句子语境中选择恰当的词汇完成表达。

替换题要求学生选出替换句子划线部分的最佳选项，主要考查学生词义辨析能力；完形填空题要求学生选出填入句子空白处的最佳选项，主要考查学生根据语境选择适当词汇的能力。这两种类型的选择题样题与编写原则也需明晰(详见表 4)。

4.2. 指令设计

有效指令是引导 AI 产出高质量试题的关键。实验中的指令设计遵循以下原则：(1) 明确角色定位，赋予 AI 特定的身份；(2) 设定具体的命题要求，涵盖题型、数量、格式和约束条件等；(3) 提供充分的输入材料，包括教材、测试规范和样题等。根据原则，初版指令设计如下(详见表 5)。

4.3. 试题分析与修改

将初版指令语和相关材料输入后，DeepSeek 能在短时间内生成初版词汇试题。首先，笔者对该试题进行初步审查，关注该试题是否符合替换题和完形填空题的样式，是否考查目标单元词汇，是否与目标单元主题相关等基本编写要求。审查发现，DeepSeek 能够较好地理解提示语中的显性要求，生成的试题在题型样式、数量等方面总体符合预设要求，且试题内容均围绕目标单元展开，说明初版指令说明能够

有效约束 AI 的基本生成方向。但该试题出现了几点其他问题，包括部分试题考查相同词汇、试题题干过于简单，以及正确答案分布不均衡，这说明初版指令存在不完整的问题。

Table 4. Sample items and item generation principles
表 4. 样题与编写原则

类型	样题	编写原则
替换题	It rained <i>continuously</i> for two whole days. (A) A) Without stopping B) Heavily C) Regularly D) At odd moments	1) 题干中斜体部分应为本单元的考查词汇，除此之外不得出现任何考查词汇； 2) 选项应使用学生已掌握的简单的词汇，不得出现任何本单元的考查词汇。
完形填空题	I saw a nasty...between two cars this morning. (D) A) Happening B) Danger C) Damage D) Accident Ann...me of a girl I used to know. (B) A) Recalls B) Reminds C) Remembers D) Recollects	1) 题干中不得出现任何本单元的考查词汇； 2) 每题四个选项中仅设置其中一个为本单元的考查词汇(可作为正确选项或干扰项)，其他选项不得使用考查词汇。
		1) 选项在语法上是正确的； 2) 选项处于同一难度； 3) 选项的长度相同； 4) 选项的词性相同； 5) 题干与选项都处于同一话题； 6) 避免使用同义词作为干扰项； 7) 避免使用反义词作为干扰项； 8) 避免在选项中重复使用相同的词语； 9) 采用用法不同的同义词作为有效的干扰项； 10) 采用与题目语境相关的词汇作为有效的干扰项。

Table 5. Composition of the initial prompt
表 5. 初版指令构成

功能	描述	补充
第一段	角色定位	你是一名经验丰富的英语词汇测试题编写专家。
	教材内容和考查词汇要求	现在你需要仔细阅读高中必修一教材第 35~46 面 Unit 3 Sports and Fitness 的教学内容和词汇表中该单元的词汇。
	测试规范要求	仔细阅读词汇题测试的构念和命题细目表。
	题型与分值要求	根据构念和细目表，编写 20 道选择题，第 1~10 题为替换题，第 11~20 题为完形填空题，每题 5 分，总共 100 分。
第二段	题目编写原则	注意参考所给样题编写，并遵循题目的编写原则。
	答案与解析要求	注意每道题有且仅有一个正确选项，并在完整试题后，给出所有参考答案和解析。
第三段	主题要求	1) 词汇试题的内容要与 Sports and Fitness 主题有关。
	原创性要求	2) 如果编写的题目有出处的话需要在最后标明(如参考了教材、高考试卷、报刊等)。

