语言学
2023 年 6 月第 5 卷第 2 期

人工智能工具辅助下英语写作水平及创造力的研究

周清扬　信以恒　邓祥腾　张镕鲲　俞韫烨

东南大学外国语学院，南京

摘　要 | Generative Pre-trained Transformer 3（GPT-3）是人工智能领域中的新产物。本项目基于GPT-3，对人工智能工具与英语写作水平和创造力之间的联系进行一定的探索，开展了一项包含15名被试的实验，被试在两次议论文写作中分别进行人工智能辅助和无辅助写作。实验后，我们对被试开展了问卷调查和访谈，并从语法搭配、文章结构、论点新颖性、论证充分性和语句合理性五个方面对文章评分。实验结果表明，语法搭配一项在辅助后有了显著提高（p=0.0062），其余方面虽均有提高，但未达到显著水平，这表明人工智能工具辅助对英语写作水平和创造力有一定的积极作用，能给未来写作发展研究带来一定的启发。

关键词 | 英语写作；人工智能；GPT-3；人机交互

1　导言

近些年来，随着人工智能的迅猛发展，许多惠及语言学习者的程序和软件

基金项目：中央高校基本科研业务费专项资金资助。
作者简介：周清扬、信以恒、邓祥腾、张镕鲲，东南大学外国语学院本科生；俞韫烨，博士，东南大学外国语学院讲师，研究方向：技术写作、教育心理测量。
文章引用：周清扬，信以恒，邓祥腾，等．人工智能工具辅助下英语写作水平及创造力的研究［J］．语言学，2023，5（2）：96-114.
https://doi.org/10.35534/lin.0502009

如雨后春笋般涌出。通过分析用户的学习数据和习惯，人工智能工具可以个性化推荐学习内容和方法，使学习更加高效和有针对性，其主要类型包括英语语音测评、智能批改和习题推荐、分级阅读、智能学情分析和智能情绪识别等（李春琳，2019）。一些辅助英语学习工具，如语音识别、在线翻译等，也采用了人工智能的技术，能帮助学习者更快捷地获取重要信息，提升阅读速度。

人工智能技术的发展使许多方面受益，其中就包括了英语写作。前人的研究进展表明，人工智能技术在在个人辅助写作、人机对话写作、机器自动写作或者群体共同写作方面，已有较成熟的应用（何高大、罗忠民，2008）。人工智能技术在英语学习和教学中也具有广泛应用的潜力和前景。

英文写作和机器辅助写作的研究主要集中在作文特征提取等方面。在“国内二语写作研究回顾与前景展望”一文中，常畅和常海潮（2020）总结了国内研究者对二语写作的研究。二语写作中书面语发展方面，主要从词语、句法准确度、流利度和复杂度等维度，揭示书面语的变化和写作能力的动态发展。杨丽萍和辛涛（2021）在对 AES（自动评分系统）的综述中，构建了“原始特征—高级特征—‘6+1’模型”的理想 AES 特征体系，这两篇文献为我们作文评分标准的制定提供了思路。

创造力相关理论包含创造力的概念和分类，以及创造力评价等方面。创造力理论研究中，倪传斌（2012）总结了前人对创造力不同表达的分类，并梳理已探明的影响因素。武欣（1997）在其综述中认为，创造力是一种一般的认知加工过程，是由认知、人格、动机及社会等因素相互作用而形成的复杂结果，与认知经验紧密相关。本文也基于这一创造力的定义开展研究。魏耀章和苑冰（2009）研究了创造力与英语隐喻生成能力的关系，分别测试被试的英语能力与创造力。我们参照其思路，同时考察被试在写作时这两方面的表现。谢元花等（2018）研究了口头叙事任务和创造力的关系，将创造力定义为一个人在完成所给任务时提出大量新颖、统计意义上稀少的解决方法的能力，并考察创造力流利度、独创性和精准性三个维度和口头叙事任务的关系。基于此，由于本课题研究的是写作创造性，我们使用统计方法划分作文论点的新颖程度，进而对作文创造性评分。

在创造性思维的形成方面，有学者将创造力划分为 divergent thinking（DT，思

https://doi.org/10.35534/lin.0502009

维发散）和 convergent thinking（CT，思维收敛）两部分（Rafner et al.，2020），Jeon 等学者（2021）在此基础上认为创造力形成分为 extending（DT），constraining（CT）和 blending（DT and CT）三步。此外，他们将创造力辅助工具（CST）的介入分为形成想法、参与实施和最后评价三个阶段。在本研究中，考虑到 GPT-3 主要强项在于依靠提示词理解和生成文本，并非专门的作文评价工具；此外，将生成的文本直接用于作文中，可能对写作水平及创造力的考察造成负面影响，因此，我们将其介入置于形成想法阶段。也有研究将形成想法概况为 6 个阶段：定义、收集、浏览、连接、搭建和呈现，并指出不能忽略灵感对想法生成的影响（Koch et al.，2019）。Chung 等学者（2021）在关于 CST 的文献综述中，总结了以上理论，将 CST 从可应用阶段、实施方法等方面进行了分类。我们在阅读上述文献后，将人工智能辅助的方式缩小至提供阅读文本和提供分论点两种，且由于后者被用于制定评分标准，最终只提供机器生成文本给被试阅读。

