arXiv:2303.03283v2 [cs.HC] 7 Nov 2023

人工智能代笔效应：用户虽不认为自己拥有人工智能生成文本，却自称是作者

Fiona Draxler和Anna Werner，德国慕尼黑大学Florian Lehmann，德国拜罗伊特大学Matthias Hoppe和Albrecht Schmidt，德国慕尼黑大学Daniel Buschek，德国拜罗伊特大学
Robin Welsch，芬兰阿尔托大学

文本生产中的人机交互增加了作者身份问题的复杂性。在两项实证研究（n1 = 30 和 n2 = 96）中，我们考察了个性化语言生成中的人机协作所涉及的作者身份和所有权。我们发现了人工智能代笔效应：用户不认为自己是人工智能生成文本的所有者和作者，却不愿公开声明人工智能作者身份。人工智能生成文本的个性化并未影响人工智能代笔效应，而参与者对文本的影响程度越高，其所有权感越强。与人工智能代笔人相比，参与者更可能将所有权归于假定为人类的代笔人，因此，对人类代笔人的所有权与作者身份之间的差距更大。参与者对人工智能代笔人和人类代笔人的作者身份所作的合理化解释相似。我们讨论了研究发现与心理所有权及人机交互之间的关系，为调整文本生成任务中的作者身份框架和人工智能用户界面奠定基础。

CCS 概念：• **计算方法** → **自然语言生成**；• **应用计算** → **出版**；文本编辑；
• **社会与专业议题** → 知识产权。

补充关键词：所有权、作者身份、大语言模型、文本生成

ACM 引用格式：
Fiona Draxler、Anna Werner、Florian Lehmann、Matthias Hoppe、Albrecht Schmidt、Daniel Buschek 和 Robin Welsch。2023年。人工智能代笔效应：当用户并未意识到 AI 生成文本的所有权，却自称为作者。载于 。 ACM，纽约州纽约，美国，40 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

想象一下，你在纽约的短暂旅程即将结束。与其花时间亲自写明信片，你也许会请 GPT-3 为你写一张个性化明信片。你会在明信片上署上自己的名字吗？在人机协作中使用个性化人工智能（AI），可能会对作者身份这一概念产生重大影响。这种协作的一个例子是使用 GPT-3 等语言生成模型来协助完成写作任务。以往的人机交互研究考察了用户对协作式创意写作[40, 71, 99], 风格迁移 [88], 文本摘要 [43], 以及对人工智能建议的感知作者身份 [72]的看法。然而，这些研究尚未探讨人们如何将生成文本归于作者名下并声明其作者身份。以往的社会科学研究考察了代笔现象的普遍程度及其合理化解释 [18, 38, 89]。代笔是指使用他人创作的文本却不注明其贡献者，这种做法在某些学术领域（如医学）十分常见[45, 108], , 在撰写自传或政治演讲稿时也很常见 [14]。使用代笔人的原因包括

允许为个人或课堂使用而制作本作品部分或全部内容的数字或纸质副本，无需付费，但前提是不得为牟利或商业利益制作或分发副本，且副本首页须附有本声明和完整引文。必须尊重本作品第三方组件的版权。其他用途请联系所有者/作者。© 2023年 版权所有归所有者/作者所有。
手稿已提交至 ACM

经济或政治利益 [45], 缺乏时间和写作专业知识 [14],以及渴望获得满足感（例如取得好成绩）的压力 [74]。个性化大语言模型（LLMs）正逐渐向公众开放，并有望在包括文本生产在内的日常任务中得到普遍应用。因此，我们有必要了解人们如何声明个性化人工智能所生成文本的作者身份，以及人工智能在多大程度上被当作人工代笔人使用。当文本根据用户的经历和个人写作风格进行个性化处理时，用户与人工智能贡献之间的界限开始变得模糊，这一问题尤其值得关注。

本文从作者身份的角度探讨个性化大语言模型的人机交互。因此，我们关注使用个性化人工智能进行文本生成时，在归属所有权和作者身份的过程中涉及的心理过程；这里的个性化人工智能指针对单个用户微调的模型。请注意，这与当前关于人工智能生成材料的法律与伦理讨论重点不同；后者旨在形成规范性框架，即规定应如何声明作者身份，并且主要关注学术和专业情境中的非个性化材料 [91]。因此，我们研究个人写作情境中的作者身份；在这种情境中，声明的作者身份反映的是个人决定，而非明确的指南，并且针对个人写作风格进行个性化仍是一种新兴实践。

为探讨个性化大语言模型中的所有权和作者身份问题，我们开展了两项实证研究，揭示了个性化人工智能的人工智能代笔效应。我们将这一效应定义为使用个性化生成式人工智能，却不注明人工智能的贡献。由此，人工智能扮演了代笔人的角色。总体而言，我们发现，用户并不认为自己是AI生成文本的作者，但往往不会公开声明人工智能的作者身份。在研究1中，我们比较了个性化和不同交互方法对感知所有权与声明的作者身份的影响。在研究2中，我们在一个大型样本中复现人工智能代笔效应，并将其与人工代笔人进行比较。

我们表明，个性化并非人工智能代笔效应的关键因素（研究1）。对交互过程和内容的主观控制感会增强所有权感。此外，人们会对人工智能代笔人和人工代笔人采用类似的合理化解释（研究2，已预注册于 https://aspredicted.org/RKV_ZXX）[1]。基于此，我们探讨如何从以用户为中心的视角拓展常见的作者身份声明框架（例如 CRediT 分类法[2], [1, 3, 13]），以纳入对人工智能的支持。理解控制感、所有权感与作者身份之间的相互作用，也有助于指导未来人工智能辅助文本生成系统的交互设计。

2 相关工作

本节从技术和以人为中心的视角，概述文本生成模型的现状，以及使用文本生成系统的情况。此外，本节还讨论所有权和作者身份的概念，以及它们目前如何随自动化生成而变化。

2.1 生成模型与个性化

近年来，各种生成模型已出现，可用于文本到文本 [15]和文本到图像[3] 生成、基于提示词的音乐创作[4]以及文本简化 [111]等任务。其他模型则以先前作品的风格生成图像 [90], 、诗歌 [102],、说唱歌词 [85], 或音乐 [57] 。生成模型通常采用包含编码器和解码器的 Transformer 架构，并利用注意力机制，使模型能够回溯参考

[1]请注意，两项研究均在 Nature 和 arXiv 提出人工智能署名政策之前开展，参见第 2.3 节。[2] https://credit.niso.org [3]https://openai.com/dall-e-2/ [4]https://mubert.com

之前的内容 [105]。这些模型通常经过训练，能够根据先前的文本预测最有可能出现的下一个词元。因此，它们学会了后续词语之间的联系，非常适合进行文本生成。

在这项工作中，我们聚焦于使用 GPT-3 进行文本生成。GPT-3 于 2020年 发布，目前是一个规模特别大、能力强大的语言模型 [25, 27]。它采用深度学习方法，例如注意力机制和 Transformer 架构，并通过自回归预训练进行训练 [15]。由于该模型能够特别关注关键文本特征，并在较长的文本段落中连接相互关联的词语，这种架构提升了长篇连贯文本的生成能力[103] [32]。GPT-3 生成的文本[5] 在统计层面上与给定的自然语言提示相契合；它利用句法能力关联词语，却并不理解查询的语义和上下文 [34]。生成文本的质量在语法、连贯性和流畅性方面通常可与人类撰写的文本相媲美，但也可能生成荒谬或错误的内容 [32, 103]。与许多其他机器学习系统一样，GPT-3 也可能再现其训练数据中存在的偏见，例如种族偏见或性别偏见 [15, 76]。GPT-3 擅长的一种应用是“少样本学习”，即在提示中加入该任务的输入输出对示例 [15]。这些上下文示例的选择会显著影响生成文本的内容和质量 [75]。

同样的理念也可以用于个性化：系统的输出会根据有关用户 [111] 及其情境 [29] 的信息进行调整，以使生成文本更符合写作者的期望和需求。根据 Wang 等人 [106], 的观点，个性化可以分为两个层面：其中一类包括个人数据、用户属性和用户偏好等事实性知识。第二类是风格修饰项，可分为情境性的或个人化的。例如，*Gmail* 智能撰写 会在基于个人已发送邮件训练的全局模型和个人模型之间进行插值 [19]。相关人工智能辅助场景中的研究表明，个性化和适配可以提升结果的感知质量 [65, 83] 和用户参与度 [56]。文本生成很可能也有类似效果。然而，要让生成文本符合写作者的价值观或观点颇具挑战性，而这会影响其接受度 [12]。总体而言，大语言模型可视为一种强大且用途广泛的写作技术，并且可以通过个性化进一步改进。

2.2 与文本生成系统的交互

在人机交互研究中，文本生成也被研究并应用于文本输入的用户界面。最初，这类“智能”或“预测性”文本输入方法被开发为辅助与替代沟通（AAC）工具，以减少认知或运动障碍人士的手动操作负担。具体而言，这包括提供完整单词的建议，用户随后可以选择这些单词，而不必逐字母输入 [36, 52]。后来，类似的方法被应用于效率这一宏大目标 [69]：理想化文本输入方法的瓶颈在于构思内容的认知过程，而不是输入内容的物理过程。在 Google 的 智能回复 等系统中，简短的文本建议旨在使用户有可能完全跳过手动写作 [61]。

随着大语言模型的兴起，我们如今看到，越来越多的系统被构想、呈现和／或反思为写作助手，它们或许能提供一定程度的灵感 [10, 21, 71, 97, 112]。正如 Swanson 和 Gordon 所提出的，生成模型也能提升交互式故事讲述系统的多功能性 [99]。这类系统的用户界面如今开始探索超越在文本框中的用户光标处或键盘上方的“建议栏”（参见 [4, 11, 16, 86]）的设计：例如，通过侧边栏 [112], 、视觉故事弧线 [22], 、可操控隐喻 [39], 以及文本之外的模态（如图像、声音） [97]来控制和呈现生成的灵感内容。

[5]主要用于英语

总而言之，对写作辅助的更广泛界定，将传统的人机交互问题（如用户界面设计因素）与新兴的作者身份问题联系起来（参见 [72]）。

事实上，许多研究者在交互研究中（含蓄地）触及了这类问题，却没有在研究设计中明确探讨，因为这并非他们关注的重点。例如，Lee 等人 [71] 将其系统命名为“CoAuthor”，并引入了“互惠性”和“平等性”这两个测量指标，用于衡量用户与人工智能在文本创作中的贡献。类似地，*WordCraft*被描述为与辅助语言模型系统 [112]的“协作”和“对话”。Buschek 等人 [16] 聚焦于用户在此类系统中的输入行为，通过定量研究识别出九种模式，从忽略建议到连续串联多个建议短语。Arnold 等人 [4] 指出，与被视为预测的后续内容的词语相比，短语更多被看作创意。与此相关，Singh 等人 [97] 描述了作者如何主动“跳跃”，围绕建议短语来调整故事。类似地，Dang 等人 [26] 发现，作者会修改原文以影响人工智能生成的摘要，并将这些摘要解读为文本的一种外部视角。Bhat 等人 [10] 从 Hayes [51], 的写作认知过程模型出发，定性剖析人工智能对作者及其文本的影响。该模型区分了提出想法、将想法转换为文本，以及最后通过特定输入法将文本录入这三个过程。建议可能会影响这三个过程。近期研究表明，在此背景下，人工智能可能会影响作者的观点 [59]。

一个在实践中已从充足的人工智能支持中受益的具体写作应用场景是交流（例如聊天消息、电子邮件）。针对这一场景中的人工智能支持的研究提出了“AI介导的沟通”（AI-MC）这一术语 [47]。我们的明信片写作任务也属于类似场景。不过，与部分关于AI-MC的研究不同，我们不考察人工智能对消息接收者或发送者与接收者的关系所产生的影响（参见 [53, 60, 78]）。相反，我们研究发送者对文本所有权的（自我）感知。在这方面，我们的研究与 Mieczkowski 和 Hancock [77], 关于AI对写作者感知到的能动性影响的AI-MC研究相关。他们通过出声思维研究探讨能动性，重点关注使用人工智能写作的过程；而我们研究的是创建消息后相关的“最终”感知，即感知所有权。

总而言之，目前处于人机交互与自然语言处理交叉领域的文本输入研究，正明显转向为人类写作提供更全面的人工智能支持。尽管许多研究会关注写作过程中（感知到的）作者身份等方面，但目前仍缺少专门探讨作者身份如何被归属的研究。这促使我们在研究1中，围绕多个与心理学相关的变量（例如所有权感）开展细粒度分析，并比较不同的交互设计。

2.3 面对生成式人工智能时声明的作者身份、版权与代笔

作者身份和版权的概念与代笔实践形成对比。直观而言，任何文本或艺术作品的创作者都应被声明为其作者并拥有版权，但情况并非总是如此。代笔是指出版物不署名内容创作者的做法 [23, 108]。代笔的原因包括缺乏时间、背景细节或修辞技巧 [14, 89], 在某些情况下，必须以公众人物形象的姓名作为作者姓名以进行营销 [18, 38], 同辈压力，促使其提升在科学领域的声誉或可信度 [46, 81], 或获取经济利益 [45, 101]。诸如代笔之类的不道德作者身份实践推动了作者身份标准的发展 [23]。例如，国际医学期刊编辑委员会（ICMJE）规定，若某人满足以下条件，则应被声明为作品的作者：对作品的构思或作品数据的分析作出重大贡献、起草作品或对其进行重要的批判性修改、批准拟出版的版本，并且对作品各个方面承担责任

作品[6]。其他组织，例如 ACM[7]，也制定了类似标准。CRediT 分类法采取了略有不同的方法，将不同类型的贡献进行分类，以明确每位个人贡献者的角色 [1]，而 Bd 等人提出的概念 [8] 则允许进行交互式角色声明。当作者身份声明变得复杂时，这一点尤为重要，例如涉及多个参与方，或作品源自先前的作品时 [28]。尽管如此，部署平台有时甚至不支持详细的归因 [20]。

将个性化生成式人工智能模型纳入其中，使作者身份声明更加复杂。一方面，版权和知识产权成为备受争议的法律问题 [6, 37, 98]，计算机算法或其开发者，以及训练数据的创作者，都可能拥有内容的权利 [5, 37, 58, 91]。正如各国立法存在显著差异所表明的，权利分配确实可能完全取决于主观判断 [91]。另一方面，生成式人工智能的开发者将所有权利赋予用户。OpenAI 表示，使用文本生成模型 GPT-3 或图像生成模型 DALL·E 的客户，对生成的内容[8]拥有完整权利，包括商业用途。根据其指南，GPT-3 生成的文本必须署上用户或其公司的名称。必须注明内容由人工智能生成，但 OpenAI 保留对语言模型本身的一切权利、所有权和权益。

出版实践中存在一些实例，可以为我们了解不同利益相关者对生成式人工智能作者身份的看法提供线索。在新闻业中，人工智能贡献有时会被声明，但并非总是如此 [44, 64, 79]。例如，Graefe 在其《自动化新闻指南》中 [44] 指出，人工智能作者身份声明对于透明度十分重要。Montal 和 Reich [79] 建议，使用自动生成文章的署名栏来注明软件供应商，并在披露声明中提及由算法生成的文本部分。在少数情况下，人工智能贡献者也曾被列为科学出版物的作者 [70, 82, 110]。这促使多个科学协会发布相关政策，反对或禁止将文本生成工具列为作者，例如 Nature[9] 和 arXiv[10]；它们建议将这些工具列入方法部分。需要注意的是，这些政策并未区分文本生成式人工智能与个性化人工智能。因此，关于将生成式人工智能应用于文本生成的指南，主要针对专业情境，尚不够细致，无法充分考虑个性化生成式人工智能。

随着大语言模型的发展，针对特定应用场景的规范框架正在逐步形成，但针对个性化人工智能文本生成的专门框架仍然缺失。尤其尚不清楚，文本生成式人工智能的使用与代笔有何不同。通过研究用户对个性化文本生成式人工智能的作者身份声明，并将其与常见的代笔实践进行比较，可以为这一讨论提供实证数据。

