Advances in Education 教育进展, 2026, 16(11), 1092-1100
Published Online November 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.16112478

Hans汉斯

生成式人工智能赋能英语口语教学的实证研究

叶倾宸，刘思睿*，姚佳欣，尹　浩，古　明

中国矿业大学(北京)文法学院，北京

收稿日期：2026年10月7日；录用日期：2026年11月6日；发布日期：2026年11月17日

摘　要

研究以60名英语专业学生为对象，考察生成式人工智能在口语训练中的用法及其对口语产出的影响。实验组(30名)使用豆包AI辅助训练，对照组(30名)采用自主录音回听训练，借助雅思口语Part 2、Part 3任务开展8周共4次训练，从词汇复杂度、句法复杂度、准确度与流利度四个维度分析口语产出，并结合访谈、反思日志与有声思维报告作主题分析。结果显示：实验组词汇复杂度随训练显著上升($F = 13.63, p < 0.001$)，准确度显著下降($F = 91.80$，$p < 0.001$，偏$\eta^2 = 0.760$)，实验组句法复杂度先升后降，两组流利度均无显著改善，实验组在复杂度与准确度之间的取舍更明显。质性分析归纳出学习陪伴者、反馈纠错者、学习示范者、输出与反馈重述闭环、学习监督者五类人机交互模式，前四类已经出现，监督者角色尚未形成，说明AI主要提供练习场合、错误反馈和语言范例，训练仍由学习者推进。研究认为，AI能否改善口语教学，主要取决于训练环节如何组织，而不取决于工具本身。

关键词

生成式人工智能，英语口语教学，人机交互模式，复杂度和准确度权衡

Empowering English Speaking Instruction with Generative AI: An Empirical Study

Qingchen Ye, Sirui Liu*, Jiaxin Yao, Hao Yin, Ming Gu

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

Received: October 7, 2026; accepted: November 6, 2026; published: November 17, 2026

Abstract

This study examines how generative artificial intelligence (GAI) is used in speaking practice and how it affects the quality of learners' oral production. Sixty English majors were assigned to an

*通讯作者。

文章引用: 叶倾宸, 刘思睿, 姚佳欣, 尹浩, 古明. 生成式人工智能赋能英语口语教学的实证研究[J]. 教育进展, 2026, 16(11): 1092-1100. DOI: 10.12677/ae.2026.16112478

experimental group (n = 30, Doubao AI-assisted training) and a control group (n = 30, self-recording and playback training). Over eight weeks and four training sessions based on IELTS Speaking Parts 2 and 3, oral productions were analysed in terms of lexical complexity, syntactic complexity, accuracy and fluency, and interviews, reflective journals and think-aloud reports were analysed thematically. Lexical complexity rose significantly across sessions in the experimental group ($F = 13.63$, $p < 0.001$), while accuracy fell significantly ($F = 91.80$, $p < 0.001$, partial $\eta^2 = 0.760$). Syntactic complexity rose and then declined in the experimental group, and fluency did not improve significantly in either group; the experimental group showed a clearer trade-off between complexity and accuracy. Five patterns of human-AI interaction were identified: companion, corrector, demonstrator, an output-feedback-restatement loop, and supervisor. The first four were observed, whereas the supervisory role did not emerge, indicating that AI mainly provides practice settings, error feedback, and model language, while the training is still driven by learners. The study concludes that whether GAI improves speaking instruction depends mainly on how training is organised rather than on the tool itself.

Keywords

Generative Artificial Intelligence, English Speaking Instruction, Human-AI Interaction Pattern, Complexity-Accuracy Trade-Off

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. 研究背景

英语口语能力是学习者在具体语境中运用英语表达思想、协商意义、实现交际意图的综合能力。研究表明，英语口语能力的内涵已从“口头产出”拓展为“交际互动”视角下的综合能力。郭雄如[1]将其界定为运用英语表达思想、开展口头交际的能力；张琳、金艳[2]进一步从交互观出发，强调交际双方动态建构意义、协商互动的过程同样重要。在测评维度上，吴坚豪等[3]构建了涵盖词汇复杂度、句法复杂度、准确度、流利度的量化框架，通过词汇密度、从句占比、语法正确率、非流利停顿频次等指标实现口语表现的可量化评估；Canale 和 Swain [4]则提醒，完整的口语能力还应包括语篇、社会语言与交际策略能力。

