Advances in Education 教育进展, 2026, 16(3), 189-199
Published Online March 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.163470

Hans汉斯

人工智能在高中英语听力命题中的应用研究

邓文珺，段少敏

黄冈师范学院外国语学院，湖北 黄冈

收稿日期：2026年1月25日；录用日期：2026年2月24日；发布日期：2026年3月3日

摘　要

随着人工智能技术的飞速发展，其在教育领域的应用日益广泛和深入。人工智能技术在英语听力方面的研究多集中在口语测试方面，因此开展人工智能在高中英语听力命题的应用研究是有必要的。本实验分别采用传统人工和人工智能设计的听力短对话试题进行测试，再对比分析这两者的难度和区分度指标。结果表明，人工智能生成的题目在区分度上略低于命题组设计的试题，整体区分度尚可，但人工智能在题目难度的设计上偏难，难度也存在两极分化的情况。本研究为人工智能技术赋能英语听力试题的设计提供了实践依据和优化方向。

关键词

人工智能，高中英语，听力命题

Research on the Application of Artificial Intelligence in High School English Listening Test Design

Wenjun Deng, Shaomin Duan

School of Foreign Languages, Huanggang Normal University, Huanggang Hubei

Received: January 25, 2026; accepted: February 24, 2026; published: March 3, 2026

Abstract

With the rapid development of artificial intelligence (AI) technology, its application in the field of education has become increasingly widespread and in-depth. Research on AI technology in English listening comprehension has predominantly focused on oral testing. Therefore, it is necessary to conduct research on the application of AI in generating English listening test items for high schools. In this experiment, tests were conducted using short-dialogue listening items designed by traditional

文章引用: 邓文珺, 段少敏. 人工智能在高中英语听力命题中的应用研究[J]. 教育进展, 2026, 16(3): 189-199.
DOI: 10.12677/ae.2026.163470

methods and by AI, respectively. A comparative analysis of the difficulty and discrimination indices of these two sets of items was then performed. The results show that the AI-generated items had a slightly lower discrimination level than those designed by the human test-setting group, although the overall discrimination was acceptable. However, the AI tended to design items that were overly difficult, and their difficulty levels also showed a pattern of polarization. This study provides empirical evidence and directions for optimization for the AI-empowered design of English listening test items.

Keywords

Artificial Intelligence, High School English, Listening Test Design

1. 引言

近年来，人工智能技术的快速发展为教育领域带来了深刻变革。移动互联网、大数据、云计算等新兴信息技术的广泛应用正在重塑各行业的生产模式，同时也对我国外语专业人才培养提出了新的要求(郑燕虹等，2025) [1]。

在英语测评领域，试题自动生成(Automatic Item Generation, AIG)技术正逐渐成为研究热点。基于语音技术的听力、口语上机考试已得到普遍应用，手写识别、自然语言理解等与人工智能有关的技术也逐步融入教育考试评卷过程(何屹松等，2018) [2]。然而，现有应用多集中于文本类题型，如阅读理解和完形填空。在高中英语听力这种同时涉及音频与文本双模态的任务中，关于人工智能生成题目质量的教育测量学验证，特别是经典测试理论(Classical Test Theory, CTT)中的难度与区分度等关键指标，仍缺乏充分的实证研究。作为英语语言能力的核心维度，听力测试的命题质量直接影响教学诊断的准确性和学习评价的公平性。

本文聚焦人工智能在高中英语听力短对话题型中的应用效果，在经典测试理论(CTT)的框架下，通过系统比较其与传统人工命题在难度和区分度上的差异，旨在明确人工智能命题的优势与不足，为相关实践提供理论参考和方法借鉴。

2. 文献综述

试题自动生成(Automatic Item Generation, AIG)的概念于 20 世纪 60 年代出现于西方学术界与考试机构，其核心是综合运用认知心理学、心理计量学模型与计算机编程技术，以实现特定测量目标的试题自动化构建(Kurdi *et al.*, 2020) [3]。

AIG 的早期实践主要依赖模板法(template-based)与规则法(rule-based)。模板法通过在预设试题框架中变更参数来规模化生产试题，尤其适用于数学这类构念明确的领域(Bejar *et al.*, 2002) [4]。规则法则是依据文本的句法或语义特征，建立规则集来执行题眼定位、题型匹配和问题生成等关键命题步骤(Kurdi *et al.*, 2020) [3]。

