Modern Linguistics 现代语言学, 2026, 14(5), 687-694
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145448

Hans汉斯

大语言模型在BEC高级写作评分中的效能对比研究

——基于DeepSeek与豆包的实证分析

张楠竹，李华东

上海海事大学外国语学院，上海

收稿日期：2026年4月8日；录用日期：2026年5月12日；发布日期：2026年5月27日

摘　要

为检验国产大语言模型在BEC高级商务英语作文评分中的实用性，本研究以DeepSeek-V3.2与豆包为对象，选取了90篇BEC高级作文，设置无提示词、仅提供评分标准、同时提供评分标准和人工打分范文三种提示词场景开展对比实验。结果显示，DeepSeek-V3.2整体评分准确性、稳定性均优于豆包，从作文类型来看，两款模型均对商务报告评分最准确，商务信函评分能力较弱。仅提供评分标准会降低模型评分效果，搭配人工范文可明显提升评分质量。两款模型均可用于作文初评，但与专业人工评分仍有差距，暂不能完全替代人工。本研究为商务英语写作的人机协同评分提供了参考。

关键词

大语言模型，BEC高级写作，作文自动评分，提示词干预

A Comparative Study on the Efficacy of Large Language Models in BEC Higher Writing Scoring

—An Empirical Analysis of DeepSeek and Doubao

Nanzhu Zhang, Huadong Li

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: April 8, 2026; accepted: May 12, 2026; published: May 27, 2026

文章引用: 张楠竹, 李华东. 大语言模型在 BEC 高级写作评分中的效能对比研究[J]. 现代语言学, 2026, 14(5): 687-694.
DOI: 10.12677/ml.2026.145448

Abstract

To examine the practicability of domestic large language models (LLMs) in the scoring of BEC Higher business English writing, this study selects DeepSeek-V3.2 and Doubao as research objects, and employs a dataset of 90 BEC Higher writing scripts to carry out controlled comparative experiments under three prompt scenarios: no prompt, only scoring criteria provided, and both scoring criteria and human-scored sample essays provided. The findings reveal that DeepSeek-V3.2 surpasses Doubao in both overall scoring accuracy and stability. In terms of writing genres, both models deliver the highest scoring accuracy for business report, while their scoring performance for business letter is relatively weaker. Providing only scoring criteria undermines the models' scoring efficacy, whereas the combination of scoring criteria and annotated sample essays can markedly improve scoring quality. Although both LLMs are applicable to the preliminary evaluation of writing scripts, a distinct gap remains between LLMs scoring and human scoring, indicating that they cannot fully replace human raters for the time being. This research offers implications for the implementation of human-machine collaborative scoring in business English writing assessment.

Keywords

Large Language Models, BEC Higher Writing, Automatic Writing Scoring, Prompt Intervention

1. 引言

随着人工智能技术在语言测评领域的深度渗透，自动作文评分(Automated Essay Scoring, AES)已从传统规则驱动模型向大语言模型(LLMs)驱动模式转型。剑桥商务英语高级(BEC Higher)作为全球认可度最高的商务英语能力测评之一，其作文评分需兼顾语言规范性与商务场景适配性，即不仅要求评估语法、词汇等基础语言维度，还需判断文本是否符合商务沟通逻辑，例如报告结构完整性、邮件语用得体性、论证数据支撑力度等等，这对大语言模型评分系统的专业化能力提出更高要求。

当前国内自动作文评分研究多聚焦于通用英语写作如大学英语四六级、雅思和托福写作等，而针对BEC 高级这类专业化商务测评场景的研究较少[1]。现有商用人工智能工具如批改网、Grammarly，在商务术语识别、跨文化商务语用评估等维度表现不足，而国产大语言模型如 DeepSeek、豆包凭借本土化训练优势，在专业领域的适配性逐渐凸显，但尚未有研究系统验证其在 BEC 高级作文评分中的效能。