生成式人工智能(GAI)的发展为语言教学带来新的可能。GAI 依托大语言与多模态模型，可自动生成文本、音频、图像等内容并提供个性化服务。卢宇等[5]指出，它突破了传统 AI 只能识别、分类的局限，实现了人机自然交互，为教育带来新机遇。GAI 兴起后，外语学界围绕其教育应用形成三大方向。其一，书面技能应用。潘海英等[6]的综述发现，现有研究重心集中于写作、阅读，多为工具介绍与可行性分析；郑春萍等[7]指出研究还涉及 AI 辅助语法校对、文本润色、翻译及技术伦理挑战；王立群[8]强调需构建适配智能时代的测试伦理体系。其二，口语教学短期效用。吴坚豪等[3]证实 GAI 能营造低焦虑环境、提供即时反馈，缓解口语焦虑、提升练习频次；Qiao 和 Zhao [9]得出类似结论，但干预周期短、未揭示内在机制。其三，人机协同构想。孙璐[10]提出人机协同教学模式，但探索仍集中于写作与翻译，口语尚处起步。就数量而言，潘海英等[6]纳入的 32 项实证研究中，写作类 13 项(占 40.6%)，口语类仅 4 项；郑春萍等[7]对 135 项研究的梳理，也将“听说能力研究仍需加强”列为主要短板。

总体来看，现有研究局限明显：一是研究多集中于写作、阅读等书面技能，口语专项实证偏少；二是以短期干预为主、量化研究居多，缺少纵向长期追踪，质性材料较为匮乏，难以完整揭示学习者使用

AI 过程中口语能力的动态演变路径。基于此，本研究聚焦英语专业大学生，专门测评其口语能力；参照吴坚豪等[3]的框架从词汇复杂度、句法复杂度、准确度与流利度四个维度对口语产出进行量化评估(见表1)。

Table 1. Dimensions for assessing speaking proficiency
表 1. 口语能力评价维度

维度	指标	计算公式	说明
词汇复杂度	词汇密度	实词数/总单词数	实词包括名词、动词、形容词等；比值越高信息承载量越大
句法复杂度	结构复杂度	(从句 + 并列结构 + 特定短语)/总句数	比值越高句法越复杂
准确度	语法正确率	无错误从句数/总从句数	反映目标语言规则的掌握程度
流利度	语速、停顿频率与时长、非流利型停顿与修正频次	语速(音节数/秒)、停顿频率(次/分钟)、平均停顿时长(秒)、>2~3 秒停顿/填充词/重复/自我修正	语速越快、停顿越少越短、非流利现象越少，流利度越高

本研究采用混合研究设计与 8 周干预：量化数据用于检验口语产出的变化，访谈、反思日志与有声思维报告用于解释变化的发生方式，两类数据交叉印证，以增强结论的可信度。

2. 研究设计

2.1. 研究问题

本研究回答以下两个研究问题：

1) 在生成式人工智能的训练下，英语专业学生口语产出的词汇与句法复杂度、准确度与流利度呈现怎样的变化趋势？

2) 学生与 AI 工具之间形成了怎样的人机交互模式，各类模式具备何种特征、对教学有何提示？

2.2. 研究对象

本研究选取 60 名英语专业大二、大三在读学生，实验组男生 8 名、女生 22 名，对照组男生 7 名、女生 23 名，年龄 20 至 22 岁，均具备一定的口语基础(高考英语最低分 119 分)。60 名被试被随机分为实验组 30 人(豆包 AI 平台辅助训练)与对照组 30 人(自主录音回听训练)，两组在年龄、性别比例与英语水平上严格匹配。