随着深度学习技术在自然语言处理领域取得巨大成功(LeCun *et al.*, 2015) [5]，AIG 的发展进入了新阶段。生成式人工智能作为一种基于深度学习的技术，能够根据自然语言提示自动生成响应内容，在数据

分析方面展现出强大能力(刘邦奇等，2024) [6]。该技术采用语料库语言学研究方法，通过对海量语言数据的分析，深入把握语言的使用特征和结构规律，从而实现高效准确的语言生成(张智义，2024) [7]，并能够根据教材编写者提供的主题，生成符合教学目标的语篇内容(贾蕾，马颖，2025) [8]。研究者已将人工智能技术有效应用于语言测试中复杂题型(如选择题)的自动化命题中，极大地提升了命题效率(陈大建，胡杰辉，2025) [9]。

人工智能在测评应用方面，研究表明人工智能能够对学科题库进行智能分析，为教师命题提供优化建议(刘一凡，2018) [10]。目前，“机器批改 + 人工批改”模式已成为外语作文评阅的主流方式，广泛应用于写作教学和考试评价(甘容辉，何高大，2017) [11]；高校在英语作文阅卷中运用人工智能评测技术，能够有效优化传统阅卷的不足，节省时间和人力，实现大规模阅卷(黄岚，2024) [12]。在口语测评领域，机器评卷可以大幅减轻人工工作量，其稳定性和客观性优势明显，但仍存在一定局限性(吕鸣，2015) [13]；人工智能技术能够提供精准高效的测评结果，实现个性化学习指导，丰富测评形式和学习资源(朱柯睿，2025) [14]。教育实践表明，基于大数据的智能分析系统可以精准诊断学生学习问题，提供针对性练习建议(栾爱春，2019) [15]。

对人工智能自动生成试题的评价可从语言质量、内容效度及测量学表现三个维度展开。首先，在语言学与内容质量层面，评价内容包括语言流畅度与准确性、答案在给定材料中的存在性与唯一性。此外，题目的难度需与目标考生的认知水平相匹配，且干扰项应具备足够的迷惑性，同时在结构或语义上与正确项保持相似，以有效甄别考生掌握程度(陈大建，胡杰辉，2025；王鸿滨，吕海辉，2025；段晨曦，2024) [9] [16] [17]。其次，教育测量学评价从专业角度评估试题是否达成测量目标，核心在于一组试题能否均衡覆盖预设的微技能考查点(如“理解主旨”“推断意义”)以保证构念效度，并通过专家可用性评分来判断其直接应用的潜力。最后，通过实证测试与数据分析进行最客观的评价，其核心指标包括评估测量结果稳定性的信度(Reliability)，以及检验题目实际表现是否符合模型预期的题目拟合度(Item Fit) (陈大建，胡杰辉，2025) [9]。

尽管人工智能自动生成试题技术前景广阔，但在当前应用中仍面临技术、实践和伦理层面的严峻挑战。技术层面的局限性体现在对复杂语言环境的适应性不足，以及难以灵活匹配不同学生的语言水平(刘一凡，2018；朱柯睿，2025) [10] [14]。更致命的缺陷是内容的准确性与“幻觉”问题，即模型可能捏造事实，生成看似合理但错误的信息(王鸿滨，吕海辉，2025；张晖，郭宇航，2025) [16] [18]。即便是经过专用数据集优化的模型，也无法完全避免在微技能分布、难度控制等方面的质量问题(陈大建，胡杰辉，2025) [9]。实践应用层面，师生的数字素养与接受度是一大障碍，部分师生因不了解、不信任或担心公平性而对人工智能技术产生抵触，，其有效应用也受限于使用者的数字素养(朱柯睿，2025；张晖，郭宇航，2025；李婷，2020) [14] [18] [19]。同时，过度依赖于人机协同的难题日益凸显，研究普遍认为，从素材筛选到最终审核，人类专家的介入是保证试题质量的必要环节(陈大建，胡杰辉，2025；王鸿滨，吕海辉，2025；段晨曦，2024) [9] [16] [17]。安全与伦理层面，人工智能测评系统收集的学生数据存在泄露和滥用的风险，可能引发隐私安全问题(朱柯睿，2025；王鸿滨，吕海辉，2025) [14] [16]。算法偏见与价值导向同样值得警惕，模型的训练数据和设计可能带有特定价值观或意识形态，使用者需保持高度辨别力，确保教学内容的正确导向(郑燕虹，罗常军，蒋洪新，2025) [1]。

