大语言模型支持下高校英语写作形成性评价机制研究——基于教师反馈与多模型反馈的比较

任晨扬，魏宏*
温州大学教育学院，浙江温州
DOI:10.62836/jer.v4n5.1099

摘要：在高校英语写作教学中，形成性评价长期面临反馈成本高、反馈频率低与个性化支持不足等现实困境。为探讨大语言模型支持下写作形成性评价的优化路径，本文以96名高校一年级研究生为对象，在记叙文、应用文、议论文和说明文四类写作任务中，比较教师反馈、GPT-5 Fast、GPT-5 Thinking与GPT-5 Thinking+小样本提示四种方式在反馈特征、评分一致性、反馈实施、作文进步及学习体验等方面的差异。研究发现：优化后的推理模型在具体建议、可操作反馈和评分一致性上最接近教师反馈；四种反馈方式均能促进学生作文修订，但总体呈现“教师反馈≈GPT-5 Thinking+小样本提示＞GPT-5 Thinking＞GPT-5 Fast”的格局；高质量反馈还伴随写作自我效能感提升与写作焦虑下降。研究表明，大语言模型可以成为高校英语写作形成性评价的有效辅助工具，但其教学价值不取决于是否使用人工智能本身，而取决于提示设计、评价标准以及教师主导下的人机协同机制。本文据此提出“教师把关—模型细化—学生实施”的高校英语写作形成性评价优化路径。

关键词：大语言模型；高校英语写作；形成性评价；反馈实施；人机协同

A Study on the Formative Assessment Mechanism of College English Writing Supported by Large Language Models —A Comparison of Teacher Feedback and Multi-Model Feedback

Chenyang Ren, Hong Wei*
Department of Education, Wenzhou University, Wenzhou, Zhejiang

Abstract: Formative assessment in college English writing has long been constrained by high feedback costs, limited feedback frequency, and insufficient personalized support. To explore an optimized path for formative assessment supported by large language models, this study investigated four feedback modes—teacher feedback, GPT-5 Fast, GPT-5 Thinking, and GPT-5 Thinking with few-shot prompting—among 96 first-year graduate students across four writing genres: narration, practical writing, argumentation, and exposition. The analysis focused on feedback features, scoring consistency, feedback uptake, writing improvement, and learning experience. The results showed that the optimized reasoning model with few-shot prompting was the closest to teacher feedback in terms of specific suggestions, actionable feedback, and scoring consistency. All four feedback modes promoted students' writing revision, but the overall pattern was: teacher feedback≈GPT-5 Thinking with few-shot prompting>GPT-5

Thinking>GPT-5 Fast. High-quality feedback was also associated with increased writing self-efficacy and reduced writing anxiety. The findings suggest that large language models can serve as effective auxiliary tools for formative assessment in college English writing; however, their educational value depends not simply on the use of AI itself, but on prompt design, evaluation criteria, and a teacher-led human–AI collaborative mechanism. Based on these findings, the study proposes an optimization path of formative assessment in college English writing characterized by teacher guidance, model-supported refinement, and student implementation.

Keywords: large language models; college English writing; formative assessment; feedback uptake; human-AI collaboration

1 引言

在高等教育数字化转型不断深化的背景下，课程评价正在从以结果判定为主的终结性评价，逐步转向强调过程支持、持续改进与学习促进的形成性评价[1]。相较于单纯以考试分数衡量学习效果，形成性评价更关注学生在学习过程中的问题诊断、反馈获取、策略调整与能力发展，其核心不在于判定学得如何，而在于支持学生学得更好[2]。在这一转型趋势下，如何通过更高频、更精准、更具可操作性的反馈提升课程教学质量，已成为高校教学改革中的重要议题[3]。

英语写作是高校公共英语课程中最能体现形成性评价价值的教学环节之一。一方面，写作能够较为全面地反映学生在词汇、语法、篇章组织、逻辑表达与任务完成等方面的综合语言能力[4]；另一方面，写作又是教师教学负担最重、反馈成本最高的环节。尤其在公共英语课程中，学生数量较多、基础差异明显、教学时间有限，教师往往难以持续提供高频、个性化、细致化的书面反馈，导致写作训练频次不足、反馈滞后、修订指导不充分等问题长期存在[5]。由此可见，写作教学中的核心矛盾并不仅是学生不会写，更在于教师难以及时、系统地支持学生改进写作。

随着生成式人工智能尤其是大语言模型的发展，英语写作评价与反馈获得了新的技术可能[6]。相较于传统自动写作评价系统主要聚焦拼写、语法和表层纠错，大语言模型在文本理解、自然语言生成和解释性反馈方面表现出更强潜力，能够在一定程度上提供更完整、更自然、也更贴近教学语言的评价意见[7]。这意味着，大语言模型有望突破传统自动写作评价“重纠错、轻解释”“重速度、轻深度”的局限，进入更具教学价值的形成性反馈场景。对于高校英语写作教学而言，这不仅是技术工具的更新，更可能带来反馈机制、教学方式与评价流程的重构。

然而，大语言模型进入高校写作教学并不意味着形成性评价问题已经自动解决。首先，模型能否生成反馈，并不等于其能够生成真正具有教学价值的反馈。写作反馈是否有效，不仅取决于反馈数量，更取决于反馈是否具体、是否可操作、是否真正指向学生的后续修订[8]。其次，不同模型能力和不同提示设计可能导致明显不同的反馈质量[9]。如果不区分快速响应模型、推理型模型以及经过示例优化的模型反馈，就很难准确判断大语言模型的反馈在教学中究竟发挥了怎样的作用。再次，在真实课堂情境中，学生并不会机械地接受所有反馈，而是会基于来源信任、任务理解和修改成本对反馈进行选择性采纳[10]。因此，评价大语言模型支持下的形成性评价机制，不能只停留在模型输出了什么，还必须进一步考察学生实施了什么、哪些反馈真正转化为了学习改进。