2.3. 研究工具与干预方式

本研究以豆包 AI 平台为实验组的主要工具，模拟雅思口语 Part 2 与 Part 3 的对话场景；对照组采用自主录音回听训练，即用手机或录音设备自行录制、再回听自评。研究周期 8 周，两组训练频次与时长一致，均按“提问、作答、反馈”的流程进行：实验组由 AI 扮演虚拟考官，每次训练后由系统生成对话文本并提供结构化反馈；对照组由被试自主选题录制，研究者提供标准答案范例，被试自行回听并自评。第 1 次训练的数据作为基线，第 2 至 3 次为干预过程，第 4 次作为后测，四次训练统一使用雅思口语 Part 2、Part 3 任务范式以控制任务差异。

数据收集包括四类。一是口语录音，实验组为与 AI 互动的语音，对照组为自主录音，每次训练后转写、清洗，由四名研究者独立校对后编码，共形成 240 组数据(两组各 120 组)。二是半结构化访谈，两组各抽取 4 名代表性被试，于第 1 至第 4 次训练后各访谈 1 次，围绕使用方式、词汇与句法变化、准确度

与流利度变化及现存困难展开。三是反思日志与屏幕录制，记录每次训练的使用方式、存在问题与改进计划，并捕捉实时交互行为。四是有声思维报告，要求被试在训练中口述策略选择与决策逻辑[11]，以印证自我报告数据。

2.4. 分析方法

为回答第一个研究问题，本研究对每次训练的口语录音转写、清洗后，依据四个维度计算各指标。口语音频由平台生成文本后经四名研究者独立校对，再依评价框架对句法、错误与非流利现象编码，分歧经讨论统一后计算指标。流利度细化为语速、停顿频率、停顿时长，以及非流利型停顿与修正频次。使用 SPSS 26.0 开展描述性统计、单因素方差分析与重复测量方差分析。

为回答第二个研究问题，本研究对两组访谈录音、反思日志、屏幕录制与有声思维报告进行主题分析，提炼口语训练的行为类别，归纳人机交互模式的类型与特征，并与量化结果交叉印证。屏幕录制用于观察实时操作行为(如是否反复回听、是否查阅资料)，录音用于分析学习者与 AI 或录音设备的互动方式。

3. 研究结果

3.1. 实验组与对照组口语产出变化趋势

本研究对两组 4 次训练的产出进行量化分析，形成 240 组数据(两组各 120 组)。评价维度为词汇复杂度、句法复杂度、准确度与流利度(含语速、停顿频率与时长、非流利型停顿与修正频次)，其中非流利型停顿与修正频次数值越低表示越流利。实验组各维度按训练次数分组的描述统计见表 2。

Table 2. Descriptive statistics for each dimension of the experimental group by training session
表 2. 实验组各维度按训练次数分组的描述统计

次数	词汇复杂度 M	词汇复杂度 SD	句法复杂度 M	句法复杂度 SD	准确度 M	准确度 SD	流利度 M	流利度 SD
第 1 次	0.4144	0.0870	1.7050	0.8514	0.8626	0.0368	10.0000	4.5461
第 2 次	0.4527	0.0374	2.3650	1.5478	0.7429	0.1459	15.0000	9.8995
第 3 次	0.5107	0.0419	1.9175	0.9559	0.5287	0.0468	20.6250	23.5071
第 4 次	0.5008	0.0682	1.8825	0.8154	0.5563	0.0585	18.2500	18.8745

由表 2 可见，实验组词汇复杂度由第 1 次的 0.4144 升至第 3 次的 0.5107，第 4 次略降至 0.5008，总体上升；句法复杂度在第 2 次达到最高(2.3650)，第 3、4 次回落，呈先升后降；准确度由 0.8626 持续降至 0.5287，第 4 次略回升至 0.5563；流利度(数值越高越不流利)由 10.0000 升至 20.6250，第 4 次略降至 18.2500。对照组数据见表 3。

Table 3. Descriptive statistics for each dimension of the control group by training session
表 3. 对照组各维度按训练次数分组的描述统计

