

DeepSeek 赋能高中英语读后续写评价的实证研究

代星彦 王洪渊

四川轻化工大学 外语学院

DOI:10.32629/er.v9i8.7343

[摘要] AI 技术深度融入高中英语教学是教育数字化发展的必然趋势。读后续写作为高中英语核心题型，其智能化评价是教学数字化转型的重要方向。本研究以 DeepSeek-V3.2 大语言模型为评价工具，选取四川省某中学 150 名高二学生的读后续写样本，基于高考评分标准与内容、语言、连贯、结构四维量表，对比 DeepSeek 与教师评分的一致性。结果显示，二者总分呈显著正相关且高度一致；语言、连贯维度一致性良好，内容、结构维度一致性中等，差异源于 DeepSeek 侧重语言形式规范，教师更关注情节逻辑、语境契合等深层维度。研究验证了 DeepSeek 在读后续写评价中的可靠性，揭示其优势与局限，为人机协同实施高中英语读后续写评价提供实证参考。

[关键词] DeepSeek；读后续写；写作评价；评分一致性

中图分类号：G633.4 文献标识码：A

An Empirical Study on DeepSeek-empowered Evaluation of High School English Continuation Writing

Xingyan Dai, Hongyuan Wang

School of Foreign Languages, Sichuan University of Science and Engineering

Abstract: The in-depth integration of artificial intelligence into high school English teaching represents an inevitable trend of educational digitalization. As a core task in senior high school English examinations, the intelligent evaluation of continuation writing has become a key direction of teaching digital transformation. Adopting the DeepSeek-V3.2 large language model as the evaluation tool, this study collects continuation writing samples from 150 senior two students in a middle school of Sichuan Province. Based on the national college entrance examination scoring criteria and a four-dimensional evaluation scale covering content, language, coherence and structure, this paper compares the scoring consistency between DeepSeek and professional teachers. The results indicate that the total scores of AI and teacher evaluation show a significant positive correlation with high consistency. The two achieve favorable consistency in language and coherence dimensions and moderate consistency in content and structure dimensions. The scoring discrepancy mainly stems from different evaluation focuses: DeepSeek prioritizes formal linguistic standardization, while teachers attach greater importance to in-depth factors such as plot logic and contextual appropriateness. This study verifies the reliability of DeepSeek in continuation writing evaluation, clarifies its strengths and limitations, and provides empirical support for the human-machine collaborative evaluation of high school English continuation writing.

Keywords: DeepSeek; continuation writing; writing evaluation; scoring consistency

引言

《普通高中英语课程标准（2017 版，2025 修订）》明确指出，评价活动应贯穿教学的全过程，为检测教学目标服务，以发现学生学习中的问题，并提供及时帮助和反馈，促进学生更有效地开展学习（教育部，2025）。

读后续写作为高中英语核心题型与高考必考内容，兼具语言运用、逻辑思维与情节创造属性，其评价需兼顾内容契合度、语言准确性、篇章连贯性与结构完整性。传统人工评分存在耗时费力、主观性较强、评分标准难以统一等问题，

而 AI 大语言模型凭借语言处理与文本分析优势，为写作评价智能化提供了新路径。

现有研究多聚焦 AI 在通用英语写作中的应用，针对 DeepSeek 等前沿大模型在高中英语读后续写评价中的实证检验仍较为缺乏。基于此，本研究以 DeepSeek-V3.2 为评价工具，对比 AI 与教师在读后续写总分及多维指标的评分一致性，分析影响一致性的核心因素，验证 AI 赋能读后续写评价的有效性，为高中英语读后续写智能化评价与教学优化提供实践依据。

1 文献综述

国内外学者对人工智能在语言评价中的应用进行了大量的研究,使其在广度和深度都得到了不断地发展:一方面,对智能语言评价系统发展的研究,例如,智能写作评价系统快速发展阶段的国外 IEA 和 E-Rater 等系统无法评估论述的观点,也无法评估内容的逻辑性和连贯性,此外, E-Rater 只提供作文的分数,没有详细的反馈内容(张荔、盛越, 2015);在智能写作评价系统广泛应用阶段时,国内外相继出现了 Criterion, My Access, Writing Roadmap, 批改网、iWrite 等系统, My Access 系统能够为学生提供即时的反馈信息,包括词汇、语法、句式等方面的建议(Dikli & Semire, 2010), 批改网的即时批改、自动反馈等功能优势可以弥补人工批改的不足(庄翠娟, 2022);近年来人工智能大语言模型高速发展,依托 ChatGPT 开展读后续写反馈不仅有助于优化英语写作的互动质量,而且有助于拓宽英语写作的元语用教学资源,对于推动英语学科教学的数智化转型也具有一定实践价值(王一航等, 2024)。另一方面,人工智能技术也被拓展至特定领域的语言评价研究。例如,程秀峰等(2024)和王琳(2024)探索了大语言模型在学术论文评价中的应用,以提高教师反馈的效率和精准度。