这一部分的实证研究主要关注机器在创造力活动中的辅助功能。Clark 等学者（2018）尝试使用人工智能进行短篇小说和标语的写作。类似的工作还有将人工智能用于生成游戏（Kreminski et al.，2019）、辅助作家写作（Osone，Lu and Ochiai，2021）等，但几乎没有辅助学生英语写作的应用。对创造力活动较为常见的评判标准有：结果的实用性和质量、交互程度、实验者对过程和结果的满意度、结果的创新性。其他一些学者采用的标准还有：实验前后分数的变化（李婷，2020）、实验者在创作过程中的乐趣、表达容易度和作品所有感（Kantosalo and Riihiaho，2019）等。这些研究为我们设计评分标准、问卷和访谈提供了思路。

本研究基于人工智能模型 Generative Pre-trained Transformer 3（GPT-3），使用雅思作文范文微调模型，搭建辅助写作平台，开展了一项包含 15 名被试的实证研究，对比有辅助和无辅助状态下的英语写作，以探索人工智能工具与英语写作水平和创造力的关系。被试作文参照雅思考试作文评分标准，从语法搭配、文章结构、论证充分性、语句合理性等多个方面评估和分析写作水平，并采用统计的方法评判观点新颖性。实验后，我们开展了问卷和访谈调查，询问被试对人工智能辅助的态度与写作体验。本研究致力于证明人工智能工具在英语写作中的辅助作用，旨在为未来的英语写作发展提供一定的启示和指导。

2 研究方法

2.1 研究问题

本论文的研究问题是：

（1）人工智能辅助工具能否提高使用者的英语写作水平？

（2）人工智能辅助工具能够提高使用者的创造力？

为此，我们检索了英文写作、创造力和人机交互等领域的理论和实证研究。

2.2 研究被试

参与本研究的 15 名被试均为某双一流大学本科生，2 名被试为非英语专业，其余 13 名为英语专业学生，涵盖了大一到大四的各学习层次，其中女生 4 名，男生 11 名。

2.3 实验工具

本研究使用的人工智能辅助工具是在 GPT-3 基础上搭建的写作平台。GPT-3 为基于 Transformer 的生成式预训练人工智能模型，是一种自回归的语言处理模型，可以完成以生成人类语言为基础的一系列相关任务。GPT-3 含有 1750 亿个可学习参数，训练集包含约 5000 亿个 token（Brown et al.，2020），是当时（2020 年）最大的密集语言模型。

在实验中直接使用 GPT-3 较为困难，一方面是虽然它有多个功能强大的自带模型，但由于其生成对象是多种不同的文本，所以对于特定类型的文本，生成效果不是很好，具体表现为缺少议论文段落结构、表述语言不够正式等。另一方面，GPT-3 官方网站的注册、登录和使用均较为复杂。因此，我们微调了原模型，并建立了实验网站方便被试使用。

微调训练集的材料来自互联网（IELTS Mentor，2023），为雅思写作 task 2 的作文。我们共搜集了 1200 篇，经过人工校对和格式调整，作为训练集。我们参照 GPT-3 官方文档（OpenAI API，2023），基于 babbage 模型微调，其余参数均选用默认参数。