2.4 人工智能支持与所有权意识

作者身份不仅是声明问题，也涉及个人判断。一个人是否觉得自己是作者，取决于其感知控制和能动性等因素。对有形物品以及思想等无形实体的所有权感受，都可以通过创作和控制发展起来（心理所有权；[66, 84]）。例如，在视频再混合情境中，Diakopoulos 等人 [28] 发现，创作者通常觉得自己必须对既有作品进行重大修改，才能将其视为自己的作品。这再次表明，个人投入对于感知所有权至关重要 [84]。关于生成式人工智能的感知作者身份研究仍然十分有限。值得注意的是，Lehmann 等人 [72] 在交互设计中发现，作者身份与控制感存在相关性：在他们的研究中，相较于接受文本建议，参与者自行撰写文本时的感知作者身份更强。在一项发表于

[6]https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html [7] https://www.acm.org/publications/policies/roles-and-responsibilities#h-criteria-for-authorship [8] https://openai.com/api/policies/sharing-publication/#content-co-authored-with-the-openai-api-policy [9] https://www.nature.com/nature/for-authors/initial-submission [10]https://info.arxiv.org/help/moderation/index.html#policy-for-authors-use-of-generative-ai-language-tools

2017年，新闻机构成员并不认为算法是作者 [79]。相反，他们提到了相应的新闻机构，或提到了参与创建算法的人员，这表明他们对作者身份持有人格化的理解。因此，写作过程和最终产品必须分开看待。因此，在本文中，我们从感知所有权和作者身份声明的角度，探讨这两个维度之间的相互作用。

2.5 相关工作总结

围绕模仿代笔实践的文本生成式人工智能中的作者身份归因问题，目前仍存在持续的争论。对于个性化生成式人工智能以及非专业领域而言，这一争论更加复杂。人机交互可以通过研究与人工智能交互的设计如何影响作者身份声明及相关心理现象来促进这一讨论，其中最重要的心理现象是所有权。需要实证数据，以从以用户为基础的视角为这一讨论奠定实证基础，并为文本生成式人工智能界面的设计者提供指南。

3 研究问题与假设

我们提出以下研究问题：

研究问题1：个性化AI生成文本的所有权感是否与作者身份声明一致？**研究问题2**：人机交互中的影响程度是否会影响所有权感？**研究问题3**：所有权感是否取决于个性化质量？**研究问题4**：人工智能代笔效应 与人工代笔者有何不同？

本文旨在从人机交互的视角，探究哪些因素会影响个性化大语言模型生成文本的作者归属。首先，我们识别人机交互中进行作者归属时涉及的心理过程。需要注意的是，我们考察的过程包括作者归属（从所有权感到作者身份声明）、人机交互中感知到的控制程度、个性化的主观质量，以及引领人机交互的感受。换言之，所有权感指的是主观层面：一个人是否觉得自己是文本的作者，并且拥有该文本？这也包括对产出的控制感，因为这两个概念密切相关（参见第2.4节）。另一方面，作者身份声明指的是文本被归属的主体，例如页眉或署名栏中所标示的主体。

如果人工智能系统（此处为 GPT-3）的使用方式与人工代笔人非常相似，那么我们假设存在一种人工智能代笔效应：

H1.1: 参与者会认为人工智能（而非自己）是个性化AI生成文本的所有者，但**H1.2:** 在出版个性化AI生成文本时，他们不会声明人工智能作者身份或人工智能支持。

其次，我们希望探究常见的交互方式如何影响作者身份归属过程。这一点很重要，因为界面设计者面临着大量关于作者身份的规范性陈述，却很少了解用户界面如何影响作者身份声明，以及它们在人机交互过程中会带来怎样的用户体验。

为了在设计基于大语言模型的交互式应用时充分考虑作者身份，我们需要了解不同的交互方法如何影响个性化AI生成文本的作者身份感。Lehmann 等人 [72]发现，作者身份感与用户对人工智能贡献的影响程度呈正相关。在此，客观控制是指用户对AI生成文本的影响程度，例如采用归类为写作、编辑、选择和获取的交互方法；而感知控制是指用户主观上认为自己能够预先

影响生成文本的程度。主导权则指用户感知到的塑造结果的主动性，即交互过程中的能动性。

因此，我们还评估感知控制和主导权，并提出以下假设：

H2.1: 不同交互方法所体现的对生成文本的影响程度，会影响控制感；而**H2.2:** 对生成文本的影响程度会影响领导感。

第三，我们探究人工智能个性化如何影响这一过程。在此，个性化是指通过微调AI模型，使其适应个人用户的特定偏好和写作风格。我们特别研究个性化文本生成的情形，因为在这种情况下，AI模型与用户之间的界限变得模糊。然而，近期研究发现，微调后的AI系统的可用性和用户体验并不取决于个性化和适配的质量 [104]。即使是非自适应系统，只要被介绍为自适应和个性化的，也会被认为具有自适应性和个性化 [68]。如果这一点对个性化大语言模型也成立，那么仅仅将人工智能标记为个性化（即安慰剂个性化），就应能对所有权感产生相当的影响。

我们考察两种个性化质量水平——通过微调实现个性化和安慰剂个性化——并提出以下假设：

假设3： 所有权感不受AI模型个性化质量的影响。

第四，我们探讨使用人工智能与雇用人工代笔人的常见做法在性质上有何相似之处或不同之处。虽然代笔在文学、演讲稿乃至学术界等多个领域都很常见，但人工智能的出现改变了这一过程的动态，因为广大用户都能轻松使用生成式AI模型。

研究表明，相较于人类，人们更倾向于利用算法或AI模型 [62]。这一现象会影响作者归属、所有权感，以及这些人工智能辅助文本向公众呈现的方式。我们提出以下假设：

假设4： 与人工智能代笔人撰写的文本相比，人们对人工代笔人撰写的文本的所有权感更低。**假设5：** 人们更可能将（共同）作者身份归于人工代笔人，而非人工智能代笔人。**假设6：** 使用人工智能生成文本时，即使人们没有所有权感，也并不总会声明人工智能是（共同）作者（人工智能代笔效应）。本假设旨在复现假设1，但不考虑个性化。

我们在研究1中检验假设1至假设3。通过采用不同的交互方法（被试内因素），我们改变了参与者在撰写并公开发布一张来自纽约的明信片这一任务中对AI生成文本的影响程度。基于 [72, 84], 提出的影响程度与所有权感之间的相关性，我们按用户影响程度从高到低定义了四种方法：写作、编辑、选择和获取。我们生成了经过个性化或安慰剂个性化处理的文本（针对各个参与者进行微调的模型与基础模型；被试间因素）。撰写明信片的情境代表了一种人们通常熟悉的个人写作任务。明信片由个人撰写，但往往描述相似的经历，例如著名景点或当地美食。这使我们能够应用安慰剂个性化处理。我们选择在博客式网站上发布明信片，以减弱潜在明信片收件人可能对作者身份声明抱有的社会期望所产生的影响。

为回答研究问题4，并复现和拓展研究1，我们开展了研究2（已预注册，见 https://aspredicted.org/RKV_ZXX[11]），比较了人工智能辅助写作与人类作者支持写作任务的情形。

总而言之，研究1旨在考察人工智能文本生成背景下控制感、个性化和所有权感的影响。其目的是引入人工智能代笔效应，该效应指的是，对于个性化人工智能生成文本，用户不会将作者身份归因于人工智能，也不会声明获得任何形式的人工智能支持。本研究的结果为研究2提供了依据；研究2深入探讨人工智能代笔与人工代笔之间的动态关系，进一步加深我们对人机交互中作者身份归属实践的理解，并将其与写作中的人人交互进行比较。因此，研究2复现了研究1，并阐明了在声明人工智能代笔人与人工代笔者的作者身份方面存在的定量差异。

4 研究1：所有权感与声明的作者身份

在明信片写作任务中，我们通过采用不同的交互方法，改变了参与者对人工智能生成文本的影响程度。在研究1中，我们的主要目标是评估参与者如何将所有权归属于个性化人工智能生成文本、交互中的控制感是否会影响这一归属，以及参与者如何为不同的交互方法声明作者身份。该研究于2022年7月进行。

4.1 方法

我们实施了四种交互方法，这些方法涉及用户对输出内容的影响程度不同，从手动写作文本（完全由用户控制）到获取一篇完全由人工智能撰写的文本（用户没有影响）。首先，30名参与者填写了一份问卷，使我们能够微调文本生成模型的参数，从而使生成文本符合每位参与者的写作风格。随后，同一批参与者分别与实际个性化模型或安慰剂个性化模型进行交互，使用全部四种交互方法对其进行评估，并接受指示将这些文本发布到线上。

大约两天后，我们探究了参与者是否仍能区分通过不同交互方法生成的文本。加入这一环节是为了评估生成文本的自我关联性。如果参与者能够持续区分自己的文本与他人的文本，这可能表明，他们与文本之间的关系会因交互方式不同而产生明显差异。需要注意的是，为了简洁起见，并且由于讨论个性化人工智能生成文本的自我关联性会偏离本文的重点，我们将研究1这一探索性部分的结果作为补充材料报告。

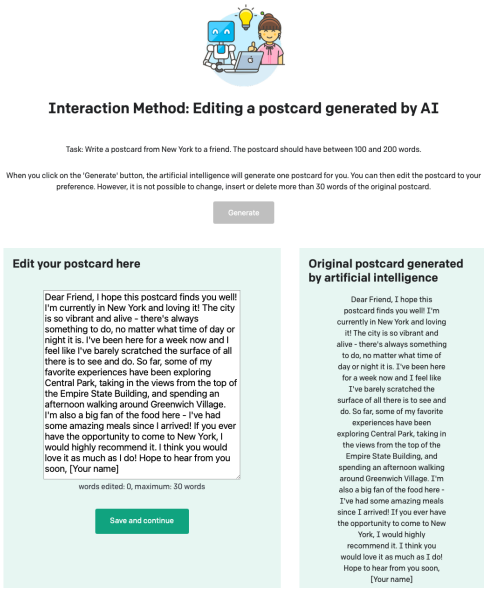
因此，这项研究采用了 4（交互方式）$\times 2$（个性化与安慰剂个性化）设计。在个性化条件下，GPT-3 根据用户的输入进行了微调；在安慰剂个性化条件下，技术层面并未进行微调。在两组中，参与者都被告知他们使用的是个性化 AI模型。

*4.1.1 交互方法的实现。*我们在一个基于网络的文本生成平台上实现了四种交互方法（交互方法）。每种交互方法代表参与者对文本影响力的不同层级。请参见图1，其中展示了网络原型的截图。下面将更详细地介绍这四种交互方法。控制层级源自 [30]和 [71]。

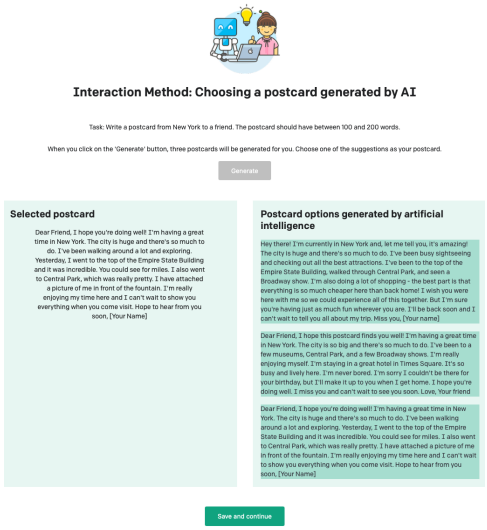
[11]请注意，我们对措辞做了细微调整，以便与手稿相符。



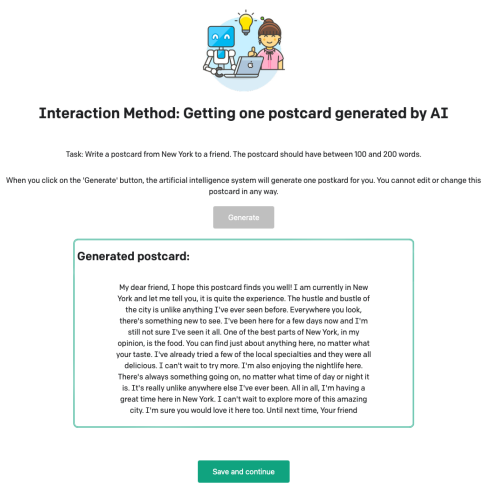
(a) 写作



(b) 编辑



(c) 选择



(d) 获取

图1. Sc 研究1中交互方式的屏幕截图。人机协作图标已获 Iconfinder 授权 [107].

(1) 写作明信片。在这一基线条件下，参与者无需人工智能辅助，手动写一张明信片，也就是说，他们完全掌控写作结果。他们需要输入 100 至 200 个词。这相当于 等级 *0*：完全由人类撰写的语言，见 [30]。

(2) 编辑明信片。参与者会收到一张人工智能生成的明信片，但可以更改（插入、删除或替换）最多 30 个词。通过这种方法，参与者可以手动影响文本。人工智能提供了明信片的初稿。具体的编辑限制根据预试进行了调整。这种方法代表了典型的人工智能辅助场景，如 [71]中的编辑阶段。

(3) 选择一张明信片。参与者会看到三张人工智能生成的明信片，并被要求从中选择一张。参与者无法通过这种方法手动影响文本。不过，他们仍然可以选择自己认为最合适的明信片版本。这对应于 [30] 中的等级 *3*：条件自动化，并类似于 [71]中的获取建议阶段。

(4) 获取明信片。参与者只会看到一张人工智能生成的明信片。与其他方法不同，参与者无法影响明信片文本。相反，这代表了作为最终明信片生成器的人工智能影响程度最高。该方法采用第 *4* 级：高度自动化，即在没有人工监督的情况下生成 [30]。

这些文本在研究展示之前就已生成。不过，参与者点击“生成”按钮后 3 秒，文本才会呈现，以营造明信片实际上是按需生成的假象。

4.1.2 程序。研究分为三个主要部分：(1) 用于个性化的微调调查，随后进行后台处理以生成个性化文本；(2) 采用交互方式制作明信片文本（见第 4.1.1 节），随后填写关于作者身份和交互的问卷（见表 6，第 2 部分）；(3) 最后填写有关文本识别、文本评价和总体结论的问卷。该最终问卷于两天后发放。根据《赫尔辛基宣言》，我们向参与者告知其权利和研究流程，并且只有在他们同意参与后才开始研究。

第*1*部分：微调调查与后台处理。为收集微调输入，我们首先要求参与者针对18个不同的提示词，分别撰写约330—800个字符的回答。这些提示词的风格和主题类似于我们在第2部分中用于明信片生成的提示词（见下文）。提示词示例包括“请用大约5句话介绍一下自己。”和“你最喜欢哪次度假？”随后，我们要求参与者上传他们此前撰写的五篇文本（例如博客文章或给朋友的电子邮件），并为每篇文本匹配一个合适的提示词。提示词的完整列表见表6。

完成微调问卷后，我们进行了后台处理。我们将一半参与者随机分配到个性化条件，另一半分配到安慰剂个性化条件。我们使用个性化组参与者的数据对 Davinci 模型[12] 进行个性化，并针对明信片提示词生成了五张个性化明信片（“写一张从纽约寄给朋友的明信片。明信片应包含 100 至 200 个单词。”；选择生成三张明信片，其他每种交互方式各生成一张）。对于

[12]https://beta.openai.com/docs/guides/fine-tuning

安慰剂个性化组，我们在三个不同的文本集中随机交替使用，每个文本集包含 15 篇预生成文本中的 5 篇[13]。

所有参与者无论被分配到哪个个性化组，都回答了微调问题，并且都必须等待参加第 2 部分的邀请。这有助于减轻人机交互中的应答偏差和安慰剂效应 [17, 68]。

*第 2 部分：写作与上传、作者身份评估。*一天后，参与者开始了研究的第 2 部分。在这一部分中，他们的任务是创作四张写给朋友的明信片，假装自己当时正在纽约游览，并将创作的明信片上传到一个博客式网站。他们将明信片文本粘贴到文本框中，并填写标题、日期和作者声明的元数据，以及访问密钥（研究 ID）。作者声明是一个文本框，参与者可以在其中列出任何他们希望列出的实体。每张明信片均采用四种交互方法中的一种进行创作（即不同程度的影响）。交互方法的顺序经过平衡化处理。在完成所有交互方法后，我们呈现了一份最终问卷，用于了解参与者的交互体验、感知控制，以及他们对生成的明信片的喜爱程度。对于涉及特定交互方法的问题，我们提供了相应用户界面的屏幕截图。