基于 DeepSeek 的互动性特点，笔者针对上述问题输入相应修改指令，形成第二版试题，并通过再次审查发现未出现其他问题，遂进入该试题的质量评估环节。本研究依据前文建立的词汇测试分析框架，从测试内容适切性、语言情境质量和题目设计质量三个维度对该试题进行总体分析(详见表 6)。

综上，AI 生成试题在教材内容覆盖和基本测试目标一致性方面表现较好，语境化程度也较为明显，但在教材词汇覆盖度、情境的真实性与支持程度，以及正确选项和干扰项的区分度方面仍存在进一步优化空间。

Table 6. Overall test item analysis
表 6. 试题总体分析

一级维度	二级指标	试题表现	存在问题	修改方向
测试内容 适切性	教材词汇 覆盖度	试题围绕目标单元中的词汇展开，且与单元主题高度契合。	词汇覆盖主要集中于名词和动词；对教材词汇整体覆盖面的广度缺少明确比例控制。	后续编题可依据教材词汇表明确核心词汇覆盖率，并适当增加不同词性的代表性词汇，避免部分词汇类型集中。
	测试目标 一致性	第 1~10 题要求考生选择与目标词意义最接近的选项，主要调查词义识别与理解；第 11~20 题将词汇置于具体句子中，考查词汇在语境中的理解和使用，整体与词汇测试目标一致。	/	/
	词汇知识 层次	试卷体现出词汇形式，意义和使用三个层次：第 1~10 题主要考查词汇形式识别与意义理解，第 11~20 题进一步考查词汇在具体语境中的使用及搭配。	/	/
语言情境 质量	语境真实 性	大部分题目设置了较为明确的体育运动情境，如跑步、足球比赛、体育训练等，能够将目标词汇置于较真实的体育语境中。	个别语境存在真实性不足的问题。	应进一步检查题干中的现实逻辑关系，提高语境真实性。
	语言自然 度	题干表达完整、清晰，符合基本英语表达习惯；第 11~20 题尤其注重目标词与具体语言环境的结合。	/	/
	语境支持 程度	大多数题目能够通过句子信息推断出所需的目标词。	个别题对目标词的语境支持较弱。	适当增强句子信息线索，使学生能够根据语境进行合理推断，同时避免线索过强而直接暴露答案。
题目设计 质量	题干设计	题目整体结构统一，指令明确。第 1~10 题统一要求选择意义最接近的选项，第 11~20 题统一要求选择最佳选项完成句子，题干完整，任务要求清晰。	/	/
	正确选项 质量	正确答案总体能够与题干中的目标词形成较明确的语义或搭配关系。	个别正确选项并非最佳的语义对应。	对正确答案进行唯一性检验，重点检查近义词、上下义词和语境相关词之间的竞争；必要时重新设计题干或替换干扰项。
	干扰项有 效性	部分题目的干扰项具有一定迷惑性，能够考查学生对近义词和相关词的辨析。	部分干扰项与正确答案的语义距离过近，容易造成多答案倾向；另一些干扰项距离过远，基本没有实际迷惑作用。因此干扰项质量不够均衡。	根据目标词的语义特征设计干扰项，优先选择同词性、同语义场且具有一定合理性的选项，同时保证只有一个选项能够在具体语境中成立。