次数	词汇复杂度 M	词汇复杂度 SD	句法复杂度 M	句法复杂度 SD	准确度 M	准确度 SD	流利度 M	流利度 SD
第 1 次	0.3982	0.0715	1.5825	0.7236	0.8450	0.0421	11.5000	3.8730
第 2 次	0.4125	0.0583	1.6400	0.8150	0.8250	0.0512	12.2500	4.5000
第 3 次	0.4538	0.0489	1.6975	0.9025	0.7950	0.0485	12.7500	5.1236
第 4 次	0.4721	0.0556	1.7150	0.8562	0.7800	0.0523	13.2500	5.6782

由表 3 可见，对照组词汇复杂度由 0.3982 升至 0.4721，句法复杂度由 1.5825 升至 1.7150，准确度由 0.8450 微降至 0.7800，流利度由 11.5000 变为 13.2500。与实验组相比，对照组各维度变化幅度较小，准确度下降有限，说明自主录音训练的“复杂度与准确度的权衡”效应不如 AI 辅助训练明显。

从流利度的细分指标看，实验组平均语速由第 1 次的 2.85 音节/秒升至第 4 次的 3.12 音节/秒，停顿频率却由每分钟 8.5 次增至 12.3 次，平均停顿时长由 1.8 秒延长至 2.4 秒；对照组语速由 2.78 升至 2.91 音节/秒，停顿频率由 9.2 次微增至 10.1 次/分钟，平均停顿时长稳定在 1.9 至 2.0 秒。可见实验组语速的提升与停顿的增加同时出现，说明学习者在单位时间内试图使用更多、更复杂的语言材料，而在线加工能力尚未同步跟进；对照组的流利度变化平缓。

以“训练次数”为自变量进行单因素方差分析，词汇复杂度($F = 15.60, p < 0.001$)与准确度($F = 106.57, p < 0.001$)四次训练间差异显著，句法复杂度($F = 2.01, p = 0.116$)与流利度($F = 2.47, p = 0.065$)不显著；以训练次数为被试内因素的重复测量方差分析结果一致：词汇复杂度($F = 13.63$，$p < 0.001$，偏 $\eta^2 = 0.320$)与准确度($F = 91.80$，$p < 0.001$，偏 $\eta^2 = 0.760$)显著，句法复杂度($F = 1.75, p = 0.162$)与流利度($F = 2.45, p = 0.069$)不显著。偏 η^2 为效应量，取值 0 至 1，达到 0.14 即属大效应，0.760 说明训练次数对准确度有很大影响。

综上可见，AI 辅助训练带来的变化主要集中在词汇复杂度上，同时带来准确度的明显下降，句法复杂度与流利度没有实质改善，对照组四项指标的变化都较小。两组都出现复杂度上升而准确度下降的迹象，实验组的幅度更大，这说明 AI 让学习者更多尝试复杂表达，但语法监控没有同步跟上。

3.2. 人机交互模式及其特征

为回答第二个研究问题，本研究对两组各 4 名代表性被试(共 8 名)的访谈、反思日志与有声思维报告作主题分析，并结合屏幕录制与录音交叉验证，归纳出五类人机交互模式。

1) 学习陪伴者模式：低焦虑练习环境的主导角色

学生与 AI 工具之间首先形成了学习陪伴者模式，即 AI 作为虚拟口语考官或对话练习伙伴，为学习者提供安全、低焦虑、无评判的口语练习环境，支持多轮次、重复性口语输出。

屏幕录制显示，被试均沿“读取话题、开口作答、等待追问、继续作答”的路径完成多轮对话，单次有效作答轮次逐次增加，开口前的犹豫时长随训练递减。被试者 2 称“豆包对我这种哑巴英语挺友好，至少不会让我觉得丢脸”，被试者 3 称“对着 AI 练习不会紧张”。这一模式的特征是 AI 只搭建环境、不主动引导，低焦虑与可重复是核心优势，也是后续纠错、示范与闭环得以发生的前提。

