



The Impact of LLMs Feedback on L2 Writing Development

Yuhua Zhou

Faculty for College English, Zhejiang Yuexiu University of Foreign Languages, Shaoxing, China
Email: 2656275290@qq.com

How to cite this paper: Zhou, Y.H. (2026)
The Impact of LLMs Feedback on L2
Writing Development. *Open Access
Library Journal*, 13: e15998.
<https://doi.org/10.4236/oalib.1115998>

Received: September 7, 2026

Accepted: October 5, 2026

Published: October 8, 2026

Copyright © 2026 by author(s) and Open
Access Library 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

Abstract

This paper empirically investigates the impact of feedback from Large Language Models (LLMs) on the development of writing proficiency among university-level English learners. Employing a quasi-experimental pre-test/post-test design, the study assigned 52 non-English major freshmen to either an experimental group, which received LLM-generated feedback, or a control group, which received traditional teacher feedback, over a 15-week writing intervention. Writing samples were analyzed quantitatively using the core metrics of Complexity, Accuracy, and Fluency (CAF). The results indicate that the experimental group achieved significantly greater gains in writing accuracy compared to the control group. However, the differences in improvement regarding complexity and fluency between the two groups were not statistically significant. The study concludes that LLM feedback demonstrates a distinct advantage in enhancing the accuracy of linguistic forms, though its effectiveness in promoting linguistic complexity and fluency is less clear within the current model of application. These findings provide concrete empirical support for the use of LLMs in Automated Writing Evaluation (AWE) and offer new perspectives on how L2 writing instruction can effectively harness this technology to realize innovative human-computer collaboration.

Subject Areas

Linguistics

Keywords

Large Language Models (LLMs), Writing Feedback, L2 Writing Development

1. 引言

教师书面纠错反馈(written corrective feedback, 下称 WCF)是二语写作教

学的重要环节,有效的 WCF 能够帮助学生提高写作水平和语言质量[1],可以为二语学习者的写作发展提供积极的影响,尤其是写作的准确性[2][3]。然而,Hyland 和 Anan [4]指出与英语为母语的教师相比,英语为非母语的教师更注重局部反馈(local feedback),对文章内容和组织等方面关注较少。对于二语写作教师的挑战是没有足够的时间对每一位学生进行细致的写作反馈。而且,教师因教学任务繁重、学生人数众多等客观因素限制,其书面反馈存在质量不高、针对性不强等问题[5]。近年来,以 Chat GPT 为代表的大语言模型(LLMs)的兴起,为突破上述困境提供了新的可能。基于自然语言生成的大语言模型可以理解并生成类似人类的自然语言,颠覆了以教师为核心的传统写作反馈模式[6]。魏爽等[6]通过对比 ChatGPT 反馈和教师反馈发现,ChatGPT 既能提供内容与结构的整体反馈,又能针对语言层面给与局部反馈,可以弥补教师由于实际教学任务重或学生人数较多等限制。尽管 LLMs 展现出巨大潜力,但其反馈对二语学习者写作能力发展的实际影响,仍需严谨的实证检验。因此,本研究旨在通过对比分析 LLMs 反馈与教师反馈,系统研究 LLMs 反馈对学生写作复杂度、准确度和流畅度(CAF)所产生的具体影响,以为 LLMs 在二语写作教学中的应用提供实证依据。

2. 大语言模型二语写作反馈研究

LLMs 在写作反馈中的应用,与传统的自动写作评估(automated writing evaluation, AWE)系统存在本质的区别。AWE 系统主要依托自然语言处理(Natural Language Processing, NLP)和人工智能技术,虽能模仿人类评分者进行作文评分,但其准确性常受到质疑,且提供的反馈往往冗长且难以理解[7][8]。与 AWE 系统相比,LLMs 是生成式人工智能,通过算法和自然语言处理技术生成类人的语言输出[9],这使得其不仅能总结文本、回答问题,还能提供具有时效性、针对性和有用性的写作反馈,从而有助于提升学生的写作能力[10][11]。