具体来看，本文选取替换题与完形填空题中的典型问题进行分析与修改(详见表 7)。这些题能够较好地体现正确答案质量不足、干扰项有效性不足，以及语境真实性不足的问题。

Table 7. Item analysis and revision
表 7. 试题分析与修改

试题内容	问题分析	修改方式	修改后试题
The young swimmer was extremely proud when she received a gold medal for her excellent performance in the final. A) Trophy B) Prize C) Ticket D) Gift	分析依据：正确选项质量 原题将 medal 与 prize 对应。prize 是奖品、奖赏的上位概念，而 medal 是一种具体的荣誉奖牌，二者并非完全同义。虽然在语境中 prize 可以解释为一种奖励，但作为意义最接近的选项仍不够精确。	将正确选项调整为能够更直接体现奖牌/荣誉奖章含义的表达，同时重新设置干扰项。	The young swimmer was extremely proud when she received a gold medal for her excellent performance in the final. A) Great comment B) Expensive present C) Award for success D) Event invitation
Many great players believe that learning from failure is the first step towards future success in sports. A) Loss B) Mistake C) Defeat D) Effort	分析依据：干扰项有效性 原题答案为 defeat，但 failure 与 defeat 虽然语义接近，却并非在所有语境下可以互换；同时 loss 也可能与体育比赛中的失败形成较强竞争关系。因此，题目对近义词细微差异的要求较高，而题干提供的语境不足以完全排除竞争选项。	强化语境，使目标词的具体意义更加明确，并将干扰项调整为语义场相关但能够被语境排除的词。	Many great players believe that learning from failure is the first step towards future success in sports. A) Effort B) Mistake C) Defeat D) Change
Thousands of excited fans gathered in the huge _____ to watch the live broadcast of the football final. A) Stadium B) Field C) Court D) Park	分析依据：语境真实性 原题希望通过 “thousands of excited fans gathered in the huge...” 考查 stadium，但聚集在体育场观看比赛直播存在一定情境逻辑问题，通常是在体育场观看现场比赛，而不是直播。该语境可能使学生关注情境逻辑，而不是目标词本身。	将 live broadcast 删去，使场所与活动之间形成自然对应。	Thousands of excited fans gathered in the huge _____ to watch the football final. A) Stadium B) Field C) Court D) Park
After years of practice, she finally became the _____ of the national tennis tournament. A) Champion B) Winner C) Leader D) Master	分析依据：正确选项质量、干扰项有效性 原题考查 champion，但 winner 同样可以自然表示 “比赛的获胜者”，因此 A、B 之间存在较明显的竞争。此外，选项 D 也是本单元考查词汇，不符合题目选项的编写原则。	增加能突出 champion 特征的语境，如强调获得金奖，从而降低 winner 的竞争性；同时优化干扰项。	After years of practice, she finally became the _____ of the national tennis tournament and received the gold prize. A) Champion B) Winner C) Leader D) Manager

4.4. 指令语优化

通过对试题的审查与分析可以发现，AI 生成试题质量不仅受到模型语言生成能力影响，也与输入指令的具体程度密切相关。因此，本研究根据上述试题的问题进一步对指令进行优化。优化后的终版指令如下(详见表 8)。

Table 8. Composition of the final prompt wording
表 8. 终版指令语构成

阶段	功能	描述	补充
第一阶段： 基本任务设定	角色定位	你是具有英语词汇测试编写经验的高中英语词汇测试题编写专家。	/
	教材内容要求	现在你需要仔细阅读高中必修一教材第 35~46 面 Unit3 Sports and Fitness 的教学内容和词汇表中该单元的词汇。	Unit 3 考查词汇
	测试规范要求	仔细阅读词汇题测试的构念和命题细目表。	构念和命题细目表
	题型与分值要求	根据构念和细目表，编写 20 道选择题，第 1~10 题为替换题，第 11~20 题为完形填空题，每题 5 分，总共 100 分。	/
第二阶段： 题目设计要求	题目编写原则	试题应参照样题编写，并遵循题目的编写原则。	样题和编写原则
	答案与解析要求	注意每道题有且仅有一个正确选项，正确选项应在 A、B、C、D 之间均衡分布，避免出现明显的答案规律。在完整试题后，给出所有参考答案和解析。	新增约束
第三阶段： 质量控制要求	主题要求	试题内容应与单元主题相关。	/
	原创性要求	题目应基于教材内容进行原创设计，如参考教材、高考试卷或其他资料，应在最后注明来源。	/
	词汇覆盖要求	尽可能覆盖目标词汇，同一目标词不得重复作为不同题目的核心考查词，以提高有限题量下的词汇覆盖率。	新增约束
	语境真实性要求	题干应提供完整、自然且符合真实语言使用特点的语境，避免孤立考查词义；上下文应能够为正确答案提供必要但不过度的语义线索。	新增约束
	语境支持程度要求	题干应提供一定的信息线索，同时避免线索过强而直接暴露答案。	新增约束
	题干难度要求	题干内容不能过于简短，要符合高一学生的词汇水平。	新增约束
	正确答案唯一性要求	每道题有且仅有一个最佳答案。尤其要检查近义词、上下义词和语境相关词之间是否存在答案竞争，确保正确答案能够依据题干和测试目标得到明确判断。	新增约束
	干扰项有效性要求	干扰项应与正确答案具有一定的语义关联、词性一致性和合理迷惑性，避免明显错误或与题干完全无关的选项，同时避免干扰项在特定语境下也能够成立。	新增约束