*第 3 部分：文本识别与评分。*在第 2 部分完成两天后，Prolific 发出了第 3 部分的邀请。在最终问卷中，我们询问参与者对明信片文本的记忆程度，以及这些文本是通过哪种交互方式创作的。设置这一延迟是为了确定文本个性化对记忆巩固可能产生的影响 [92], ，同时我们也测量了他们的文本识别表现。最后，我们询问了参与者对自动文本生成的看法，以及是否应该强制标记借助人工智能生成的文本。我们将这一部分的结果作为补充材料提供。

*4.1.3 实验设备与测量指标。*研究1设计为一项基于浏览器的在线实验，包含三份问卷、第 4.1.1 节所述的定制交互方式，以及一个定制的博客式网站。该网站包含一个展示示例明信片的首页（参见图11）和一个用于输入明信片文本并添加元数据的上传表单（参见图12）。各个组件由一个研究框架协调，该框架负责处理研究组件之间的平衡化、日志记录和参与者流程。在微调问卷中，我们记录了参与者撰写的文本和提示词。目前尚无公认的方法来测量所有权感、作者身份、主导权或个性化程度。因此，在第 2 部分和最终问卷（第 3 部分）中，我们采用了定制的视觉模拟量表，具体如下文所述。

*作者身份。*我们测量了作者身份声明，并将其操作化为参与者在博客上传表单中列为署名作者的实体。研究1采用开放式格式，这也启发了我们在研究2中设置选项供参与者选择。该测量为研究问题1和研究问题4提供信息，探讨人工智能生成文本中作者身份的感知与归因。我们使用该测量来检验 H1.2、H5 和假设6。

所有权。 所有权感表示参与者在多大程度上认为自己是文本的作者，以及相应地，他们在多大程度上认为人工智能是作者。“这段文本应归谁所有”通过视觉模拟量表（范围为 -50 我到 +50 人工智能）进行回答。该测量指标同样为研究问题1和研究问题4提供信息，并用于检验我们关于所有权的假设（H1.1、假设3、假设4和假设6）。

[13]技术细节和文本集作为补充材料提供。

控制感。 通过控制感，我们评估了参与者是否觉得自己能够控制生成的文本。该指标通过“谁控制了明信片的内容？”这一条目进行测量，参与者通过视觉模拟量表（范围为 -50 我到 +50 人工智能）作答。这与我们关于控制感的研究问题相关，并用于检验交互方法对控制感影响的假设，即检验假设 H2.1。

主导权。 领导感 表示参与者认为在写作过程中由谁主导。我们通过“谁主导了写作？”这一题项进行测量，参与者在视觉模拟量表上作答（范围为 -50（我）至 +50 人工智能）。这有助于回答RQ2，并检验H2.2。

个性化。 我们评估个性化的两项指标。个性化的客观质量 用于比较经微调用的文本与经过或未经个性化处理的生成文本。个性化的主观质量 反映参与者对交互方式生成的文本与其写作风格契合程度的印象。后者有助于回答RQ3并检验假设3。

用户体验。 最后，我们通过 *AttrakDiff* 问卷评估 交互方式 的用户体验[50] ，并提出开放式问题。AttrakDiff 长版通过成对的反义形容词评估用户体验的四个方面：（1）实用质量，即产品的可用性；（2）享乐质量，即愉悦程度，包括刺激性，即能否激发灵感并促进个人发展；（3）身份，即能否传达期望的身份；以及（4）总体吸引力 [49, 50]。这有助于深入了解交互方式，并为未来的大语言模型用户界面设计提供启示，还能揭示用户体验与控制感如何关联到RQ2的研究方法。

表 6 列出了这些构念及辅助测量指标的完整清单（例如，为了更全面地了解用户接受度而测量的感知诡异感）。人口统计信息从 Prolific 导出。

4.1.4 参与者。我们通过 Prolific 招募了 43 名英语母语者，其中 30 人正确完成了研究的所有必需步骤。7 名参与者退出了微调问卷。在其余 36 名参与者中，有 1 人从未开始第 2 部分，另有 1 人遇到了技术问题。第三名参与者在完成第 2 部分的第一步和最后一步之间间隔了两天以上，因此这会影响其在第 3 部分中的回答。因此，我们邀请 33 名参与者参加第 3 部分，其中 32 人完成了该部分。其中两人上传了重复文本，因此不得不舍弃他们的数据。在最终的 30 名参与者中，13 人认同为女性，17 人认同为男性，没有人认同为非二元性别或其他性别。他们的平均年龄为 37.2 岁（$SD = 11.8$, $min = 20$, $max = 70$）。他们列出的居住国家包括英国（17 人）、美国（9 人）、南非（2 人）、加拿大（1 人）和津巴布韦（1 人）。5 名参与者此前没有使用过文本生成系统的经验。19 人（63%）使用过单词或句子建议，20 人（67%）在自己的设备上使用过自动更正功能。半数参与者此前曾借助自动补全撰写文本，8 人（27%）使用过智能回复选项。此外，两名参与者表示，有时会由其他人代他们撰写文本，例如工作中的助理。参与者因参加第 1 部分获得 18 英镑，第 2 部分获得 13.50 英镑，第 3 部分获得 3 英镑；完成所有步骤还可获得 6 英镑奖金，即他们花费约四小时总共获得 40.50 英镑。

4.2 结果

我们首先考察文本个性化是否成功。随后，我们分析所有权感和作者身份声明（H1.1 和 H1.2），包括个性化的影响（假设3）。最后，我们检验交互方式是否操纵了控制感和领导感（H2.1 和 H2.2）。我们进一步分析控制感和领导感与所有权感之间的关系，

表1 ．参与者对采用“获取”交互方式的文本的感知（1 非常不同意—7 非常同意）

	个性化		安慰剂个性化	
	MD	***SD***	***MD***	***SD***
这张明信片大多包含我通常用英语写作时会使用的词语和/或短语。	5.07	1.67	4.20	1.90
我自己也会写出一张类似的明信片。	4.40	1.99	3.53	2.10

并对其进行扩展。对其余测量指标（包括开放式问题）进行深入分析超出了本文的范围。我们通过 OSF 仓库提供其他结果的简要概述[14]。具体而言，该仓库包含文本记忆分析、关于 交互方式的定性陈述，以及与所有权、领导感和控制感相关的其他测量指标概述。

分析。 下文中，我们使用带有Bonferroni校正 *t*-检验的方差分析（ANOVA）进行事后比较，分析两个个性化组和四种交互方法在作者身份方面的差异。对于重复测量名义数据，我们使用Cochran’s Q-检验。为了估计分层重复测量中的回归系数，我们使用线性混合模型（LMM），并采用Kenward-Roger估计自由度。对于二水平变量，我们使用效应编码。在研究1中比较交互方法时，我们将各条件分别与写作条件进行对比。采用秩对齐重复测量方差分析 [63]检验非参数序数数据；在此基础上，后续分析采用Bonferroni校正的ART-C对比 [31]。我们使用BayesFactor软件包 [80]计算贝叶斯因子，以评估与主要研究问题相关的零假设。重复测量方差分析模型对残差正态性违背具有稳健性 [41, 48]。因此，为求简洁，对于d’Agostino正态性检验 [24] （Shapiro-Wilk检验过于敏感 $n > 50$）表明正态性违背，且非参数模型与参数模型结果一致的情况，正文仅报告参数模型（而不同时报告非参数模型）。当球形性假设不成立时，我们采用Greenhouse-Geisser校正。统计推断中， α 设为5%。我们使用显著性标记 * 表示$p < 0.05$， ** 表示 $p < 0.01$， *** 表示 $p < 0.001$。最后，我们使用R中的tidytext软件包 [96]计算余弦相似度，以比较文本。

操纵检验：生成文本的个性化。为检验个性化是否成功，我们计算了每位参与者在微调调查中的回答合并文本与第2部分向该参与者展示的人工智能生成文本合并文本之间的余弦相似度。我们预期，真实个性化组的文本相似度应高于收到非个性化文本的安慰剂个性化组。确实，真实个性化组的平均余弦相似度（$M = .77$, $SD = .04$）高于安慰剂个性化条件（$M = .72$， $SD = .06$， $t(25。72) = -2.89$， $p = .008$， $d = -1.06$）。表1证实，参与者认为，与安慰剂个性化文本相比，个性化文本往往更符合他们的个人写作风格。

H1.1 和假设*3*：所有权感。接下来，我们检验所采用的交互方法是否改变了人机交互中生成文本的所有权感。在一项重复测量方差分析中，我们分析了个性化（被试间因素，2个水平：真实个性化与安慰剂个性化）和交互方式

[14]https://osf.io/n4svx/?view_only=916873e81d244be6be8e0790531b1197

表 2。参与者对选定李克特矩阵问题的回答（1 强烈不同意–7 强烈同意）

	写作		编辑		选择		获取	
	平均值	标准差	平均值	标准差	平均值	标准差	平均值	标准差
1：我感觉人工智能就像一个我可以控制的工具。	4	2.15	5	1.52	5	1.75	4	2.27
2：我感觉人工智能系统就像一个 Ghostwriter，负责写作代表我写明信片。	2.5	2.18	5	1.63	6	1.71	6.5	2.32
3：我感觉是我在写作文本，而人工智能在协助我。能。	1.5	2.44	3	1.99	2.5	2.16	2	2.03
4：我感觉是人工智能在写作文本而我在协助它。	1	1.73	5.5	1.91	5	1.82	4	2.13

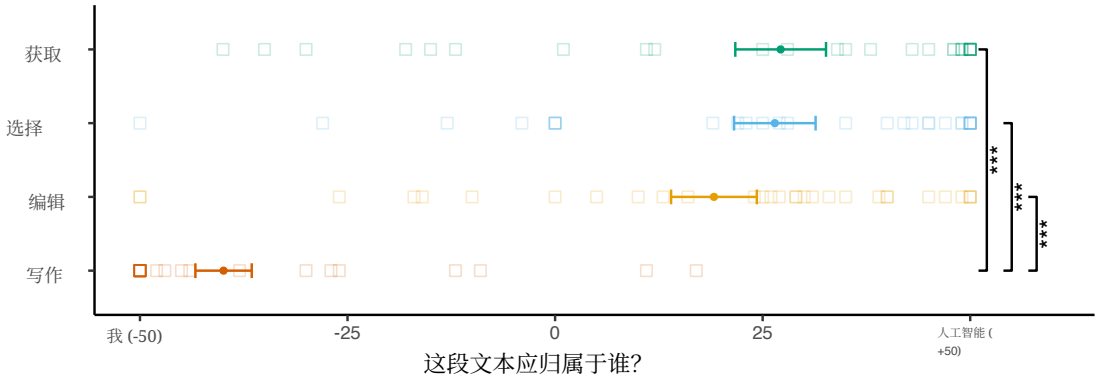


图2。每种交互方式下对“这段文本应归属于谁”的均值评分，该项目采用视觉模拟量表作答（范围为-50（我）到 +50 人工智能），各个数据点以方框表示。误差线表示 ± 一个均值标准误。

（被试内因素，4个水平：写作、获取、选择、编辑），我们分析了参与者通过视觉模拟量表作答的题目“这段文本应归谁所有”（范围为-50（我）至 +50 人工智能）[15]。由于违反了球形假设，$W > .45$，$p < .001$，我们对所有被试内因素应用了Greenhouse-Geisser校正。

如表3所示，我们发现交互方式存在显著影响。所有与人工智能相关的交互方法均与写作条件存在差异（均 p <.001），但所有与人工智能相关的交互方法彼此之间均无差异（均 $p > .05$），另见图2。我们未发现个性化的影响，也未发现交互方式 × 个性化的交互效应。针对个性化这一不显著的主效应，我们进行了独立贝叶斯 t检验，发现：在给定先验的情况下，假设不存在差异的模型比假设均值存在差异的模型更有可能，前者的可能性是后者的2.38倍[16]。

因此，我们可以得出结论：（安慰剂式）个性化人工智能生成的文本改变了所有权感（H1.1）。具体而言，所有与人工智能相关的方法所引发的所有权感均不同于写作。然而，个性化并未影响参与者是否归属作者身份（假设3）。请注意，在图2中，可以看到一些参与者

[15]残差正态性假设不成立（p <.001）。因此，我们进行了非参数方差分析 [63]。与参数模型一样，交互方式存在效应，$F(3,84) = 45.24$，$p < .05$。事后检验结果与参数模型相似；其他所有效应均为 $p > .05$[16]。此处先验设为默认值，标准化柯西先验采用 $\sqrt{2}/2$，具体实现见 [80]。先验敏感性检验显示，使用更宽的先验时，假设均值之间没有差异的模型变得更可能。

表3。交互方式、个性化及其交互的重复测量方差分析结果。当球形假设不成立时，我们对所有被试内因素应用了Greenhouse-Geisser校正。校正后的结果以 ^{G}标记。显著性标记： ** 表示 $p < 0.01$， *** 表示 $p < 0.001$

测量指标	效应	F	df	p	η_p^2
所有权感：这段文本应归属于谁	交互方式	55.75^{G}	(2.02, 56.58)	< **.001*****	.67
属于?	个性化	0.52	(1，28)	.477	.02
	交互方式。×个性化。	0.48^{G}	(2.02, 56.58)	.625	.02
控制感：谁控制了明信片的内容?	交互方式	28.18^{G}	(2.25, 62.95)	< **.001*****	.50
	个性化	1.05	(1，28)	.315	.04
	IM. ×Pers.	0.60^{G}	(2.25, 62.95)	.570	.02
领导感：写作中由谁主导?	交互方式	28.18^{G}	(2.25, 62.95)	< **.001*****	.50
	个性化	1.05	(1，28)	.315	.04
	IM. ×Pers.	0.60^{G}	(2.25, 62.95)	.570	.02
AttrakDiff：实用质量	交互方式	4.20	(3，84)	**.008****	.13
	个性化	1.85	(1，28)	.185	.11
	交互方式. ×个性化.	3.46	(3，84)	.020	.11
AttrakDiff：享乐质量 - 身份	交互方式	8.50^{G}	(2.23, 62.32)	< **.001*****	.23
	个性化	0.76	(1，28)	.391	.03
	IM. ×个性化	2.24^{G}	(2.23, 62.32)	.109	.07
AttrakDiff：享乐质量 - 刺激性	交互方式	4.61	(3，84)	**.005****	.14
	个性化	0.34	(1，28)	.562	.01
	IM. ×Pers.	2.29	(3，84)	.084	.08
AttrakDiff：吸引力	交互方式	8.07	(3，84)	< **.001*****	.22
	个性化	0.97	(1，28)	.332	.03
	IM. ×Pers.	2.27	(3，84)	.086	.08

在写作条件下将所有权感归属于人工智能。这可能是因为参与者沿用了其他条件下人工智能生成文本的表述，也可能是由于回答不认真。

H1.2：作者身份声明。 所有权感是否与在网站上将人工智能声明为作者相关？为此，我们考察了作者 字段中的自由文本条目，该字段属于我们的博客上传表单（参见第4.1.2节）。我们将这些条目归类为指向参与者本人（即姓名或缩写）、人工智能、非个人化归因（“人类”“随机”）或多个类别的组合（例如“人工智能和凯尔”）。

有趣的是，尽管参与者的所有权感很弱，但他们在将明信片发布到网站时仍会署上自己的名字，见 图 3。不过请注意，我们的结果只能从相关性视角对此进行探讨。我们发现，在 写作 条件下，人们会自称为作者，但对于（安慰剂）个性化AI生成文本，他们也会自称为作者。我们使用 Cochran Q检验检验了各比例与随机概率之间的差异。 交互方式 对提及“人工智能”的频率有显著影响， $Q(3) = 18.54$， $p < .001$。图 图 3 右侧的计数汇总了在声明明信片作者身份时，人工智能被列出或未被列出的频率。每种 交互方式 下（写作除外），只有六到七名参与者提及了人工智能。在 编辑 条件下，所有提及人工智能的情况都将其描述为与参与者协作，或作出非个人化的归因。例如，他们写上自己的名字和“人工智能”。在 选择 条件下，六次提及中有三次只列出了“人工智能”；另外三次则将作者身份归于人工智能与参与者的协作，或归于“人类”。在 获取 条件下，七次提及中只有一次是这种情况；其余情况都将“人工智能”列为唯一的

作者。写作条件与其他条件之间的计划比较均达到显著水平（全部 $p < .03$）；不过，编辑、获取和选择条件之间的两两比较均未达到显著水平（全部 $p > .05$）。因此，我们认为，尽管人工智能在通过人工智能交互生成的明信片中被列为作者，但并非所有参与者都如此认为（部分支持 H1.2）。