本研究选取 DeepSeek-V3.2 (以下简称 DeepSeek)与豆包两款当前应用广泛、公开可及的国产大语言模型开展实证分析，尝试构建适用于 BEC 高级作文的评分效能评估框架。本研究的模型选择具有一定的随机性，仅基于模型的公开可用性与本土化适配优势，不构成对其他大语言模型的优劣评判，亦不具有排他性。研究旨在初步探索国产模型在商务英语写作评分中的应用可行性，补充商务英语写作测评智能化方向的实证资料，以期为同场景下大语言模型的落地应用、提示策略优化提供一定参考与借鉴。同时，通过量化对比两款模型的评分准确性与效率，为 BEC 培训机构及企业商务英语考核提供可落地的 AI 辅助评分方案，进一步明确大语言模型与人工评分的适用边界，为人机协同评分模式的构建提供实践依据，缓解 BEC 作文评分中专业考官资源紧张、评分效率偏低的现实问题。具体研究问题如下：

1) DeepSeek 与豆包在 BEC 高级英语作文评分中，哪一款模型与人工评分相比的整体评分一致性更高？

2) 两款模型对于不同类型的写作评分存在哪些差异，不同提示词设置对评分效果有何影响？

3) 两款模型是否具备替代人工进行 BEC 高级作文评分的可行性？

2. 文献综述

国外人工智能作文评分研究起步较早。早期研究以 ETS 的 e-rater 为代表，通过提取句法复杂度、词汇丰富度等人工特征构建评分模型，在 TOEFL 和 GRE 写作评分中实现了二次加权卡帕(Quadratic Weighted Kappa, QWK)值 0.8 以上的一致性。随着大语言模型的发展，GPT 系列、Claude 系列在专业化测评场景的应用成为研究热点。Kim (2025)指出，GPT-4 具备较高的自身评分一致性，与人工分数呈中等程度的正相关，且在分级一致性上优于人工评分者平均水平[2]。Suhan & Wolf (2026)对比 GPT-4 与人工对 EFL 学生作文的评分，发现 GPT-4 的评分准确率达到 79.3% [3]。Lan 等(2025)指出，大语言模型在专业化写作评分中，需通过领域语料微调提升场景适配性，否则易出现通用语言评分偏差[4]，例如过度关注语法错误，忽视商务场景特有的沟通有效性。

国内研究中，李颖(2021)针对 iWrite 系统的研究显示，国产 AI 工具在邮件写作的内容维度上评分一致性较高，但在结构维度上一致性较低[5]。但是大语言模型对专业场景例如商务英语评分的情况仍存在研究空白，针对专业文体评分是否一致仍待研究。刘玉屏(2025)对生成式 AI 在国际中文教学测评中的研究发现，大语言模型如 ChatGPT、文心一言和星火大模型在对作文进行评分时，有时候不只输出分数，还会附加一些详细的评分信息，诸如作文的结构特点、语法和词汇使用情况，甚至对话语流利度也会做出判断[6]，但是能否通过提示词干预提升评分准确性尚未得到验证。

综合现有文献，当前研究存在三大缺口。一是多数研究聚焦通用英语写作，针对 BEC 高级这类专业化商务英语写作测评的研究较少。二是模型对比对象较为单一，国外研究多围绕 GPT、Claude 展开，国内对 DeepSeek、豆包等国产模型的对比研究较为缺乏。三是评估维度不完整，现有研究多关注整体评分一致性，缺乏对 BEC 核心的商务写作不同类型细分验证，难以全面判断模型评分对于人工评分进行替代的可行性。

3. 研究方法

本研究针对上述不足，以 BEC 高级作文为研究对象，对比 DeepSeek 与豆包的评分效能，从整体与不同作文类型双层面验证其与人工评分的一致性，明确模型优势与局限。

