

生成式人工智能与教师英语写作任务评改一致度研究

刘秀梅, 董 静, 田静莹, 古 明*

中国矿业大学(北京)文法学院, 北京

收稿日期: 2026年6月3日; 录用日期: 2026年8月14日; 发布日期: 2026年8月25日

摘 要

本研究旨在探讨生成式人工智能与教师在评改英语写作任务时的一致程度。共15名大学生参加了为期三个月的雅思写作任务2教学实验, 由两位教师和两种生成式人工智能(KIMI与ChatGPT 4.0)分别对120份作文进行独立评分并提供文字版评改建议。通过计算生成式人工智能与教师评分的一致性, 发现两种生成式人工智能评分的内部一致性高于两位教师的内部一致性; 人工智能与教师间的一致程度一般。从成绩分布情况看, 生成式人工智能成绩分布呈现比较集中的趋势, 且成绩评定偏高于学生实际英语写作能力。通过编码分析教师与生成式人工智能对学生英语作文的评改建议, 发现两种生成式人工智能在对120份作文的评改建议中稳定且一致地遵守了雅思作文2的评改维度; 而两位教师则有时会有主观倾向, 评改建议不能覆盖雅思作文2的四个评分维度。研究还发现, 教师与生成式人工智能在评改建议提出方式方面也有差别。研究结论表明目前教师尚不能完全依赖人工智能评改作文, 但可与人工智能协作开展写作任务评改, 提高效率的同时, 有效避免因教师人为因素产生的一致性。

关键词

生成式人工智能, 教师, 英语写作, 评改一致度

Consistency between GAI's and Teachers' Rating and Corrective Feedback in English Writing Task

Xiumei Liu, Jing Dong, Jingying Tian, Ming Gu*

School of Law and Humanities, China University of Mining and Technology, Beijing

Received: June 3, 2026; accepted: August 14, 2026; published: August 25, 2026

*通讯作者。

文章引用: 刘秀梅, 董静, 田静莹, 古明. 生成式人工智能与教师英语写作任务评改一致度研究[J]. 创新教育研究, 2026, 14(8): 376-383. DOI: 10.12677/ces.2026.148616

Abstract

This study aims to explore the consistency between generative artificial intelligence (GAI) and teachers in assessing English writing tasks. A total of 15 university students participated in a three-month teaching experiment focused on IELTS Writing Task 2. Two teachers and two generative AI systems (KIMI and ChatGPT 4.0) independently scored 120 essays and provided written feedback. By calculating the consistency between the AI and teacher scores, it was found that the internal consistency of the two AI systems was higher than that of the two teachers, while the consistency between the AI and teachers was moderate. In terms of score distribution, the AI systems exhibited a relatively concentrated trend, with scores generally higher than the students' actual English writing proficiency. Through coding and analysis of the feedback provided by teachers and AI on the students' essays, it was observed that the two AI systems consistently adhered to the assessment criteria of IELTS Writing Task 2 across all 120 essays. In contrast, the teachers occasionally displayed subjective tendencies and did not fully cover all four scoring dimensions of the task. Additionally, differences were noted in the manner in which teachers and AI delivered their feedback. The study concludes that, at present, teachers cannot fully rely on AI for essay assessment. However, collaboration between teachers and AI in evaluating writing tasks can enhance efficiency while effectively mitigating inconsistencies and subjectivity arising from human factors.

Keywords

Generative Artificial Intelligence (GAI), Teachers, English Writing, Consistency in Rating and Feedback

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

根据输出假说, 写作在二语教学中扮演着不可或缺的角色[1]。外语写作活动一方面可以拓宽学生语言学习深度和广度、提高学习者语言准确性、流利性和创造性, 促进学习者语言能力的全面发展。同时, 外语写作活动也是帮助大学生重构语言能力、培养跨文化交际意识、提升自主学习能力、提高思维表达水平的关键途径。