图 3 进一步展示了所有使用人工智能生成的个性化文本的条件下，所有权感与声明的作者身份之间的关系，并将所有权感与计数数据进行了对比。尽管参与者在大多数情况下认为文本属于人工智能，但他们仍在大多数情况下在明信片下方署上自己的名字。我们将这种差异称为人工智能代笔效应。

问卷结果和参与者陈述为理解人工智能代笔效应的原因提供了见解。例如，当参与者不是在写自己的文本时，他们往往觉得人工智能是一种工具，并且扮演着代笔人的角色（见表 2）。参与者与人工智能之间的协作程度（问题 3）在写作 和 获取这两种条件下最低，即没有人工智能支持和完全由人工智能支持的条件。17 名参与者（56.7%）表示，在他们看来，不应强制要求标注借助人工智能创作的文本。他们提到的理由包括，人工智能“其实只是一种工具，用来表达作者想说的话”，以及“只要使用人工智能的人在发送文本前审阅过，那么这样就没问题。”一些参与者还认为，披露人工智能支持不会受到认可，例如：“如果人们以为我懒得自己花时间发短信，会觉得被冒犯。”其他人则明确将人工代笔人作为比较：“在机器生成文本出现之前，曾有一个由人类组成的完整行业，以类似的方式填充内容空间。”（安慰剂个性化也产生了影响：安慰剂个性化组的一位参与者指出，没有必要进行标注，“尤其是当它反映了你是谁，而且你选择了最适合自己的那条信息时。”

相反，13 位参与者（43.3%）认为应强制披露人工智能贡献。相关理由包括伦理和透明度，例如：“了解自己是在与人类还是算法交互很重要”以及“这样才能知道是否有人说了些不太合乎情理的话。”其他人则提到了自我保护，并表示否则“感觉有点像在作弊。”

H2.1：控制感。我们分析了参与者对“谁掌控着明信片的内容？”这一题项的回答。该题项使用视觉模拟量表作答（范围为 -50（我）到 +50 （人工智能）），并采用重复测量方差分析模型进行分析。同样，球形假设未得到满足，$W > .60$，$p < .02$，因此我们对所有被试内因素及其交互项应用了 Greenhouse–Geisser 校正。残差的正态性未被违反（$p = .14$）。我们发现，交互方式影响了控制感。经 Bonferroni 校正的 t检验显示，除 $p < .05$比较获取与选择（$p > .99$）外，所有比较均存在显著差异（均为 $p < .05$）；另见图 4）。因此，我们可以认为，对交互方式的操纵成功改变了人机交互中的控制感。

如表3所示，个性化既没有产生效应，也没有交互效应。通过贝叶斯 t-检验比较个性化[17]，我们可以量化得出：假设组间不存在差异的模型，比假设组间存在差异的模型更有可能，可能性高出1.98倍。

我们通过拟合线性混合模型扩展了方差分析模型，以根据控制感和交互方式预测所有权感。模型将参与者作为随机效应纳入，以解释数据中重复测量的结构。模型的总体解释力较强（条件 $R^2 = 65\%$；

[17]先验设定为　默认值，即标准化柯西先验上的 $\sqrt{2}/2$，如 [80]中所实现的那样。同样，先验敏感性分析表明，当先验范围更宽时，假设均值之间不存在差异的模型变得更有可能。

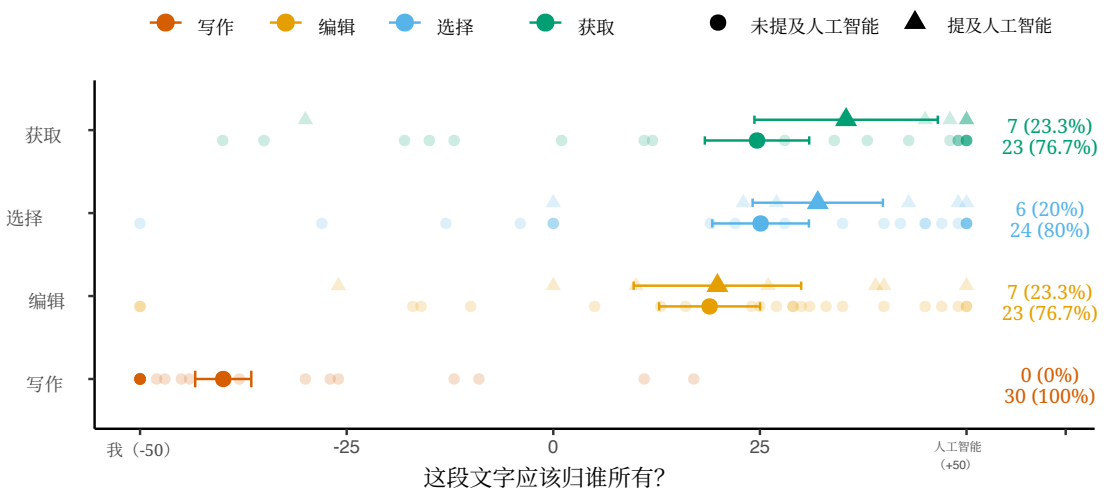


图 3. AI生成文本的所有权感（横轴）和作者身份声明（标记形状：圆形与三角形）随交互方式的变化。误差线表示 ± 一个标准误。右侧数字表示在我们的样本中（$n = 30$），按交互方式划分、将任何提及人工智能为作者的情况二值化后的作者身份声明数量。请注意，所有参与者都体验了所有交互方式。一位参与者在所有涉及AI生成文本的交互方式中，都将明信片的撰写者称为“随机”。这些情况被归入未提及人工智能类别。

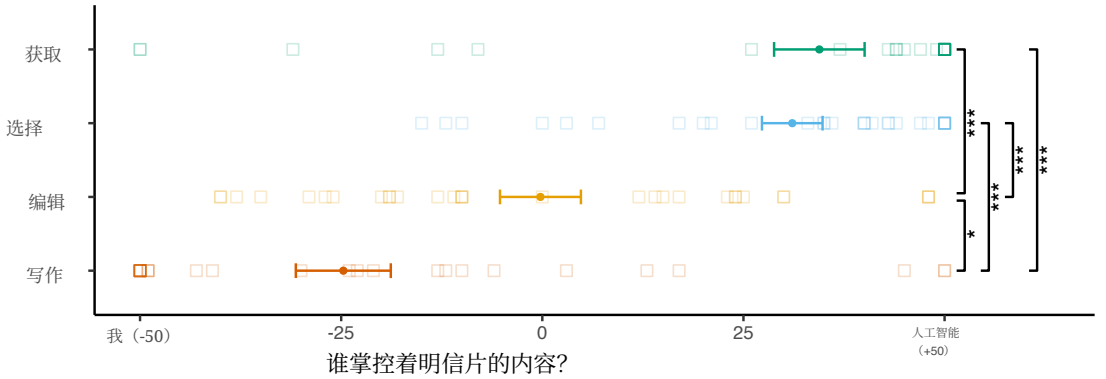


图 4. AI生成文本的控制感随交互方式的变化。误差线表示 ± 一个标准误。

边际 $R^2 = 55\%$）。与此前对作者身份的分析一样，我们发现交互方式对所有权感具有显著影响，$F(3, 86.02) = 28.00, p < .001$。控制感同样能够预测所有权感，$F(1, 101.55) = 6.47, p = .02$。该效应在统计上显著且为正向（$\beta = 0.21$，95% CI [0.05，0.37]，$t(109) = 2.57, p = 0.012$；$\beta_z = 0.20$，95% CI [0.05，0.36]）。关于控制感对所有权感影响的预测图及原始数据，请参见图5A。个性化的主效应以及个性化与交互方式的交互效应均不显著（两者均为 $p > .05$）。

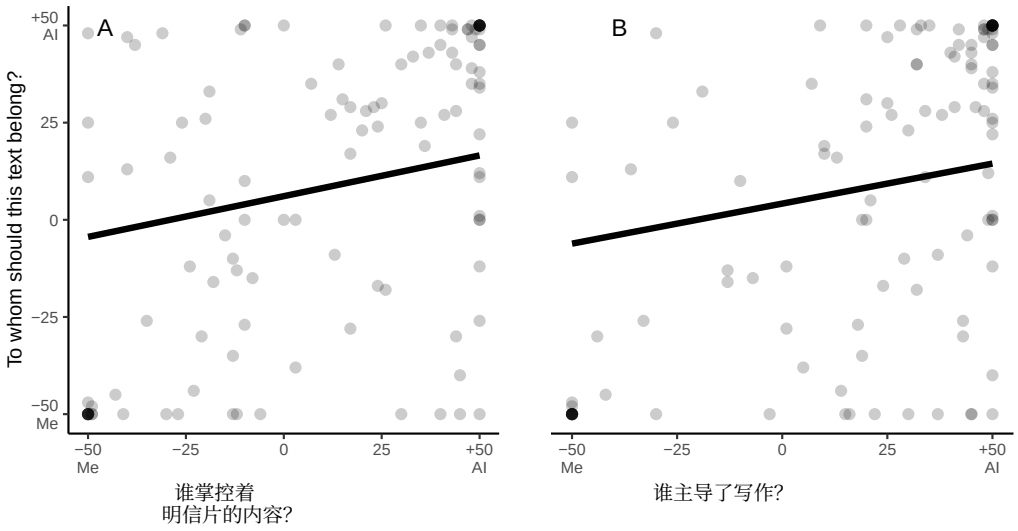


图 5. A：人工智能生成文本的控制感与所有权感的关系。B：人工智能生成文本的领导感与所有权感的关系。圆点表示原始数据，线条表示在不考虑交互方式影响的情况下，控制感对作者身份感的边际效应。

H2.2：领导感。参与者对“写作时是谁主导的？”这一问题的回答（同样使用视觉模拟量表作答，范围为 -50（我）至 +50 人工智能）通过混合方差分析模型[18]进行了分析。如表3所示，交互方式的效应显著，但个性化没有效应，也不存在交互效应。请注意，贝叶斯*t*检验（采用与控制感相同的模型）表明，假设个性化不存在差异的模型，比假设存在差异的模型可能性高 2.52 倍。事后检验显示，除获取与选择之间的差异外，所有交互方式之间的差异 $p > .99$均显著，参见图6。

我们再次通过拟合线性混合模型扩展方差分析，以根据领导感预测所有权感。我们发现结果模式相似。交互方式对所有权感有显著影响，$F(3, 84.91) = 23.71, p < .001$ 而个性化没有主效应或交互效应（两者均为 $p > .05$），但领导感有显著主效应，$F(3, 95.65) = 5.93, p = .016$。所有权感与领导感之间呈正相关（$\beta = 0.21$, 95% CI [0.04, 0.37], $t(109) = 2.46, p = 0.015$；$\beta_z = 0.19$，95% CI [0.04, 0.34]），另见图5B。模型的总体解释力相当可观（条件 $R^2 = 64\%$；边际 $R^2 = 55$）。

交互方法影响了对内容的控制感以及交互中的领导感。这些因素预测了所有权感，进而影响人们认为文本应归属于谁的感受（H2）。

[18]残差正态性假设再次被违反（$p = 0.03$）。因此，我们进行了非参数方差分析 [63]。与参数模型类似，交互方式产生了效应，$F(3,84) = 30.26, p < .05$。非参数事后检验的结果与参数分析相似。

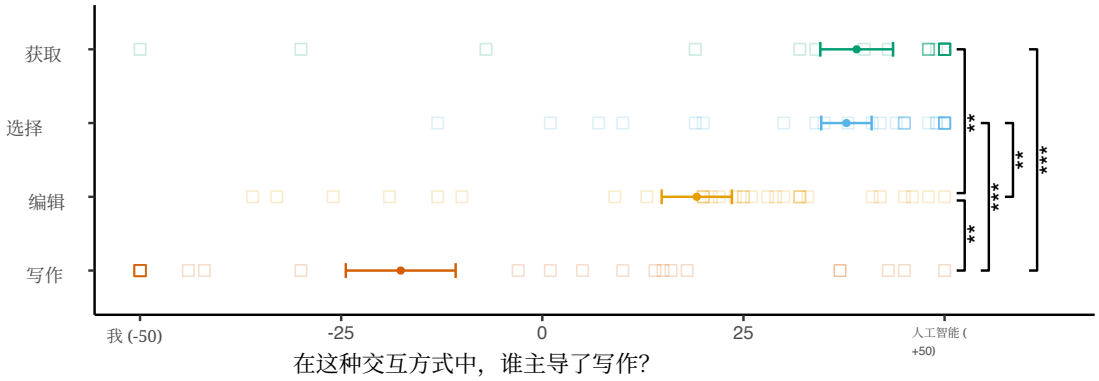


图 6. 人工智能生成文本的领导感随交互方式而变化。误差线表示 ± 一个均值标准误。

4.3 用户体验

我们使用带有 Greenhouse-Geisser 校正的重复测量方差分析模型，对 AttrakDiff 的所有子量表进行了分析。所有方差分析结果均见表3。

对于享乐质量—认同，个性化没有影响，但交互方式有显著影响。针对交互方式进行的Bonferroni校正的事后比较表明，与写作（$M = 4.78$、$SD = 0.88$）、选择（$M = 4.47$、$SD = 1.21$）和编辑（$M = 4.81$、$SD = 1.06$）相比，获取（$M = 3.74$、$SD = 1.28$）交互的感知更差（所有其他比较 $p > .05$）。个性化 × 交互方式的交互项未达到显著水平。因此，获取交互可视为自我关联性最低的交互方式。

对于享乐质量 – 刺激性的方差分析，我们发现了类似的结果模式：交互方法存在显著差异，但个性化以及个性化 × 交互方式之间的交互作用不存在显著差异。仅针对交互方式因素进行Bonferroni校正的比较表明，获取（$M = 3.91$、$SD = 1.33$）与编辑（$M = 4.81$、$SD = 1.06$）条件之间存在差异（所有其他比较 $p > .05$）。不出所料，因此认为获取交互的刺激性低于编辑方法。

吸引力分析呈现出相同的模式。同样，获取（$M = 3.99$、$SD = 1.66$）的吸引力显著低于写作（$M = 5.24$、$SD = 0.89$）、选择（$M = 4.94$、$SD = 1.39$）和编辑（$M = 5.13$、$SD = 1.38$；所有其他比较 $p > .05$）。

对实用性特征进行的方差分析显示，结果模式有所不同。个性化仍然没有显著影响，但交互方式产生了影响。不过需要注意的是，这一结果受到显著的个性化× 交互方式交互项的影响。对个性化 ×交互方式各组均值进行Bonferroni校正后发现，在未获得个性化文本的参与者中，获取条件下的实用性特征（$M = 4.32$, $SD = 1.04$）与选择条件（$M = 5.14$, $SD = 0.84$）显著不同，但与写作条件（$M = 5.70$, $SD = 0.74$）也有差异（其余比较 $p > .05$）；而在获得个性化文本的组中，情况亦然。这表明，尽管获取显然是交互模式时间效率最高的交互方式，但参与者主观上并不认为它特别实用。

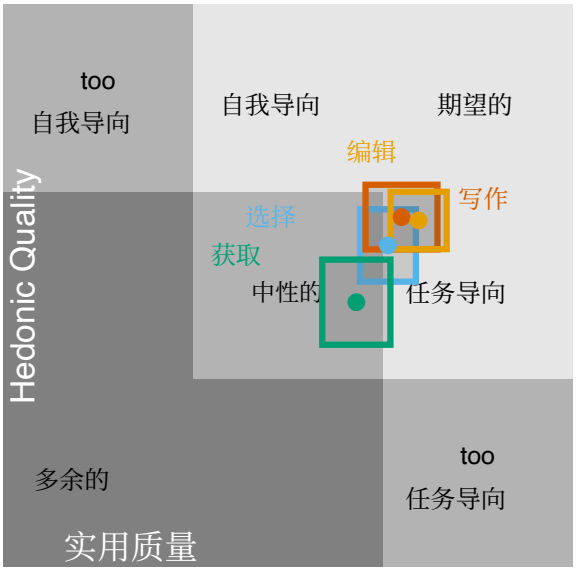


图 7. 研究1中各交互方式的 AttrakDiff 结果。图中显示了两个维度的汇总 AttrakDiff 分数的均值（圆点）和标准误（矩形）：实用质量 和享乐质量。