近期研究已经从不同角度验证了这一潜力。在反馈层面,多项研究发现 LLMs 的反馈比传统教师反馈更全面。例如,Dai 等人[12]在研究中使用 ChatGPT 为 102 篇学习者作文提供反馈,并将其反馈与教师的反馈进行了对比分析。研究发现,与教师的反馈相比,ChatGPT 提供的反馈更为详尽且易于理解,但在某些写作要点上的反馈还很难与教师反馈保持一致。魏爽等[6]的研究也指出,Chat GPT 反馈和教师反馈各具特点,但是 ChatGPT 在提供写作反馈方面具有显著优势,特别是在初阶修改(如语法和标点修正)、中阶润色(包括词汇替换、语义通顺和风格调试)以及高阶输出(如梗概撰写)等方面。在评估任务的可行性上,Mizumoto 和 Eguchi [13]在实验研究中,采用雅思写作评分标准,利用 OpenAI 的 text-davinci-003 模型(GPT-3.5 的前身)对 12,100 篇托福作文进行了评分,并与人类评分者的结果进行了比较。研究得出结论,像 Chat GPT 这样的生成型 AI 工具能够根据评分标准评估作文,从而证实了生成型 AI 工具在 AES 中的可行性和有效性。然而,反馈内容的优势是否能直接转化为学习成果的提升,目前尚无定论。Escalante 等人[14]的研究提供了一个更为复杂的视角。他们发现,接受 AI 生成写作反馈的学生与接受人类反馈的学生在学期结束时的学习成果和表现上并无显著差异。此外,同时接受 AI 和

人类写作反馈的学习者对两种反馈方式的偏好并无明显倾向,认为两者各有优势。他们对 AI 反馈的清晰度和具体性给予了积极评价,而对人类反馈则认为其在提供后续问题和更高层次的互动性方面更为出色。这一发现说明 LLMs 反馈的“潜力”与其对学习者的能力发展的“实际效果”之间可能存在差距。

综上,尽管学界普遍看好 LLMs 在写作反馈中的应用前景,但关于其对学生写作能力(如复杂度、准确度、流畅度)的纵向发展究竟能有多大影响,尤其是与传统教师反馈进行严格对比的实证研究尚显不足。正是基于这一研究缺口,本研究旨在系统研究 LLMs 反馈对非英语专业大学生英语写作发展的具体作用。

3. 研究设计

本研究采用准实验设计(Quasi-experimental Design),包含一个实验班和一个控制班,通过前测-后测来考察 LLMs 反馈对学生英语写作能力的影响。

3.1. 数据收集

本研究的数据收集跨越一个学期,共 15 周。研究对象为浙江越秀外国语学院两个非英语专业自然班的 52 名大一新生,其中一个班($n = 26$)为实验班,另一个班($n = 26$)为控制班。研究设计需要的写作样本来自这两个班的大学英语课堂。所有学生每周完成一篇作文(120~150 字),题型为大学英语四级作文。实验之前,实验班的 26 名学生进行了一节课(45 分钟)的训练。研究者向学生介绍了本次实验使用的 LLM 工具(千问 3.7),并提供了统一的提示语模板(详见附录 A),指导学生如何获取反馈。具体的干预流程如下:学生完成每周的作文初稿后,需使用“千问 3.7”和标准提示语获取反馈。“千问 3.7”会针对文章的内容、结构和语言提供综合性修改建议。学生被要求根据反馈至少进行一轮完整修改,并可就具体问题进行最多两次的自由追问。整个过程学生独立完成,教师不进行干预,以确保反馈完全来自 LLMs。控制班的 26 名学生仍采用教师反馈。为确保与实验组的可比性,教师反馈遵循以下规范:1) 反馈范围:与千问反馈类似,同时覆盖内容、组织和语言等局部与整体特征;2) 反馈形式:采用书面形式在学生作业上进行标注和评语;3) 反馈时长:教师在收到作业后 3~4 天内返回反馈;4) 反馈与修改:教师提供一轮反馈,并要求学生根据反馈完成一轮修改后提交。除反馈方式不同外,两组学生接受了完全相同的课堂写作指导和修改时间。