从现有研究看，围绕大语言模型与英语写作

的研究虽然增长较快，但仍存在三方面不足。其一，已有研究较多关注模型用于作文评分、语法纠错或写作辅助的可行性，较少在同一框架下系统比较教师反馈、快速模型反馈、推理模型反馈及提示优化模型反馈之间的差异[11]。其二，已有研究更重视反馈生成本身，而对反馈如何通过学生实施行为转化为写作改进关注不足[12]。其三，相关研究对学习者情感体验的考察仍然较为薄弱，尤其缺少把反馈特征、学生实施、自我效能感和写作焦虑纳入同一分析框架的实证研究[13]。换言之，当前研究尚未充分回答这样一个更具教育学意义的问题：大语言模型究竟能否作为教师形成性评价的有效辅助工具，以及它通过何种机制促进学习改进。

基于此，本文以高校英语写作教学为具体场景，围绕教师反馈、GPT-5 Fast反馈、GPT-5 Thinking反馈与GPT-5 Thinking+小样本提示反馈四种方式展开比较研究，试图回答以下三个层面的问题：第一，教师、基准模型、加入了深度推理的模型、加入了小样本提示的模型这四种评价反馈方式在评分可靠性、反馈具体性和可操作性上有何差异；第二，这些差异是否进一步体现在学生的反馈实施行为和作文修订成效上；第三，不同反馈方式是否会影响学生的写作自我效能感与写作焦虑，并由此作用于学习体验[14]。通过对上述问题的系统考察，本文希望在技术应用与教育机制之间建立更清晰的连接，进而为高校英语写作形成性评价改革提供实证依据。

本文的基本观点是：大语言模型可以成为高校英语写作形成性评价的有效辅助工具，但其教学价值并不取决于是否使用AI，而取决于模型能力、提示设计以及反馈是否具体、可操作并能被学生实施[14]。也就是说，真正值得关注的不是大语言模型能不能给反馈，而是什么样的大语言模型反馈更能促进学习改进。在这一意义上，本文不仅关注技术工具的使用效果，更关注其背后的教学逻辑与机制重构，力图从形成性评价的视角说明：在高校课程教学中，教师主导、模型辅助、学生实施的评价路径，或许比“教师与AI谁替代谁”的讨论更具实践价值[15]。

2 方法

2.1 理论基础

形成性评价理论强调，评价的核心功能不只是对学习结果进行甄别和排序，更重要的是通过持续性的反馈帮助学习者发现问题、调整策略并实现改进。与以分数、等级和一次性测验为代表的终结性评价不同，形成性评价更关注学习过程中的信息回流与持续优化，其本质在于以评促学、以评促教[16]。在这一理论视角下，评价并不是教学结束后的附属环节，而是教学过程本身的重要组成部分。

基于这一理论，本文不把大语言模型简单视为一个作文评分工具，而是将其放在形成性评价机制之中加以考察。本文关注的重点并不是模型是否能够生成评价意见本身，而是其能否作为教师反馈的有效补充，进入“问题识别—反馈生成—学生修订—学习改进”的完整链条，并最终服务于学生写作能力的发展。

如果说形成性评价理论回答的是为什么要重视反馈，那么反馈干预理论进一步回答的是什么样的反馈更可能真正产生作用[17]。该理论指出，反馈并非天然有效，反馈是否促进学习改进，取决于反馈的内容焦点、呈现方式、信息清晰度以及学习者如何加工这些信息。换言之，评价者提供了反馈并不自动等于学习者获得了提升，反馈必须通过学习者的理解、判断与实施，才可能转化为实际表现改善[18]。

因此，本文在比较不同反馈方式时，特别关注具体建议、可操作反馈和反馈实施率这几项指标[19]。原因就在于，反馈干预理论提示我们：影响学习结果的关键，不是反馈说了多少，而是反馈是否真正帮助学生知道下一步该怎么改。也正是在这一理论支持下，本文将反馈实施行为视为连接反馈质量与写作改进的重要中介机制。

在传统写作评价研究中，评价效果常常以分数的变化进行衡量。但从真实教学过程看，写作反馈的影响并不只体现在文本层面，还体现在学习者的

主观体验层面。尤其对于高校英语学习者而言，写作既是一项语言任务，也是一项伴随较高认知负担与情绪压力的学习活动[20]。因此，如果只观察作文成绩变化，而忽视学生的能力感知、焦虑体验和反馈接受意愿，就很难完整理解不同反馈方式为何会产生不同效果。

与此同时，写作焦虑研究也表明，英语写作中的紧张、不安、回避与自我怀疑，会直接影响学生的写作投入与表现。造成写作焦虑的重要原因之一，就是任务的不确定性：学生不知道自己问题在哪里，不知道如何达到评价标准，也不知道是否有能力完成修改。由此可见，反馈在情感层面同样具有重要作用。如果反馈能够有效降低任务不确定性、增强学生控制感，那么它就可能在促进文本改进的同时，缓解写作焦虑[21]。

2.2 分析框架

本文的分析框架展示了大语言模型如何嵌入高校英语写作形成性评价这一问题。

首先，不同反馈方式会在反馈质量层面形成差异。教师反馈、快速模型反馈、推理模型反馈以及小样本提示优化后的模型反馈，在评分可靠性、反馈具体性和可操作性等方面并不相同。这意味着，学生面对的并不是抽象意义上的反馈，而是结构、质量和信度的多种反馈信息[22]。