总体而言，图 7 显示，获取绝对而言是最中性的交互方式，且写作和编辑的用户体验相当。 选择与AI模型的交互最少，因此，相较于获取，可能会被评为更实用。

5 用户研究2：人类与AI代笔者的声明作者身份

通过研究1，我们证明，与自行撰写的文本相比，个性化AI生成文本会降低所有权感。这与作者身份声明形成了反差。在大多数情况下，参与者在线出版文本时，并未将人工智能系统声明为文本作者。我们将这一现象概括为人工智能代笔效应。我们发现，交互方式会影响人机交互中的控制感，而控制感的提升会增强所有权感。我们发现，个性化并未提升所有权感或控制感，也未在描述性层面上提升作者身份声明。因此，真实个性化并非人工智能代笔效应的必要条件。这一结果模式使我们能够针对个性化开展绿野仙踪研究。因此，在研究2中，我们将明信片描述为个性化明信片，但实际上随机向参与者分配非个性化文本。

在研究2中，我们针对先前研究的四项重要局限性进行了处理：首先，我们发现，参与者将作者身份感归因于人工智能，即使对于自行撰写的文本也是如此。这可能是由于我们的实验设计所致：参与者经历了所有交互方法，之后才被要求说明自己对四张明信片的所有权感。参与者可能将自行撰写的文本中出现的信息，带入了此前的人工智能生成文本中。为解决这一局限性，我们只采用一种交互方法。为了全面检验人工智能代笔效应，我们选择了获取，因为这种方法的所有权感最低，且参与者最常将人工智能提及为作者。其次，为了将任何作者身份归因于人工智能生成文本，参与者必须为人工智能想出一个标签。这导致一名参与者将该明信片标记为“随机”。在研究2中，我们向参与者提供六个

预先设定的作者身份标签，供他们选择，这一做法受到研究1的启发。第三，可以认为，不声明协作并非人机交互所特有，在人人交互中也会发生，即存在人工代笔人的情况下。因此，我们加入了一名虚构的人工代笔人。第四，较小的样本量无法准确估计将人工智能声明为作者的相对频率。更大规模的研究将能够更准确地估计这一效应。正如第3节中所述，我们假设能够复现人工智能代笔效应，并且该效应超过人人交互中的代笔效应。研究2于2022年12月开展。

5.1 方法

5.1.1 程序。 该研究由两部分组成：(1) 安慰剂个性化调查；(2) 与生成的明信片进行交互。在开始第1部分之前，我们向参与者说明了研究程序及其权利，并取得了他们的同意。

第*1*部分：个性化问题。 首先，我们要求参与者上传五篇他们过去撰写的、每篇至少500字的文本。我们告知他们，这些文本最好是非正式文本，例如个人电子邮件；他们可以删除或修改任何不想分享的细节。这一部分旨在让参与者相信系统会为他们生成文本。由于先前的研究显示个性化没有效果，因此这些文本实际上并未被使用。不过，参与者输入了一个姓名或化名，我们通过将所提供的姓名用作明信片中的问候语来实施安慰剂个性化。参与者还回答了有关其对人工智能态度的问题（参见 表7）。

第*2*部分：作者归属。 一天后，我们随机向参与者展示研究1中使用的15篇非个性化人工智能生成的明信片文本之一。我们告知他们，该文本要么由人类撰写，要么由人工智能代笔人撰写，并且文本是根据他们在第1部分中提供的信息进行个性化的。随后，他们看到一个用于上传生成文本的博客式上传表单，并被要求选择一张图片并提供元数据：标题、日期和作者归属。这次，作者从一个随机化列表中选择，选项包括人工智能、*GPT-3*、人类、*Sasha*（虚构代笔人的姓名）、其他 + 文本字段，以及第1部分中提供的参与者姓名、姓名首字母或化名 。上传后，参与者立即回答了有关其感知到的能动性以及他们对最终生成的明信片喜爱程度的问题。对于第二种代笔人类型，即人工智能或人类，也重复了这一过程。条件顺序经过平衡化处理。在与两种代笔人类型完成交互后，我们询问了参与者对作者归属的推理，并在研究参与结束后重新评估了他们对人工智能的态度。

5.1.2 实验设备与测量指标。本研究通过两份 Qualtrics 问卷实施。与研究1一样，我们测量了参与者的所有权感、控制感和主导权。此外，我们统计了人工智能被提及为作者的次数。所有记录的测量指标概览见表7。

5.1.3 参与者。 我们通过 Prolific 招募了107名英语母语者。由于未遵循说明，一名参与者在第1部分结束后被排除。其余参与者均受邀参加第2部分，其中100人参加。在这些参与者中，有4人未通过注意力检测。分析仅包括剩余的96名正确完成两部分的参与者。他们报告的性别为男性（53人）或女性（43人）；非二元性别、其他或未披露选项均无人选择。他们的年龄介于20岁和75岁之间（M = 36.3, SD = 13.3）。他们当前的居住国家包括英国（41人）、南非（25人）、加拿大（15人）、爱尔兰（5人），以及各有1至3人列出的墨西哥、澳大利亚、西班牙、以色列、葡萄牙和美国。与研究1一样，我们询问了他们使用文本生成的经历。55人表示此前使用过智能回复

表4. Welch 校正的 t检验结果，比较人类与人工智能代笔人的表现，并报告效应量的 Cohen's d。当正态性假设不成立时，我们还采用了非参数 Wilcoxon 检验。显著性标记：* 表示 $p < 0.05$

测量指标	Welch 校正的 t检验			Wilcoxon 检验		
	$t(95)$	p	d	W	p	r
所有权感：这段文字应该属于谁?	−2.05	**.043***	−0.21	1095	**.018***	.10
控制感：谁掌控了这张	0.28	.784,	0.03			
明信片的内容?						
领导感：谁主导了写作?	−1.10	.274	−0.11	876	**.021***	$r = .12$

功能，81人使用过单词或句子建议，63人使用过自动补全，86人使用过自动纠错。13名参与者有时会请他人代为写文本。整个研究大约耗时40分钟，参与者获得了5英镑的报酬。

5.2 结果

本节评估假设4—6，并进一步对控制感与主导权进行相关性分析。OSF 仓库[19] 提供了有关控制感、用户体验以及参与者对人工智能态度的测量指标的更多结果。

分析。下文使用配对t检验，分析 Ghostwriter 两组在所有权方面的差异。当达戈斯蒂诺检验 [24] 显示违反正态性假设时，我们还会计算非参数 Wilcoxon 检验。与研究1一样，我们使用 BayesFactor 软件包 [80] 计算贝叶斯因子，以评估与主要研究问题相关的零假设。我们使用 McNemar χ^2 检验，比较在被呈现为人类撰写的文本中添加人类作者标签（*Sasha*/人类）的频率，与在人工智能撰写的文本中添加人工智能作者标签（*GPT-3*/人工智能）的频率。我们采用斯皮尔曼秩相关，评估所有权感与作者身份声明之间的相关性（0 = 未提及 Ghostwriter，1 = 提及 Ghostwriter）。同样，显著性水平 α 设为5%。

假设*4*：所有权感的差异。我们比较了人类和人工智能条件下的所有权感，发现存在显著差异，参见表4。非参数 Wilcoxon 检验也显示显著差异（不符合正态性，$p < .01$）。与代笔者被描述为人类时相比，被描述为人工智能时，人们认为自己拥有更多所有权，见图8。

假设*5*：作者身份声明的差异。我们发现，作者身份声明存在差异（假设3）。总体而言，虽然人们更常将人工智能列为作者（研究2中约为71.3%，而研究1中最高为23.3%；我们将在第6.5节讨论这一点），但我们发现，他们提及人工智能的频率显著低于提及“Sasha”或任何其他非自身实体的频率，$\chi^2(1) = 31.92$，$p < .001$，参见表5。

这一趋势也体现在一个问题上：是否应该强制标注明确指出文本是在人工/人工智能的帮助下创作的。对于人类条件，63 名参与者（65.6%）回答是；而人工智能条件下，回答是的有 59 人（61.5%）（其余参与者均选择了否）。在说明人类贡献时，参与者列举的理由包括情感投入，例如“我们不想伤害他们的感情，对吧？”以及“如果换作是我，我希望自己的贡献能得到认可，”还有“这是他们的知识产权。”另一个论点则源于对……的感知真实性

[19] https://osf.io/n4svx/?view_only=916873e81d244be6be8e0790531b1197

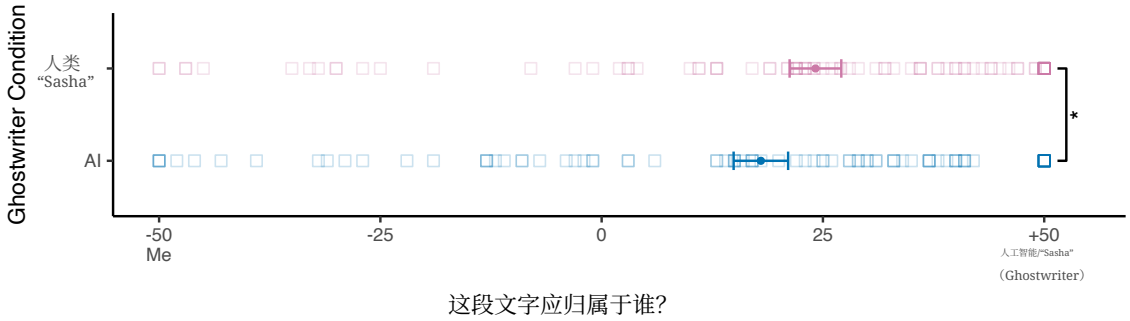


图 8. 所有权感随 Ghostwriter 的变化。误差线表示 ±一个均值标准误。

表 5. 各 Ghostwriter 条件下，对于是否提及人工智能/人类为作者，将声明的作者身份二值化后的计数。请注意，所有参与者均经历了所有代笔者条件。

	AI	人类
提及 Ghostwriter	71（73.9%）	79 (82.3%)
未提及 Ghostwriter	25 (26.0%)	17 (17.7%)

人工撰写的文本，例如，“如果这确实是一个人向你讲述其亲身经历，那就应该署上他们的名字”以及“因为它真实地体现了一个人的价值。”一位认为只有人工代笔者才应获得署名的参与者解释说，“这个人不是 *[一个]* 工具，而人工智能系统是一个可以 *[被]*使用的工具。”不署名感谢人工代笔人的理由，则提到了现有的代笔实践，例如，“以此为生的人”以及“这就相当于私人助理，他们不一定会署上自己的名字。”一位参与者再次提到了个性化：“我有点犹豫，但既然我给出了自己写作风格的例子，那就不用了。”

与研究1一样，参与者认为出于伦理和透明度方面的考虑，应对人工智能贡献予以说明，例如：“申明这一点似乎合乎伦理”，而且这“有助于*[s]* 让接收者了解背景情况。”此外，不申明贡献“可能会产生法律后果。”有趣的是，一些参与者还补充说，承认人工智能的贡献能够肯定其潜力，例如：“这也有助于树立人工智能的正面形象。”以及“这让人们了解人工智能已经发展到什么程度。”那些认为标注人工智能并非强制要求的人，常常将人工智能描述为一种工具，例如：“它表达的是所有者的感受和观点。它只是所有者使用的一种工具。”以及“人工智能是一种我们可以使用的工具，就像我们不会为了说明某项作品是自己创作的，而标注使用了 *Word* 或 *Google* 文档一样。”同样，有一人表示“人工智能是代表个人使用的。是否披露人工智能的使用情况，应由出版者决定。”语境也起到了一定作用，例如：“接收者可能会认为，寄出这张明信片时并没有投入多少心思。”

假设*6*：人工智能代笔效应。为评估假设6，我们计算了所有权感与将人工智能声明为作者之间的斯皮尔曼秩相关。所有权感与将人工智能声明为作者之间的联系具有显著性，但非常微弱，$r_s = .24$，$p = .017$。请注意，人工代笔人之间的联系同样如此，$r_s = .26$，$p = .011$。因此，所有权感与将人工智能声明为作者之间不存在直接的一对一对应关系，也不存在强相关，这支持了假设6。

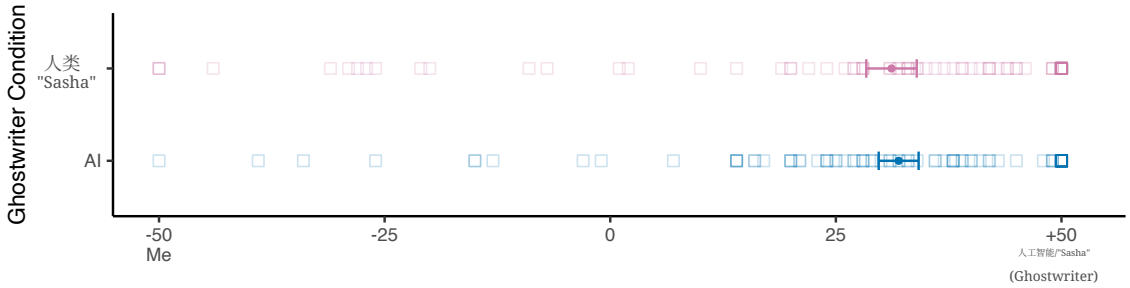


图9. 控制感随 Ghostwriter 的变化。误差线表示 ±一个均值标准误。

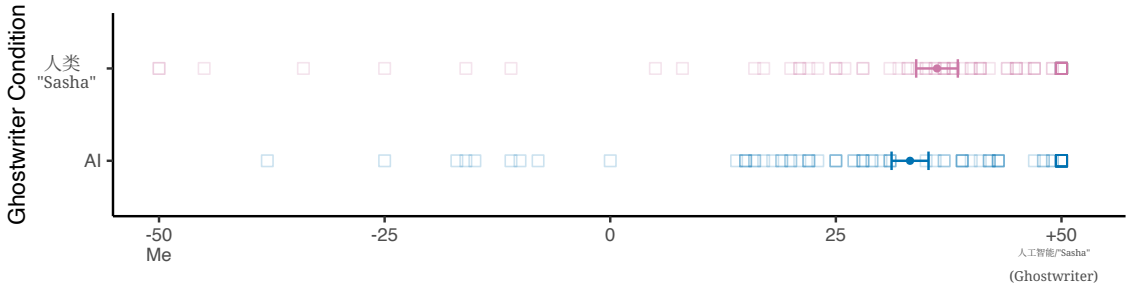


图10. 领导感随 Ghostwriter 的变化。误差线表示 ±一个均值标准误。

控制感与领导感。 我们发现，两种条件下的控制感没有差异[20]，参见 表4和 图9。对于领导感，我们发现，经过 Welch 校正的 t检验显示均值没有显著差异，参见 表4。然而，使用 Wilcoxon 检验比较秩时，发现了显著效应，这可能是由于数据存在较大的偏斜（正态性不满足， $p < .01$）。如 图10所示，参与者对被呈现为人工智能的代笔人的领导感更高。

6 讨论

我们的研究表明，在使用个性化 AI 文本生成时，所有权感与作者身份声明之间存在差异。这就是人工智能代笔效应。在文本生成情境中，人们通常会声明自己是作者，而不是为其生成文本的人工智能。即使他们觉得文本的所有权属于人工智能，情况也是如此。

研究1在明信片写作场景中初步揭示了人工智能代笔效应： 在参与者获得人工智能支持的所有交互方法中，他们往往认为文本应归人工智能所有（H1.1）。尽管如此，声明人工智能是明信片作者的参与者不到五分之一（H1.2）。研究还表明，控制感和主导权与其对文本生成的影响程度呈正相关；影响程度从接收不可编辑的人工智能文本（获取）到在没有人工智能支持的情况下撰写文本不等

[20]采用默认先验的贝叶斯 t检验表明，假设不存在差异的模型比假设存在差异的模型更有可能成立，可能性约为后者的8.5倍。

（写作；H2）。参与者收到个性化还是安慰剂个性化的生成文本，并不会影响其所有权感和作者身份声明（假设3）。因此，研究2中无需将个性化纳入自变量。

通过预注册研究2，我们得以在更大的样本中复现人工智能代笔效应，并将其与人工代笔进行比较。在绿野仙踪设计中，同一批安慰剂个性化文本被呈现为由“人工智能”生成，或由一位名叫“Sasha”的代笔人撰写。与收到人工智能撰写的文本相比，收到据称由人类撰写的文本的参与者，其所有权感显著较低。结合参与者陈述，这表明人工智能被视为一种工具，而非独立的共同作者（假设4）。与人们更愿意利用人工智能而非人类的倾向一致 [62], 人类代笔人也比人工智能代笔人更常获得署名（假设5）。研究2向参与者提供了一个多项选择列表，让他们从中选择，而非采用开放式作者声明。因此，他们可以立即看到，人工智能也可以被选为作者。尽管如此，仍有一些参与者没有承认人工智能代笔人或人工代笔人（假设6）。