为了评测大语言模型反馈在多大程度上影响学生的写作表现,实验前后分别安排前测(第一周)和后测(第 15 周),两次测试材料均为同等难度的大学英语四级真题作文。评分根据大学英语四级作文评分标准,检查作文的局部特征(local aspects,包括语言的准确性)和整体特征(global aspects,包括内容和组织)。

3.2. 数据分析

本研究共收集 104 篇有效作文文本(52 篇前测,52 篇后测)。数据分析包含整体评分和 CAF 细粒度分析两个层面。

1) 整体评分

邀请两位参加过大学英语四六级改卷的教师作为评分员，依据大学英语四级评分标准对所有作文的整体特征(内容、组织)和局部特征(语言准确性)进行独立打分。采用皮尔逊相关系数检测评分者信度，若信度达标($r > 0.8$)，则取两位评分员的平均分作为每篇作文的最终得分。

2) CAF 细粒度分析

对于学生文本中局部特征，本研究采用 Ellis 和 Barkhuizen [15] 的 CAF 细粒度分析，对文本的复杂度、准确度和流畅度进行测量，来评估学生在其语言使用中的写作发展。首先，将学生的写作文本编码成 T 单元和从句。根据 Hunt [16] 的定义，T 单元是一个独立从句，后面可以跟随附加的或嵌入的从属从句。本研究将把包含可见主语(除非是命令句)和谓语动词的结构识别为从句(无论是独立从句还是从属从句)。为确保 CAF 编码的可靠性，本研究邀请了另一位有经验的研究员作为第二编码员。两位编码员首先共同编码 10% 的样本以统一标准，然后独立对另外 20% 的数据(约 21 篇作文)进行编码。T 单元、从句和错误类型的切分与分类的一致性通过组内相关系数(ICC)进行检验，结果显示信度良好($ICC > 0.90$)。表 1 是 CAF 的具体实施：

表 1. CAF 测量指标

CAF	测量	计算
复杂性 (Complexity)	句法复杂度 词汇复杂度	$\frac{\text{从句数量}}{\text{T 单元数量} + \text{从句类型/词汇总量}}$
准确性 (Accuracy)	无错误从句(EFCs)的 百分比 加权从句比率(WCR)	$\frac{\text{EFC 数量}}{\text{从句数量}}$ 每个从句的严重性等级(具体见表 2)
流畅性 (Fluency)	每篇文本的总字数	每篇文本的总字数

其中，准确性的加权从句比率(WCR)借鉴 Foster 和 Wigglesworth [17] 的错误严重性，对每个从句进行评分。对于加权从句比率(WCR)的计算，当一个从句包含多种不同等级的错误时，评分将依据其中最严重的错误等级来确定(见表 2)。

表 2. 加权从句比率的错误严重性

错误严重性	从句的分值	描述
无错误	1	语法无错误
1 级	0.8	拼写和标点(包括逗号连接)，大写，冠词和介词，代词，词性和主谓一致等小错误
2 级	0.5	动词形式，词序，词汇选择，连词成句，句子片段(缺少动词/主语/宾语)但是不影响理解
3 级	0.1	连词成句，句子片段，严重的词序问题，影响句意理解

如表 2 所示, 根据包含的错误严重性给每一从句评分。没有错误的分句给满分, 即 1 分。包含小错误的分句(如语法、拼写、标点)、词性、限定词和介词为 1 级(Level 1)评分为 0.8。二级错误包括动词词性、词汇顺序、词汇选择和句子结构, 但这些错误并不影响语言的理解性, 评分为 0.5。较严重的三级错误包括句子片段、严重的词序和句子结构问题以至于句子难以理解, 评分为 0.1。写作准确性由所有从句的评分总和除以所有从句的数量而得到(如公式 1)。

公式 1: WCR 计算

$$\text{加权从句比率} = \frac{\text{所有从句的评分总和}}{\text{所有从句的总数量}}$$

数据分析采用 SPSS 27.0 软件, 来分析学生在前测和后测的 CAF 得分。首先, 先用独立样本 t 检验(若数据不符正态分布, 则用 Mann-Whitney U 检测)比较实验组和控制组在前测中的各项得分, 以确保两组在试验前无显著差异。在确认两组起点对等后, 为检测干预效果, 本研究将计算每位学生的增益分数(见表 3, 增益分数 = 后测分数 - 前测分数)。最后, 再次使用独立样本 t 检验实验组和控制组在各项 CAF 指标及总分上的平均增益分数是否存在显著差异, 从而判断 LLMs 反馈是否比教师反馈更有效地促进了学生写作能力的发展。