然而, 在大学英语教学实践中, 英语写作教学面临着诸多挑战。其中尤为突出的是教师对学生作文的评改效率低、质量不足、针对性不强[2]。批改网等自动写作评估系统(AWE)通过为学习者提供纠错反馈, 在一定程度上帮助教师和学生优化了写作过程。但批改网反馈多集中于语法和词汇层面的纠错, 而对写作结构、论证逻辑等高阶写作技能的指导则显得力不从心[3] [4]。

随着生成式预训练变换器(GPT)等大语言模型(LLMs)的问世, 如 OpenAI 的 GPT-4, 基于自然语言处理的模型以其高效、大量的反馈能力备受关注, 为外语写作教学提供了更多的可能性[5]。较之批改网等 AWE, 生成式人工智能在英语作文评改中展现出显著优势。它不仅能够准确解读作文的主题与内容, 还能深入考量相关语境、逻辑结构和论证方法, 提供更为全面的反馈。与批改网相比, 生成式人工智能在处理复杂句子和上下文关联时表现更加精准, 有效避免了机械性误判。此外, 生成式人工智能可以根据学生的写作风格与语言水平, 提供个性化的指导, 助力学生在语法、词汇、结构等多方面能力的提升。通过动态互动, 它还能模拟对话式学习, 帮助学生深入理解错误原因并探索改进方法, 从而促进其获得

写作能力的长足发展。此外，生成式人工智能还能够就写作结构、论证逻辑等高阶技能提供比较详细的分析与指导，有效弥补了批改网这方面的不足。

鉴于生成式人工智能在英语作文评改方面展示出的巨大潜力，大学英语教师是否可以完全或部分利用 AI 来评改学生英语作文成为业内关注的热点问题，并展开一系列研究[6]。本研究将回答以下两个问题。

- (1) 生成式人工智能与大学英语教师对英语作文评定分数的一致程度如何？
- (2) 生成式人工智能与大学英语教师在评改建议的评价维度和评改方式上有何异同？

2. 研究设计

研究采用实验法，研究团队在某 211 高校公开招募 15 名同学志愿参加为期三个月的雅思作文任务 2 写作实验。要求实验参与者之前没有参加过任何形式的雅思写作培训或雅思考试。15 名同学性别构成为女生 9 名、男生 6 名；大一同学 9 名、大二 1 名、大四 1 名和研一同学 4 名；实验参与者的学科构成为人文社科 7 名和理工科 8 名。实验者均自愿参与，并知情研究内容且同意本次实验相关保密协议。实验参与者具体信息见表 1。

Table 1. Basic information of experimental participants
表 1. 实验参与人员基本信息

变量	类别	人数	百分比(%)
性别	男	6	40
	女	9	60
年级	大一	9	60
	大二	1	7
	大四	1	7
	研一	4	27
学科	人文社会学科(英语)	7	47
	理工科	8	53

实验分为准备与实施两个阶段。准备阶段持续两周，研究人员组织教师对 15 位同学进行雅思作文 2 基本要求集中讲解，每周一次课(约 50 分钟)，帮学生熟悉雅思作文 2 基本要求与时间分配；以真题为例讲解雅思作文 2 题目类型及破题思路；结合高分和低分作文样例解析雅思作文 2 评分标准。经过两周培训，实验参与者可顺利开展雅思作文 2 写作任务。研究团队还对实验参与人员进行 ChatGPT 4.0 和 KIMI 的使用培训，确保每位同学都可熟练使用 ChatGPT 及 KIMI 获取对写作任务的评分、意见反馈以及进行个性化追问。

实验实施阶段历时 8 周，参与者每周完成一个完整的雅思作文 2 写作任务，写作过程采用“初稿写作 - 评改反馈 - 二稿修订”的循环模式。实验参与者在规定时间内(40 分钟)完成一篇雅思作文 2 初稿，该初稿将同时提交给两位资深雅思写作教师以及两种生成式人工智能系统(KIMI 与 ChatGPT 4.0)进行独立评改。为确保评改结果的可比性，研究团队为教师与人工智能系统提供了统一的雅思作文 2 评分标准，并要求教师和人工智能对每篇作文给出总体评分、四个子维度评分(任务响应、连贯性与衔接、词汇资源、语法范围与准确性)和文字反馈建议。实验参与者基于教师与人工智能的反馈建议，对初稿进行修改完善后形成二稿。实验参与者共完成 8 个完整的写作任务。研究团队收集到的数据包括 4 位评分员(2 位教师、