我们的研究结果拓展了关于写作中人机协作的研究。具体而言，我们关注其对认知与行为的影响以及交互设计。这些结果也为人工智能文本生成界面的设计，以及拓展作者身份框架以支持文本生产，提供了启示。

6.1 写作中的认知与行为

当人工智能系统融入写作过程时，可能会影响文本生产中的认知与行为模式，例如控制感的变化，以及既有写作过程动态的改变；这些变化进而会影响用户的所有权和作者身份。

*参与度与心理所有权。*随着交互式人工智能引入自动化，对交互的影响力成为一个核心问题 [9, 94]。已有研究（例如 Lehmann 等人 [72], ）发现，感知到的作者身份会因对文本的控制程度而变化；例如，使用人工智能生成的短语撰写文本，以及引导人工智能自动写作时便是如此。我们通过更详细地考察作者身份和所有权、使文本生成个性化，并将研究范围从句子层面的交互拓展出去，延伸了这项工作。

请注意，在研究1中，影响程度并未直接影响所有权感；但所有权感会随对内容的控制感以及交互中的主导权感而变化，而这些又受到我们的交互方式影响。回归模型考虑了个体在控制感方面的差异，显示人们对于用户参与度在人机交互中的心智模型存在很大差异。这与Pierce等人 [84]的发现一致：所有权感会随着个人的主观参与度提高而增强。因此，要理解用户如何从所有权角度看待人工智能生成内容，关键的不仅是客观控制（例如由不同交互方式带来的控制），还有人机交互中对控制的主观体验。

*人工智能支持与写作认知。*在写作任务中加入文本建议工具，也会影响写作所涉及的认知过程。最初，广为采用的 Flower和Hayes模型 [35] 将写作描述为一个迭代过程，由不同粒度层次上的规划、转换和修订等子过程构成，且这些过程都发生在写作者内部。Hayes修订模型 [51] 对该模型进行了扩展和适配，例如引入了“任务环境”这一要素。该模型也承认技术在转录中的作用，但仍将其定位为写作者掌控的一种工具。Bhat 等人 [10] 认为，如今建议系统已超越这种被动角色。在

事实上，他们发现，即使建议并不反映人们自己的观点，建议系统仍会影响他们写下的内容（另见 [59]）。因此，他们提出对Hayes模型进行适配，在系统与迄今写成的文本之间建立相互联系。人工智能系统所扮演的主动角色也会影响作者身份：如果文本受到系统影响，作者还能否承担全部责任？

算法利用。 我们发现，与人人交互相比，人机交互中作者身份声明存在较大差距（研究2）。与人类写作者相比，我们的参与者较少提及人工智能支持。这与近期关于算法利用的研究结果一致。Karpus等人 [62]发现，尽管自动驾驶车辆的用户非常了解交通规则以及自动驾驶车辆在道路上的作用，但与人类驾驶员驾驶的汽车相比，驾驶员更倾向于利用自动驾驶汽车。因此，我们发现，就隐形代笔而言，算法利用同样存在于借助人工智能进行的文本生产中。然而，算法利用是如何形成的，以及如何减轻其影响，仍需进一步深入研究。

6.2 以感知到的作者身份为导向及与之协同的设计

人工智能代笔效应 对交互设计具有重要影响：人工智能写作工具的以人为中心的设计应将（感知到的）作者身份视为设计维度。下一步的关键是深入了解相关的用户界面特征。这些特征可能包括控制层级 [95], 写作过程中的角色 [72], 支持创作者的功能[39], 用于反思所生成文本的功能、人工智能的呈现方式（例如是否拟人化，同时考虑研究2中人工智能代笔人与人类代笔人的结果），以及用户在完成写作后如何归属作者身份（参见第6.3节）。

例如，正如我们在研究1中的用户界面对比所显示的那样，如果某一特定使用场景的设计目标是促进用户感知到的作者身份和控制感，那么用户界面至少需要支持对人工智能撰写的文本进行手动编辑。虽然 Getting 让模型完全掌控文本写作，但提供编辑选项后，感知控制和主导权便转向了用户，描述性分析还显示，感知所有权也略有提升。除本文所研究的用户界面外，如果与编辑功能结合，展示多个建议也可能影响这种感知，因为这样一来，用户总体上便能表达更多“编辑决策”。更广泛地说，可以研究功能丰富的图形用户界面组件，让用户能够提前或间歇性地对人工智能做出决策：例如，图形用户界面可以提供选项选择器，以调整文本的情感倾向、范围或长度（参见*Grammarly*[21]的用户界面）。未来的研究可以考察，对生成过程（而非最终结果）的此类控制，会如何影响本文观察到的效果。

6.3 作者身份声明的用户界面设计启示

除了人工智能辅助写作的用户界面设计启示外，我们的研究结果也促使我们更深入地审视如何设计用于归属作者身份的用户界面。在研究中，我们分别使用了自由文本字段（研究1）和下拉列表（研究2），但设计空间要广阔得多：例如，假设设计目标是让人工智能文本生成的使用情况更加透明。在这种情况下，我们或许可以借鉴早期移动电子邮件中的“免责声明”（例如“随时随地撰写，可能含有错别字”，或仅仅是“从我的 iPhone 发送”）。类似地，作者可以在文末添加“使用模型X撰写”之类的注释，从而揭示整篇文本受到的人工智能影响。

在更深层次上，例如在句子层面，遵循信息可视化中“按需细节”理念的交互 [93] 或许能揭示更多信息：例如，新闻文章通常会列出作者姓名。可以对其进行扩展，使得点击某个姓名时，突出显示由该人撰写的文本部分。“人工智能生成”届时可能会成为

[21] https://app.grammarly.com/

除人类姓名之外的另一种选项。请注意，这并不一定意味着人工智能被呈现为与人类作者平起平坐。因此，在写作过程中声明人工智能时，应基于用户的心智模型进行设计。

6.4 拓展人机协作的贡献框架

我们发现，当作者被明确告知可以选择添加人工智能（研究2）时，他们比没有这一选项时（研究1）更有可能这样做。与此相呼应，近期出版物中的研究人员也在为如何报告文本生成而苦恼 [70]，相关政策从禁止文本生成到强制标注人工智能文本生成不等 [33]。我们的数据并不支持这种简单的解决方案。例如，我们发现，使用 AI模型时，感知所有权会随主观控制感而显著变化。因此，可能需要采用更连续的人工智能声明方式，以契合写作者的心智模型。用于界定作者贡献角色的 CRediT 分类法 [2, 3, 13] 旨在细致说明作者在科学论文撰写不同阶段的参与情况（例如，A 撰写初稿，B 负责监督研究）。调整 CRediT 等框架以适用于算法的贡献，可以从立足于常见实践的人本视角，丰富相关讨论。

从更广泛的角度来看，人们认为此类作者身份声明有多重要，可能会因文化以及围绕人工智能使用而新兴的社会规范和期望而有所不同。未来进一步研究这些方面，或可将本文的工作与人工智能介导的沟通及此类技术相关的社会与文化维度等研究方向联系起来（例如， [42, 47, 78]）。

例如，在纸质个人通信中，手写字迹通常标示着作者身份——即便只是写在预印贺卡上的一则手写便条。明信片应用等服务的出现，使寄件人输入文本后由服务打印并寄出（例如，MyPostcard[22]），这意味着接收者已经无法完全追溯作者身份。我们预计，寄件人将越来越多地借助人工智能支持来生成此类文本。这些不断变化的做法将引发关于个人通信情境中作者身份声明与验证的讨论。

6.5 局限性

在解读研究结果时，需要考虑情境、个性化程度以及研究中所用界面所带来的若干局限。首先，写明信片这一情境可能会被认为有些人为且简单；理想情况下，未来研究应在其他需要撰写更长、更细腻文本的写作任务中重现人工智能代笔效应 。不过，先前研究通常采用学术写作任务（例如， [1, 81]），而在这类任务中，个性化写作可能不那么重要，且其情境未必能推广到作者身份问题上——因为学术环境中已有相应的惯例和规则。与现实生活中书写明信片不同，本研究中的明信片并未寄给任何人，而是在线上发布的。参与者写的也不是他们实际经历过的旅行。这可能影响了参与者决定是否声明作者身份的方式。尽管如此，我们认为，明信片没有实际寄出，反而提高了参与者声明人工智能为作者的可能性，因为个人利害关系较低。

其次，在研究1中，我们仅为一半的参与者对生成文本进行了个性化处理，而在研究2中则完全没有进行个性化处理。此外，用于微调的提示—补全对数量低于 OpenAI[23]建议的数量。因此，有人可能会断言，真实个性化可能会产生不同的结果。然而，在所有权感和声明的作者身份方面，我们没有发现真实个性化与安慰剂个性化之间存在差异。这表明，人工智能代笔效应并不取决于个性化质量，而只取决于参与者能否与生成的文本产生联系。同样，一些参与者对个性化程度并不完全满意，

[22]https://www.mypostcard.com/en/ [23]
https://platform.openai.com/docs/guides/fine-tuning

不过，随着人工智能生成文本的个性化质量提高，人工智能代笔效应的规模应会进一步扩大。因此，我们认为，选择安慰剂个性化并不会限制我们的结果，反而说明该效应的存在并不取决于个性化质量。

第三，我们发现，两项研究中的作者声明存在显著差异。在研究1中，只有约20%的参与者在明信片下方的自由文本字段中提到了人工智能。在研究2中，约70%的参与者在选择菜单中进行选择时提到了人工智能。未来的研究可以深入探讨如何设计作者身份声明界面，以便申明人工智能系统的贡献。此外，未来的研究还可以考察，在撰写科学论文时，GPT或其他大语言模型是否可能被披露为所使用的工具，例如在论文的方法部分中。

第四，我们的研究只能描述人们实际上如何声明个性化AI生成文本的作者身份。认知科学与人机交互领域的跨学科合作或许能够制定规范性框架，建议人们在人工智能的辅助下应该如何声明作者身份；然而，人们将如何运用这些规范性框架，仍有待探究。

第五，为求简洁，本研究分析的数据纯属定量数据，因此只能描述个别的判断、印象和行为。若采用更具定性的写作研究方法，遵循 Bhat 等人的做法，[10], 或许能更细致地呈现参与者在为个性化文本的写作贡献中，决定是否认可人工智能时所依据的推理。

第六，我们聚焦于（安慰剂式）个性化AI生成文本，因为它们反映了代笔过程。因此，我们的结果可能存在偏差，更倾向于将作者身份归于自己。对于非个性化人工智能写作情境中的交互是否也存在这种差异，仍有待研究。

7 结论与展望

基于两项实证研究（共126名参与者）的发现揭示了人工智能代笔效应；我们表明，人们在写作中使用个性化AI生成文本时，尽管会将所有权归于人工智能，却往往不会将人工智能披露为作者。这种效应不受交互方式影响；将开放式作者声明字段改为预设回复建议后，该效应有所减弱。比较人类与AI代笔者后，我们发现，在人工智能支持下写作时，人们更常将作者身份归于自己。鉴于感知所有权与作者声明之间存在差异，我们呼吁开发能够将用户及其与生成模型之间的关系纳入考量的作者身份归属框架。

参考文献

[1] Liz Allen、Alison O'Connell 和 Veronique Kiermer。2019年。我们如何确保研究贡献的可见性和多样性？贡献者角色分类法（CRediT）如何助力从作者身份向贡献者身份的转变。Learned Publishing 32，第1期（2019年1月），71–74。 https://doi.org/10.1002/leap.1210[2] Liz Allen、Alison O'Connell 和 Veronique Kiermer。2019年。我们如何确保研究贡献的可见性和多样性？贡献者角色分类法（CRediT）如何助力从作者身份向贡献者身份的转变。Learned Publishing 32，第1期（2019年1月），71–74。https://doi.org/10.1002/leap.1210[3] Liz Allen、Jo Scott、Amy Brand、Marjorie Hlava 和 Micah Altman。2014年。出版：应得的署名。Nature 508，第7496期（2014年4月），312–313。 https://doi.org/10.1038/508312a[4] Kenneth C. Arnold、Krzysztof Z. Gajos 和 Adam T. Kalai。2016年。在移动文本创作中，短语建议与词语预测之比较。载于第29届用户界面软件与技术年度研讨会论文集。ACM，日本东京，603–608。 https://doi.org/10.1145/2984511.2984584[5] Clark D. Asay。2020年。人工智能世界中的独立创作。FIU Law Review 14，第2期（2020年1月），201–222。 https://doi.org/10.25148/lawrev.14.2.5[6] Rosa Maria Ballardini。2019年。人工智能生成内容：人工智能时代的作者身份与发明人身份。载于《欧盟在线内容分发》。Edward Elgar Publishing，117–135。 https://doi.org/10.4337/9781788119900.00015

[7] Christoph Bartneck、Dana Kulic 和 Elizabeth Croft。2017年。测量机器人的拟人化、生命性、可喜欢性、感知智能和感知安全性。（2017年），311700 字节。https://doi.org/10.6084/M9.FIGSHARE.5154805 作品大小：311700 字节。出版商：figshare。[8] Ac Bd、Christine Bauer 和 Afsaneh Doryab。2016年。解决第一作者身份之争：利用交互技术突出贡献。载于2016年CHI计算系统中的人为因素会议扩展摘要论文集。ACM，美国加利福尼亚州圣何塞，609–620。https://doi.org/10.1145/2851581.2892582[9] Joanna Bergström、Jarrod Knibbe、Henning Pohl 和 Kasper Hornbæk。2022年。能动感与用户体验：二者是否存在关联？ACM计算机与人类交互汇刊 29，第4期（2022年8月），1–22。https://doi.org/10.1145/3490493[10] Advait Bhat、Saaket Agashe、Niharika Mohile、Parth Oberoi、Ravi Jangir 和 Anirudha Joshi。2022年。研究作者—建议交互：一项旨在理解作者如何与对齐／未对齐的下一短语建议交互的定性研究。https://doi.org/10.48550/arXiv.2208.00636 arXiv:2208。00636 [cs]。[11] Xiaojun Bi、Tom Ouyang 和 Shumin Zhai。2014年。既完整又正确？触摸屏键盘的多目标优化。载于SIGCHI计算系统中的人为因素会议论文集（加拿大安大略省多伦多）（CHI '14）。计算机协会，美国纽约州纽约，2297–2306。https://doi.org/10.1145/2556288.2557414[12] Oloff C. Biermann、Ning F. Ma 和 Dongwook Yoon。2022年。从工具到伙伴：故事作者希望人工智能写作者尊重其个人价值观和写作策略。载于交互系统设计会议。ACM，澳大利亚线上活动，1209–1227。https://doi.org/10.1145/3532106.3533506[13] Amy Brand、Liz Allen、Micah Altman、Marjorie Hlava 和 Jo Scott。2015年。超越作者身份：归因、贡献、协作与署名。Learned Publishing 28，第2期（2015年4月），151–155。https://doi.org/10.1087/20150211[14] Deborah Brandt。2007年。"谁是总统？"：读写能力中的代笔与价值观变迁。College English 69，第6期（2007年），549–571。出版商：JSTOR。[15] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习器。载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M. F. Balcan 和 H. Lin（编），第33卷。Curran Associates, Inc.，1877–1901。https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf[16] Daniel Buschek、Martin Zürn 和 Malin Eiband。2021年。多个并行短语建议对英语母语写作者和非英语母语写作者电子邮件输入及创作行为的影响。载于2021年CHI计算系统中的人为因素会议论文集（日本横滨）（CHI '21）。计算机协会，美国纽约州纽约，第732篇文章，13页。https://doi.org/10.1145/3411764.3445372[17] Ana Caraban、Evangelos Karapanos、Daniel Gonçalves 和 Pedro Campos。2019年。23种助推方式：人机交互中技术介导助推的综述。载于2019年CHI计算系统中的人为因素会议论文集（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约，1–15。https://doi.org/10.1145/3290605.3300733[18] John Carvalho、Angie Chung 和 Michael Koliska。2021年。挑战透明度：从爵士时代到社交媒体的代笔。Journalism 22，第3期（2021年3月），709–725。https://doi.org/10.1177/1464884918804700[19] Mia Xu Chen、Benjamin N. Lee、Gagan Bansal、Yuan Cao、Shuyuan Zhang、Justin Lu、Jackie Tsay、Yinan Wang、Andrew M. Dai、Zhifeng Chen、Timothy Sohn 和 Yonghui Wu。2019年。Gmail 智能撰写：实时辅助写作。载于第25届ACM SIGKDD知识发现与数据挖掘国际会议论文集。ACM，美国阿拉斯加州安克雷奇，2287–2295。https://doi.org/10.1145/3292500.3330723[20] Hyerim Cho、Chris Hubbles 和 Heather Moulaison-Sandy。2022年。对电子游戏责任主体的探讨：编目实践、用户需求和作者身份理论研究。Journal of Documentation 78，第6期（2022年10月），1420–1436。https://doi.org/10.1108/JD-10-2021-0198[21] John Joon Young Chung、Shiqing He 和 Eytan Adar。2021年。创意支持工具中用户、角色、交互和技术的交汇。载于2021年交互系统设计会议。ACM，美国线上活动，1817–1833。https://doi.org/10.1145/3461778.3462050[22] John Joon Young Chung、Wooseok Kim、Kang Min Yoo、Hwaran Lee、Eytan Adar 和 Minsuk Chang。2022年。TaleBrush：使用生成式预训练语言模型勾画故事。载于2022年CHI计算系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约，第209篇文章，19页。https://doi.org/10.1145/3491102.3501819[23] Larry D. Claxton。2005年。科学作者身份。Mutation Research/Reviews in Mutation Research 589，第1期（2005年1月），31–45。https://doi.org/10.1016/j.mrrev.2004.07.002[24] Ralph B. D'Agostino。1971年。适用于中等和大样本的正态性综合检验。Biometrika 58，第2期（1971年），341–348。https: //doi.org/10.1093/biomet/58.2.341[25] Robert Dale。2021年。GPT-3：它有什么用？Natural Language Engineering 27，第1期（2021年1月），113–118。https://doi.org/10.1017/S1351324920000601[26] Hai Dang、Karim Benharrak、Florian Lehmann 和 Daniel Buschek。2022年。超越文本生成：利用连续自动文本摘要支持作者。载于第35届ACM用户界面软件与技术年度研讨会论文集（美国俄勒冈州本德）（UIST '22）。计算机协会，美国纽约州纽约，第98篇文章，13页。https://doi.org/10.1145/3526113.3545672[27] N Dehouche。2021年。大规模生成式预训练变换器（GPT-3）时代的剽窃。Ethics in Science and Environmental Politics 21（2021年3月），17–23。https://doi.org/10.3354/esep00195[28] Nicholas Diakopoulos、Kurt Luther、Yevgeniy (Eugene) Medynskiy 和 Irfan Essa。2007年。混音社会中作者身份的演变。载于第18届超文本与超媒体会议论文集——HT '07。ACM Press，英国曼彻斯特，133。https://doi.org/10.1145/1286240。