5. 讨论与反思

5.1. AI 辅助命题的优势与局限

本研究以 DeepSeek 辅助高中英语词汇选择题编写为实践案例。通过试题生成过程和审查分析可以发

现, DeepSeek 在命题中具有较为明显的辅助价值, 其优势主要体现在生成效率、语言加工和持续修订层面, 而在试题质量控制方面仍具有明显局限。

首先, AI 具有较高的试题生成效率。在输入材料后, DeepSeek 能够在较短时间内完成 20 道选择题的初步生成, 较大程度上降低教师从零开始设计题目的时间成本。对于需要进行课后练习、单元检测或形成性评价的一线英语教师而言, 这种效率优势具有较强的实践意义。

其次, AI 在语境创设和语言表达方面具有较强的辅助能力。本研究中的多数题目能够围绕单元主题构建相对完整的语言情境, 将目标词放入体育比赛、训练等场景中。相比单纯罗列词义或直接进行英汉匹配, AI 生成的语境能够为教师提供更多情境化命题思路, 也与《课标》强调在语境中理解和运用语言知识的要求一致。

再次, AI 具有较强的互动修订能力。当研究者发现初版试题存在词汇重复、题干简单和答案分布不均衡等问题后, 可以通过追加指令要求进行修改, 并较快获得新的试题版本。这种交互式特点使命题过程由一次性生成转变为多轮调整, 为教师不断完善题目提供了便利。

然而, AI 的诸多优势并不意味着其能够独立完成高质量试题开发, 本研究发现 AI 的局限性主要体现在以下三方面。

第一, AI 对试题中考查内容的覆盖仍然有限。DeepSeek 虽能依据指令从单元词汇表中选取目标词汇, 但其对考查内容的覆盖呈现出明显的随机性和不均衡性。这种不均衡反映了 AI 在内容采样机制上的局限, 它缺乏对哪些词汇更重要或更需要诊断的价值判断能力。

第二, AI 生成语境的真实性与支持程度仍需要人工判断。尽管 DeepSeek 能够生成语法正确、形式完整的英语句子, 但这并不等同于语境真实。这说明 AI 的语言生成能力主要保证了句子的表层可接受性, 而真实语言使用所涉及的情境逻辑、交际合理性和语用适切性仍需要教师结合教学经验进行判断。此外, AI 生成的语境在提供充分且唯一的答题线索时表现不稳定, 部分题干的语境线索不足, 导致多个选项均可成为正确答案。

第三, AI 对正确答案唯一性和干扰项有效性的判断不够稳定。本研究发现, 部分生成题目存在两个选项均具有一定合理性的情况, 而部分干扰项与正确答案的语义距离过远, 几乎不能发挥区分作用。这与选择题应保证正确答案唯一并设置具有合理迷惑性的干扰项的要求存在一定差距。