2 种人工智能)分别对 15 名同学 8 次写作任务共 120 份作文的评分结果以及文字评改建议约 14 万字。

就量化的数字数据，研究者采用 MATLAB (R2016b)对 2 位教师评分员(T1, T2)和 2 种生成式人工智能(KIMI, ChatGPT 4.0)所评定分数的一致程度进行统计分析。对质性的文字评改建议，研究团队采用主题编码和文献阅读方式分析数据，比较教师和生成式人工智能在英语作文评改建议方面的异同点。

3. 研究结果与讨论

为更好反映教师和生成式人工智能对写作任务分数评定的一致性，研究者分别计算了教师内部、人工智能内部、以及教师与人工智能间的评分一致性。

3.1. 教师评分内部一致性

首先，我们对两位教师(T1, T2)对 15 位实验参与者 8 次写作任务共 120 份作文评分结果的一致性进行分析，通过计算 Cronbach’s Alpha 系数以及 Pearson 系数，结果分别为 0.857 以及 0.757(见表 2)，两类系数均满足一致性检验范围，表明两位教师评分结果具有较强的内部一致性(Cronbach’s Alpha 系数介于 0.7~0.9 时，表明两者内部一致性良好；Pearson 系数介于 0.7~0.9 时，表明两者呈强相关)。其次，我们计算了两位教师之间的 MSE 系数与 MAE 系数，发现均方误差和平均绝对误差都超过 0.6。雅思写作任务满分为 9 分，两位评分员分差大于 0.5 分时即表明两位教师的评分结果存在较大差异。这一研究结果与之前学界观点一致，“写作评分作为一种主观性判断行为可能产生误差”[7]。因此，引入生成式人工智能与教师协作进行写作评分，不失为降低评分员主观性、提高写作评分稳定性的方法之一。在教学实践中，生成式人工智能可以监控教师评分稳定性，也可把人工智能评分作为附加评价结果，为教师评分提供一个稳定且量化的参考分数。

Table 2. Consistency and differences between teachers’ and GAI’s rating
表 2. 教师与生成式人工智能评分一致性与差异性

对象	一致性指标		差异性指标	
	Cronbach’s Alpha 系数	Pearson 系数	MSE	MAE
T1, T2	0.857	0.756	0.615	0.621
ChatGPT, KIMI	0.840	0.737	0.335	0.429
教师均值, AI 均值	0.746	0.669	0.630	0.638

3.2. 人工智能评分内部一致性

同样，我们对 ChatGPT、KIMI 两种生成式人工智能对 120 份作文的评分结果进行了一致性分析，Cronbach’s Alpha 系数以及 Pearson 系数分别为 0.840 与 0.737，两种系数同样满足一致性检验范围，这说明两种生成式人工智能评分结果同样具有较强的内部一致性。其次，我们计算了 KIMI 与 ChatGPT 之间的 MSE 系数及 MAE 系数，发现误差均低于 0.5，表明两种生成式人工智能的对雅思作文 2 的评分差异性在可接受范围，人工智能打分波动较小，比教师更加稳定。这一研究结果表明，生成式人工智能较好克服了之前批改网等自动评分系统在稳定性及公平性方面的风险[8]。值得注意的是，稳定性只是衡量写作评价结果的指标之一，评分结果的准确性同样重要。如果生成式人工智能在多个写作任务中都给出了相似的评分，但这些评分不是学生真实写作能力的反映，那么这种稳定性就弊大于利。生成式人工智能评分的稳定性还带来另一个潜在风险，即其反馈和改进建议有可能呈现机械和僵化趋势，对写作任务中学生某些类型的不足与问题缺乏敏感性，削弱了评价结果对写作教学的正向反拨作用。