其次，反馈特征会进一步影响学生的实施行为。若反馈更具体、更明确、更易于理解，学生就更可能真正将其转化为修订行动；若反馈过于笼统、重点不清或与教师标准脱节，学生的采纳程度就可能降低。也就是说，学生实施行为是反馈质量转化为学习结果的关键中间环节[23]。

再次，学生实施行为将直接作用于写作结果。只有当反馈被真正落实到文本修订中，才可能最终体现为作文质量提升。与此同时，高质量反馈和有效实施还可能通过增强能力感与控制感，带来自我效能感提升与写作焦虑下降等更广义的学习体验改善[24]。

在这一框架下，教师反馈与大语言模型反馈的比较，不再只是谁更像老师或谁给的建议更多的问题，而是一个关于教学机制是否更有利于形成性改进的问题[25]。据此，本文构建了如**图1**所示的分析框架。

通过对上述问题的考察，本文希望说明：在高校英语写作教学中，大语言模型的真正价值不在于简单替代教师批改，而在于通过与教师标准和教学目标相结合，成为形成性评价机制中的扩展性支持

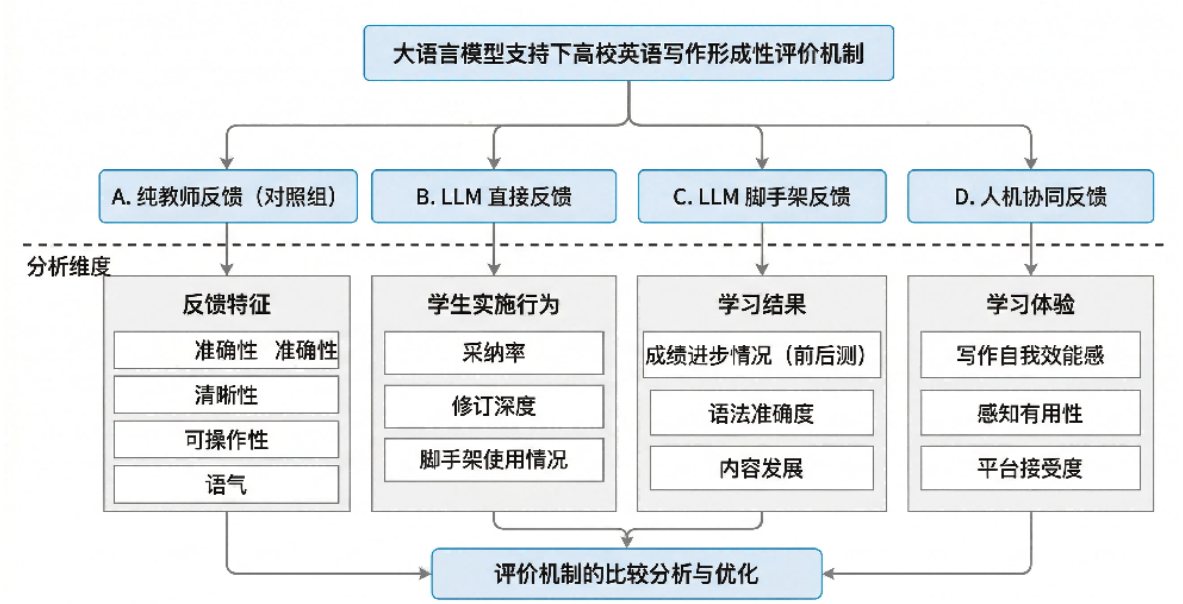

图1. 分析框架图

力量[26]。也就是说，本文试图从教育机制而非单纯技术性能的角度，重新理解大语言模型在课程评价改革中的位置。

2.3 研究设计

本研究以某高校4个英语公共课班级的96名一年级研究生为研究对象，其中男生65人、女生31人。选择高校英语公共课作为研究场景，主要基于两点考虑：其一，公共英语课程中的写作任务具有较强代表性，既覆盖常见英语写作文体，也能较真实地反映高校课堂中形成性评价的实施需求；其二，公共课班级人数较多、学生基础存在差异，更容易体现传统教师反馈在时效性、频率和个性化支持方面的现实压力，因此适合作为考察大语言模型辅助形成性评价的典型场景[27]。

为提高比较的内部效度，本研究在每个班级内部再将学生随机均分为4组，分别接受教师反馈、GPT-5 Fast反馈、GPT-5 Thinking反馈和 GPT-5 Thinking+小样本提示反馈。按最终分组统计，教师反馈组25人，GPT-5 Fast组24人，GPT-5 Thinking组23人，GPT-5 Thinking+小样本提示组24人。这样的分组方式避免了“某一文体完全对应某一种反馈来源”的混淆，使不同反馈条件能够在相近教学情境下进行比较[28]。

本研究围绕高校英语写作教学中的四类常见文体展开，即记叙文、应用文、议论文和说明文。四种文体在表达目标、结构要求和评价重点上存在明显差异，因此能够较好反映不同反馈方式在多样化写作任务中的适用情况。所有写作任务均依托课程教学正常实施，由教师通过线上平台统一发布，学生在规定时间内独立完成并提交作文初稿。

需要说明的是，四种反馈方式既用于比较反馈特征与评分一致性，也用于学生实际接收反馈后的作文修订。为进行评分一致性分析，所有作文初稿均分别接受四种方式的独立评分；但在学生实际修订环节，每位学生仅接收来自其所在组别的一份反馈，以避免多源反馈同时作用于同一篇作文而干扰修订结果。