1286272[29] Shiran Dudy、Steven Bedrick和Bonnie Webber。2021年。重新聚焦相关性：自然语言生成中的个性化。载于《2021年自然语言处理实证方法会议论文集》。计算语言学协会，线上及多米尼加共和国蓬塔卡纳，5190–5202。https://doi.org/10.18653/v1/2021.emnlp-main.421[30] Justin Edwards、Allison Perrone和Philip R. Doyle。2020年。语言生成的透明度：自动化水平。载于《第二届对话式用户界面会议论文集》。ACM，西班牙毕尔巴鄂，1–3。https://doi.org/10.1145/3405755.3406136[31] Lisa A. Elkin、Matthew Kay、James J. Higgins和Jacob O. Wobbrock。2021年。用于多因素对比检验的对齐秩变换程序。载于《第34届美国计算机协会用户界面软件与技术年度研讨会》。ACM，美国线上活动，754–768。https://doi.org/10.1145/ 3472749.3474784[32] Katherine Elkins和Jon Chun。2020年。GPT-3能通过作家的图灵测试吗？《文化分析期刊》5，2（2020年9月），16。https: //doi.org/10.22148/001c.17212[33] Annette Flanagin、Kirsten Bibbins-Domingo、Michael Berkwits和Stacy L. Christiansen。2023年。非人类“作者”及其对科学出版诚信和医学知识的影响。《美国医学会杂志》329，8（2023年2月），637–639。https://doi.org/10.1001/jama.2023.1344 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2801170/jama_flanagin_2023_ed_230004_1676659370.84192.pdf[34] Luciano Floridi和Massimo Chiriatti。2020年。GPT-3：其性质、范围、局限与后果。《心灵与机器》30，4（2020年12月），681–694。https://doi.org/10.1007/s11023-020-09548-1[35] Linda Flower和John R. Hayes。1981年。写作的认知过程理论。《大学写作与交流》32，4（1981年12月），365。https://doi.org/10.2307/356600[36] Andrew Fowler、Kurt Partridge、Ciprian Chelba、Xiaojun Bi、Tom Ouyang和Shumin Zhai。2015年。语言建模及其个性化对触摸屏打字表现的影响。载于《第33届美国计算机协会计算系统中的人为因素年度会议论文集》（韩国首尔）（CHI ’15）。计算机协会，美国纽约州纽约，美国，649–658。https://doi.org/10.1145/2702123.2702503[37] Giorgio Franceschelli和Mirco Musolesi。2022年。生成式深度学习中的版权。《数据与政策》4（2022年），e17。https://doi.org/10.1017/dap. 2022.10[38] Tiffany Derville Gallicano、Kevin Brett和Toby Hopp。2013年。代笔博客是否类似于演讲稿撰写？从业者关于代笔博客伦理的调查。《公共关系期刊》7，3（2013年），1–41。[39] Katy Ilonka Gero和Lydia B. Chilton。2019年。Metaphoria：用于隐喻创作的算法助手。载于《2019年CHI计算系统中的人为因素会议论文集》（苏格兰格拉斯哥）（CHI ’19）。计算机协会，美国纽约州纽约，美国，1–12。https://doi.org/10.1145/3290605.3300526[40] Maliheh Ghajargar、Jeffrey Bardzell和Love Lagerkvist。2022年。一个红发女子走进酒吧：使用人工智能创作小说的体验。载于《第25届国际学术Mindtrek会议》。ACM，芬兰坦佩雷，230–241。https://doi.org/10.1145/3569219.3569418[41] Gene V Glass、Percy D. Peckham和James R. Sanders。1972年。未满足固定效应方差与协方差分析所依据假设的后果。《教育研究评论》42，3（1972年9月），237–288。https://doi.org/10.3102/00346543042003237[42] Emma Goldenthal、Jennifer Park、Sunny X. Liu、HannahMieczkowski和Jeffrey T.Hancock。2021年。并非所有人工智能都平等：探索人工智能介导的沟通技术的可及性。《人类行为中的计算机》125（2021年），106975。https://doi.org/10.1016/j.chb.2021.106975[43] Tanya Goyal、Junyi Jessy Li和Greg Durrett。2022年。GPT-3时代的新闻摘要与评估。（2022年）。https://doi.org/10.48550/ ARXIV.2209.12356 出版者：arXiv 版本号：1。[44] Andreas Graefe。2016年。《自动化新闻指南》。（2016年）。https://doi.org/10.7916/D80G3XDJ 出版者：哥伦比亚大学。[45] Peter C Gøtzsche、Asbjørn Hróbjartsson、Helle Krogh Johansen、Mette T Haahr、Douglas G Altman和An-Wen Chan。2007年。行业发起的随机试验中的隐名作者。《PLOS医学》4，1（2007年1月），e19。https://doi.org/10.1371/journal.pmed.0040019[46] Peter C Gøtzsche、Jerome P Kassirer、Karen L Woolley、Elizabeth Wager、Adam Jacobs、Art Gertel和Cindy Hamilton。2009年。应采取何种措施应对医学文献中的代笔？《PLOS医学》6，2（2009年2月），e1000023。https://doi.org/10.1371/journal.pmed.1000023[47]Jeffrey T Hancock、Mor Naaman和Karen Levy。2020年。人工智能介导的沟通：定义、研究议程与伦理考量。《计算机媒介传播期刊》25，1（2020年1月），89–100。https://doi.org/10.1093/jcmc/zmz022 arXiv:https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf[48] Michael R. Harwell、Elaine N. Rubinstein、William S. Hayes和Corley C. Olds。1992年。方法学研究中的蒙特卡罗结果汇总：单因素与双因素固定效应方差分析情形。《教育统计学期刊》17，4（1992年12月），315–339。https://doi.org/10. 3102/10769986017004315[49] Marc Hassenzahl、Michael Burmester和Franz Koller。2003年。AttrakDiff：感知享乐与实用质量测量问卷。载于《人类与计算机2003》，Gerd Szwillus和Jürgen Ziegler（编）。第57卷。Vieweg+Teubner Verlag，威斯巴登，187–196。https://doi.org/10.1007/978-3-322-80058-9_19 丛书名称：德国计算机协会分会报告。[50] Marc Hassenzahl和Andrew Monk。2010年。从美感推断感知可用性。《人机交互》25，3（2010年7月），235–260。https://doi.org/10.1080/07370024.2010.500139[51] John R. Hayes。2012年。写作的建模与再建模。《书面交流》29，3（2012年7月），369–388。https://doi.org/10.1177/ 0741088312451260 出版者：SAGE出版公司。

[52] D. Jeffery Higginbotham。1992年。五种辅助沟通技术中的击键节省评估。《Augmentative and Alternative Communication》8，4（1992年1月），258–272。https://doi.org/10.1080/07434619212331276303[53] Jess Hohenstein和Malte Jung。2020年。人工智能作为道德缓冲区：人工智能介导的沟通对归因和信任的影响。《人类行为中的计算机》106（2020年），106190。https://doi.org/10.1016/j.chb.2019.106190[54] Joo-Wha Hong、Ignacio Cruz和Dmitri Williams。2021年。人工智能，你能驾驶我的车吗：我们如何评价人类驾驶员与自动驾驶汽车。《人类行为中的计算机》125（2021年12月），106944。https://doi.org/10.1016/j.chb.2021.106944[55] Joo Wha Hong、Qiyao Peng和Dmitri Williams。2021年。你准备好迎接人工智能莫扎特和Skrillex了吗？一项检验期望违背理论与AI音乐的实验。《New Media & Society》23，7（2021年7月），1920–1935。https://doi.org/10.1177/1461444820925798[56] Anna Y.Q. Huang、Owen H.T. Lu和Stephen J.H. Yang。2023年。人工智能驱动的个性化推荐对翻转课堂中学习者学习投入、动机和学习成果的影响。《Computers & Education》194（2023年3月），104684。https: //doi.org/10.1016/j.compedu.2022.104684[57] Cheng-Zhi Anna Huang、Ashish Vaswani、Jakob Uszkoreit、Noam Shazeer、Ian Simon、Curtis Hawthorne、Andrew M. Dai、Matthew D. Hoffman、Monica Dinculescu和Douglas Eck。2018年。Music Transformer。（2018年）。https://doi.org/10.48550/ARXIV.1809.04281 出版者：arXiv 版本号：3。[58] Jani Ihalainen。2018年。计算机创造力：人工智能与版权。《Journal of Intellectual Property Law & Practice》13，9（2018年9月），724–728。https://doi.org/10.1093/jiplp/jpy031[59] Maurice Jakesch、Advait Bhat、Daniel Buschek、Lior Zalmanson和Mor Naaman。2023年。与带有观点的语言模型共同写作会影响用户的看法。https://doi.org/10.1145/3544548.3581196 arXiv:2302.00560 [cs].[60] Maurice Jakesch、Megan French、Xiao Ma、Jeffrey T. Hancock和Mor Naaman。2019年。人工智能介导的沟通：个人资料文本由人工智能撰写这一感知如何影响可信度。载于《2019年CHI计算系统中的人为因素会议论文集》（苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约，美国，1–13。https://doi.org/10.1145/3290605.3300469[61] Anjuli Kannan、Karol Kurach、Sujith Ravi、Tobias Kaufmann、Andrew Tomkins、Balint Miklos、Greg Corrado、Laszlo Lukacs、Marina Ganea、Peter Young和Vivek Ramavajjala。2016年。智能回复：电子邮件自动回复建议。载于《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》（美国加利福尼亚州旧金山）（KDD '16）。计算机协会，美国纽约州纽约，美国，955–964。https://doi.org/10.1145/2939672.2939801[62] Jurgis Karpus、Adrian Krüger、Julia Tovar Verba、Bahador Bahrami和Ophelia Deroy。2021年。算法利用：人类热衷于利用善意的人工智能。《iScience》24，6（2021年6月），102679。https://doi.org/10.1016/j.isci.2021.102679[63] Matthew Kay、Lisa A. Elkin、James J. Higgins和Jacob O. Wobbrock。2021年。mjskay/ARTool：ARTool 0.11.0。https://doi.org/10.5281/ZENODO.594511[64] Tom Kent。2019年。机器人新闻的伦理检查清单。https://medium.com/@tjrkent/an-ethical-checklist-for-robot-journalism-1f41dcbd7be2[65] Dongwhan Kim和Joonhwan Lee。2019年。为个性化和交互式新闻阅读设计算法驱动的文本生成系统。《International Journal of Human–Computer Interaction》35，2（2019年1月），109–122。https://doi.org/10.1080/10447318.2018.1437864[66] Sangmi Kim、Seong-Gyu Kim、Yoonsin Jeon、Soojin Jun和Jinwoo Kim。2016年。合适还是混搭？社会认可和心理所有权对在线社区分享意图的影响。《人机交互》31，2（2016年3月），97–132。https: //doi.org/10.1080/07370024.2015.1022425[67] Joanna Kisker、Thomas Gruber和Benjamin Schöne。2021年。虚拟现实体验促进自传体记忆提取机制：实验室体验与虚拟体验的电生理学相关性。《Psychological Research》85，7（2021年10月），2485–2501。https://doi.org/10.1007/s00426- 020-01417-x[68] Thomas Kosch、Robin Welsch、Lewis Chuang和Albrecht Schmidt。2023年。人机交互中的人工智能安慰剂效应。《ACM计算机与人类交互汇刊》29，6，文章56（2023年1月），32页。https://doi.org/10.1145/3529225[69] Per Ola Kristensson和Keith Vertanen。2014年。无黏性文本输入速率及其作为移动文本输入终极目标的应用。载于《第16届人机交互移动设备与服务国际会议论文集》（MobileHCI '14）。计算机协会，加拿大安大略省多伦多，335–338。https://doi.org/10.1145/2628363.2628405[70] Tiffany H. Kung、Morgan Cheatham、ChatGPT、Arielle Medenilla、Czarina Sillos、Lorie De Leon、Camille Elepaño、Maria Madriaga、Rimel Aggabao、Giezel Diaz-Candido、James Maningo和Victor Tseng。2022年。ChatGPT在美国执业医师资格考试中的表现：使用大语言模型开展AI辅助医学教育的潜力。预印本。《Medical Education》。https://doi.org/10.1101/2022.12.19.22283643[71] Mina Lee、Percy Liang和Qian Yang。2022年。CoAuthor：设计人机协作写作数据集以探索语言模型能力。载于《CHI计算机系统中的人为因素会议》。ACM，美国路易斯安那州新奥尔良，1–19。https://doi.org/10.1145/3491102. 3502030[72] Florian Lehmann、Niklas Markert、Hai Dang和Daniel Buschek。2022年。建议列表与连续生成：在移动设备上使用生成模型进行写作的交互设计会影响文本长度、措辞和感知到的作者身份。载于《Mensch und Computer 2022》。ACM，德国达姆施塔特，192–208。https://doi.org/10.1145/3543758.3543947[73] Angelica Lermann Henestrosa、Hannah Greving和Joachim Kimmerle。2023年。自动化新闻：人工智能作者身份和评价性信息对科学新闻文章感知的影响。《人类行为中的计算机》138（2023年1月），107445。https://doi.org/10.1016/j. chb.2022.107445[74] Lisa Lines。2016年。代笔人保证成绩？澳大利亚高等教育学生可获得的在线代写服务质量。《Teaching in Higher Education》21，8（2016年11月），889–914。https://doi.org/10.1080/13562517.2016.1198759