3.3. 教师与人工智能评分一致性

最后,为了分析教师与生成式人工智能之间评分一致性与差异性,我们将两位教师评分与 ChatGPT、KIMI 两种人工智能评分都进行均值化处理。教师评分及 AI 评分的内部一致性为均值化处理提供了保证,分别的内部均值可代表教师与 AI 的评分结果。统计结果表明,教师均值与 AI 均值之间的 Cronbach's Alpha 系数以及 Pearson 系数分别为 0.746 与 0.669 (见表 2)。可以发现不管是教师内部一致性还是 AI 的内部一致性,都高于教师与 AI 间的一致性。表 2 同时显示教师评分与 AI 评分之间存在中等程度的一致性。其次,我们计算教师均值与 AI 均值之间的 MSE 系数与 MAE 系数,分别为 0.630、0.638,教师与 AI 之间的误差值高于教师内部及 AI 内部的误差值。这一结果表明,教师与生成式人工智能评分结果虽基本一致,但这种一致性仍有较大提高空间。

通过对比教师评分分布和 AI 评分分布(见图 1),我们发现两位教师评分结果正态分布更加匀称,而 KIMI 和 ChatGPT 评分结果正态分布的曲线非常陡峭。其中 ChatGPT 在 120 个写作任务评分中,5 分和 6 分评分结果多于 90 个,占总数 75%以上。KIMI 在 120 个写作任务中,6 分的评分结果占 50%以上。两种生成式人工智能的评分分布偏离了典型的正态分布,显示出高度集中的趋势,且对于初学雅思写作的实验参与者而言,成绩评定偏高。因此,生成式人工智能评分虽然稳定,但并不能非常精确地反映学生真实写作水平。图 1 同时显示,分别对教师和 AI 评分结果均值化处理后,教师评分结果正态分布更合理,符合学生真实写作水平,而 AI 的评分结果仍然体现出明显的集中趋势,且 6 分得分者超过总人数的 50%,与学生真实雅思写作能力有较大偏差。以上结果表明目前教师尚不能完全依赖生成式人工智能对英语写作任务进行评价;同时,本研究启示教师需要接受 AI 评分工具的使用培训,充分了解人工智能的工作原理、优势和局限性,以更好地辅助教学实践。