实验网站（见图 1）内嵌了 GPT-3 API 的部分代码，主要包括两个功能：“生成新文本”（根据提示词生成一篇文章）和“头脑风暴”（根据提示词生成相

https://doi.org/10.35534/lin.0502009

关论点）。在实验中，仅开放了前者给被试，被试在生成文本后可在“历史文本”中查看。“头脑风暴”功能主要用于给文章评分。

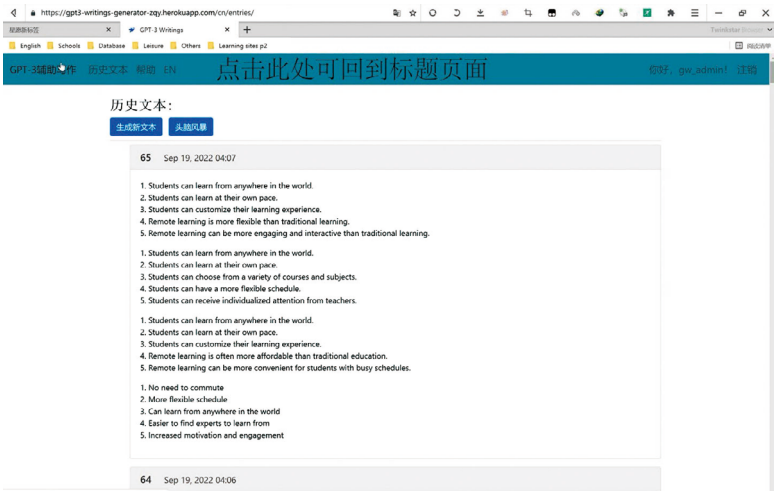

图 1　实验网站截图

Figure 1　A screenshot of the interface of the AI-assisted writing platform

文本生成的设定为：temperature = 0.5; max_tokens = 450; presence_penalty = 1; frequency_penalty = 1; prompt = “Write an argumentative essay about xxx. Word limit: 350–450.”（其中 xxx 部分可替换为作文主题，如 the advantages of remote learning 等，下同）；头脑风暴的设定是：temperature = 0.7; max_tokens = 200; presence_penalty = 1; frequency_penalty = 1; top_p = 1; prompt = “List five advantages of xxx:\n1.”。

2.4　实验设计

我们随机将被试分为 A、B 两组，其中 A 组 7 人，B 组 8 人。每次实验分为实验组与对照组，实验组使用实验网站辅助写作，对照组不使用，两组写作题目相同。实验于同一教室同时进行，被试均使用教室电脑写作。期间，实验组除实验网站外不可查阅其他资料、词典等，而对照组则仅依靠自身能力写作。实验分两次进行，中间间隔一周。我们提前告知被试实验流程，被试需事先阅读及确认知情同意书、熟悉实验网站操作，但无需做其他准备。

正式写作时长为 60 min，作文不得少于 250 字，并只能选择一个立场阐述观点。被试完成作文并填写问卷后即可离开。

第一次实验，A 组为对照组，B 组为实验组，实验后 B 组填写问卷 1。间隔一周后，A 组为实验组，B 组为对照组，实验后 A 组填写问卷 1，且 A 组和 B 组均填写问卷 2。确认填写完毕后，发放实验报酬。

两次的作文题目分别为：“Good teaching is more important for academic success than individual ability. To what extent do you agree or disagree?”以及“Some people believe that educational qualifications will always bring success in life. Other people say that educational qualifications do not necessarily bring success. To what extent do you agree or disagree?”

2.5 评分标准与评分过程

实验作文评分标准主要参考雅思作文（IELTS，2022）的评分标准，分为五个评分项，分别为（具体评分表如表 1 所示）：

表 1 评分标准表

Table 1 The writing rubric

语法搭配	取 Grammarly 得分（自动生成），该项不用评分手册人工评分				
人工评分各项得分	5	4	3	2	1
文章结构	文章有完整的标准考试议论文结构（明确分开的开头、主体、结尾），能够进行充分合理的分段；首段有明确指出作者的立场，结尾有对上文的总结，主体段部分每一段只含一个论点，条理清晰，衔接流畅	文章有完整的标准考试议论文结构，能够进行充分合理的分段；首段立场不太明确，结尾较为空泛，主体段部分每一段只含一个论点，条理清晰，衔接流畅	文章有完整的标准考试议论文结构；使用段落写作但未能保持段落间逻辑；首段立场不太明确，结尾较为空泛，主体段部分每段可能包含多个论点，条理较为清晰，衔接较为流畅	文章缺少部分标准考试议论文结构，分段不足 / 段落使用造成疑惑；首段立场不明确，结尾空泛，主体段部分每段可能包含多个论点，条理不太清晰，衔接不太流畅	文章缺少较多标准考试议论文结构，未使用段落写作；立场不明确，结论空泛，各论点间条理不清晰，逻辑不通顺，衔接不流畅

https://doi.org/10.35534/lin.0502009

续表

语法搭配	取 Grammarly 得分（自动生成），该项不用评分手册人工评分				
人工评分各项得分	5	4	3	2	1
论点新颖性	各论点和文章主题、立场联系紧密，全文论点数（包括正反方）小于等于 3 个，其中出现在 cliché1 区的个数为 0，出现在 2 区的个数小于等于 1 个，有论点不在 1 区和 2 区中（1 区和 2 区依据出现频次进行划分）	各论点和文章主题、立场联系紧密，全文论点数小于等于 3 个，其中出现在 cliché1 区的个数小于等于 1 个，有论点不在 1 区和 2 区中	个别论点和文章主题、立场联系不太紧密，全文论点数共 4 个，有论点不在 cliché1 区和 2 区中 Or 全文论点数小于等于 3 个，在 cliché1 区的个数大于 1 个，有论点不在 1 区和 2 区中	部分论点和文章主题、立场联系不太紧密，全文论点数大于等于 5 个，有论点不在 cliché1 区和 2 区中 Or 全文论点数小于等于 4 个，所有论点均在 1 区和 2 区中	大部分论点和文章主题、立场联系不太紧密，全文观点数大于等于 5 个，所有观点均在 cliché1 区和 2 区中
论证充分性	每一论点均有合理且详细的数据、例子及其他方式进行充分说明，论证令人信服	部分论点缺少合理且详细的数据、例子及其他方式进行充分说明，论证令人信服	部分论点缺少合理的数据、例子及其他方式进行充分说明，论证不太令人信服	绝大部分论点缺少合理且详细的数据、例子及其他方式进行充分说明，论证不太令人信服	所有论点缺少数据、例子及其他方式进行充分说明，论证不令人信服
语句合理性	使用广泛的词汇，可以自然和熟练的掌握词汇的特征，句型选用合理、不滥用（指长句、短句及一些句型），可以使用一些恰当且新颖的修辞，表达清晰	使用还算广泛的词汇，可以较灵活的使用不常见的词汇表达准确的意思，句型选用较为合理，很少出现滥用现象，可以使用少量恰当的修辞，表达比较清晰	使用足够的词汇，能有意识的使用不常见的词汇，但在用词选择所表达出的意思不太准确切题，句型选用较为合理，对部分句式可能依赖过多，可以使用少量的修辞，表达比较清晰	使用还可以的词汇，能有意识的使用比较不常见的词汇，但表达的意思不准确却并不影响阅读，句型选用有较明显的依赖，基本没有修辞出现，表达不太清晰	只使用可以重复使用的基本词汇，只使用非常有限的结构，句型重复，没有使用修辞，表达不清晰