综上所述，现有研究已清晰展示了试题自动生成(AIG)的技术发展与评价框架，并证实了其在教育测评中的应用潜力。然而，AIG 在处理高中英语听力等复杂特定场景时，仍面临技术适应性不足和实践应用困难等挑战，相关研究尚不充分。因此，本研究旨在探索将人工智能应用于高中英语听力命题的具体方法，为构建保障试题质量的人机协同模式提供实践参考。

3. 研究设计

3.1. 研究对象

本研究选取湖北省某县级市重点高中高二年级 2 个平行班共 103 名学生，开展听力测试。对照班(1班)和实验班(2 班)的男女生比例分别为 29:23，27:24。

为保障教学实验的效度，对两个班级在实验前连续三次英语测试的总分及听力成绩进行了正态性检验与组间差异显著性检验，分析结果如下(表 1~7)。

Table 1. Results of the normality test for test 1
表 1. 测试一正态性检验分析结果

	班级	柯尔莫戈洛夫－斯米诺夫(V)[a]			夏皮洛－威尔克		
		统计	自由度	显著性	统计	自由度	显著性
总分	1	0.062	52	0.200*	0.975	52	0.335
	2	0.076	50	0.200*	0.981	50	0.595
听力	1	0.144	52	0.009	0.920	52	0.002
	2	0.139	50	0.017	0.945	50	0.022

*这是真显著性的下限。[a] 里利氏显著性修正。

Table 2. Results of the independent samples t-test for the total scores of test 1
表 2. 测试一总分独立样本 t 检验分析结果

		莱文方差等同性检验		平均值等同性 t 检验						
		F	显著性	t	自由度	Sig. (双尾)	平均值差值	标准误差差值	差值 95%置信区间	
									下限	上限
总分	假定等方差	0.059	0.809	0.497	100	0.620	1.180	2.374	−3.529	5.890
	不假定等方差			0.497	98.584	0.621	1.180	2.377	−3.537	5.898

Table 3. Results of the non-parametric test for the listening section of test 1
表 3. 测试一听力非参数检验分析结果

	听力
曼－惠特尼U	1145.500
威尔科克森W	2420.500
Z	−1.045
渐近显著性(双尾)	0.296

分组变量：班级。

第一次测试结果显示，Shapiro-Wilk 正态性检验表明两个班级的总分成绩均服从正态分布(对照班：$P = 0.335 > 0.05$；实验班：$P = 0.595 > 0.05$)，而听力成绩均不服从正态分布(对照班：$P = 0.002 < 0.05$；实验班：$P = 0.022 < 0.05$)。基于此，对总分进行独立样本 t 检验，Levene 方差齐性检验表明两组数据满足方差齐性要求($P = 0.809 > 0.05$)，t 检验结果提示两组总分无显著统计学差异($P = 0.62 > 0.05$)；对听力成

绩采用 Mann-Whitney U 检验，结果亦显示无显著差异(P = 0.296 > 0.05)。

Table 4. Results of the normality test for test 2
表 4. 测试二正态性检验分析结果

	班级	柯尔莫戈洛夫－斯米诺夫(V)[a]			夏皮洛－威尔克		
		统计	自由度	显著性	统计	自由度	显著性
总分	1	0.091	52	0.200*	0.955	52	0.048
	2	0.109	51	0.181	0.948	51	0.026
听力	1	0.214	52	0.000	0.870	52	0.000
	2	0.216	51	0.000	0.903	51	0.001

*这是真显著性的下限。[a] 里利氏显著性修正。

Table 5. Results of the non-parametric test for total scores and listening scores of test 2
表 5. 测试二总分及听力非参数检验分析结果

	总分	听力
曼－惠特尼U	1268.500	1154.000
威尔科克森W	2646.500	2532.000
Z	−0.379	−1.163
渐近显著性(双尾)	0.704	0.245