从 Messick 的效度理论的视角审视上述局限, 可以发现这些问题会影响测试的内容适切性、实质性等方面[11]。例如, 词汇覆盖的随机性可能使测试内容偏离预设的构念范围, 削弱内容效度; 语境支持不足或干扰项失效则可能导致考生通过非目标策略(如排除明显错误或无关选项、依赖语法线索等)作答, 使得测试任务未能真正触及词汇理解与应用这一实质性构念, 从而降低测试的构念效度。由此可见, AI 生成试题的上述局限并非彼此孤立的技术缺陷, 而是可能从多个层面共同影响测试整体效度的关键因素。

因此, AI 在命题过程中更适合被定位为高效的命题助手和语言资源生成工具, 而教师的主要价值不再只是负责逐字编写每一道题, 而更多体现在构念确定、命题规范设定、AI 输出评价以及最终质量控制等环节。

5.2. 人机协作模式的反思

本研究表明, 人机协同命题的核心是重新划分人与 AI 在命题过程中的功能。AI 更擅长快速生成语言材料、提供多样化语境以及根据明确指令进行修改; 教师则更擅长理解课程标准、判断测试目标、以及评价题目的教学适切性。因此, 两者具有一定的互补关系。

从本研究的实际过程来看, 人机协作可以形成一个较为清晰的模式: 教师确定测试构念与命题标准、教师制定命题细目表、教师设计初始指令、AI 生成试题、教师依据分析框架审查、教师提出针对性修改

要求、AI 再次生成、教师最终审核与人工修订。这一协作模式首先要求教师具备较强的测试素养。如果命题者自身缺乏对词汇测试理论和选择题编写原则的理解,即使能够向 AI 提出详细指令,也很难判断 AI 生成结果是否真正有效。因此, AI 的使用并没有降低教师专业能力的重要性,反而对教师提出了新的要求。教师不仅需要会使用 AI, 还需要会评价 AI, 能够从教材内容、测试构念、语言质量和题目设计等多个层面识别 AI 输出中的潜在问题。

其次, 人机协作要求教师从传统的题目生产者逐渐转向命题设计者和质量审核者。从认知负荷理论 (Sweller, 1988) 的视角分析, AI 的有效介入可能有助于优化教师的认知资源分配[14]。在传统命题中, 教师需要同时处理测试构念设计、语境创设、选项编写、语言打磨等多项认知任务, 这些任务共同构成了较高的内在认知负荷。而 AI 能够承担语言素材生成、语境初步创设等相对程序化的工作, 将教师的认知资源从这些较低的生成性任务中解放出来, 使其能够更多地投入到评价性和决策性等高阶认知任务中, 从而降低工作记忆负担, 提升整体命题质量。换言之, AI 的引入并不是减少教师专业工作的价值, 而是促使教师将专业能力更多用于高层次的设计和判断。

因此, 未来的高中英语测试开发应坚持教师主导、人机协同的基本原则, 在充分发挥 AI 生成优势的同时, 建立稳定的人工审查机制。只有将 AI 的生成能力与教师的测试专业判断结合起来, 才能实现命题效率和试题质量之间的平衡。

5.3. 研究局限与展望

本研究作为一项探索性案例研究, 虽揭示了 DeepSeek 在高中英语词汇选择题编写中的基本应用路径与问题, 但仍存在一定局限, 这也为后续研究指明了方向。

首先, 在研究的广度上, 本研究仅聚焦于 DeepSeek 这一特定模型、人教版必修一 Unit 3 一个单元主题, 以及替换题和完形填空两种选择题题型。因此, 未来研究可扩大范围, 将实验拓展至不同的 AI 模型 (如豆包、文心一言等)、不同的学科主题单元, 以及不同的题型 (如阅读理解题、简答题或写作任务), 以系统考察本研究发现的共性与差异。

其次, 在研究设计的深度上, 本研究侧重于质性过程描述与分析, 未设置系统的对比实验。后续研究可以设计严格的对比实验, 从命题效率、试题质量和效度证据等多个维度, 定量比较纯人工命题、纯 AI 命题与人机协同命题三种模式的实际效果, 从而为“教师主导、人机协同”模式提供更具说服力的实证数据。