综上所述,现有研究存在明显的局限性:首先,针对 ChatGPT, DeepSeek 等前沿大语言模型在语言评价中的应用探索仍显不足;其次,人工智能技术在读后续写这一特定写作形式的评价反馈机制研究尤为匮乏。基于此,本研究提出将 DeepSeek 大语言模型应用于读后续写评价反馈系统具有重要的实践意义,能够减轻教师评阅学生读后续写任务的负担,同时通过及时、高效的反馈反过来又能提升学生的写作能力。

2 研究设计

2.1 研究问题

2.1.1 基于高考英语读后续写评分标准,人工智能评分系统(DeepSeek)的总分评价结果与教师人工评价的一致性程度如何?(总分评分一致性)

2.1.2 在内容、语言、连贯和结构四个具体维度上, DeepSeek 的自动化评分结果与教师人工评价的一致性程度如何?(多维度评分一致性)

2.2 研究对象

本研究选取四川省 Z 市 Z 中学高二年级三个班共 150 名学生作为研究对象,其英语水平根据高二上学期期末考试和开学考试成绩确定为中等,可代表典型的高中生群体。该年级学生经过一年半读后续写相关训练,已具备一定的读后续写能力;同时高二学生相较于高三学生课业压力相对较小,有更多的时间和精力投入实验。

2.3 数据收集

本研究以 2025 年高考英语新课标全国 II 卷中的读后续写题为写作任务,收集 150 名学生的读后续写文本,教师分别依据高考英语读后续写评分标准及读后续写多维度评价量表对文本进行评价。随后,将部分教师评价结果输入开源大语言推理模型 DeepSeek-V3.2,以训练该模型的评价准确性。DeepSeek-V3.2 根据高考英语读后续写评分标准对学生的写作成果进行评分,并采用相同的评价量表对各篇读后续写在多个维度上进行分项评分与反馈,以分析其评价与人工评价之间的一致性。此外,基于高考英语读后续写评分标准和续写能力表(鲁周焕, 2021)制定读后续写多维度评价量表。量表内容如下:读后续写多维度评价量表共设置内容、语言、连贯、结构四个评价维度,各维度满分均为 25 分,总分 100 分,划分为优秀(21-25 分)、良好(11-20 分)、较差(0-10 分)三个等级。

内容维度:优秀等级要求续写与原文契合度高,情节新颖严谨、内容完整,与段落开头语衔接合理,细节饱满,前后呼应;良好等级强调与原文联系紧密,情节合理、续写较为完整,可与开头语形成一定衔接,细节稍显单薄,无重大逻辑漏洞;较差等级体现为与原文关联薄弱,情节逻辑混乱残缺,开头语衔接不畅,存在抄袭原文、脱离语境、缺乏有效细节等问题。

语言维度:优秀等级要求词汇语法丰富准确,行文流畅,错误极少,句式灵活,具备书面表达美感;良好等级的语法词汇可满足写作任务,表达较流畅,少量错误仅局部干扰表意,不影响整体理解;较差等级存在句式单调、词汇贫乏问题,语言错误多,严重阻碍语义传递,表达生硬,写作技巧不足。

连贯维度:优秀等级可自然使用衔接手段,上下文衔接紧密,逻辑贯通,篇章结构紧凑;良好等级能够有效运用连接成分,上下文基本衔接,逻辑较为通顺,无明显内容断层;较差等级衔接手段匮乏,上下文脱节,逻辑混乱,篇章连贯性缺失。

结构维度:优秀等级段落清晰,两段篇幅均衡,字数合规,与开头语衔接自然,首尾相互照应;良好等级段落框架基本清晰,篇幅比例、字数接近标准,与开头语有一定衔接,结构较为完整;较差等级段落排布不合理,字数严重偏离要求,衔接失效,结构松散,难以构成完整篇章。