3.1. 研究对象

本研究选取了 90 篇 BEC 高级考试考生模拟测试作文作为研究样本，样本来源为周之南主编的《BEC 写作全攻略(高级)》和盛梅主编的《BEC 写作高分快训(高级)》，这两本书提供了这些样本作文的人工打分，可用来与模型打分进行对比分析。选取文本为 BEC 高级写作考试第二部分题目，即选择以下三种类型的作文中的一类来完成写作任务，分别为商务信函、商务报告和商务建议。本研究在每种类型中各选择 30 篇，共计 90 篇文本作为研究样本进行大语言模型评分实验。

研究样本的筛选遵循以下标准，一是分数分布趋近于真实情况。鉴于参加 BEC 考试的考生大多数都已有良好的写作基础和英语水平，本研究参照 BEC 官方评分等级(0~5 分)进行样本控制，其中 3~4 分样本占比最高，0~2 分样本相对较少，以求更接近真实考试情况。二是错误类型全面，涵盖商务术语误用(如“profit margin”误写为“profit rate”)、语用不当等典型问题，确保评估的全面性与有效性。

3.2. 研究流程

本研究选取 DeepSeek 和豆包两款大语言模型作为评分工具，数据收集截止时间为 2025 年 11 月 10 日。

研究之初将文本转化为纯文本 txt 格式，然后手动将文本数据输入大语言模型进行打分。评分过程共分为三轮，第一轮评分时只给出打分指令，即“请对以下 BEC 高级考试作文打分，打分区间为 0~5 级，5 级为最好，不用给出打分理由，只输出分数。”第二轮评分除以上内容外，再加入 BEC 高级作文的官方评分标准，第三轮评分则同时给出评分标准和带有人工打分的范文(range finders)。范文数量为 6 篇，每种作文类型各两篇，以供大语言模型参考。

得到大语言模型的打分数据后，参照语言测评领域的通用标准，与这些样本的人工打分进行对比，计算二次加权卡帕(Quadratic Weighted Kappa, QWK)，Spearman 等级相关系数(Spearman’ s Rank Correlation)、相邻一致率(Adjacent Agreement Rate)和均方根误差(Root Mean Square Error, RMSE)等核心指标，对模型评分表现进行分析。

4. 结果与分析

下面从整体评分一致性对比和不同类型作文的评分一致性差异两个方面展示研究结果。

4.1. 整体评分一致性对比

研究结果显示，DeepSeek 在作文评分任务中的整体表现显著优于豆包，且两款模型均受提示词干预影响，但影响趋势与程度存在明显差异。人工打分参考范文的引入对模型评分的相关性提升具有积极作用，而评分标准的单独提供未能实现模型评分效能的有效提升(表 1)。

Table 1. Score data comparison between DeepSeek and Doubao
表 1. DeepSeek 和豆包评分数据统计

	无 prompt		评分标准		评分标准 + 人工打分范文	
	DeepSeek	豆包	DeepSeek	豆包	DeepSeek	豆包
QWK	0.752	0.609	0.637	0.560	0.668	0.577
Spearman’ s Rank Correlation	0.750	0.653	0.749	0.750	0.900	0.780
Adjacent Agreement Rate	1.000	0.970	0.922	0.933	0.989	0.960
RMSE	0.636	0.752	0.651	0.663	0.689	0.656

从核心一致性指标 QWK 的表现来看，在三种实验场景中，DeepSeek 的 QWK 值始终保持对豆包的绝对领先，且整体数值区间为 0.637~0.752，豆包则处于 0.560~0.609 的较低区间，二者的差距在无 prompt 场景下达到最大(0.752 vs. 0.609，差值 0.143)，反映出在无外部信息辅助的情况下，DeepSeek 对 BEC 高级作文评分的内在逻辑与标准的理解远优于豆包。从场景变化趋势来看，DeepSeek 的 QWK 值表现为无 prompt 时 > 提供评分标准加人工打分范文 > 仅提供评分标准。DeepSeek 评分在无 prompt 场景下达到峰值 0.752，在仅提供评分标准时出现明显下降(0.637)，引入人工打分范文作为参考后略有回升(0.668)。豆包的 QWK 值表现为相同趋势，无 prompt 场景下为 0.609，仅评分标准时降至最低 0.560，引入人工打分范文作为参考后回升至 0. 577。这一趋势表明，单纯的评分标准提示词不仅未对两款模型的评分一致性产生正向引导，反而造成了评分逻辑的干扰，而人工打分范文参考的引入能够在一定程度上修正模型的评分偏差，提升评分与真实标准的吻合度，但其修正效果有限，未使模型恢复至无 prompt 的原始表现。