2) 反馈纠错者模式：错误识别与反馈供给

在此基础上，学生与 AI 之间形成了反馈纠错者模式，即 AI 在口语输出结束后，对学习者的语法、词汇、搭配、发音等错误进行识别与反馈，为学习者自我纠错提供依据。

被试者 2 的做法是“先自己说一遍，再让豆包标错，最后自己重说一遍”；被试者 3 提到 AI“完成用词、语法纠错，提供地道替换表达”，但也指出“仅纠正单句层面错误，识别不到整体逻辑连贯性问题”；被试者 1 在日志中把高频错误整理后“反复看”。有声思维显示，被试的注意焦点集中于时态、搭配等局部错误。可见 AI 是错误识别者而非错误消除者：它能指出错误，纠错记忆与重述练习仍由学习者完成，顽固错误不会因反馈自动消失。

3) 学习示范者模式：语料供给与范例借鉴

学习示范者模式指 AI 输出高级词汇、复杂句式与完整范例，供学习者积累语料。屏幕录制显示，被试收到高阶表达后多将其复制到自建语料文档并作改写，而非直接照读；被试者 2 提到“借鉴豆包高级表达，但是拒绝背诵模板”，被试者 4 称“自己整理成属于自己的表达”。这一模式的特征是 AI 为语料

供给源而非模板提供者：四名被试都明确拒绝背诵模板，却也承认“记住 AI 给的表达，自己口语中无法调取”、“照搬示范句式容易引发复杂句错误增多”，说明语言甄别与提取能力仍在形成之中。

4) 输出、反馈、重述闭环模式：学习者主导的训练闭环

输出、反馈、重述闭环模式指学习者先独立输出，再接收 AI 反馈，然后自主标记错误、重述修正，形成“输出、反馈、重述、积累”的闭环。被试者 2 称“列出关键词完整口述，请 AI 核查时态、搭配，完成后重新复述一遍，积累好用短语”，并在日志中记录“回听录音、标记错误，再次重说”。有声思维显示，被试在重述前多会低声复述错误点作为自我提示。其核心不是 AI 自动教学，而是学习者主动利用反馈进行自我调节，闭环的关键环节由学习者而非 AI 驱动。

5) 潜在学习监督者模式：尚未实现的理想角色

学习监督者模式尚未形成，即 AI 主动督促训练、推送任务、追踪历史错题、提醒复习。四次训练全程均未出现 AI 推送任务、弹出练习提醒或显示错题追踪面板的画面。被试者 2 称“练习频次、训练时长、复习计划、错题回顾全部由学习者自我规划”，被试者 4 的日志中也“不存在 AI 主动提醒练习、自动追踪历史错题的记录”。监督功能完全由学习者自我承担，AI 目前只是被动工具而非主动管理者；而受访者普遍期待 AI 能定制“个性化学习路径”，说明这一角色存在真实需求。

4. 讨论

4.1. 口语产出的变化：“复杂度与准确度”权衡现象

本研究发现，随训练次数增加，实验组词汇复杂度显著上升($F = 13.63$, $p < 0.001$)，对照组亦有所上升但幅度较小；实验组句法复杂度先升后降，两组流利度均未显著改善，但两组在准确度变化上呈现明显差异：实验组准确度显著下降，而对照组下降幅度较小。这可从“复杂度与准确度的权衡”角度解释：实验组被试者在 AI 辅助下更倾向于尝试更复杂的表达，但语法监控能力未能同步跟上，错误率随之上升；对照组被试者因缺乏即时反馈激励，表达策略相对保守，复杂度提升有限，准确度下降也较少。因此，准确度下降未必意味着口语能力整体倒退，更可能反映学习者在“复杂度与准确度”之间的阶段性取舍。