表 3. 增益分数的计算

	增益分数(Gain Score)
计算	后测分数 - 前测分数
效果	干预

4. 研究结果

4.1. 整体写作分数的评分者信度

为确保作文整体分数的可靠性, 本研究邀请两位参加过大学英语四六级改卷的教师对全部 104 篇作文进行独立评分。通过计算皮尔逊相关系数检验评分者信度, 结果显示两位评分者的一致性极高($r = 0.92, p < 0.01$)。这表明评分结果是客观且可靠的。表 4 中列出的是两位评分者分数的平均值(M)作为每组作文的整体写作分数和标准差(SD)。

表 4. 两组学生前后测的作文成绩

变量	组别	前测(Pre-test)	后测(Post-test)
		M (SD)	M (SD)
整体写作分数(满分 15 分)	实验组(n = 26)	9.85 (1.12)	12.15 (0.98)
	控制组(n = 26)	9.78 (2.20)	10.88 (1.15)

4.2. CAF 测量

学生写作反馈的结果包括复杂性(句法和词法), 准确性(EFCs 和 WCR)以及流畅性。

表 5 列出了实验班和控制班在干预前后(前测和后测)各项测量指标上的平均数(M)和标准差(SD)。

表 5. 两组学生前后测 CAF 指标

变量	组别	前测(Pre-test)	后测(Post-test)
		M (SD)	M (SD)
句法复杂度(从句/T 单元)	实验组(n = 26)	1.41 (0.22)	1.48 (0.25)
	控制组(n = 26)	1.43 (0.24)	1.46 (0.23)
词法复杂度(TTR)	实验组(n = 26)	0.52 (0.08)	0.55 (0.07)
	控制组(n = 26)	0.51 (0.09)	0.53 (0.08)
准确性(EFCs)	实验组(n = 26)	0.45 (0.11)	0.68 (0.10)
	控制组(n = 26)	0.44 (0.13)	0.53 (0.12)
准确性(WCR)	实验组(n = 26)	0.66 (0.10)	0.87 (0.08)
	控制组(n = 26)	0.65 (0.12)	0.74 (0.10)
流畅性(总字数)	实验组(n = 26)	135.4 (10.2)	145.8 (9.5)
	控制组(n = 26)	134.8 (11.5)	142.1 (10.8)

注：TTR = 词汇类型/词汇总量；EFCs = 无错误从句百分比；WCR = 加权从句比率。

4.3. 干预效果分析

为了检验 LLMs 反馈的干预效果，本研究对表 5 中两组学生在各项指标上的增益分数(后测分数 - 前测分数)进行独立样本 t 检验，如表 6。

表 6. 两组学生平局增益分数的独立样本 t 检验结果

变量	组别	平均增益(M)	t 值	p 值(双尾)
句法复杂度(从句/T 单元)	实验组(n = 26)	0.07	0.85	0.401
	控制组(n = 26)	0.03		
词法复杂度(TTR)	实验组(n = 26)	0.03	1.12	0.268
	控制组(n = 26)	0.02		
准确性(EFCs)	实验组(n = 26)	0.23	5.51	0.000*
	控制组(n = 26)	0.09		
准确性(WCR)	实验组(n = 26)	0.21	6.72	0.000*
	控制组(n = 26)	0.09		
流畅性(总字数)	实验组(n = 26)	10.4	1.43	0.159
	控制组(n = 26)	7.3		

注：* $p < 0.05$ 。自由度(df)为 50。

表 6 所列结果清晰地表明，实验组的进步幅度在不同维度上存在差异，尤其在准确性上远超控制组。

1) 实验组在准确性(EFCs)指标上的提升幅度较大(实验组：0.23，控制组：

0.09), 且两组之间的差异达到统计学上的显著水平($t = 5.51, p < 0.001$)。这说明 LLMs 反馈能极其有效地帮助学生识别并修正句子层面的基础语法、拼写、搭配错误, 从而产生出更多完全正确的句子。