Spearman 等级相关系数显示，DeepSeek 具有更优的稳定性与提升潜力，而豆包在提示词影响下波动较大。无 prompt 场景下，DeepSeek 与豆包的 Spearman 系数较为接近(0.750 vs. 0.653)，DeepSeek 小幅领先。仅提供评分标准时，豆包的 Spearman 系数出现大幅跃升，从 0.653 升至 0.750，与 DeepSeek (0.749)基本持平，这是豆包在所有指标中唯一与 DeepSeek 持平的表现，表明评分标准的引入虽干扰了豆包的评分一致性(QWK 下降)，但显著提升了其对作文质量等级的区分能力，使其排序逻辑更贴合人工评分标准。而 DeepSeek 在该场景下系数基本保持稳定，反映出其对评分标准的理解已形成较为稳定的内在逻辑，外部评分标准的引入未产生明显干扰。引入人工打分范文作为参考后，两款模型的 Spearman 系数均出现显著提升，且差距再次拉大，DeepSeek 从 0.749 跃升至 0.900，实现了质的提升，豆包则从 0.750 提升至 0.780，提升幅度远低于 DeepSeek。这一结果充分说明，人工打分参考范文的引入是提升模型评分区分度的有效手段，尤其对 DeepSeek 而言，人工打分的示范作用能够使其快速贴合真实人工评分的排序逻辑，实现评分区分度的大幅提升。而豆包虽能从中获益，但可能受限于自身对评分标准的理解能力，提升效果有限。同时，DeepSeek 在该场景下 0.900 的 Spearman 系数，表明其在评分标准和人工打分的双重辅助下，已具备较高的作文质量排序能力，基本能够准确区分不同质量等级的作文。

两款模型在相邻一致性率指标上均表现出较高水平，整体数值均在 0.922 以上，说明二者在作文评分中均能有效控制严重偏差，评分结果的稳健性较好，但 DeepSeek 仍保持领先优势。无 prompt 场景下，DeepSeek 的相邻一致性率达到满分 1.000，展现出完美的偏差控制能力，所有评分结果均未出现跨等级偏差，而豆包为 0.970，虽表现优异，但仍存在 3%的跨等级偏差情况。仅提供评分标准时，DeepSeek 的该指标出现明显下降，从 1.000 降至 0.922，豆包则小幅回升至 0.933，二者差距大幅缩小，这一结果与 QWK 的变化趋势一致，再次印证了单纯评分标准的引入对 DeepSeek 的评分稳健性产生了负面影响，使其出现了一定的跨等级偏差，而豆包则在评分标准的引导下，小幅优化了偏差控制能力。引入人工打分参考范文后，两款模型的相邻一致性率均出现明显回升，DeepSeek 从 0.922 升至 0.989，豆包从 0.933 升至 0.960，恢复至较高水平，且 DeepSeek 再次拉开差距。这表明，人工打分参考的引入能够有效修正模型的评分偏差，减少跨等级评分情况的发生，提升模型评分的稳健性，这与人工打分参考为模型提供了具体的评分示范，使其更清晰地把握不同等级作文的评分边界密切相关。整体而言，两款模型均具备较强的评分稳健性，不易出现严重的评分偏差，而 DeepSeek 在偏差控制的精准度上更具优势，人工打分参考则是提升二者稳健性的有效干预手段。