为较完整地揭示不同反馈方式在形成性评价中的作用机制，本文同时采集文本评分数据、反馈文本数据、修订结果数据和问卷数据。

其次，在反馈分析方面，本文对反馈文本进行内容编码。结合写作反馈研究的常见分类方式，本文将反馈划分为“总结”“询问”“赞扬”“具体建议”“一般建议”五类，并将“具体建议”和“一般建议”进一步归并为“可操作反馈”。其中，“具体建议”主要指能够明确指出问题并提供较清晰修改方向的反馈；“一般建议”则指提出修改方向但操作路径相对概括的反馈。这样的处理方式，有助于从“反馈是否能被落实”这一角度比较不同反馈方式的教学价值[29]。

再次，在修订结果方面，本文关注两个层面的变量：一是反馈实施行为，即学生对可操作反馈的采纳情况；二是作文进步情况，即学生修改前后作文得分的变化。具体而言，学生对可操作反馈的处理被划分为“完全实施”“部分实施”和“拒绝实施”三类，并进一步计算实施率；作文进步则以绝对进步和相对进步率两个指标进行衡量。

最后，在学习体验方面，本文采用两套量表考察不同反馈方式对学生情感体验的影响：一是写作自我效能感量表[30]，用于测量学生对自身写作能力及写作任务完成能力的判断；二是外语写作焦虑量表[31]，用于测量学生在英语写作中的紧张、不安与回避倾向。两套量表均在正式分析前进行了信效度检验，以保证测量结果的可靠性。

2.4 研究实施过程

第二阶段为评分与反馈生成。所有作文初稿分别由教师、GPT-5 Fast、GPT-5 Thinking和 GPT-5 Thinking+小样本提示四种方式独立评分，以形成评分一致性分析所需数据。同时，学生按照分组接收对应来源的一份反馈建议。

第三阶段为学生修订与再评分。学生根据收到的反馈对作文进行修改，形成修改稿。之后，由教师对所有修改稿进行统一评分，以保证后测评分口径一致。研究者同时对学生对可操作反馈的采纳情况进行统计，以形成反馈实施率数据。

第四阶段为问卷测量与综合分析。学生在写

作活动前后分别完成写作自我效能感和写作焦虑量表，以考察不同反馈方式对学习体验的影响。随后，研究者对评分数据、反馈文本数据、修订结果数据和问卷数据进行综合分析。研究实施流程如**图2**所示。

2.5 数据分析方法

第一，在反馈特征比较方面，首先对四种反馈方式的反馈总量、反馈类型和可操作反馈数量进行描述统计；在此基础上，结合数据分布特点采用非参数检验比较四组之间的差异，以判断不同反馈方式在反馈结构上的表现差异。

第二，在评分质量比较方面，采用组内相关系数（ICC）比较三种模型评分与教师评分之间的一致性，并进一步从总分、分项维度和不同文体三个层面展开分析，以判断模型反馈在写作评价中的可靠程度。

第三，在学习结果比较方面，通过比较学生修改前后作文得分，分析不同反馈方式对写作改进的作用；同时结合反馈实施率，考察反馈质量如何通过学生实施行为转化为学习成效。

第四，在学习体验分析方面，对写作自我效能感与写作焦虑量表的前后测结果进行描述统计与组间比较，以考察不同反馈方式对学生主观体验的影响。

本文试图通过“反馈特征—评分一致性—反馈实施—写作改进—学习体验”的层级，解释不同反馈方式在高校英语写作形成性评价中的作用差异。

本文的研究目标并非简单比较“教师与AI谁更优”，而是从形成性评价视角探讨：在高校英语写作教学中，大语言模型是否能够作为教师反馈的有效补充，以及在何种条件下更可能发挥作用。因此，本文既关注技术层面的反馈质量，也关注教学层面的实施行为和学习体验，力图在“技术可行