具体步骤为:(1)收集 150 名学生的读后续写文本,由 3 名英语教师依据高考读后续写评分标准与多维度量表完成评分与反馈。(2)将教师评分样本导入 DeepSeek-V3.2 进行模型训练,使其熟悉评分规则与量表标准。(3)指令 DeepSeek 严格按照高考评分标准与多维量表完成总分评分

与四维分项评分，每篇作文重复评分3次，取三次得分的平均值以保证信度。(4)将教师与DeepSeek评分数据录入SPSS 29.0软件，用于一致性统计分析。

2.4 数据分析

本研究的数据分析旨在通过定量统计方法，系统检验DeepSeek-V3.2大语言模型基于高考读后续写评分标准及多维度评价量表与教师评分的一致性。分析过程以SPSS 29.0为主要工具，综合运用描述性统计、信度分析、相关性分析、一致性检验及差异检验等方法，量化DeepSeek与教师评分在总分及四个维度（内容、语言、连贯、结构）上的吻合程度，为回答研究问题提供数据支撑。数据分析分为两部分：一是基于高考评分标准的总分评分一致性分析，二是基于多维度量表的分项评分一致性分析。

2.4.1 DeepSeek与教师总分评分一致性分析

首先，通过描述性统计计算教师与DeepSeek评分的均值、标准差和极值范围，结果见表1。结果表明二者在总分分布上较为接近，DeepSeek评分均值略低于教师评分。

表1 描述性统计表

	N	最小值	最大值	平均值	标准差
教师1评分	150	12	23	18.23	2.703
教师2评分	150	13	23	18.25	2.583
教师3评分	150	14	23	18.28	2.575
DeepSeek1评分	150	10	22	17.80	2.742
DeepSeek2评分	150	10	22	17.79	2.884
DeepSeek3评分	150	11	23	17.98	2.673
有效个案数（成列）	150				

其次，采用克隆巴赫 α 系数检验教师间及DeepSeek三次评分间的内部一致性。结果显示（见表2），教师评分信度 α 为0.981，DeepSeek评分信度 α 为0.955，表明教师间评分高度一致，DeepSeek自身评分稳定性良好。

随后，通过Pearson相关系数 r 和组内相关系数ICC评估DeepSeek与教师总分评分的一致性与相关性。结果显示，Pearson相关系数 $r=0.925$ （ $p<0.01$ ），表明二者呈极高正相关；ICC值为0.955，表明评分者间一致性极高。

最后，运用配对样本 t 检验检测二者评分的统计差异。结果显示， $p<0.001$ ，表明DeepSeek与教师评分均值差异具有统计学意义。结合表1中教师平均分（约18.25）与DeepSeek平均分（约17.85）的对比，DeepSeek评分显著低于教师评分，差值均值为0.40分。

表2 教师与DeepSeek总分评分一致性比较

指标	结果	分析
教师评分信度 α	0.981	信度极佳

DeepSeek评分信度 α	0.955	信度良好
Pearson相关 r	0.925	极强正相关
ICC（一致性）	0.955	高度一致
配对 t 检验 p 值	<0.001	差异显著

上述结果表明，DeepSeek在模拟高考读后续写评分标准时，与教师评分在总分上具有极高的一致性和相关性，但存在系统性压分倾向，可能反映其在某些主观维度上的评价差异。

2.4.2 DeepSeek与教师多维度评分一致性分析

为探究DeepSeek与教师在不同评价维度上的评分一致性，本研究基于内容、语言、连贯、结构四个维度，分别计算教师评分信度、DeepSeek评分信度、皮尔逊相关系数、组内相关系数（ICC）及配对样本 t 检验，结果汇总于表3。

表3 教师与DeepSeek多维度评分一致性比较

指标	内容维度	语言维度	连贯维度	结构维度
教师评分信度 α	0.983	0.982	0.982	0.982
DeepSeek评分信度 α	0.959	0.959	0.958	0.959
Pearson相关 r	0.761	0.790	0.831	0.820
ICC（一致性）	0.665	0.788	0.773	0.536
配对 t 检验 p 值	<0.001	<0.001	<0.001	<0.001

在信度方面，教师评分在各维度的 α 系数均为0.982或0.983，DeepSeek评分在各维度的 α 系数均为0.958或0.959，表明二者在各维度的内部一致性均良好。