（1）语法搭配

语法搭配主要由机器评分，我们选用 Grammarly，其设定为：domain—

general; intent—describe & convince; audience—knowledgeable; formality—neutral。

（2）文章结构

这一项主要参考雅思写作中 task response 和 coherence and cohesion 部分内容，考察文章是否符合考试议论文的写法，即开头、主体和结尾段分明，且首段明确表达观点，主体部分分段明确、无观点堆砌，结尾做出总结。此外，也考察文章是否衔接通顺，包括连接词等连接手法的使用情况。

（3）论点新颖性

这一点考察文章的论点是否较为新颖、不俗套陈旧。评判方法是使用“头脑风暴”功能每次生成 5 个论点，一共生成 20 组，统计论点出现频次并排序，分出“cliché 1 区”和“cliché 2 区”（落在这两个区域内的论点会酌情扣分，1 区为出现次数大于等于 12 次，2 区为大于等于 6 次；若论点和主题无关，也不算作新颖论点）。比较罕见但合理的论点，视为新颖性较高。

（4）论证充分性

我们认为论证的充分性和其创新性在一定程度上也是相关的。一些过于普适的“例证”，如名人名言、俗语等，说服力其实较低。但同时，论证的创新性难以衡量，不同评判者会给出不同标准。故论证充分性这一块，我们仅考察其说服力的高低。在具体操作中，主要是看论点后是否有相应的论据支撑，且这些论据必须是合理且具体的，而非简单地对论点进行补充解释。

（5）语句合理性

这一点对应了雅思写作中 lexical resource 和部分 grammatical range and accuracy 的内容，主要考察文章词汇使用的丰富度以及难易度，考察句型选取与修辞使用。目的是测试文章作者对语言的把握能力。

对于文章评分，语法搭配一项单独计分，后四项由人工打分后，相加得到最终得分。人工打分评分员为 2 名高校英语教师，具有丰富的英语写作教学经验和大规模标准化考试评分经验，评分开始前，评分员先阅读了评分标准，保证评分标准的统一；评分过程中两人各自评分、互不影响，保证评分独立性。我们取两位评分员评分的均值作为各维度得分；每个维度取值为 1 ~ 5 分，5 分为最高，1 分为最低。人工评分中，若出现某篇作文某个维度上 2 位评分员分差

https://doi.org/10.35534/lin.0502009

>2 分，则进行商议、复核后重新评分，以保证评分的一致性，保证研究的信度。

2.6 问卷与访谈

本实验问卷在形式上主要参考了 Kantosalo 和 Riihiaho（2019）的方式，即每次写作后询问实验组被试对人工智能辅助方法的感受，使用问卷 1；并在两次写作后询问两组两次不同方法的对比，使用问卷 2。此外，问卷 1 也加入了 9 个聚焦于想象力过程特征和被试个体差别的问题（Wu，Yu and An，2022）。

问卷 1 以选择题为主，除上述问题外，还根据评分标准的 5 个方面，询问人工智能辅助（即实验网站）对写作的影响，采用里克特五度量表。问卷 2 要求被试对比两次写作（即有辅助和无辅助）的感受，做出体验偏好选择，并说明理由。

访谈共采访 11 人，问题参考评分标准制定。

3 研究结果

3.1 被试对人工智能辅助工具的态度、心理特点

对于问卷 1（被试对人工智能辅助方法的感受），我们采用里克特五度量表进行测量和统计，计算平均值和标准差。如表 2 所示，总体来说，被试对人工智能辅助的效果不太认可。15 个问题中，除“AI 辅助使我感到悲观”一项被试不太同意、“我认为 GPT-3 极大地帮助我拓展了思路”和“AI 辅助在考虑单词方面为我提供了很多细节”两项被试无明显偏向外，其余 12 个回答均较负面。被试尤其不赞同人工智能辅助可以提升语法准确性和行文逻辑性，并认为 AI 辅助使自己的想象力不活跃。