2) 实验组在准确性(WCR)指标上的提升幅度也较大(实验组: 0.21, 控制组: 0.09), 两组之前的差异也达到显著水平($t = 6.72, p < 0.001$)。这进一步证实了 LLMs 反馈的优势不仅在于修正小错误, 还能够有效帮助学生避免或修正那些对句子结构和意义影响更大的严重错误。

3) 实验组在句法结构上的提升幅度略高于控制班, 但差距很小(实验组: 0.07, 控制组: 0.03)。独立样本 t 检验也显示, 两组之间的差异未达到显著水平($t = 0.85, p = 0.401$)。这可能表明无论是 LLMs 还是教师反馈, 在本研究的周期内, 都未能显著推动学生使用更复杂的句式。

4) 在词汇多样性方面, 两组的进步幅度同样非常微小且彼此接近(实验组: 0.03, 控制组: 0.02)。独立样本 t 检验结果($t = 1.12, p = 0.268$)显示两组之间的差距未达到显著水平。这说明两种反馈方式在鼓励学生使用更丰富、更多样的词汇方面, 效果都比较有限。

5) 在流畅度方面, 实验组平均多写了 10.4 个词, 略高于控制组(多写了 7.3 个词)。尽管实验组在流畅性上的平均增益看起来比控制组高, 但两组之间的差异未达到统计学上的显著水平($t = 1.43, p = 0.159$)。这可能说明 LLMs 的即时反馈在一定程度上能激发学生的写作信心, 使其更愿意多写一些。

4.4. 研究结果总结

综合以上统计分析, 本研究得出如下研究结果:

- 1) 前测结果显示, 实验组与控制组在英语写作的各项指标上水平相当, 无显著差异。
- 2) 经过为期一学期的干预, 与接受教师反馈的对照组相比, 接受 LLMs 反馈的实验组在整体写作表现和写作准确性(包括无错误从句比例和加权从句比率)上取得了较大的进步。
- 3) 在写作复杂度(包括句法和词法)和写作流畅度方面, 两组学生的进步幅度没有表现出显著差异。

这些结果表明, 本研究中 LLMs 反馈对提升学生的写作准确性和整体分数具有显著优势, 但在促进语言复杂度和流畅度方面的效果尚不明确。

5. 讨论

5.1. 主要研究发现

本研究最重要的发现是 LLMs 反馈对写作准确度的有效促进作用。在为期 15 周的干预后, 接受 LLMs 反馈的实验组在“无错误从句比例(EFCs)”和“加权从句比率(WCR)”两项准确性指标上的进步幅度, 均显著超过接受教师反馈的控制组。这一结果有力地证明了 LLMs 在帮助学生修正语言形式错误方面的有效性。其原因可能有以下两个方面:

- 1) 与教师反馈的延迟(通常需要数天)且次数有限不同, LLMs 能够为学

生的每一次修改提供即时且无限次的反馈。这种“即写即改、即改即评”的高效循环，能够极大地提高纠错效率。

2) LLMs 能够精准定位到每一个拼写、语法、标点等表层错误，并提供具体的修改意见。学生可以在无压力环境中反复试错，而不必担心因犯错而收到教师的负面评价，这有助于鼓励学生更主动地去修改错误，从而内化语言知识。

但是，本研究未发现 LLMs 反馈能有效促进写作复杂度和流畅度的发展。在句法复杂度(从句/T 单元)和词法复杂度(TTR)上，实验组的进步幅度与控制组相比并无显著差异。同样，在流畅度(总字数)上，尽管实验组的平均增幅略高，但是这种差异不具有统计学意义。其原因可能是因为：