在相关性一致性方面，各维度呈现出明显差异，具体表现如下。

语言维度：Pearson相关 $r=0.790$ ，ICC=0.788，配对 t 检验 $p<0.001$ ，表明DeepSeek与教师评分在语言维度呈中高度相关、一致性良好，但均值差异显著。

连贯维度：Pearson相关 $r=0.831$ ，ICC=0.773，配对 t 检验 $p<0.001$ ，表明一致性良好，相关性较高，但均值差异显著。

内容维度：Pearson相关 $r=0.761$ ，ICC=0.665，配对 t 检验 $p<0.001$ ，表明一致性中等，相关性中等偏强，均值差异显著。

结构维度：Pearson相关 $r=0.820$ ，ICC=0.536，配对 t 检验 $p<0.001$ ，表明相关性较高，但一致性水平较低，均值差异显著。

3 结果与讨论

3.1 DeepSeek与教师总分评分一致性分析

研究问题一聚焦于DeepSeek与教师在总分评价上的一致性程度。描述性统计（表1）表明二者在总分分布上较为接近，但DeepSeek评分均值略低于教师评分。

信度分析(表2)显示,教师评分信度 $\alpha = 0.981$,DeepSeek 评分信度 $\alpha = 0.955$,表明教师间评分高度一致,DeepSeek 自身评分稳定性良好,均满足教育测量对评分信度的基本要求。

相关性分析显示,Pearson 相关系数 $r = 0.925$ ($p < 0.01$),表明 DeepSeek 与教师总分评分呈极高正相关,二者在评分趋势上高度一致。组内相关系数(ICC) = 0.955,进一步验证了二者评分的一致性处于极高水平。

然而,配对样本 t 检验结果显示, $p < 0.001$,表明 DeepSeek 与教师评分均值差异具有统计学意义。结合均值对比,DeepSeek 评分显著低于教师评分,差值均值为 0.40 分。

上述结果表明,DeepSeek 在模拟高考读后续写评分标准时,与教师评分在总分层面具有极高的一致性和相关性,但存在系统性压分倾向。这一发现与已有研究(宛平、顾小清,2024)一致,即 AI 评分系统在规则明确的任务中具有较高可靠性,但可能因对语言形式规范要求更严格而导致评分偏低。结合后续多维度分析,这种压分倾向可能源于 DeepSeek 在语言和连贯等维度上对形式规范的高度敏感,而教师则更关注情节逻辑与语境契合度等深层维度。

3.2 DeepSeek 与教师多维度评分一致性分析

研究问题二进一步探究 DeepSeek 与教师在内容、语言、连贯、结构四个维度上评分的一致性。信度分析(表3)显示,教师评分在各维度的 α 系数均为 0.982 或 0.983,DeepSeek 评分在各维度的 α 系数均为 0.958 或 0.959,表明二者在各维度的内部一致性均良好。

在相关性与一致性方面,各维度呈现出明显差异:

语言维度: Pearson $r = 0.790$, ICC = 0.788, 配对 t 检验 $p < 0.001$ 。表明 DeepSeek 与教师在语言维度呈中高度相关、一致性良好,但均值差异显著。这一结果说明,DeepSeek 在语法、词汇等语言形式判断上与教师较为接近,但可能存在系统性的评分偏差。结合配对 t 检验结果,DeepSeek 可能对语言错误的扣分更为严格,而教师对不影响表意的语言问题更具宽容度。

连贯维度: Pearson $r = 0.831$, ICC = 0.773, 配对 t 检验 $p < 0.001$ 。表明一致性良好,相关性较高,但均值差异显著。连贯性评价涉及篇章逻辑和衔接手段的使用,DeepSeek 在识别显性连接词和段落衔接方面表现较好,但在语义连贯性和逻辑流畅性的深层判断上可能与教师存在差异。

内容维度: Pearson $r = 0.761$, ICC = 0.665, 配对 t 检验 $p < 0.001$ 。表明一致性中等,相关性中等偏强,均值差异显著。内容评价涉及情节发展的逻辑性、与原文的契合度以及细节丰富度等,这些维度需要较强的语境理解和创造性判断能力,DeepSeek 在此类任务上仍存在局限。