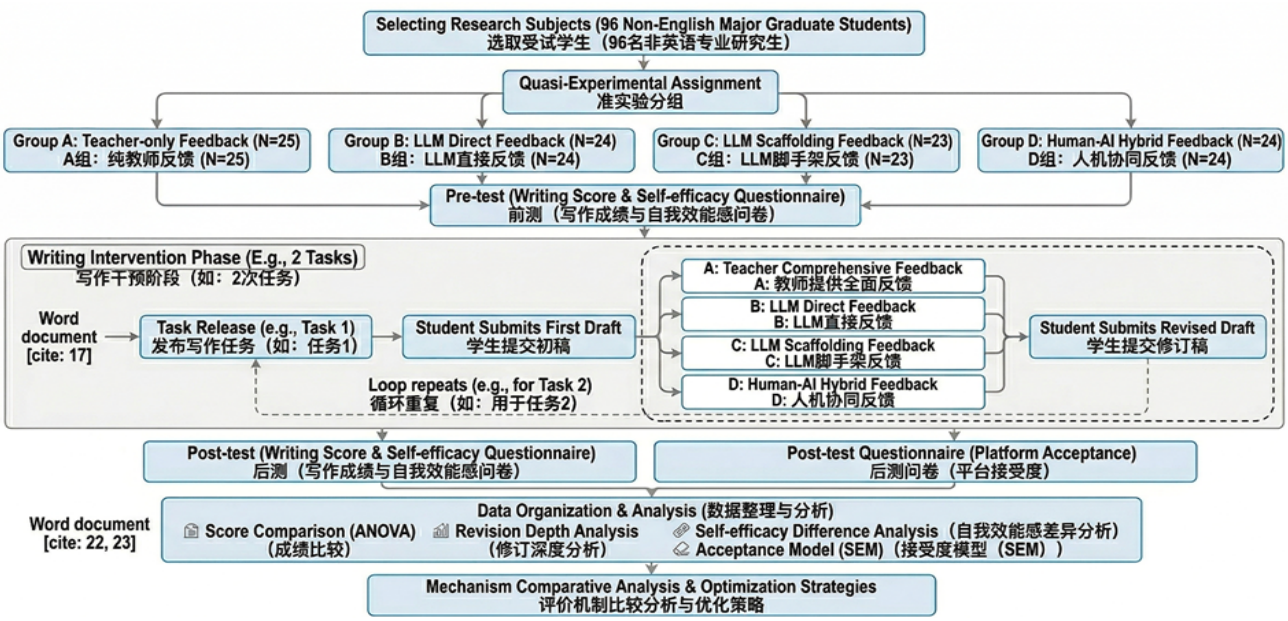

图2.研究实施流程图

表1. 研究对象与研究设计概况

项目	内容
研究对象	某高校一年级研究生96人
课程场景	高校英语公共课
写作任务	记叙文、应用文、议论文、说明文
反馈条件	教师反馈、GPT-5 Fast、GPT-5 Thinking、GPT-5 Thinking+小样本提示
主要数据	评分数据、反馈文本、修订结果、问卷数据
核心指标	评分一致性、反馈特征、实施率、作文进步、自我效能感、写作焦虑

性”与“教育有效性”之间建立更清晰的联系。研究样本、任务分布与分组情况见**表1**。

3 结果

从反馈特征看，四种评价反馈方式在反馈总量、反馈结构以及可操作反馈数量上呈现出明显差异。总体而言，GPT-5 Thinking+小样本提示反馈在反馈总数、具体建议和可操作反馈数量上均表现较为良好，GPT-5 Fast反馈虽然总量较多，但总结性内容占比较大，GPT-5 Thinking反馈则整体介于快速模型与优化模型之间。也就是说，不同反馈方式之间的差异，并不只是反馈多或少的区别，更体现为反馈是否具体、是否便于转化为后续修订行动。

从评分可靠性看，三种模型反馈与教师评分之间也表现出较清晰的层次差异。总体上，GPT-5 Fast的总分均值明显高于教师评分，且一致性相对较低，表现出较明显的偏高评分倾向；GPT-5 Thinking的总分均值已较接近教师评分，一致性显著提升；GPT-5 Thinking+小样本提示的评分一致性最高。进一步从分项维度看，快速模型在若干维度上虽然能够给出较高分数，但与教师评分之间的稳定性不足；推理模型在逻辑水平、段落结构等较高层次指标上的表现更为稳健；经过小样本提示优化后，模型在多个关键维度上进一步贴近教师评分标准。

结果表明，不同反馈方式的首要差异不只是来源差异，而是反馈质量差异。相较于单纯依赖快速生成的反馈，经过提示优化的推理型模型更能够提供接近教师评价逻辑的评分与建议，为其进入高校英语写作形成性评价场景奠定了基础。

四种反馈方式在反馈特征、实施率及评分一致性上的核心表现见**表2**。

从学习结果看，四种反馈方式均在一定程度上促进了学生作文修改后的成绩提升，但提升幅度并不一致。总体样本中，96名学生中有88名在修改后作文得分提高，占比91.67%，表明无论是教师反馈还是模型反馈，都能够在“反馈—修订—再评价”的闭环中发挥一定促进作用。但进一步分析发现，不同反馈方式之间的教学效应存在明显层级差异。

从组内变化看，教师反馈组和GPT-5 Thinking+小样本提示组的绝对进步和相对进步率最为明显，且两组均实现了全员进步；GPT-5 Thinking组整体表现居中，虽优于GPT-5 Fast，但仍未达到前两组水平；GPT-5 Fast组虽然也表现出显著提升，但其进步幅度在四组中最低。换言之，四种反馈方式在促进写作改进方面总体呈现出“教师反馈≈GPT-5 Thinking+小样本提示>GPT-5 Thinking>GPT-5 Fast”的格局。

四组学生作文前后测均值及进步情况见**表3**。

进一步结合反馈实施情况来看，这一结果并非偶然。前述分析已经表明，GPT-5 Thinking+小样本提示在具体建议和可操作反馈上最具优势，而教师反馈虽然反馈总量不占最高，但其建议通常更聚

表2. 四种反馈方式在反馈特征、实施率与评分一致性上的比较

指标	教师反馈	GPT-5 Fast	GPT-5 Thinking	GPT-5 Thinking+小样本提示
反馈总数（M）	18.68	24.38	16.65	21.63
具体建议（M）	7.60	6.63	7.04	11.50
可操作反馈（M）	10.92	12.13	12.09	14.54
实施率	0.76	0.61	0.67	0.71
总分ICC（与教师）	—	0.41	0.68	0.78

表3. 四组学生作文前后测与进步比较

组别	前测均值	后测均值	绝对进步	相对进步率
教师反馈	82.91	94.19	11.28	13.61%
GPT-5 Fast	75.54	81.04	5.50	7.28%
GPT-5 Thinking	72.89	79.86	6.97	9.56%
GPT-5 Thinking+小样本提示	78.63	89.50	10.87	13.82%

焦关键问题。与之相应，这两类反馈在实施率上也处于较高水平。相比之下，GPT-5 Fast 反馈虽然篇幅较大，但由于总结性内容较多、建议不够聚焦，其实施率相对较低。由此可以看出，反馈作用于写作改进并不是直接发生的，而是需要通过学生的理解、采纳与落实才能转化为文本提升。换言之，反馈实施行为是连接反馈质量与学习结果的关键中间环节。