在 RMSE (均方根误差)指标上，两款模型的整体数值均在 0.636~0.752 之间，偏差程度相对可控，但豆包在部分场景下实现了对 DeepSeek 的小幅反超，反映出豆包虽在一致性、区分度上不及 DeepSeek，但在部分场景下的评分数值精准度存在一定优势。无 prompt 场景下，DeepSeek 的 RMSE 为 0.636，显著低于豆包的 0.752，说明其评分数值与真实评分的偏差更小，数值精准度更高。仅提供评分标准时，DeepSeek 的 RMSE 小幅上升至 0.651，豆包则大幅下降至 0.663，二者差距大幅缩小，豆包仅以 0.012 的微弱差距落后于 DeepSeek。引入人工打分参考后，DeepSeek 的 RMSE 继续上升至 0.689，豆包则进一步下降至 0.656，豆包实现了对 DeepSeek 的小幅反超。这一趋势表明，随着外部提示词干预的逐步深入，DeepSeek 的评分数值偏差呈逐步扩大趋势，而豆包则呈逐步缩小趋势，说明豆包的评分数值精准度更易受外部提示词的正向影响，在评分标准和人工打分的双重辅助下，其数值偏差能够持续优化，而 DeepSeek 的原始评分数值精准度较高，外部提示词的引入反而使其偏离了原始的精准评分逻辑，导致数值偏差略有扩大。同时，两款模型的 RMSE 整体波动幅度较小，均保持在 0.7 以下(除无 prompt 场景下的豆包)，说明二者的评分数值偏差均处于可控范围，即使是偏差最大的场景，数值误差也未超过 0.8 分，反映出两款大语言模型在作文评分的数值精准度上均具备一定的基础能力。

4.2. 不同类型作文的评分一致性差异

对 DeepSeek 与豆包在商务信函、商务报告、商务建议三类作文的评分表现展开分析，结果表明 DeepSeek 在三类商务作文评分中整体保持领先，豆包则展现出差异化提升特征。提示词干预对两款模型的影响存在题型差异，无 prompt 场景下模型评分一致性更优，人工打分参考对豆包的提升作用更显著，且商务报告类作文成为两款模型评分效能的最优适配题型。

从题型整体表现来看，商务报告是两款模型评分表现最稳定、效能最高的题型(表 2)，DeepSeek 在此类作文中的 QWK 值始终保持 0.73 以上，豆包也稳定 0.63~0.69 区间，远高于其在商务信函、商务建议类的表现。商务信函则是两款模型的共同短板(表 3)，尤其是豆包，其 QWK 值在全场景均低于 0.57，最低至 0.44，评分一致性与真实标准偏差较大。商务建议类作文则处于中间水平(表 4)，两款模型评分表现波动适中，且豆包在此类题型中展现出明显的提升潜力。

Table 2. Data analysis of scoring for business report writing tasks
表 2. 商务报告类作文打分数据分析

	无 prompt		评分标准		评分标准 + 人工打分	
	DeepSeek	豆包	DeepSeek	豆包	DeepSeek	豆包
QWK	0.77	0.67	0.76	0.63	0.73	0.69
Spearman’ s Rank Correlation	0.78	0.71	0.75	0.84	0.77	0.89

Table 3. Data analysis of scoring for business letter writing tasks
表 3. 商务信函类作文打分数据分析

	无 prompt		评分标准		评分标准 + 人工打分	
	DeepSeek	豆包	DeepSeek	豆包	DeepSeek	豆包
QWK	0.79	0.57	0.56	0.45	0.59	0.44
Spearman’ s Rank Correlation	0.80	0.58	0.74	0.62	0.64	0.62

Table 4. Data analysis of scoring for business proposal writing tasks
表 4. 商务建议类作文打分数据分析