这一结果与吴坚豪等[3]并不完全一致。该研究以 100 名非英语专业学生开展 16 周教学实践，发现实验组在词汇复杂度、句法复杂度与流利度上均显著优于对照组，准确度未见显著差异。这一分歧可能与干预强度、任务量和测量方式有关。从认知加工的角度看，准确度下降并不意外。工作记忆容量对复杂度、准确度和流利度同时构成限制[12] [13]，学习者想在一句话里放进更多内容时，留给语法监控的注意力就会减少；Skehan [14]的权衡假说与 Mehnert [15]关于准备时间的研究都表明，三者共享有限的注意力资源，此长彼消是常见现象。本研究的数据也支持这一解释：实验组准确度在第 3 次降到最低(0.5287)，第 4 次回到 0.5563，同期词汇复杂度由 0.5107 微降至 0.5008，两项指标在第 4 次同时趋于平稳。问题在于，第 4 次只回升到 0.5563，仍明显低于 0.8626 的起点。如果训练中不给准确度留出修补的环节，这种取舍就可能固定下来。Swain [16]指出，注意到问题是语言学习的第一步，之后还要经过假设检验和反思才能修正；AI 指出的错误，如果没有重述和再产出，只完成了第一步。本研究的访谈正说明了这一点：被试者 3 提到 AI“仅纠正单句层面错误”，被试者 1 将错题整理放在训练之外自己完成。

句法复杂度在第 2 次达到最高(2.3650)，之后回落到 1.9 左右。访谈中的说法比较一致：被试在最初几次愿意尝试长句，发现自己说到一半容易卡住，此后又回到较短的说法。口语的句法复杂度更多体现在从句和限定成分的嵌套上[17]，而 AI 给出的范例多是书面长句，口头复现的难度较大，这可能是句法复杂度没有继续上升的原因之一。流利度没有改善，则与练习方式有关。AI 作为考官几乎不打断，也不要求澄清，学习者很少遇到需要当场修补的交际困难；而按照互动假说，流利度的发展在很大程度上来

自意义协商中的调整[18]。本研究中实验组停顿频率由每分钟 8.5 次升到 12.3 次，也说明在没有时间压力的情况下，学习者将更多注意力放在了选词和组织内容上，说得慢一些，内容反而更完整。

4.2. 人机交互的角色结构与教学主体性

五类交互模式显示，AI 在口语训练中承担的是三项功能：提供练习场合(陪伴者)、指出错误(纠错者)、给出范例(示范者)。训练闭环由学习者自己推动，监督者角色没有出现。这与潘海英等[6]借助生态系统理论[19]描述的微系统情况相符：AI 介入之后，教师不再是唯一的知识来源，低焦虑和即时反馈提高了参与度，学习者对技术的依赖也随之增加。本研究看到的具体表现是：学习者得到了大量练习机会，却没有相应的练习安排。练几次、什么时候复盘、错题怎么处理，全由自己决定，自觉的学生受益，不自觉的学生只是把“不敢开口”换成了“练得不多”。

这与郑春萍等[7]的判断一致：现有系统在反馈质量和策略指导上仍有不足，能判断对错，但不会安排学习过程。Kohnke [20]也主张将聊天机器人当作教学的补充工具；Huang 等[21]的综述也发现，聊天机器人的效果取决于任务设计和教师的中介。本研究为此提供了口语方面的证据：AI 可以让学生开口，但难以让学生说得更好。后者需要目标、顺序、反馈标准和评价办法，这些属于教学设计的工作。就此而言，AI 改变的是练习的组织方式，而不是教学本身的分工。

监督者角色之所以缺位，一方面与平台功能有关：豆包是通用对话模型，没有内置的练习计划、错题本和复习提醒；另一方面与研究安排有关，本研究只要求完成四次训练，没有规定间隔和复盘方式。即便换一个带任务推送功能的平台，如果课程不规定练几次、什么时候练，学生仍可能将它当成随手可用的问答工具。

4.3. 口语教学方法改进的三条路径

基于数据，本研究提出三条教学方法改进的路径。

第一，重排任务顺序，将话题清单变成难度阶梯。本研究的任务按话题平行铺开，难度没有编排，学习者在 Part 2 和 Part 3 之间来回切换，注意力难以积累。可行的做法是沿两个维度排序：认知维度(此处与彼处、现在与过去、具体与抽象、陈述与论证)和语用功能维度(描述、比较、解释、评价、说服)，每次训练只聚焦一个功能，稳定之后再进入下一层。文秋芳[22]的产出导向法提供了可用的框架：以“驱动”呈现真实任务、制造产出缺口，以“促成”提供语言和内容支架，以“评价”检验产出是否达标。例如，前两次只要求描述熟悉的地方，第三次则要求就同一话题作比较和评价，追问控制在两轮以内，避免难度一次抬高。AI 可以在“促成”环节生成范例和即时反馈，但目标和达标标准要由教师定[23]。