1) 学生可能更倾向将 LLMs 定位为一个高效的纠错工具，而不是一个能够启发思路、优化表达的协作伙伴，可能较少主动寻求提升句式结构和词汇多样性的高级帮助。

2) LLMs 的默认建议往往更倾向于规范和清晰，这种通过算法生成的“安全”表达在一定程度上限制了其语言的复杂度。

5.2. 研究的教学启示

本研究表明，LLMs 在提升写作准确性方面具有显著优势，这为构建一种人机协同的写作反馈新生态提供了实证支持。教师可利用 LLMs 处理学生作文中大量的语言形式错误，引导学生进行有效的自我修正。这种模式不仅能极大地提升反馈效率与个性化程度，更关键的是，它可将教师从重复地批改工作中解放出来，使其更专注于指导学生的篇章结构、逻辑论证以及批判性思维等高阶能力。然而，为了实现这一目标，教师的角色必须经历深刻转变：从传统的知识传授者与评估者，向人机交互的引导者与教学设计者演进。鉴于学生自发使用 LLMs 可能仅停留在表层纠错，教师的核心任务可以转向设计精准的教学干预，培养学生的 AI 素养。具体而言，教师需要通过示范与练习，指导学生掌握有效的指令策略，以引导 LLMs 生成旨在提升句法复杂度、词汇多样性与论证深度的深层次反馈，从而最大限度地发挥人机协作在促进学生写作能力全面发展中的作用。

6. 结语

本研究通过准实验设计，对 LLMs 反馈在大学英语写作教学中的应用效果进行了实证考察。研究发现，相较于传统教师反馈，LLMs 反馈能够显著地提升学生写作的准确度，但在促进句法复杂度和写作流畅度方面的效果尚不明确。此发现解释了 LLMs 作为教学辅助工具的巨大潜力与当前应用中的局限性，为人机协同的写作教学模式提供了有价值的参考。

同时，本研究亦存在一些局限。首先，本研究采用准实验设计，将两个自然班级分别分配给不同干预条件，这意味着研究结果可能受到班级固有差异(如课堂氛围、任课教师风格等)的混杂影响，无法完全将干预效果与班级效应剥离开来。因此，研究结论应更严谨地被视为“在这两个特定班级中，LLMs 反馈条件下的学生进步优于教师反馈条件”，其向其他群体的推论需谨慎。其次，研究样本仅限于一所高校的 52 名非英语专业学生，样本规模和单一情

境也限制了研究结论的外部效度。再次，15 周的干预周期对于考察写作这一复杂技能的长期发展可能仍显不足。最后，本研究以定量方法为主，聚焦于反馈的最终效果，未能深入探究学生与 LLMs 反馈进行交互的具体认知过程，这使得对于“为何复杂度未显著提升”等问题的解释仍带有推测性。

为了弥补上述不足并深化该领域的研究，未来的探究可在多个方向上展开。其一，后续研究应扩大样本规模，纳入不同语言水平、不同专业背景的学习者，以增强研究结果的普适性。其二，建议采用纵向研究设计，追踪更长时间内 LLMs 对学生写作能力，特别是复杂度和流畅度等维度发展的持续影响，以揭示学生在处理 LLMs 反馈时的决策与认知机制，从而更深刻地理解 AI 技术辅助下语言学习的内在规律。