这一发现对理解大语言模型支持下的形成性评价具有重要意义。若仅从技术角度看，模型“能够生成反馈”似乎已足够令人乐观；但从教学效果角度看，真正关键的不是反馈是否生成，而是其是否能被学生实施。也正是在这一点上，教师反馈和GPT-5 Thinking+小样本提示表现出更强的形成性支持价值。

不同反馈方式下学习体验相关指标的变化情况见**表4**。

除作文成绩变化外，本文还考察了不同反馈方式对学生学习体验的影响，重点关注写作自我效能感和写作焦虑两个变量。结果显示，四组学生的写作自我效能感总体上均呈上升趋势，而写作焦虑总体上呈下降趋势。这说明，高质量反馈不仅有助于改善文本本身，也可能通过增强任务确定性和能力感，改善学生对写作任务的主观体验。

在写作自我效能感方面，不同反馈方式的提升重点并不完全一致。总体上，教师反馈在写作任务效能感上的促进作用较为突出，表明教师反馈更容易帮助学生建立完成这类写作任务的信心；GPT-5 Thinking与GPT-5 Thinking+小样本提示则在写作技能效能感上表现出更明显提升，说明高质量模型反馈在帮助学生理解具体问题、掌握修订技巧方面具有一定优势。这一结果提示，不同反馈方式并不只是效果强弱的差异，也可能在影响学生的路径层面存在区别。

在写作焦虑方面，四组学生总体焦虑水平均有所下降。这表明，写作反馈在本研究中并未成为额外负担，反而普遍起到了减轻写作不确定感的作用。值得注意的是，高质量反馈之所以能够缓解焦虑，并不只是因为它指出了问题，更因为它使学生认识到问题是可以被识别和修正的。当学生能够根据反馈完成有效修订，并看到作文表现改善时，其面对写作任务时的控制感和可预测性也会相应增强。

若将学习体验结果与前述作文改进结果结合起来看，可以发现：教师反馈和GPT-5 Thinking+小样本提示不仅在成绩提升上表现更优，也更可能在能力感提升和焦虑缓解方面形成正向支持。这说明，大语言模型支持下的形成性评价若要真正具备教学价值，就不能只追求输出速度和反馈数量，还必须关注其是否能够改善学生的学习体验。对高校英语写作教学而言，能够同时实现文本改进与体验改善的反馈机制，才更具有持续应用价值。

总体而言，本研究结果表明，大语言模型在高校英语写作形成性评价中已经具备一定应用潜力，但这种潜力并不是自动实现的。只有当模型反馈在评分可靠性、具体性和可操作性上接近教师标准，并能够真正被学生实施时，其教学价值才会较充分地显现出来。由此也为后文进一步讨论大语言模型如何赋能形成性评价机制优化提供了实证基础。

总体而言，大语言模型在高校英语写作形成性评价中已经具备一定应用潜力，但这种潜力并不是自动实现的。只有当模型反馈在评分可靠性、具体性和可操作性上接近教师标准，并能够真正被学生实施时，其教学价值才会较充分地显现出来。由此可以进一步看出，反馈特征差异通过学生实施行为

表4. 不同反馈方式下学习体验指标变化比较

指标	教师反馈	GPT-5 Fast	GPT-5 Thinking	GPT-5 Thinking+小样本提示
自我效能感总分差值	0.47	0.40	0.55	0.34
写作技能效能感差值	0.31	0.38	0.66	0.61
写作任务效能感差值	0.71	0.43	0.38	0.47
总体写作焦虑差值	2.02	2.73	2.58	1.13

进一步作用于作文进步和学习体验，构成了本文所强调的形成性评价作用机制。

4 讨论

本文结果首先表明，大语言模型已经能够在高校英语写作形成性评价中发挥一定支持作用。无论是快速模型、推理模型还是经过小样本提示优化后的模型反馈，都在不同程度上促进了学生作文修改后的成绩提升，这意味着生成式人工智能已经不再只是英语写作教学中的技术展示工具，而具备了进入真实课堂评价环节的初步条件[29,30]。

本文最值得强调的发现之一，是反馈作用差异的关键并不主要来自教师还是大模型这一来源区别，而更来自反馈是否具体、是否可操作、是否能够被学生真正实施。研究结果显示，GPT-5 Thinking+小样本提示在具体建议和可操作反馈方面整体上最接近教师反馈，而教师反馈与该组在作文改进效果上也最为接近；相反，GPT-5 Fast虽然反馈总量较高，但其总结性内容更多、评分一致性较弱，最终在实施率和作文进步上也表现较弱。

在本文所有比较条件中，GPT-5 Thinking+小样本提示之所以表现最接近教师反馈，关键不只是因为其使用了推理型模型，更因为其通过小样本提示将教师评价逻辑部分嵌入到了模型输出之中。研究结果表明，单纯依赖模型推理能力虽能改善快速模型反馈质量，但这种改善仍有限；当提示中加入少量高质量教师评分与反馈示例后，模型在评分一致性、具体建议和作文促进效果上都进一步提升。

本文结果还表明，最值得强调的并不是某一种模型优于教师或接近教师，而是高校英语写作形成性评价可以在教师反馈与模型反馈之间形成一种更有效的协同结构。具体而言，教师反馈在任务权威性、重点把握和学生信任方面仍然具有明显优势，而经过优化的模型反馈则在细致性、即时性和高频可用性方面显示出较强潜力。与此同时，学生访谈结果也表明，学习者并不期待模型完全替代教师，而更倾向于接受一种教师把关、模型辅助的评价模式[31]。

第一，英语写作形成性评价改革应从增加写作任务转向优化反馈机制。在现实教学中，学生写作训练不足往往并不只是任务量不够，而是由于缺乏及时且可实施的反馈，导致训练难以真正转化为能力提升。大语言模型的引入，使高频反馈成为可能，但前提是反馈质量必须得到保障。