第二，改变反馈方式，从改错走向分层反馈加复述。访谈显示，学习者从 AI 那里得到的大多是单句层面的语法和搭配修改，这类反馈对准确度的帮助有限。可以将反馈分成四层并逐层落实：语音层(重音、停顿、连读)、语块层(可以迁移的替换表达)、语用层(得体性与语域)、篇章层(衔接与逻辑)；同时，在流程上固定一个重述环节，同一话题改完再说一遍，两次录音作对比。以一次 Part 2 作答为例，先由 AI 标出时态错误，学习者将这些句子重说一遍并录音，第二次录音中同类错误应当明显减少。这一设计对应 Swain [16]所说的假设检验，也回应了郑春萍等[7]关于反馈策略指导不足的判断。焦建利与陈婷[24]列举的大语言模型教学场景中，“提供反馈”和“模拟对话”最容易落地，但都需要教师事先规定反馈的层级和格式，否则 AI 会回到“润色句子”这一最省力的做法。

第三，分阶段撤除支持，避免练得越多越离不开。潘海英等[6]主张 AI 的使用应遵循“规范使用、提高任务真实性、逐步撤除支架”的顺序。本研究第 4 次训练的词汇复杂度趋于平稳、准确度只小幅回升，部分原因可能是学习者一直在有 AI 保底的条件下练习，缺少独立产出的检验。可行的安排是将八周分成

两段：前四周以 AI 示范和即时反馈为主；后四周改为先独立产出，再用 AI 核对，最后的录音任务完全脱离 AI，只在同伴互评和教师抽评中取得反馈。Lin 与 Mubarok [25]发现可视化反馈能提高元认知调控，这提示脱离 AI 的阶段仍可以保留数据回看，只是将解释权交回学习者。

综合来看，AI 工具在口语训练中起多大作用，取决于任务、反馈和练习安排是否到位。如果只是将话题交给 AI、让学生自由对话，词汇量可能上去，但准确度和流利度未必跟得上，本研究的结果已经说明了这一点。

5. 结语

本研究以 60 名英语专业学生 8 周 4 次人机交互训练为样本，考察了生成式人工智能对口语产出的影响及其发生方式。结果表明，AI 能够推动词汇复杂度上升，但不足以同时带动准确度和流利度改善；它在训练中扮演的是陪伴者、纠错者和示范者，而不是监督者，也不是教学的主体。这一结果说明，AI 能否改善口语教学，主要取决于任务、反馈和练习安排如何组织，而不取决于工具本身的强弱。就方法而言，可以用任务难度阶梯、分层反馈加复述、分阶段撤除支架，替代“自由练习加 AI 改错”的默认做法。本研究也存在局限：样本来自一所高校，训练只有 4 次，测量没有涉及语用得体性、互动策略和韵律特征，结论的推广需要谨慎。后续研究可以扩大样本、延长训练时间，并在设计中加入教师介入的条件，例如，分层反馈训练、支架撤除方案和班级监督安排，以检验这些做法能否带来额外的学习效果。

基金项目

本文受中国矿业大学(北京)大学生创新训练项目(项目编号：202608016)及中国矿业大学(北京) 2025 年教学改革与研究项目(项目编号：J250806)资助。

参考文献

[1] 郭雄如. 英语口语教学的语言环境和教学法[J]. 韶关大学学报(社会科学版), 1996, 17(1): 108-111.

[2] 张琳, 金艳. 基于交互观的口语互动能力构念界定[J]. 外语学刊, 2016(2): 103-108.

[3] 吴坚豪, 周婉婷, 曹超. 生成式人工智能技术赋能口语教学的实证研究[J]. 中国电化教育, 2024(4): 105-111.