分组变量：班级。

第二次测试中，正态性检验显示总分和听力成绩均不服从正态分布(所有 P < 0.05)，故均采用 Mann-Whitney U 检验，结果显示总分(P = 0.704 > 0.05)和听力(P = 0.245 > 0.05)均无显著差异。

Table 6. Results of the normality test for test 3
表 6. 测试三正态性检验分析结果

	班级	柯尔莫戈洛夫－斯米诺夫(V)[a]			夏皮洛－威尔克		
		统计	自由度	显著性	统计	自由度	显著性
总分	1	0.480	52	0.000	0.182	52	0.000
	2	0.097	51	0.200*	0.965	51	0.131
听力	1	0.299	52	0.000	0.771	52	0.000
	2	0.216	51	0.000	0.792	51	0.000

*这是真显著性的下限。[a] 里利氏显著性修正。

Table 7. Results of the non-parametric test for total scores and listening scores of test 3
表 7. 测试三总分及听力非参数检验分析结果

	总分	听力
曼－惠特尼U	1268.500	1270.500
威尔科克森W	2646.500	2596.500
Z	−0.379	−0.374
渐近显著性(双尾)	0.704	0.709

分组变量：班级。

第三次测试结果与第二次一致，总分和听力均为非正态分布(所有 $P < 0.05$)，Mann-Whitney U 检验表明总分($P = 0.704 > 0.05$)和听力($P = 0.709 > 0.05$)均未呈现显著差异。

综上所述，在实验开始前的连续三次测试中，实验班与对照班在总分及听力成绩上均未表现出统计学上的显著差异(所有 $P > 0.05$)，表明两个班级的英语学业水平在实验干预前处于均衡状态，满足进行对比实验研究的前提条件。

为确保命题质量，研究组建了一个由三名教师组成的命题小组，这些教师均具备十年以上高中英语教学经验，并曾参与市级统考命题工作。该小组负责人工命题的设计、筛选与优化，以保障试题的专业性与科学性。

3.2. 研究问题

本文主要探讨以下三个方面的问题：一是人工智能和人工生成的高中英语听力短对话题目，在难度与区分度上存在怎样的差异；二是相较于传统人工命题，人工智能命题存在哪些显著优势与突出不足；三是如何利用人工智能提升高中英语听力命题的质量和效率。

3.3. 研究工具

本研究用到的人工智能相关应用为 DeepSeek 和 TTSmaker，DeepSeek 主要用于生成实验组的听力试题和听力文本，TTSmaker 则用于根据听力稿的对话内容生成相应的听力音频。在数据分析阶段，用到了 Excel 和 SPSS 来对数据进行统计和分析。

3.4. 研究流程

3.4.1. 命题阶段

针对听力的短对话部分，让命题组和人工智能分别设计题目，确保题目覆盖细节信息、推理判断和主旨大意这几种常考的题目类型。人工智能试题通过让 DeepSeek 分析并模仿近五年全国高考英语一卷，先总结归纳出题目的共同特点，再生成高中英语听力短对话部分的试题，并提供对应的听力文本。听力音频的指令语和对话部分是 TTSmaker 网站生成，整个音频是用剪辑软件合成。传统人工试题是学校高二年级英语组负责出卷和审卷的教师共同完成的。具体题目如下：

命题组设计的试题：

1. What is the woman probably busy doing?

A. Preparing breakfast.

B. Looking for a tie.

C. Getting the kids dressed.

2. Where does the conversation take place?

A. In a taxi.　　B. On a plane.　　C. At an airport.

3. What is the relationship between the speakers?

A. Classmates.

B. Teacher and student.

C. Shop assistant and customer.

4. How much do oranges cost per kilo?

A. $3.　　B. $6.　　C. $8.

5. What are the speakers talking about?

A. Where to eat tonight.

B. How to go back to the hotel.

C. Whether to get a Chinese takeout.

人工智能设计的试题：

1. What time does the library close today?

A. 6:00 pm　　B. 7:30 pm　　C. 8:00 pm

2. Why can’t the man help the woman immediately?

A. He is waiting for a call.

B. He has a meeting soon.

C. He needs to finish a report.

3. What are the speakers mainly discussing?

A. A damaged book　　B. A research deadline　　C. A library policy

4. What will the woman do next?

A. Check online　　B. Return tomorrow　　C. Contact the manager

5. How does the man feel about the new rule?

A. Understanding　　B. Annoyed　　C. Confused

3.4.2. 实验阶段

抽取同年级两个英语基础接近的班级，分别就命题组和人工智能所出的试题，同时进行听力测试。在听力音频播放结束后，给一分钟时间让学生把答案转写到答题卡上，并第一时间收答题卡。