第二，高校课程中使用生成式人工智能时，应将提示设计视为教学设计的一部分，而不是技术操作细节。谁来设计提示、提示如何嵌入教师标准、如何形成可复用的示例库，这些都应成为未来课程评价改革中的制度性问题，而不只是教师个人的临时尝试。

第三，评价改革不能只关注成绩结果，还应关注学生是否因此更有信心、更少焦虑并更愿意持续投入学习。本文结果显示，较高质量的反馈不仅改善了文本，也改善了学习体验。这说明，形成性评价的真正价值，在于它同时作用于学习结果和学习过程。如果评价改革只追求自动化和效率，而忽略学生的能力感与情绪体验，那么这种改革仍然是不完整的。

总体而言，本文认为，大语言模型支持下的高校英语写作评价改革，不应被理解为一次单纯的技术升级，而应被理解为一次以反馈机制优化为核心的教学改革尝试。其关键在于如何在教师标准、模型能力和学生实施之间建立稳定联结，从而真正实现从反馈生成到学习改进的闭环。

5 结论

本文立足高校英语写作教学中的形成性评价实践，以教师反馈、GPT-5 Fast反馈、GPT-5 Thinking反馈和GPT-5 Thinking+小样本提示反馈四种方式为比较对象，考察了不同反馈方式在反馈特征、评分可靠性、反馈实施、作文改进及学习体验等方面的差异。综合研究结果，可以得出以下三点结论。

第一，大语言模型能够在高校英语写作形成性评价中发挥积极作用，但不同模型反馈之间存在明显质量差异。快速模型虽然能够提供较高频、较完

整的反馈，但在评分一致性、反馈聚焦度和可操作性方面相对不足；推理型模型整体表现更稳；在推理型模型基础上加入小样本提示后，模型反馈在评分可靠性、具体建议和可操作反馈方面进一步接近教师反馈。由此可见，大语言模型进入高校写作评价场景不单是否可用的问题，而是在何种模型能力和提示条件下更有效的问题。

第二，不同反馈方式对学习改进的影响并不主要取决于反馈来源本身，而取决于反馈是否具体、可操作并能够被学生实施。研究表明，四种反馈方式都能在一定程度上促进学生作文修改后成绩提升，但整体上来说，教师反馈≈GPT-5 Thinking+小样本提示>GPT-5 Thinking>GPT-5 Fast。这一差异说明，能够推动学习改进的高质量反馈，并不是反馈篇幅最长的反馈，而是最容易被学生转化为修订行动的反馈。换言之，学生实施行为是连接反馈质量与学习结果的关键中介。

第三，高质量反馈不仅改善文本质量，也会改善学生的学习体验。研究发现，不同反馈方式下学生的写作自我效能感总体呈上升趋势，写作焦虑总体呈下降趋势。其中，教师反馈在增强任务完成信心方面优势较为明显，而高质量模型反馈在促进学生获得“我知道如何修改”的技能感方面同样具有积极作用。这表明，形成性评价的价值并不局限于结果提升，还体现在帮助学生增强能力感、降低不确定性和提升持续投入意愿上。

总体而言，本文的研究结果支持这样一个基本判断：大语言模型可以成为高校英语写作形成性评价的有效辅助工具，但其教学价值建立在模型能力、提示设计和教师把关共同作用的基础之上。

第一，形成性评价改革应把重点放在反馈机制优化上，而不仅是写作任务数量增加。长期以来，高校英语写作教学中的突出问题并不只是学生写得少，更在于反馈供给不足、反馈质量不稳和反馈难以持续。大语言模型的引入，为高频、即时和细致反馈提供了现实可能，但关键仍在于如何把模型输出转化为真正有利于修订的形成性支持。

第二，高校教师在使用大语言模型时，应把提示设计视为教学设计的一部分。本文结果表明，小样本提示对提升模型教育适配性具有重要作用。这意味着，未来高校英语写作教学中更值得建设的，不只是某一种模型平台，而是与课程目标、评分标准和教师经验相结合的提示模板、示例库和评价规则。教师的专业作用并不会因为模型进入课堂而弱化，反而将更多体现在标准设定、提示设计和反馈筛选之中。

第三，更合理的评价改革路径应是教师主导、模型辅助、学生实施的协同机制。在这一机制中，教师仍应承担评价标准制定、关键问题把关和课堂解释的主导职责；模型则承担高频、细化和即时反馈的辅助职责；学生通过实施反馈完成文本修订并获得能力提升。只有当三者形成稳定链条时，大语言模型才会从“工具使用”真正走向“教学赋能”。

需要指出的是，本文仍存在一定边界。首先，研究样本主要来自同一所高校的英语公共课班级，研究结论在不同高校、不同专业和不同英语水平学生中的适用性仍需进一步验证。其次，本文虽已从反馈特征、实施行为、学习结果和学习体验四个层面展开分析，但对更长期的写作发展效果尚未持续追踪。再次，本文聚焦于英语写作这一具体课程场景，未来仍需进一步讨论这一机制能否迁移至其他课程中的形成性评价实践。

后续研究可从三个方向继续深化。其一，扩大样本范围，比较不同院校和不同学习者群体中人机协同反馈机制的稳定性。其二，开展更长周期研究，考察模型辅助反馈对学生长期写作发展和自主修订能力的影响。其三，把研究视野从英语写作拓展到其他课程任务，进一步验证大语言模型在高校形成性评价改革中的普遍适用性。

从更广的教育改革视角看，大语言模型进入高校课堂，并不意味着教师角色被削弱，而意味着课程评价方式正在发生深刻调整。对高校英语写作教学而言，真正值得重视的，不是教师与AI谁替代谁，而是如何把教师专业判断、模型反馈能力和学生实施行为整合进一个更有效的形成性评价机制之中。本文的研究表明，这种机制不仅具有现实可行性，而且在促进学习改进和改善学

习体验方面具有明显潜力。由此，生成式人工智能不应仅被视为教学技术的新工具，更应被理解为推动高校课程评价改革和教学支持方式重构的重要契机。

参考文献

[1] 饶劲松,李薇,李珩.高校数字化转型策略研究[J].中国大学教学,2025,3:52-60.