表 2 写作者对人工智能辅助工具的使用感受

Table 2 Participants’ reflections on Ai-assistant tools

问题	平均值	标准差
我觉得 GPT-3 极大地帮助我拓展了思路	3.00	1.26
我觉得 GPT-3 极大地帮助了我选取正确的语法结构	2.40	0.88
我觉得 GPT-3 极大地帮助我拓展词汇使用量	2.73	1.53
我认为 GPT-3 在文章前后文衔接上逻辑通顺	2.40	1.25

续表

问题	平均值	标准差
我认为 GPT-3 提供的论证充分支持对应论点	2.47	1.26
我认为 GPT-3 会经常蹦出一些新颖的观点	2.87	1.15
AI 辅助为我的想象力提供了丰富细节	2.93	1.48
AI 辅助在考虑单词方面为我提供了很多提示	3.00	1.46
AI 辅助使我产生负面情绪	3.07	1.39
AI 辅助使我的想象力变得更加丰富且更加复杂	2.80	1.11
AI 辅助使我在脑海中能够更好构思文章	2.67	1.25
AI 辅助使我感到悲观	2.67	1.40
AI 辅助给我提供了更清晰的写作方向和目标	2.87	1.20
AI 辅助使我失去了写作方向和目标	3.43	1.20
AI 辅助使我的想象力变得不活跃	3.53	1.09

如表 3 所示，在每一项中，被试均更倾向于只靠自己写作（无人工智能辅助）。60% 的被试倾向于自主写作。根据实验后的访谈，被试对人工智能辅助工具的这一使用倾向大体有以下几种理由：“GPT-3 生成文章会出现逻辑混乱等问题”“自己的思考和观点是最为重要的”“自主写作思路更清晰，写作节奏更紧凑，精力比较集中”；40% 的被试倾向于 AI 辅助写作，理由包括“比较新奇”“可以节省许多用来思考的时间”；但有超过 80% 的被试认为自主写作的作文更加令人满意。

表 3　写作者对人工智能辅助下的写作过程的体验

Table 3　Participants' preferences between the two approaches

问题	GPT-3	我
您认为哪种方式写作更心潮澎湃一些	26.7%	73.3%
您更倾向于用哪种方式写作呢	40.0%	60.0%
您认为哪种方式最让您的思绪天马行空，创造力源源不断呢	33.3%	66.7%
哪种方式产出的作文最使您满意呢	13.3%	86.7%

访谈反馈的结果，被试对人工智能工具的体验也偏向负面。写作者对人工智能辅助工具的正面反馈主要是认为机器生成的文章可以提供一个可用的、较新颖的框架，之后可以在此基础上补充自己的观点。在负面评价中，一个常被提及的缺点是跑题，共有 5 名被试提到了这一点。跑题大概可以分为两种，一种是完全和作文题目不符，如“它（生成文章）讲的是教育的重要性，但那个

https://doi.org/10.35534/lin.0502009

题目是 education qualification（学历）的重要性”。第二种是论点和题目不匹配，如“它教育那个点第一段说了教育，然后第二段直接就移到一个好老师，然后开始讲一个好老师的重要性”。另一个类似的问题是生成文章的观点和被试个人的想法不合，有 4 名被试提及，他们因为事先已经想好了作文的观点，所以最后没有采用 GPT–3 的论点。

拓展思路和提供论点方面，最大的问题就是前面所述的跑题。有两名被试认为 GPT–3 提供思路是一把双刃剑，“在看了它给的文章之后，就好像再也跳不出那个圈了”，认为如果生成的文章和自己思路相同，可能会拓展自己的思路，但是如果不同，反而会受到负面影响。论据方面，被试普遍没有用到，主要原因也是跑题。还有一部分被试认为论据较难使用或说服力不足。词汇和语法方面，被试多反映生成文章中的词语词组及句型较为简单，虽有时可以起到提醒作用，但帮助不大，且语法错误较多。文章过渡衔接方面，被试对句间连接词的使用评价较高，但认为段间逻辑不是很紧密，缺乏过渡。