3.4.3. 数据分析阶段

先将两个班学生的答题情况按题目和选项形成电子数据，再统计并对比两组题目在难度和区分度上的差异。

4. 研究结果与分析

根据对学生答题情况的统计和计算，可以得到人工智能和传统人工命题的显著性差异，难度和区分度。

4.1. 显著性差异分析

Table 8. Descriptive statistics results of listening test scores for the control and experimental groups
表 8. 对照班和实验班听力测试成绩的描述性统计结果

	班级	个案数	平均值	标准 偏差	标准误差平均值
分数	1	52	3.73	0.888	0.123
	2	51	2.18	1.126	0.158

Table 9. Results of the independent samples t-test on the listening test scores between the control and experimental groups
表 9. 对照班和实验班听力测试成绩差异的独立样本 t 检验结果

		莱文方差等同性检验		平均值等同性 t 检验						
		F	显著性	t	自由度	Sig. (双尾)	平均值差值	标准误差差值	差值 95%置信区间	
									下限	上限
分数	假定等方差	0.489	0.486	7.786	101	0.000	1.554	0.200	1.158	1.950
	不假定等方差			7.768	94.964	0.000	1.554	0.200	1.157	1.952

为比较两种命题方式下的学生成绩差异，首先进行了描述性统计与独立样本 t 检验。描述性统计结果(表 8)显示，接受传统人工命题班级(班级 1)的听力成绩高于接受人工智能命题班级(班级 2)的成绩，均值差为 1.55。独立样本 t 检验结果(表 9)表明，莱文方差齐性检验不显著(F = 0.489, P = 0.486 > 0.05)，满足方差齐性假设。t 检验结果显示，两种命题方式下的学生成绩差异达到统计显著水平(t = 7.786, P < 0.001)。综上，传统人工命题和人工智能命题存在显著性差异。

4.2. 难度对比分析

由表 10 可知，在难度方面，人工智能命题的整体难度(平均值 0.44)显著高于传统人工命题(平均值 0.65)。具体而言，人工智能命题中第二、第三、第四道题的难度值均低于 0.4，表明这些题目对学生而言具有较高的挑战性，第五题的难度值甚至达到了 0.2，只有少数学生能回答正确。相比之下，传统人工命题中仅第一题难度为 0.31，其余四题难度均高于 0.65，显示出更好的难度分布控制与适中性，更符合一般学业评价的难度要求。

Table 10. Comparison of item difficulty between AI-generated and human-designed items
表 10. 人工智能命题和传统人工命题的难度对比

难度	第一题	第二题	第三题	第四题	第五题	平均值
人工智能命题	0.84	0.51	0.27	0.39	0.2	0.44
传统人工命题	0.31	0.68	0.73	0.86	0.69	0.65

4.3. 区分度对比分析

由表 11 可知，从区分度来看，两种命题方式的整体区分度相近，人工智能命题区分度为 0.25，传统人工命题区分度为 0.28，但内部分布存在差异。人工智能命题中各题区分度较为接近，其中第四题表现最佳，区分度为 0.32，但最低区分度出现在第五题，区分度为 0.18，说明该题在区分高低能力学生方面作用有限。传统人工命题则表现出更好的区分度波动范围，其中第三题区分度高达 0.40，显示出优异的区分能力，其余题目区分度也均保持在 0.2 及以上，试题整体区分度与人工智能试题相比，效果更佳。

Table 11. Comparison of item discrimination between AI-generated and human-designed items
表 11. 人工智能命题和传统人工命题的区分度对比

区分度	第一题	第二题	第三题	第四题	第五题	平均值
人工智能命题	0.25	0.25	0.25	0.32	0.18	0.25
传统人工命题	0.2	0.24	0.4	0.28	0.29	0.28

综合来看，传统人工命题和人工智能命题存在显著性差异。传统人工命题在难度控制和区分度表现上优于人工智能命题，其题目更能平衡考察目标与学生实际水平，而人工智能命题则呈现出难度偏高、部分题目区分能力不足的特点。这一结果提示，人工智能生成的题目仍需在难度调控和区分效度优化方面进一步改进，尤其需加强对高难度题目中干扰项的设计，以提升其在英语考试中的适用性。