[2] Black P, Wiliam D. Assessment and classroom learning[J]. Assessment in Education: principles, policy & practice, 1998, 5(1): 7-74.

[3] Hattie J, Timperley H. The power of feedback[J]. Review of educational research, 2007, 77(1): 81-112.

[4] 王海啸.生成式人工智能在大学英语教学改革中的应用探究——以“通用学术英语写作”课程教学改革实践为例[J].外语教育研究前沿,2024,7(04):41-50+95.

[5] Fu Q K, Zou D, Xie H, et al. A review of AWE feedback: Types, learning outcomes, and implications[J]. Computer Assisted Language Learning, 2024, 37(1-2): 179-221.

[6] Dai W, Lin J, Jin H, et al. Can large language models provide feedback to students? A case study on ChatGPT[C]//2023 IEEE international conference on advanced learning technologies (ICALT). IEEE, 2023: 323-325.

[7] Ding L, Zou D, Kohnke L. ChatGPT as an automated writing evaluation tool: how students perceive it and how it affects their writing[J]. Education and Information Technologies, 2025: 1-23.

[8] Shute V J. Focus on formative feedback[J]. Review of educational research, 2008, 78(1): 153-189.

[9] AlGhamdi E, Li Y, Gašević D, et al. Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education[J]. Computers & Education, 2025: 105511.

[10] Yan D, Zhang S. L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1-14.

[11] Zhang Y, Liu Y. Designing ChatGPT-mediated feedback activities in EFL writing: a design-based study of the dialogic feedback triangle[J]. Assessment & Evaluation in Higher Education, 2026, 51(1): 137-160.

[12] Wu J, Li J, Ge Z, et al. Effectiveness of generative AI in automated written corrective feedback with prompting[J]. Journal of Educational Computing Research, 2025, 63(6): 1493-1527.

[13] 滕琳,杨玉鑫,杨静.人工智能写作反馈模式下自我调节能力对多维反馈投入的影响研究[J].外语教育研究前沿,2025,8(03):85-96.

[14] 孙培健,许嘉宇,张军.生成式人工智能反馈对大学生英文作文质量的增值效益研究[J].外语教育研究前沿,2025,8(04):24-36.

[15] 蒋贵友,殷文轩.变革抑或危机：大语言模型赋能大学教学及其限度——基于斯坦福大学的案例考察[J].电化教育研究,2025,46(01):122-128.

[16] Nicol D J, Macfarlane‐Dick D. Formative assessment and self‐regulated learning: A model and seven principles of good feedback practice[J]. Studies in higher education, 2006, 31(2): 199-218.

[17] Kluger A N, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory[J]. Psychological bulletin, 1996, 119(2): 254.

[18] Carless D, Boud D. The development of student feedback literacy: Enabling uptake of feedback[J]. Assessment & evaluation in higher education, 2018, 43(8): 1315-1325.

[19] Lu Q, Yao Y, Xiao L, et al. Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing?[J]. Assessment & Evaluation in Higher Education, 2024, 49(5): 616-633.

[20] Teimouri Y, Goetze J, Plonsky L. Second language anxiety and achievement: A meta-analysis[J]. Studies in second language acquisition, 2019, 41(2): 363-387.

[21] 杨现民,曾佳尧,李新.人工智能与教育深度融合的场景细化及落地实践[J].开放教育研究,2025,31(1):82-92.

[22] 李焕宏,薛澜.生成式人工智能应用的使能型风险规制——以高等教育应用为例[J].清华大学教育研究,2025,46(01):68-78.DOI:10.14138/j.1001-4519.2025.01.006811.

[23] Zhang Z V, Hyland K. Student engagement with teacher and

automated feedback on L2 writing[J]. Assessing Writing, 2018, 36: 90-102.

[24]Lee S, Choe H, Zou D, et al. Generative AI (GenAI) in the language classroom: A systematic review[J]. Interactive Learning Environments, 2026, 34(1): 335-359.

[25]Zhang Z, Aubrey S, Huang X, et al. The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study[J]. Innovation in Language Learning and Teaching, 2025: 1-19.

[26]宋宇,焦丽珍,林如.创新人才培养导向下的课堂教学智能评价研究[J].全球教育展望,2025,54(03):119-134.

[27]罗杨洋,周国辉,韩锡斌.高校数字化转型如何适配有效策略？——基于技术、组织、环境协同的视角[J].现代教育技术,2025,35(06):46-55.

[28]张卓,陈晨.高等教育数字治理的“悬浮化”:表征、机理及矫治[J].西南大学学报(社会科学版),2025,51(03):240-252+334-335.

[29]宋宇,许昌良,穆欣欣.生成式人工智能赋能的新型课堂教学评价与优化研究[J].现代教育技术,2024,34(12):27-36.

[30]Cheng Y S. A measure of second language writing anxiety: Scale development and preliminary validation[J]. Journal of second language writing, 2004, 13(4): 313-335.

[31]李航.大学生英语写作自我效能感量表的编制[J].北京第二外国语学院学报,2014,36(12):70-76.