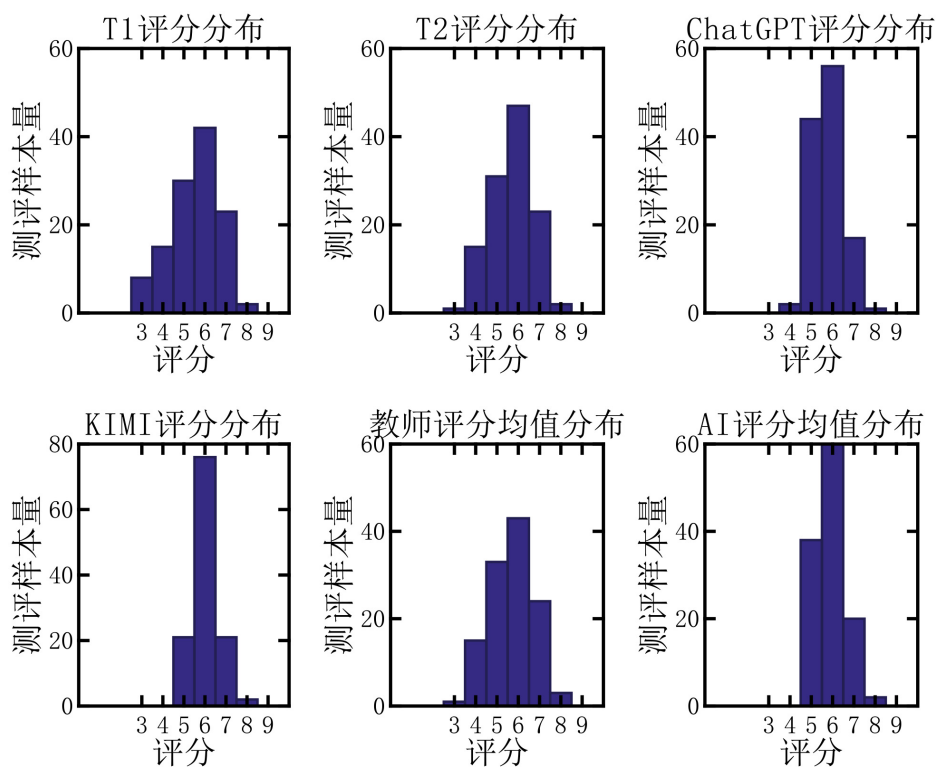


Figure 1. Distribution of teachers' and GAI's rating
图 1. 教师和 AI 评分分布图

3.4. 教师与人工智能评价维度一致性

研究团队就教师和生成式人工智能对英语作文评改建议的评价维度与评价方式两个方面进行比较。
表 3 为教师与生成式人工智能对英语作文评改建议在评价维度方面的比较。

Table 3. Comparison of evaluation dimensions for IELTS essays between teachers and GAI
表 3. 教师及生成式人工智能雅思作文评价维度比较

评分员	评分维度	点评次数	百分比(%)
T1	任务回应	109	91
	连贯性与衔接	96	80
	词汇资源	115	96
	语法范围与准确性	117	98
	其他	85	71
T2	任务回应	115	96
	连贯性与衔接	102	85
	词汇资源	108	90
	语法范围与准确性	110	92
	其他	62	52
KIMI	任务回应	120	100
	连贯性与衔接	120	100
	词汇资源	120	100
	语法范围与准确性	120	100
	其他	0	0
ChatGPT	任务回应	120	100
	连贯性与衔接	120	100
	词汇资源	120	100
	语法范围与准确性	120	100
	其他	0	0

注：“其他”包括标点使用、大小写、段落首行缩进、单词拼写和学生写作态度等。

从上表可知，两位教师与两种生成式人工智能基本都能按照雅思作文任务 2 所规定的四个评分维度进行评价反馈。与两位教师相比，KIMI 和 ChatGPT 在评改 120 份作文时，由于其算法和标准化流程，能够严格且一致地覆盖雅思作文的四个评分维度；而两位教师在少数情形下，可能受到个人主观倾向、时间压力等因素的影响，在评改建议中未全面覆盖四个评分维度。这一研究发现启示我们在教学实践中可以生成式人工智能的评改建议为补充，有效减少教师评改的人为偏差，提高评改的一致性和全面性。但我们并不能由此得出生成式人工智能的评改质量高于教师或者可以完全由人工智能替代教师提供写作任务评改的结论。因为 AI 提供的反馈有模板化倾向的风险，同时缺乏针对性和个性化，而教师则能够根据学生的具体问题提供更有针对性的建议[9] [10]，这一观点在评改形式对比中也得到验证。

表 3 同时表明，两位教师在评改学生写作任务时，都关注到雅思作文评分维度之外但体现学生基本

写作素养的问题,如标点符号使用、首字母大小写、单词拼写以及学生写作态度等。而两种生成式人工智能的评改建议中都未提及这一点。这进一步说明,生成式人工智能尽管较批改网等已有很大进步,但与教师相比,评改建议的灵活性和敏锐性仍有不足。