5. 讨论

5.1. 人工智能命题的优势与局限性

当前，人工智能在高中英语听力命题中已展现出多方面的技术优势，但其应用仍存在一定局限性。

王蕾(2023)指出，试题自动生成技术的优势在于显著降低命题成本，解决传统命题方式面临的资源短缺问题[20]。从优势来看，人工智能显著提升了命题效率，能够在极短时间内完成对历年试题的考点提取与共性分析，并快速生成整套听力测试题目，有效缓解了传统命题方式“耗时耗力”的困境。在命题质量方面，人工智能尤其擅长捕捉对话中的细节信息，如时间、地点等事实性内容。数据显示，其命题的区分度整体与人工命题相近(平均区分度分别为 0.25，0.28)，部分题目甚至表现出更优的区分能力。此外，人工智能生成的试题在形式上较为规范，题干结构和选项设置呈现出良好的标准化特征。

然而，人工智能命题仍面临一些明显挑战。人工智能技术尚未完全成熟，仍存在需要改进之处(蔡诗静等，2021) [21]。李婷(2020)的研究表明，人工智能在高中英语写作教学中的应用有利于提升学生作文水平，但仍需政府支持、家校协同，并与传统批改方式相结合[19]。中小学教师在使用生成式人工智能工具时，应当保持审慎态度，对生成内容进行必要的验证和筛选，确保教学信息的准确性(张晖，郭宇航，2025) [18]。人工智能在英语听力命题上，最突出的问题在于难度控制不够稳定，题目难度常呈现两极分化。例如，在本研究中，人工智能命制的第三题和第五题难度值分别仅为 0.27 和 0.20，成为极端难题，反映出题目内容或设问方式可能与学生的实际认知经验存在偏差。其次，人工智能在把握对话核心信息方面尚有不足，如第三题以“受损书籍”作为主题，缺乏典型性，一定程度上削弱了题目的有效性。此外，人工智能命题在语境创设方面也显得较为单一，多围绕“图书馆”等有限场景展开，未能像人工命题那样涵盖“出租车”“机场”“商店”等多样化的真实生活情境，从而限制了对学生综合听力理解能力的考察。

5.2. 高中英语听力命题的优化路径

基于经典测试理论(CTT)的分析框架，人工智能命题在区分度方面呈现出一定的均衡性特征，其生成的题目整体上分布较为集中，更适合用于学业水平考试等需广泛覆盖基础能力的测评场景。然而，在选拔性考试中，往往需要题目具有更高的区分度差异，以精准识别不同能力层次的学生。当前人工智能尚难以自主实现这种差异化命题目标，必须依赖人工干预进行针对性优化。

基于以上发现，本研究建议采用“人机协同”的命题模式，以充分发挥各自优势。在 CTT 框架下，教师的干预主要体现在对难度和区分度的精细调节上。具体而言，可首先输入考点、难度、题型等命题要求，利用人工智能，自动生成题目初稿。随后，由教师团队对试题进行优化。通过增删题干信息或调节选项迷惑性来构建更合理的难度梯度，并针对学生常见思维误区补充干扰项来优化区分度，确保题目能有效鉴别不同能力水平的学生。

这种人机协作的模式，既充分利用了人工智能的高效性，又发挥了教师在考点把握、难度控制和区分度优化方面的专业经验，能够有效提升命题的整体质量与效率。这一路径为高中英语听力命题工作的改进提供了切实可行的方向，也为未来相关研究提供了重要参考。

从更广阔的学术视野看，未来的优化路径还可引入项目反应理论(Item Response Theory, IRT)。IRT 能

够提供项目特征曲线，更精细地刻画题目在不同能力水平学生上的表现，并能评估选项的有效性，为人工智能模型优化提供更具体的数据指导。本研究因样本量和研究阶段的限制，采用了 CTT 进行分析，未来的研究可在此基础上进一步探索 IRT 框架下的 AIG 质量评估。