	无 prompt		评分标准		评分标准 + 人工打分	
	DeepSeek	豆包	DeepSeek	豆包	DeepSeek	豆包
QWK	0.69	0.55	0.59	0.58	0.68	0.61
Spearman’ s Rank Correlation	0.77	0.74	0.79	0.78	0.77	0.85

这一差异应该是源于三类商务作文的文本特征。商务报告格式规范、评分维度清晰、客观性更强，更贴合大语言模型的理解逻辑，而商务信函注重语用得体性、语境适配性，主观维度占比高，对模型的语言应用能力要求更高，商务建议类作文兼具格式规范与主观论证，成为模型能力的中间检验载体。

综合上述结果，可总结出两款模型在作文评分任务中的核心特征及提示词干预的作用规律，并得出相应的应用与优化启示。首先，DeepSeek 是更适用于作文自动评分任务的模型，其在核心一致性指标 QWK、评分区分度指标 Spearman 系数、评分稳健性指标相邻一致性率上均保持全面领先，尤其在引入人工打分参考后，Spearman 系数跃升至 0.90，展现出极高的作文质量排序能力，虽其 RMSE 随外部提示词干预逐步上升，但整体偏差仍处于可控范围，且原始无 prompt 场景下的数值精准度表现优异，充分说明 DeepSeek

对作文评分的内在逻辑、标准把握及质量区分均具备更强的能力，更适合作为作文自动评分的基础模型。而豆包虽整体表现不及 DeepSeek，但并非全无优势，其在仅提供评分标准时 Spearman 系数与 DeepSeek 持平，在引入人工打分参考后 RMSE 实现反超，说明豆包对外部提示词的敏感度更高，能够从评分标准和人工打分中逐步优化自身的评分能力，具备一定的提升潜力，但其核心一致性和区分度的短板较为明显，需进行针对性优化才能更好地应用于作文自动评分任务。

其次，提示词干预对模型评分效果的影响具有显著的差异性，单纯提供评分标准可能并非有效的优化手段，而评分标准与人工打分的组合干预能够实现模型评分能力的针对性提升。对于两款模型而言，仅提供评分标准均未实现整体评分效能的正向优化，反而导致 DeepSeek 的 QWK 和相邻一致性率下降，豆包的 QWK 也降至最低，说明单纯的文字化评分标准难以被模型有效理解和转化为评分逻辑，甚至会干扰模型原有的评分判断。而评分标准和人工打分的组合干预则展现出积极的干预效果，能够使两款模型的 QWK、Spearman 系数、相邻一致性率均出现不同程度的回升，尤其对 Spearman 系数的提升效果最为显著，说明人工打分参考作为具体的评分示范，能够将抽象的评分标准转化为具象的评分案例，帮助模型更好地理解评分标准的内涵和应用逻辑，从而提升评分的区分度和一致性。这一结果为大语言模型作文自动评分的提示词优化提供了明确方向，即相较于单纯的文字标准，文字标准与案例的组合式提示词更能有效提升模型的评分能力。

最后，两款模型均存在一定的优化空间，且优化方向需结合其自身特征制定。对于 DeepSeek，其核心优势在于评分的一致性、区分度和稳健性，但外部提示词的引入导致其评分数值偏差扩大，优化重点应在于构建适配的提示词融合逻辑，将评分标准和人工打分参考与自身原始的评分逻辑相结合，在保留高一致性和区分度的基础上，优化评分数值的精准度，减少外部干预带来的偏差。同时，可进一步强化其对不同类型、不同主题作文的评分适应性，提升模型的泛化能力。对于豆包，其核心短板在于 QWK 值偏低，评分一致性较差，优化重点应在于强化对评分标准的深度理解，通过海量的评分标准和人工打分案例进行微调，让模型更精准地把握评分标准的核心维度和权重分配，提升评分与真实标准的吻合度。同时，借助其对提示词敏感度高的优势，构建更精细化的案例提示词体系，逐步提升其评分的区分度和稳健性。