3.5. 教师与人工智能评改方式对比

通过比照教师和生成式人工智能对 120 份作文的评改建议,二者除在评改维度覆盖程度有所区别之外,在评改方式上也有所差别。首先,在评改建议呈现方式上,教师除使用文字描述外,还使用套色标记(如高亮或下划线)来引起学生注意,而 KIMI 和 ChatGPT 的评改建议仅以文本形式呈现,未采用任何视觉标记;同时教师在评改学生作文时,会就学生多次作文完成情况进行纵向对比,关注学生取得的进步和未解决的问题;教师也会横向对比不同学生的作文,发现存在的共同问题,提示学生多加注意。而生成式人工智能只关注学生本人当次提交的一份作文,不进行横向和纵向比照;另外,教师在评改建议中经常会通过举例等方式,让学生深刻理解问题和思考解决措施,而 KIMI 和 ChatGPT 只有在学生追问时才会通过举例进行分析说明。最后,在情感因素方面,生成式人工智能更能认可学生作文中表现较好的方面,对学生错误更加包容,先表扬学生后提出改进建议,而教师评改建议中则较多关注学生在写作中存在的问题,对学生表现较好的地方较少提及。

4. 结语

本研究探讨了生成式人工智能在英语写作任务评改中的潜力与不足。通过对比教师与两种生成式人工智能(KIMI 与 ChatGPT 4.0)对学生英语作文的评分与评改建议,我们发现尽管 KIMI 与 ChatGPT 4.0 在评分一致性方面表现优于教师,但其评定结果呈现明显的集中趋势,且分数评定高于学生实际写作能力。

此外,在评改维度覆盖全面程度和评改建议提出方式方面,教师与生成式人工智能也有差别。人工智能基于算法,能一致且高度稳定地覆盖给定的评改维度,但对维度之外同样重要的其他英语写作素质如写作习惯、写作态度等未作任何考量。教师在评改时,会有主观倾向性,少数评改建议中不能完全照顾各个评改维度,但教师能关注到评改维度之外的其他重要因素。在评改建议方式方面,教师除文字外,会利用视觉符号、举例等辅助讲解,同时教师还会对学生作文进行横向和纵向对比,而生成式人工智能的评改建议只有文字形式,且仅关注当人当次的写作任务。

本研究认为,在教学实践中教师可与生成式人工智能协作开展写作评改,提高效率的同时,可避免因教师人为因素产生的评改偏差。同时,教师可训练生成式人工智能进行特定写作任务的评改,提高人工智能评改的准确性和针对性。

基金项目

本研究为中国矿业大学(北京)“高质量精品课程建设与应用”项目(编号:J23ZD20)阶段性成果。

参考文献

- [1] Swain, M. (2005) The Output Hypothesis: Theory and Research. In: Hinkel, E., Ed., *Handbook of Research in Second Language Teaching and Learning*, Lawrence Erlbaum Associates, 471-483.
- [2] 张雪梅. 大学英语写作教学现状之调查[J]. 外语界, 2006(5): 28-32.
- [3] Ding, L. and Zou, D. (2024) Automated Writing Evaluation Systems: A Systematic Review of Grammarly, Pigai, and Criterion with a Perspective on Future Directions in the Age of Generative Artificial Intelligence. *Education and Information Technologies*, **29**, 14151-14203. <https://doi.org/10.1007/s10639-023-12402-3>
- [4] Li, X. and Zhang, L. (2019) A Critical Review of Pigai: Challenges and Opportunities in Automated Writing Assessment. *Computer Assisted Language Learning*, **32**, 589-605.

-
- [5] 陈莱, 吕明臣. ChatGPT 环境下的大学英语写作教学[J]. 当代外语研究, 2024(1): 161-168.
- [6] 任伟, 刘远博, 解月. 英语写作教学中 ChatGPT 与教师反馈的对比研究[J]. 外语教学理论与实践, 2024(4): 30-38, 60.
- [7] 谭智. 应用 Rasch 模型分析英语写作评分行为[J]. 外语教学理论与实践, 2008(1): 26-31.
- [8] 杨丽萍, 辛涛. 人工智能辅助能力测量: 写作自动化评分研究的核心问题[J]. 现代远程教育研究, 2021, 33(4): 51-62.
- [9] Zhang, Y. and Hyland, K. (2008) Feedback in Second Language Writing: Contexts and Issues. Cambridge University Press. <https://doi.org/10.1017/9781108635547>
- [10] Warschauer, M. and Ware, P. (2006) Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research*, 10, 157-180. <https://doi.org/10.1191/1362168806lr190oa>