[4] Canale, M. (1980) Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, **1**, 1-47. https://doi.org/10.1093/applin/1.1.1

[5] 卢宇, 余京蕾, 陈鹏鹤, 等. 生成式人工智能的教育应用与展望——以 ChatGPT 系统为例[J]. 中国远程教育, 2023, 43(4): 24-31, 51.

[6] 潘海英, 李佳佳, 马新月. 生成式人工智能在外语教学中的应用与研究: 一项系统性综述[J]. 外语教学与研究, 2026, 58(2): 250-262.

[7] 郑春萍, 于淼, 郭智妍. 人工智能在语言教学中的应用研究: 回顾与展望[J]. 外语教学, 2024, 45(1): 59-68.

[8] 王立群. 我国大规模英语测试的伦理体系建构[J]. 中国考试, 2019(11): 7-13.

[9] Qiao, H. and Zhao, A. (2023) Artificial Intelligence-Based Language Learning: Illuminating the Impact on Speaking Skills and Self-Regulation in Chinese EFL Context. *Frontiers in Psychology*, **14**, Article 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

[10] 孙璐. 生成式人工智能赋能英语专业研究性教学: 人机协同模式探索[J]. 外语电化教学, 2026(1): 10-15, 106.

[11] Ericsson, K.A. and Simon, H.A. (1993) Protocol Analysis: Verbal Reports as Data. 2nd Edition, The MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

[12] 金霞. 工作记忆容量限制对二语学习者口语产出的影响[J]. 外语教学与研究, 2012, 44(4): 523-535, 640.

[13] 韩亚文. 工作记忆容量对中国英语学习者口语准确度、复杂度和流利度的影响[J]. 外语教学, 2015, 36(5): 65-68.

[14] Skehan, P. (2009) Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics*, **30**, 510-532. https://doi.org/10.1093/applin/amp047

[15] Mehnert, U. (1998) The Effects of Different Lengths of Time for Planning on Second Language Performance. *Studies in*

Second Language Acquisition, **20**, 83-108. https://doi.org/10.1017/s0272263198001041

[16] Swain, M. (1995) Three Functions of Output in Second Language Learning. In: Cook, G. and Seidlhofer, A., Eds., *Principle and Practice in Applied Linguistics*, Oxford University Press, 125-144.

[17] Biber, D., Gray, B. and Poonpon, K. (2011) Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly*, **45**, 5-35. https://doi.org/10.5054/tq.2011.244483

[18] Long, M.H. (1996) The Role of the Linguistic Environment in Second Language Acquisition. In: Ritchie, W.C. and Bhatia, T.K., Eds., *Handbook of Second Language Acquisition*, Elsevier, 413-468. https://doi.org/10.1016/b978-012589042-7/50015-3

[19] Bronfenbrenner, U. (2005) Making Human Beings Human: Bioecological Perspectives on Human Development. Sage.

[20] Kohnke, L. (2023) A Pedagogical Chatbot: A Supplemental Language Learning Tool. *RELC Journal*, **54**, 828-838. https://doi.org/10.1177/00336882211067054

[21] Huang, W., Hew, K.F. and Fryer, L.K. (2022) Chatbots for Language Learning—Are They Really Useful? A Systematic Review of Chatbot-supported Language Learning. *Journal of Computer Assisted Learning*, **38**, 237-257. https://doi.org/10.1111/jcal.12610

[22] 文秋芳. 构建“产出导向法”理论体系[J]. 外语教学与研究, 2015, 47(4): 547-558.

[23] 张文娟. AI 赋能产出导向法教材使用的课堂实践研究[J]. 现代外语, 2025, 48(5): 679-689.

[24] 焦建利, 陈婷. 大型语言模型赋能英语教学: 四个场景[J]. 外语电化教学, 2023(2): 12-17, 106.

[25] Lin, C.J. and Mubarok, H. (2021) Learning Analytics for Investigating the Mind Map-Guided AI Chatbot Approach in an EFL Flipped Speaking Classroom. *Educational Technology & Society*, **24**, 16-35.