此外, 本研究主要从命题者视角分析试题质量, 未纳入学生实际测试表现数据。未来研究可在人机协同命题的基础上开展小规模试测, 收集学生的作答数据和反馈意见, 从受试者角度进一步验证 AI 生成试题的有效性与可用性。

6. 结论

本研究以 DeepSeek 辅助高中英语词汇选择题编写为案例, 探索了生成式人工智能在人机协同命题过程中的应用价值。

研究发现, DeepSeek 能够较高效地完成高中英语词汇选择题的初步生成, 在基本格式、教材主题契合、语境创设和语言表达等方面具有较好的辅助价值; 但 AI 生成试题仍存在考查内容覆盖不均衡、语境真实性和支持程度不足、正确答案唯一性和干扰项有效性不稳定等问题。这些问题可能对测试的内容适切性和实质性产生不良影响, 凸显了人工审查的必要性。认知负荷理论的视角进一步揭示, AI 的介入有助于将教师的认知资源从程序性任务中解放, 聚焦于更高层次的评价与决策。研究表明, 人工智能辅助命题的核心在于促进教师专业能力与人工智能生成能力的结合。

未来高中英语测试开发应坚持教师主导、人机协同原则，在保证测试效度的基础上探索人工智能赋能教育评价的新路径。后续研究需进一步扩大模型、主题和题型范围，并通过对比实验与学生实测数据，为这一模式的推广与优化提供更全面的证据。

参考文献

- [1] 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) [S]. 北京: 人民教育出版社, 2025.
- [2] 宋德龙. 高中英语词汇教学与测试策略的研究[J]. 中小学英语教学与研究, 2008, 31(8): 29-32.
- [3] Shin, D. and Lee, J.H. (2023) Can ChatGPT Make Reading Comprehension Testing Items on Par with Human Experts? *Language Learning & Technology*, 27, 27-40. <https://doi.org/10.64152/10125/73530>
- [4] Khademi, M. (2023) Can ChatGPT and Bard Generate Aligned Assessment Items? A Reliability Analysis against Human Performance. *Journal of Applied Learning & Teaching*, 6, 75-80.
- [5] Lin, Z. and Chen, H. (2024) Investigating the Capability of ChatGPT for Generating Multiple-Choice Reading Comprehension Items. *System*, 123, Article ID: 103344. <https://doi.org/10.1016/j.system.2024.103344>
- [6] Read, J. and Chapelle, C.A. (2001) A Framework for Second Language Vocabulary Assessment. *Language Testing*, 18, 1-32. <https://doi.org/10.1177/026553220101800101>
- [7] Heaton, J.B. (1975) *Writing English Language Tests*. Longman.
- [8] 陈晓扣, 郑庆珠. 论英语词汇测试中的高低、一致原则[J]. 解放军外国语学院学报, 2000(6): 64-66.
- [9] 乔辉. 情境化试题设计在高考英语中的应用探析[J]. 中小学外语教学(中学篇), 2024, 47(12): 55-61.
- [10] 周艳琼. 人机协同命题与英语选择题阅读测试效度——基于IRT的多维度比较[J]. 外语教学与研究, 2026, 58(3): 430-443.
- [11] Messick, S. (1989) Validity. In: Linn, R.L., Ed., *Educational Measurement (3rd Edition)*, Macmillan, 13-103.
- [12] Bachman, L.F. and Palmer, A.S. (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.
- [13] 中华人民共和国教育部, 国家语言文字工作委员会. GF 0018-2018 中国英语能力等级量表[S]. 北京: 高等教育出版社, 2018.
- [14] Sweller, J. (1988) Cognitive Load during Problem Solving: Effects on Learning. *Cognitive Science*, 12, 257-285. https://doi.org/10.1207/s15516709cog1202_4