6. 小结

人工智能在高中英语听力测试命题效率上优势显著，且生成的题目在形式上较为规范。从整体区分度来看，人工智能命题与人工命题结果接近，表明其已具备一定的质量基础。然而，人工智能命题仍存在明显不足。其难度控制不够稳定，在把握对话核心信息方面存在偏差，生成的题目场景设置较为单一，未能涵盖人工命题中多样化的生活情境。

针对这些问题，本文提出“人机协作”的高中英语听力命题模式，通过教师团队在难度调控和区分度优化方面进行专业干预，使试题难度和区分度达标率，同时保持了效率优势。这一协同路径为平衡命题效率与质量提供了可行的实践方案。

尽管取得了一定成果，本研究仍存在样本代表性有限、题型覆盖面不足等局限。未来研究可扩大样本范围至不同类型学校，拓展至长对话和独白等更多题型，并引入项目反应理论(IRT)等更先进的测量学模型，建立更完善的智能命题质量评价体系，探索人工智能在更复杂语境下的应用潜力，从而推动人工智能技术在教育测评领域的科学应用和健康发展。

参考文献

[1] 郑燕虹, 罗常军, 蒋洪新. 人工智能推进外语教育改革的探索[J]. 外语界, 2025(1): 8-12.

[2] 何屹松, 孙媛媛, 汪张龙, 等. 人工智能评测技术在大规模中英文作文阅卷中的应用探索[J]. 中国考试, 2018(6): 63-71.

[3] Kurdi, G., Leo, J., Parsia, B., *et al.* (2020) A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, **30**, 121-204. https://doi.org/10.1007/s40593-019-00186-y

[4] Bejar, I.I., Lawless, R.R., Morley, M.E., *et al.* (2002) A Feasibility Study of On-The-Fly Item Generation in Adaptive Testing. *ETS Research Report Series*, **2002**, i-44. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

[5] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[6] 刘邦奇, 聂小林, 王士进, 等. 生成式人工智能与未来教育形态重塑: 技术框架、能力特征及应用趋势[J]. 电化教育研究, 2024, 45(1): 13-20.

[7] 张智义. 体认语言学视阈下 ChatGPT 语言生成及性能研究[J]. 外语研究, 2024, 41(3): 20-25, 43, 112.

[8] 贾蕃, 马颖. 生成式人工智能在外语教材编写中的应用[J]. 外语研究, 2025, 42(2): 55-61, 113.

[9] 陈大建, 胡杰辉. 基于大语言模型的自动化命题研究——以英语阅读理解试题为例[J]. 外语教学, 2025, 46(2): 40-48.

[10] 刘一凡. 人工智能技术在考试中的应用探讨[J]. 科技资讯, 2018, 16(23): 200-202.

[11] 甘容辉, 何高大. 人工智能在外语教学中的应用分析[C]//教育部高等学校教育技术专业教学指导委员会. 走向智慧时代的教育创新发展研究——第 16 届教育技术国际论坛暨首届智慧教育国际研讨会论文集. 广东金融学院外国语言与文化学院, 华南农业大学外国语学院, 2017: 107-109.

[12] 黄岚. 高校英语作文阅卷中人工智能测评技术的应用研究[J]. 湖北招生考试, 2024(5): 61-63.

[13] 吕鸣. 智能测评技术在大规模英语口语考试评卷中的探索与实践[J]. 中国考试, 2015(10): 51-57.

[14] 朱柯睿. 人工智能技术在英语口语测评中的应用[J]. 现代英语, 2025(4): 118-120.

[15] 栾爱春. 人工智能视野下的英语教学: 发展趋势与应对策略[J]. 中小学教师培训, 2019(1): 60-63.

[16] 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.

[17] 段晨曦. 生成式人工智能助力高中英语试题命制的实践探究[J]. 教育评价, 2024(11): 77-81.

[18] 张晖, 郭宇航. 生成式人工智能赋能中小学英语教师的机遇与挑战[J]. 林区教学, 2025(7): 79-84.

[19] 李婷. 高中英语写作教学中的人工智能应用困境与对策——以句酷批改网为例[D]: [硕士学位论文]. 镇江: 江苏大学, 2020.

[20] 王蕾. 人工智能生成内容技术在教育考试中应用探析[J]. 中国考试, 2023(8): 19-27.

[21] 蔡诗静, 翁俐, 严静. 人工智能在中小学英语教学中的应用研究[J]. 英语广场, 2021(29): 131-133.