3.2 人工智能辅助工具对写作质量的影响

之后，我们使用 R 软件，采取单边和双边配对 t 检验对评分结果进行数据分析，α =0.05，统计了两组被试在两次写作中的表现。各项评分结果如表 4 所示。

表 4 评分结果统计表

Table 4 The grading result of participants' writing

维度		语法搭配		文章结构		论点新颖性		论证充分性		语句合理性		人工评分总分	
		Avg.	S.D.	Avg.	S.D.	Avg.	S.D.	Avg.	S.D.	Avg.	S.D.	Avg.	S.D.
A 组	第一次	52.29	18.04	3.79	0.59	3.43	0.56	3.79	0.59	3.64	0.52	14.64	1.94
	第二次	69.57	14.53	4.21	0.36	3.93	0.62	3.71	0.70	4.07	0.56	15.93	1.97
	总计	60.93	18.52	4.00	0.53	3.68	0.64	3.75	0.65	3.86	0.58	15.29	2.06
B 组	第一次	62.13	8.19	3.50	0.43	3.44	0.46	3.44	0.46	3.13	0.48	13.50	1.30
	第二次	50.75	15.32	3.69	0.35	3.44	0.39	3.13	0.33	3.19	0.35	13.44	0.85
	总计	56.44	13.54	3.59	0.40	3.44	0.43	3.28	0.43	3.16	0.42	13.47	1.10
总计	第一次	57.53	14.55	3.63	0.53	3.43	0.51	3.60	0.55	3.37	0.56	14.03	1.73
	第二次	59.53	17.66	3.93	0.44	3.67	0.57	3.40	0.61	3.60	0.64	14.60	1.93
	总计	58.53	16.21	3.78	0.51	3.55	0.55	3.50	0.59	3.48	0.61	14.32	1.86
GPT–3 辅助		65.60	12.17	3.83	0.54	3.67	0.60	3.57	0.60	3.57	0.70	14.63	2.05
无辅助		51.47	16.66	3.73	0.48	3.43	0.48	3.43	0.57	3.40	0.49	14.00	1.58

如表 5 所示，两次作文题难度相当。在 Grammarly 评分的语法搭配一项，两组经人工智能辅助后得分均有显著提高（p=0.0062）。人工评分的各项，总体来说，A 组提升明显，而 B 组则稍有提升，但可忽略不计；使用 GPT-3 辅助后，被试写作水平得分高于无辅助写作，但未达到显著水平，如表 6 所示。此外，对人工智能辅助持中立或积极态度的被试，其人工评分总分均高于无辅助时的得分。

表 5　两次作文难度比较表

Table 5　A comparison between the difficulty of the two writing tasks

	第一次平均分	第二次平均分	p 值
语法搭配	57.53	59.53	0.7513
人工评分总分	14.03	14.60	0.2969

表 6　无辅助与有辅助下写作得分比较表

Table 6　A comparison of the marks with and without ai assistance

	无辅助平均分	有辅助平均分	p 值
语法搭配	51.47	65.60	0.0062**
文章结构	3.73	3.83	0.2975
论点新颖性	3.43	3.67	0.1019
论证充分性	3.43	3.57	0.2422
语句合理性	3.40	3.57	0.1327
人工评分总分	14.00	14.67	0.1206

注：**p<0.01。

3.3　人工智能辅助下写作创造力的变化

访谈中，共有 8 名被试评价了 GPT-3 所提供论点的创造性。其中 3 名偏向正面，5 名偏负面。正面评价多认为其给出了没想到的思考方向，有参考价值。而负面评价一方面是由于跑题，另一方面，一些论点过于简单、常规，如有两名被试提到，部分论点“就是（和题目）关键词上的高度重合”“可能基本上都是一样的东西在 paraphrase”因而难以借鉴。

对于作文评分结果，使用 GPT-3 辅助后，观点新颖性一项得分也高于无辅

https://doi.org/10.35534/lin.0502009

助写作，但未达到显著水平（p=0.1019）。对人工智能辅助持中立或积极态度的被试，其观点新颖性得分均高于等于无辅助时的得分。

上述评分结果表明，人工智能工具辅助对英语写作水平和创造力有一定积极作用。

此外，我们还对实验结果进行了分析研究：

问卷访谈结果和作文评分结果不一致，我们认为有以下的原因：首先，被试在测试和使用网站时常遇到跑题现象，导致总体印象呈负面，在填写时也受其影响。阅读网站生成的文本时，部分被试提到，自己关注的重点是文章的观点及论点论据，而没太在意其语法词汇和衔接问题，在作文中也很少参考，因此这两项评价较低。其次，作文写作有一定随机性，不排除部分被试在使用人工智能辅助时发挥较好，而自己写作时则表现不佳。同时，可能被试在使用实验网站时，相比于独立写作感到有新鲜感，因此更加专注。也不能排除被试恰好对所写作文题较熟悉的可能。

关于跑题的原因，我们认为有两种可能：一是 GPT-3 等人工智能模型，在文本和逻辑组织等方面还有待提高。二是 GPT-3 对语句（尤其是长句）的理解不佳，而实验作文题目较长，且被试在输入提示词时，没有统一格式，GPT-3 与被试的理解可能出现偏差。

对于只有语法搭配一项结果有显著提高，我们认为可能和 Grammarly 评分分差较大有关。人工评分各项，两组中只有 A 组提高较大，故总体没有显著提升。而 A 组提升较大，可能是由于使用网站较晚，对其更加熟悉，也可能是因为 A 组较 B 组对人工智能评价更积极，接受度更高。

4 讨论

4.1 人工智能辅助写作过程中，需要什么样的 prompt？

如上文所述，本实验中，被试提示词的输入未加以控制。考察网站后台的输入记录，被试的输入策略主要可分为三类：依关键词生成、依题目生成和依段落（如首段或多段）生成，而很少有被试使用我们在实验前提供的参考格式。