基金项目

本课题为浙江越秀外国语学院研究基地 2024 年委托课题“大语言模型反馈对大学英语写作发展的影响”(2024WGYWH03)的部分研究成果。

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] Zhang, L.J. (2018) Appraising the Role of Written Corrective Feedback in EFL Writing. In: Leung, Y.N., *et al.*, Eds., *Reconceptualizing English Language Teaching and Learning in the 21st Century*, Springer, 161-179.
- [2] Bitchener, J. (2012) A Reflection on ‘the Language Learning Potential’ of Written Cf. *Journal of Second Language Writing*, **21**, 348-363.
<https://doi.org/10.1016/j.jslw.2012.09.006>
- [3] Kang, E. and Han, Z. (2015) The Efficacy of Written Corrective Feedback in Improving L2 Written Accuracy: A Meta-Analysis. *The Modern Language Journal*, **99**, 1-18.
<https://doi.org/10.1111/modl.12189>
- [4] Hyland, K. and Anan, E. (2006) Teachers’ Perceptions of Error: The Effects of First Language and Experience. *System*, **34**, 509-519.
<https://doi.org/10.1016/j.system.2006.09.001>
- [5] 张雪梅. 大学英语写作教学现状之调查[J]. 外语界, 2006(5): 28-32.
- [6] 魏爽, 李璐遥. 人工智能辅助二语写作反馈研究[J]. 中国外语, 2023, 20(3): 33-40.
- [7] Grimes, D. and Warschauer, M. (2010) Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *The Journal of Technology, Learning, and Assessment*, **8**, No. 6. <https://files.eric.ed.gov/fulltext/EJ882522.pdf>
- [8] Wilson, J. and Shermis, M.D. (2024) Automated Writing Evaluation. In: MacArthur, C.A., Wilson, J., Graham, S. and Allor, J.V., Eds., *Handbook of Writing Research (3rd Edition)*, The Guilford Press, 257-271.
- [9] Gatt, A. and Krahmer, E. (2018) Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *Journal of Artificial Intelligence Research*, **61**, 65-170. <https://doi.org/10.1613/jair.5477>
- [10] Biber, D., Nekrasova, T. and Horn, B. (2011) The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. *ETS Research Report Series*, **2011**, No. 1. <https://doi.org/10.1002/j.2333-8504.2011.tb02241.x>

- [11] Graham, S., Hebert, M. and Harris, K.R. (2015) Formative Assessment and Writing: A Meta-Analysis. *The Elementary School Journal*, **115**, 523-547. <https://doi.org/10.1086/681947>
- [12] Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y., Gašević, D., *et al.* (2023) Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. 2023 *IEEE International Conference on Advanced Learning Technologies (ICALT)*, Orem, 10-13 July 2023, 323-325. <https://doi.org/10.1109/icalt58122.2023.00100>
- [13] Mizumoto, A. and Eguchi, M. (2023) Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Research Methods in Applied Linguistics*, **2**, Article ID: 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- [14] Escalante, J., Pack, A. and Barrett, A. (2023) AI-Generated Feedback on Writing: Insights into Efficacy and ENL Student Preference. *International Journal of Educational Technology in Higher Education*, **20**, Article No. 57. <https://doi.org/10.1186/s41239-023-00425-2>
- [15] Ellis, R. and Barkhuizen, G. (2005) *Analyzing Learner Language*. Oxford University Press.
- [16] Hunt, K.W. (1970) Syntactic Maturity in Schoolchildren and Adults. *Monographs of the Society for Research in Child Development*, **35**, iii-iv, 1-67. <https://doi.org/10.2307/1165818>
- [17] Foster, P. and Wigglesworth, G. (2016) Capturing Accuracy in Second Language Performance: The Case for a Weighted Clause Ratio. *Annual Review of Applied Linguistics*, **36**, 98-116. <https://doi.org/10.1017/s0267190515000082>

附录1. 中文摘要和关键词

大语言模型反馈对二语写作发展的影响

摘要： 本文通过实证研究，考察了大语言模型(LLMs)反馈对大学英语学习者写作能力发展的影响。研究采用准实验前后测设计，将 52 名非英语专业大一学生分为实验组(接受 LLMs 反馈)与控制组(接受传统教师反馈)，进行了为期 15 周的写作干预。采用复杂度、准确度与流畅度(CAF)的核心指标对干预前后的写作样本进行量化分析。结果表明，实验组在写作准确度方面的进步幅度显著高于控制组；然而，在复杂度与流畅度两个维度上，两组间的进步差异未达到统计学显著水平。本研究的结论揭示，LLMs 反馈在提升学生语言形式的准确性方面具有突出优势，但在当前应用模式下促进语言复杂化和产出流畅性的作用尚不明确。这一发现为 LLMs 在自动写作评估(AWE)领域的应用提供了精确的实证支持，并为二语写作教学如何有效利用该技术、实现人机协同创新提供了新的视角。

关键词： 大语言模型(LLMs)，写作反馈，二语写作发展

附录2. 使用“千问3.7”自主修改作文的提示语模板

“请根据大学英语四级作文的评分标准，对下面作文从内容、语法和结构三个方面进行综合评分，并指出应该如何修改。”