此外，从本次实验的整体结果来看，大语言模型在作文自动评分任务中已展现出一定的应用潜力，两款模型在相邻一致性率上均保持较高水平，Spearman 系数在人工打分参考的辅助下均实现显著提升，说明大语言模型能够有效把握作文质量的基本等级，控制严重的评分偏差，具备成为作文自动评分辅助工具的基础条件。但同时，两款模型的 QWK 值均未达到 0.8 以上的高一致性水平，说明其与专业人工评分仍存在一定差距，暂无法完全替代人工评分。在实际应用中，可将大语言模型作为作文自动评分的初评工具，利用其高效、便捷的优势完成大规模作文的初筛和等级划分，再由人工对模型评分的模糊区间、高偏差样本进行复评和修正，形成模型初评与人工复评相结合的协同评分模式，既提升作文评分的效率，又保证评分的准确性和专业性。

5. 结论

本研究实证对比表明，DeepSeek-V3.2 在 BEC 高级写作评分的整体一致性、稳定性及题型适配性方面整体表现更佳，相较于豆包更适合作为商务英语作文评分的基础模型。豆包虽对外部提示敏感，可通过优化提示策略提升精度，但核心评分一致性仍有一定差距。

在提示词干预方面，研究结果表明，仅提供评分标准难以有效提升模型评分表现，而“评分标准 + 人工打分范文”的组合策略更具有评分质量优化效果，是提升两款模型评分区分度与稳定性的最优策略。

把不同作文类型分开来看，两款模型在商务报告中评分表现最佳，商务信函因侧重语用得体性成为

共同短板，商务建议类评分居中。

综合来看，两款模型均具备 BEC 高级作文初评能力，可用于大规模作文快速分级，但与专业人工评分存在差距。在实际应用中，更为可行的路径是采用“模型初评 + 人工复核”的人机协同模式，在提升评分效率的同时保证评分质量。

本研究填补了国产大语言模型在商务英语专业评分领域的空白，为 BEC 写作测评及 AI 辅助评分提供了参考。但仍存在一定局限性，样本以模拟作文为主，来源单一且规模有限，仅开展了整体评分量化分析，未按 BEC 细分维度评估，也未探究模型参数与领域微调的影响，结论泛化性与解释深度有待加强。

未来可扩大真实考场作文样本，开展多维度误差分析，优化提示词与微调策略，构建标准化人机协同评分流程，进一步完善商务英语写作自动化评分体系，为教学与测评提供更落地的方案支撑。

参考文献

[1] 韩童. 新高考背景下教育人工智能在读后续写中应用现状的调查研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨师范大学, 2023.

[2] Kim, Y. (2025) Automated Essay Scoring with GPT-4 for a Local Placement Test: Investigating Prompting Strategies, Intra-Rater Reliability, and Alignment with Human Scores. *TESOL Quarterly*, **59**, S318-S329. https://doi.org/10.1002/tesq.3405

[3] Suhan, M. and Wolf, M.K. (2025) A Comparative Study of the Human, Automated Scoring Model, and GPT-4 Ratings of Young EFL Students' Writing. *Language Testing*, **43**, 66-78. https://doi.org/10.1177/02655322251346860

[4] Lan, G., Li, Y., Yang, J. and He, X. (2025) Investigating a Customized Generative AI Chatbot for Automated Essay Scoring in a Disciplinary Writing Task. *Assessing Writing*, **66**, Article 100959. https://doi.org/10.1016/j.asw.2025.100959

[5] 李颖. iWrite 自动评分与人工评分一致性研究[D]: [博士学位论文]. 北京: 北京外国语大学, 2021.

[6] 刘玉屏, 欧志刚, 武晓琴. 生成式人工智能赋能国际中文教学的效果测评——以教学设计、HSK 模拟试题编写及作文评分为例[J]. 民族教育研究, 2025, 36(1): 156-166.