以此，出现的问题包括文体不符、文章生成不完整、GPT-3 模型无法准确理解被试的要求或写作意图等。相关研究也可以考虑从这方面入手，探究不同提示词对生成文本质量和对被试的影响。

本实验中仅利用 GPT-3 生成文章。类似于实验网站中的“头脑风暴”功能，若能提供给被试，可能得到更加正面的评价。考察 GPT-3 给出的 cliché，包括给出的各项论点后，我们发现它们都较为合理，服从当下社会主流观点的分布。部分被试在访谈中也提到了这一点，即不仅限于生成文章，而是考虑生成概括性的论点和论据，短时间内提供大量角度供使用者参考。另一种思路是生成文章的纲要，并要求其给出与论点相对应的论证。

本实验所用 GPT-3 模型为较早期版本。在微调训练时，采用了 babbage 基础模型和默认参数，未分其他基础模型和如 n_epochs、batch_size 等参数对微调模型的改变，从而影响文本生成质量。此外，GPT-3 对复杂长句理解能力较弱，未来研究可以选用如 GPT-4 等更新的人工智能模型，以获取更佳的文本生成结果。

4.2 对未来人工智能辅助写作的展望

本次写作类型针对于议论文，对象为高校学生；人工智能辅助参与的阶段，参照 Jeon 等学者的框架（2021），集中在形成想法的阶段。后续研究可探索更多文体、不同使用群体及人工智能辅助写作的方式。

同时，我们也注意到不同提示词适用的参与阶段有所不同。例如本实验中所采用的文章生成，可能更适合于参与实施阶段，让使用者直接在生成文本上修改、添加自己的想法。而形成想法阶段，更需要大量较为简洁的、有代表性的论点，以满足发散思维的目的。

如何高大和罗忠民（2008）所指出，人工智能辅助写作可通过个人、人机交互或人人合作等方式开展。也有研究采用多模态的方式辅助创造过程，并整合多种不同功能提供给使用者（Frich J et al.，2019）。未来可将生成式人工智能与传统学习方式（如词典）、算法推荐等相结合，打造功能更加强大的英语辅助写作和学习平台。

https://doi.org/10.35534/lin.0502009

4.3 本研究尚待完善之处

首先，本实验样本数仅有 15 个，被试均来自某双一流高校，且大部分为英语专业本科生，因此实验的代表性受限。

本实验中被试的英语水平不完全同质，在写作时可能对文章结构、词汇等方面的理解不同，面对 GPT–3 提供的文章时所采取的态度和策略也不同。如部分被试认为生成文章的词汇和语法比自己掌握的更加高级，在作文中乐于采纳 GPT–3 的意见，甚至大段摘录；相反，认为 GPT–3 在英语水平不如自己的被试，则更倾向于仅参考文章思路，部分被试因此在评价 GPT–3 时偏负面。

在考察微调后 GPT–3 模型的生成文本时，我们发现，虽然文章均已具有考试议论文的结构，分段明确，但内容质量不高，常出现论点无论据支撑的情况。而这和训练文本的特点是不相符的，可能与数据集规模不够有关。此外，OpenAI 的文档中已将本实验依托的 open–ended generation 类型从微调的样例中删除，其原因未知。

本研究对创造力的评判有待完善。我们主要依靠被试思维的发散程度评判创造力，而对被试想法收敛、即确定选择何种论点的过程，简化为选择不太常见的论点。这一点没有考虑到部分论点写作者由于经历背景不同，可能缺乏相关论据，且未考虑被试是否有动机从熟悉的论点转向不熟悉之处，因而结果带有随机性。

同时，我们对写作水平和创造力的考察，主要以结果为导向。此外，两组被试熟悉实验网站的时间有较大差异，且被试熟悉、使用实验网站的时间较短，我们认为其写作水平和创造力不太可能仅靠阅读文章就显著提高。一项更加专注于某一方面（写作或创造力）的历时研究可能获得更好结果。

5 结语

本实验结果对人工智能辅助英语写作研究具有一定的启示意义。我们通过实证调查发现，人工智能辅助在一定程度上能够提高写作者的英语写作水平与创造力。尤其是在语法搭配一项上，达到了显著水平，说明阅读人工智能工具

生成的文章，可以带给写作者一定的启发，如提供恰当的表达方式、拓展论证思路和丰富论点论据等。

参考文献

[1] 李春琳. 人工智能在外语教学中的应用及研究热点 [J]. 中国教育信息化，2019（6）：29–32.

[2] 何高大，罗忠民. 人工智能在外语教学中的应用——兼评《Artificial Intelligence in Second Language Learning：Raising Error Awareness》[J]. 外语电化教学，2008（3）：74–80.

[3] 常畅，常海潮. 国内二语写作研究回顾与前景展望 [J]. 外语电化教学，2020（3）：61–67，10.

[4] 杨丽萍，辛涛. 人工智能辅助能力测量：写作自动化评分研究的核心问题 [J]. 现代远程教育研究，2021，33（4）：51–62.