结构维度: Pearson $r = 0.820$, ICC = 0.536, 配对 t 检验 p

< 0.001 。表明相关性较高,但一致性水平较低,均值差异显著。结构维度评价涉及段落划分、两段比例均衡性、字数合规性等,虽然 DeepSeek 在形式识别上有优势,但教师对“结构完整”的理解更多结合情节逻辑的起承转合,而非单纯的段落形式,这可能导致 ICC 较低。

上述结果表明,DeepSeek 与教师在读后续写评价中的一致性受评价维度客观性程度的影响。在语言和连贯等偏重形式规范的维度上,DeepSeek 与教师一致性较高;而在内容和结构等需要深层理解、情节逻辑判断的维度上,一致性相对较低。因此 DeepSeek 在表层特征识别上具有优势,但在深层语义理解、语境敏感性和创造性判断方面仍存在局限。

此外,所有维度的配对 t 检验 p 值均小于 0.001,表明 DeepSeek 与教师在四个维度上的评分均存在显著系统差异。结合均值差异方向(DeepSeek 评分普遍低于教师评分),可以推断 DeepSeek 在读后续写评价中对各维度的要求可能更为严格,尤其在语言形式规范方面扣分更甚,而教师更注重整体表达效果和情节逻辑的完整性。

4 结语

本研究以 DeepSeek-V3.2 大语言模型为评价工具,通过对 150 名高二学生读后续写样本的实证分析,系统检验了 DeepSeek 与教师在总分及四个维度上的评分一致性。研究发现:总分层面,二者呈极高正相关($r = 0.925$)与高度一致(ICC = 0.955),但 DeepSeek 评分显著低于教师,存在系统性压分倾向;多维度层面,语言与连贯维度一致性良好,内容与结构维度一致性中等,差异源于 DeepSeek 侧重语言形式规范,教师更关注情节逻辑、语境契合等深层维度。研究验证了 DeepSeek 在读后续写评价中的可靠性与有效性,同时揭示了其在深层语义理解、创造性判断方面的局限。

本研究为高中英语读后续写智能化评价提供了实证支持,建议在实践中采用人机协同的评价模式,发挥 DeepSeek 在语法检测、结构识别方面的优势,同时保留教师在内容深度、逻辑连贯性等方面的专业判断。未来研究可扩大样本规模、拓展文本类型,并探索“评价-诊断-反馈”一体化系统的构建,以更科学地赋能高中英语写作教学。

[参考文献]

- [1] Dikli, Semire. The Nature of Automated Essay Scoring Feedback[J]. The Calico Journal, 2010 (28): 99-134.
- [2] 程秀峰,李嘉琪,杨金奇,严中华.大语言模型对学术论文评价的可利用性探讨[J].图书情报工作,2024(18):41-49.
- [3] 陈则航,王蔷,孙引.信息技术支持的英语“评一学一教”一体化探究[J].天津师范大学学报(基础教育版),2024(1):13-18.
- [4] 胡壮麟.ChatGPT 谈外语教学[J].学论经纬,2023(3):12-15.
- [5] 鲁周焕.基于读后续写能力培养的练习设计[J].中小

学英语教学与研究,2021(4):52-56.

[6]文秋芳.AI时代外语教育的选题框架和方法建议[J].
外语界,2025(2):2-10.

[7]文秋芳.人工智能时代的外语教育会产生颠覆性革命
吗?[J].现代外语,2024(5):722-730.

[8]文秋芳,梁茂成.人机互动协商能力:ChatGPT与外语教
育[J].外语教学与研究,2024(2):286-296.

[9]王琳.大语言模型技术背景下重塑研究生论文评价与
指导[J].学位与研究生教育,2024(12):30-37.

[10]王一航,毛延生,董亚娟.ChatGPT赋能高中英语读后
续写反馈的实证研究[J].教育测量与评价,2024(5):71-83.

[11]杨宗凯.ChatGPT/生成式人工智能对教育的影响探析

及应对策略[J].华东师范大学学报,2023(7):26-34.

[12]张荔,盛越.自动作文评阅系统反馈效果个案研究[J].
外语电化教学,2015(3):38-44.

[13]中华人民共和国教育部.普通高中英语课程标准
(2017版,2025修订)[M].北京:人民教育出版社,2025.

[14]庄翠娟.基于句酷批改网英语写作评阅系统研究[J].
现代英语,2022(5):116-119.

作者简介：

代星彦（2000-），女，汉族，四川乐山，研究生，研
究方向：中学英语教育。