[5] 倪传斌. 双语者创造力的影响因素和作用机制研究综述 [J]. 外语教学与研究，2012，44（3）：411–423，480.

[6] 武欣，张厚粲. 创造力研究的新进展 [J]. 北京师范大学学报（社会科学版），1997（1）：13–18.

[7] 魏耀章，苑冰. 创造力和中国英语学习者隐喻生成能力的相关研究 [J]. 西安外国语大学学报，2009，17（4）：80–84，92.

[8] 谢元花，周书婕. 中国学习者创造力与口头叙事任务表现的相关研究 [J]. 中国外语，2018，15（1）：69–76.

[9] Rafner J， et al. Crea. blender：A Neural Network–Based Image Generation Game to Assess Creativity [C]. Extended Abstracts of the 2020 Annual Symposium on Computer–Human Interaction in Play. Association for Computing Machinery，New York，NY，USA，2020：340–344.

[10] Jeon Y ，et al. FashionQ：An AI–Driven Creativity Support Tool for Facilitating Ideation in Fashion Design [C]. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing

https://doi.org/10.35534/lin.0502009

Machinery，New York，NY，USA，2021，Article 576，1–18.

［11］Koch J ，et al. May AI? Design Ideation with Cooperative Contextual Bandits ［C］. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery，New York，NY，USA，2019，Paper 633，1–12.

［12］Chung J J Y，He S，and Adar E. The Intersection of Users，Roles，Interactions，and Technologies in Creativity Support Tools［C］. In Designing Interactive Systems Conference 2021. Association for Computing Machinery，New York，NY，USA，2021：1817–1833.

［13］Clark E， et al. Creative Writing with a Machine in the Loop：Case Studies on Slogans and Stories［C］. In 23rd International Conference on Intelligent User Interfaces. Association for Computing Machinery，New York，NY，USA，2018：329–340.

［14］Kreminski M， et al. Cozy mystery construction kit：prototyping toward an AI–assisted collaborative storytelling mystery game［C］. In Proceedings of the 14th International Conference on the Foundations of Digital Games. Association for Computing Machinery，New York，NY，USA，2019，Article 86，1–9.

［15］Osone H，Lu J L，and Ochiai Y. BunCho：AI Supported Story Co–Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese ［C］. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery，New York，NY，USA，2021，Article 19，1–10.

［16］李婷. 高中英语写作教学中的人工智能应用困境与对策［D］. 镇江：江苏大学，2020.

［17］Kantosalo A ， Riihiaho S. Quantifying co–creative writing experiences［J］. Digital Creativity，2019，30：1，23–38.

［18］Brown T B ，et al. Language models are few–shot learners［EB/OL］. ［2022–04–16］. arXiv preprint arXiv：2005. 14165（2020）.

[19] IELTS Mentor. IELTS Writing Task 2/ IELTS Essay Sample [EB/OL]. [2023-03-14]. https://www.ielts-mentor.com/writing-sample/writing-task-2.

[20] OpenAI API. OpenAI API documentation [EB/OL]. [2023-04-24]. https://platform.openai.com/docs/introduction.

[21] IELTS. IELTS Scoring in Detail [EB/OL]. [2022-11-07]. https://www.ielts.org/for- organisations/ielts-scoring-in-detail.

[22] Wu Y, Yu Y, An P. Dancing with the Unexpected and Beyond: The Use of AI Assistance in Design Fiction Creation. [EB/OL]. [2022-11-08]. arXiv preprint arXiv: 2210. 00829 (2022).

[23] Frich J, et al. Mapping the landscape of creativity support tools in HCI [C] In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019: 1-18.

A Study on AI-assisted English Writing in Terms of Writing Proficiency and Creativity

Zhou Qingyang Xin Yiheng Deng Xiangteng
Zhang Rongkun Yu Yunye

School of Foreign Languages, Southeast University, Nanjing

Abstract: Generative Pre-trained Transformer 3 (GPT-3) is a relatively new artificial intelligence model with powerful functions. This project was an attempt to use GPT-3 in English (as a second language) writing and integrate artificial intelligence tools into instruction. It also explored the relationship

https://doi.org/10.35534/lin.0502009

between artificial intelligence tools and English writing proficiency together with creativity. An experiment involving 15 subjects was carried out on two Saturdays. The subjects performed two argumentative writing tasks, one with artificial intelligence assistance and the other without. After the experiment, they were asked to complete two questionnaires and one interview. Their writings were scored from five aspects: grammar and collocation, argumentative structure, argument novelty, proof adequacy, as well as lexical and syntactic use. The result showed that grammar and collocation improved significantly after the AI assistance (p=0.0062), and the other aspects, though not reaching the significant level, also improved to different degrees. This suggested that artificial intelligence assistance had a positive effect on English writing proficiency and creativity to some extent, which could inspire the following studies in relevant fields.

Key words: English writing; Artificial intelligence; GPT-3; Human-computer interaction