[75] Jiachang Liu、Dinghan Shen、Yizhe Zhang、Bill Dolan、Lawrence Carin 和 Weizhu Chen。2021年。GPT-3 的优质上下文示例具备哪些特点？（2021年）。https://doi.org/10.48550/ARXIV.2101.06804 出版方：arXiv 版本号：1。[76] Li Lucy 和 David Bamman。2021年。GPT-3 生成故事中的性别与表征偏差。载于《第三届叙事理解研讨会论文集》。计算语言学协会，虚拟会议，48–55。https://doi.org/10.18653/v1/2021.nuse-1.5[77] Hannah Mieczkowski 和 Jeffrey Hancock。2022年。探究 AI 介导的沟通中人工智能系统的能动性、专业知识与角色。https://doi.org/10.31219/osf.io/asnv4[78] Hannah Mieczkowski、Jeffrey T. Hancock、Mor Naaman、Malte Jung 和 Jess Hohenstein。2021年。AI介导的沟通：指称通信任务中的语言使用与人际效应。《ACM 人机交互会议论文集》5，CSCW1，文章 17（2021年4月），14 页。https://doi.org/10.1145/3449091[79] Tal Montal 和 Zvi Reich。2017年。我，机器人。你，记者。谁是作者？自动化新闻中的作者身份、署名栏与全面披露。《数字新闻学》5，7（2017年8月），829–849。https://doi.org/10.1080/21670811.2016.1209083[80] Richard D. Morey 和 Jeffrey N. Rouder。2022年。BayesFactor：常见实验设计的贝叶斯因子计算。https://CRAN.R-project.org/ 软件包=BayesFactor R软件包版本 0.9.12-4.4。[81] Magne Nylenna、Frode Fagerbakk 和 Peter Kierulf。2014年。作者身份：挪威研究人员的态度与实践。《BMC医学伦理》15，1（2014年12月），53。https://doi.org/10.1186/1472-6939-15-53[82] Siobhan O'Connor 和 ChatGPT。2023年。护理教育中的开放人工智能平台：促进学业进步的工具，还是滥用？《Nurse Education in Practice》66（2023年1月），103537。https://doi.org/10.1016/j.nepr.2022.103537[83] Fengjiao Peng、Veronica Crista LaBelle、Emily Christen Yue 和 Rosalind W. Picard。2018年。月球之旅：用于自我反思的个性化动画电影。载于《2018年 CHI 计算机系统中的人为因素会议论文集》。ACM，加拿大魁北克省蒙特利尔，1–10。https://doi.org/10.1145/3173574.3173827[84] Jon L. Pierce、Tatiana Kostova 和 Kurt T. Dirks。2003年。心理所有权的现状：整合开拓展一个世纪的研究。《普通心理学评论》7，1（2003年3月），84–107。https://doi.org/10.1037/1089-2680.7.1.84[85] Peter Potash、Alexey Romanov 和 Anna Rumshisky。2015年。GhostWriter：使用长短期记忆网络自动生成说唱歌词。载于《2015年自然语言处理实证方法会议论文集》。计算语言学协会，葡萄牙里斯本，1919–1924。https://doi.org/10.18653/v1/D15-1221[86] Philip Quinn 和 Shumin Zhai。2016年。文本输入建议交互的成本效益研究。载于《2016年 CHI 计算机系统中的人为因素会议论文集》（美国加利福尼亚州圣何塞）（CHI '16）。计算机协会，美国纽约州纽约，83–88。https://doi.org/10.1145/2858036.2858305[87] Björn Rasch 和 Jan Born。2013年。睡眠在记忆中的作用。《生理学评论》93，2（2013年4月），681–766。https://doi.org/10.1152/physrev.00032.2012[88] Emily Reif、Daphne Ippolito、Ann Yuan、Andy Coenen、Chris Callison-Burch 和 Jason Wei。2021年。利用大语言模型进行任意文本风格迁移的方法。（2021年）。https://doi.org/10.48550/ARXIV.2109.03910 出版方：arXiv 版本号：4。[89] Linda A. Riley 和 Stuart C. Brown。1996年。塑造公众形象：一项关于代笔伦理的实证研究。《Journal of Business Ethics》15，7（1996年7月），711–720。https://doi.org/10.1007/BF00381736[90] Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman。2023年。DreamBooth：针对主体驱动生成的文本到图像扩散模型微调。（2023年6月），22500–22510。https://doi.org/10.1109/CVPR52729.2023.02155[91] Pamela Samuelson。2020年。人工智能作者身份？《Commun. ACM》63，7（2020年6月），20–22。https://doi.org/10.1145/3401718[92] Benjamin Schöne、Marlene Wessels 和 Thomas Gruber。2019年。虚拟现实体验：通往自传体记忆的一扇窗。《Current Psychology》38，3（2019年6月），715–719。https://doi.org/10.1007/s12144-017-9648-y[93] Ben Shneiderman。2003年。《眼见为实：信息可视化的任务与数据类型分类法》。载于《信息可视化的艺术》，BENJAMIN B. Bederson 和 BEN Shneiderman（编）。Morgan Kaufmann，旧金山，364–371。https://doi.org/10.1016/B978-155860915- 0/50046-9[94] Ben Shneiderman。2020年。弥合伦理与实践之间的鸿沟：构建可靠、安全、可信的以人为中心的人工智能系统的指南。《ACM Transactions on Interactive Intelligent Systems》10，4（2020年12月），1–31。https://doi.org/10.1145/3419764[95] Ben Shneiderman。2020年。以人为中心的人工智能：可靠、安全且可信。《International Journal of Human–Computer Interaction》36，6（2020年4月），495–504。https://doi.org/10.1080/10447318.2020.1741118[96] Julia Silge 和 David Robinson。2016年。tidytext：在 R 中运用整洁数据原则进行文本挖掘与分析。《Journal of Open Source Software》1，3（2016年），37。https://doi.org/10.21105/joss.00037[97] Nikhil Singh、Guillermo Bernal、Daria Savchenko 和 Elena L. Glassman。2023年。偷来的大象藏在哪里：多模态机器智能推动创意写作的飞跃。《ACM 人机交互汇刊》30，5，文章 68（2023年9月），57 页。https://doi.org/10.1145/3511599[98] Bob L. T. Sturm、Maria Iglesias、Oded Ben-Tal、Marius Miron 和 Emilia Gómez。2019年。人工智能与音乐：版权法和工程实践中的开放问题。《Arts》8，3（2019年9月），115。https://doi.org/10.3390/arts8030115[99] Reid Swanson 和 Andrew S. Gordon。2012年。随便说什么：运用文本案例推理实现开放域交互式故事讲述。《ACM Transactions on Interactive Intelligent Systems》2，3（2012年9月），1–35。https://doi.org/10.1145/2362394.2362398[100] Adam Tapal、Ela Oren、Reuven Dar 和 Baruch Eitam。2017年。能动感量表：测量对自身心智、身体及周围环境的有意识感知控制。《Frontiers in Psychology》8（2017年9月），1552。https://doi.org/10.3389/fpsyg.2017.01552

[101] PLoS Medicine 编辑部。2009年。代笔：医学出版中那个刚刚变得更严重的肮脏小秘密。《PLoS Medicine》6，9（2009年9月），e1000156。https://doi.org/10.1371/journal.pmed.1000156[102] Alexey Tikhonov 和 Ivan P Yamshchikov。2018年。猜猜是谁？自动生成作者风格化诗歌的多语言方法。载于《2018年 IEEE 语音语言技术研讨会（SLT）》。IEEE，希腊雅典，787–794。https://doi.org/10.1109/SLT.2018.8639573[103] Adaku Uchendu、Zeyu Ma、Thai Le、Rui Zhang 和 Dongwon Lee。2021年。TURINGBENCH：神经文本生成时代图灵测试的基准测试环境。（2021年）。https://doi.org/10.48550/ARXIV.2109.13296 出版方：arXiv 版本号：1。[104] Kristen Vaccaro、Dylan Huang、Motahhare Eslami、Christian Sandvig、Kevin Hamilton 和 Karrie Karahalios。2018年。控制幻觉：控制设置的安慰剂效应。载于《2018年 CHI 计算机系统中的人为因素会议论文集》。ACM，加拿大魁北克省蒙特利尔，1–13。https://doi.org/10.1145/3173574.3173590[105] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。Attention is All you Need。载于《神经信息处理系统进展》，I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett（编），第 30 卷。Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf[106] Ziwen Wang、Jie Wang、Haiqian Gu、Fei Su 和 Bojin Zhuang。2018年。自动条件生成个性化社交媒体短文本。载于《PRICAI 2018：人工智能趋势》，Xin Geng 和 Byeong-Ho Kang（编）。第 11013 卷。施普林格国际出版，Cham，56–63。https://doi.org/10.1007/978-3-319-97310-4_7 丛书名称：计算机科学讲义。[107] Webalys。2023年。工作、协作、机器人图标。https://www.iconfinder.com/icons/5692280/working_协作_机器人_开发者_人工智能_创意_人类_共同_工程师_女性_协作_图标 Iconfinder 基础许可。[108] J. S. Wislar、A. Flanagin、P. B. Fontanarosa 和 C. D. DeAngelis。2011年。高影响力生物医学期刊中的荣誉作者与隐名作者：一项横断面调查。《BMJ》343，oct25 1（2011年10月），d6128–d6128。https://doi.org/10.1136/bmj.d6128[109] Paweł W. Woźniak、Jakob Karolus、Florian Lang、Caroline Eckerth、Johannes Schöning、Yvonne Rogers 和 Jasmin Niess。2021年。令人毛骨悚然的技术：它是什么，又该如何衡量？载于《2021年 CHI 计算机系统中的人为因素会议论文集》。ACM，日本横滨，1–13。https://doi.org/10.1145/3411764.3445299[110] Beta Writer。2019年。锂离子电池：当前研究的机器生成摘要。施普林格国际出版，Cham。https: //doi.org/10.1007/978-3-030-16800-1[111] Diyi Yang 和 Lucie Flek。2021年。迈向以用户为中心的文本到文本生成：一项综述。载于《文本、语音与对话》，Kamil Ekštein、František Pártl 和 Miloslav Konopík（编）。第 12848 卷。施普林格国际出版，Cham，3–22。https://doi.org/10.1007/978-3-030-83527-9_1 丛书名称：计算机科学讲义。[112] Ann Yuan、Andy Coenen、Emily Reif 和 Daphne Ippolito。2022年。Wordcraft：使用大语言模型创作故事。载于《第 27 届国际智能用户界面会议》（IUI '22）。计算机协会，美国纽约州纽约，841–852。https://doi.org/10.1145/ 3490099.3511105

致谢

我们感谢 Otso Haavisto 对早期草稿提出意见。本项目由巴伐利亚州科学与艺术部与巴伐利亚数字化转型研究所（bidt）资助；同时，欧盟地平线2020研究与创新计划也通过资助协议编号 952026，资助了 HumanE AI 网络。

研究1：完整的测量指标列表

表6。研究1第2部分和第3部分所使用的测量指标。第1部分中提出的微调问题作为补充材料发布于OSF：https://osf.io/n4svx/?view_仅=916873e81d244be6be8e0790531b1197。除非另有说明，我们均使用7点李克特量表（1—强烈不同意，7—强烈同意）。必要时，我们会对问题进行细微修改，使其适用于文本生成的语境（例如，在诡异感量表中使用“使用”而非“穿着”）。

第2a部分：系统交互之前
先前经验
• 请列出您过去使用生成文本进行写作的所有经历（无 – 使用单词或句子建议进行写作 – 使用自动补全进行写作 - 使用自动纠错进行写作 - 使用智能回复功能 - 其他（+ 文本字段））
• 是否有时会由他人代您撰写文本，例如由工作中的助理为您准备文本、演示文稿或演讲稿？（是 / 否）
Mood
• 你今天的情绪如何？（评分：悲伤表情 1 – 开心表情 5）
第2b部分：针对每种交互方式（IM）重复
所有权感
• 关于感知到的作者身份和所有权的问题，源自国际医学期刊编辑委员会所定义的作者身份标准[24] 以及 Nylenna 等人 [81]：– 我是这张明信片内容的主要贡献者。– 我对文本内容作出了实质性贡献。– 我起草了这张明信片。– 我对这张明信片进行了批判性修订。– 我已最终批准上传此文本。– 我对文本的所有方面负责。– 我对文本的至少一部分负责。– 我的名字应出现在这张明信片下方。
• 我觉得自己是这段文本的作者。 [72]
• 这段文本应归谁所有？（滑块：我 – 人工智能）
领导感
• 我感觉是我在写作，而人工智能在协助我。 [72]
• 我感觉是人工智能在写作，而我在协助它。 [72]
• 在这种交互方式中，谁主导了写作？（滑块：我 – 人工智能）

[24] https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

控制感

- 谁掌控了内容？（滑块：我 – 人工智能）
- 能动感量表 [100]– 我只是人工智能系统手中的一个工具。– 生成的文本只是自行产生的，并非出于我的意图。– 与系统交互时，我是自己行为的作者。– 与系统交互时，我觉得自己行为的后果与自己的行为在逻辑上并不相符。– 我的行为结果通常会让我感到意外。– 与系统交互时，我觉得自己就像一个遥控机器人。– 我对生成的文本负全部责任。

- 在任务过程中，我觉得自己能够控制文本。 [9]
- 我觉得人工智能系统就像一个代笔人，代表我写这张明信片。
- 我觉得人工智能系统就像一个我可以控制的工具。

文本匹配

- 我真的会把这张明信片寄给我的朋友和家人。
- 我自己也会写一张类似的明信片。
- 这张明信片主要包含我用英语写作时通常会使用的词语和/或短语。 [72]
- 我对这段文字感到满意。 [72]

用户体验

- AttrakDiff 中使用的 7 点李克特量表形容词配对，见 https://www.attrakdiff.de/index-en.html，另见 [49]：– 实用质量：人性化 – 技术化，简单 – 复杂，实用 – 不实用，笨重 – 流畅，可预测 – 不可预测，令人困惑 – 结构清晰，难以驾驭 – 易于管理 – 享乐质量 – 刺激性：有创意 – 传统，缺乏想象力 – 富有创造力，大胆 – 谨慎，创新 – 保守，乏味 – 引人入胜，要求不高 – 富有挑战性，新颖 – 平常 – 享乐质量 – 身份：有凝聚力 – 孤立，不专业 – 专业，俗气 – 时尚，高档 – 廉价，融入 – 疏离，让我与他人疏远 – 拉近我与他人的距离，体面 – 不体面– 享乐质量 – 吸引力：愉快 – 不愉快，丑陋 – 有吸引力，讨人喜欢 – 令人反感，拒人于千里之外 – 热情友好，好 – 坏，令人厌恶 – 引人喜爱，鼓舞人心 – 令人气馁

- 你有多喜欢这张明信片？（评分：1-5颗星）
- 你觉得理解和使用这种方法有多难？（非常难 - 非常容易）
- 我在与系统互动时玩得很开心。
- 关于交互方式的开放式问题– 如何改进这种交互方式？ – 这种方法在写作交互方面有哪些优点？ – 这种方法在写作交互方面有哪些缺点？

- 针对各个交互方法的具体问题– 选择：您是否希望有三篇以上的文本可供选择？ – 选择：您认为可供选择的文本数量多少比较合适？ – 选择：您为什么选择了这张明信片，而不是另外两张？ – 选择：您在挑选明信片时采用了哪些策略？ – 编辑：您是否希望能在此任务中更改更多词语？ – 编辑：更改的词语数量– 编辑：变更日志（例如，添加、删除或更改的词语或标点符号）

第2c部分：使用所有交互方法后的总体评估

Mood

- 参与者当前情绪 [92]

诡异感

- 系统总体的技术感知诡异感量表（PCTS） [109]

第 3 部分：交互两天后

记忆

- 睡眠和情绪，以确定对参与者记忆可能产生的影响 [87, 92]
- 基于 [67]的文本回忆– 你觉得这段文本是什么？（记忆犹新 - 有印象 - 不太熟悉 - 完全陌生）– 这段文本是如何创作的？（文本由我自己撰写 - 文本由人工智能生成并由我编辑 - 文本由人工智能生成并由我选择 - 文本由人工智能生成 - 我不知道）

明信片质量的回顾性评估

- 你有多喜欢这张明信片？（评分：1-5 星）

人工智能在写作中的作用

关于交互方法和人工智能技术整体作用的未解问题

- 人工智能自动文本生成的优势
- 人工智能自动文本生成的劣势/风险
- 未来使用此类技术
- 人工智能文本生成的应用场景
- 是否应强制标明文本是在人工智能帮助下创作的（是-否-其他） + 填写理由的开放式文本框

- 反馈

B 研究2：测量指标完整列表

表7。研究2中采用的测量指标。除非另有说明，我们使用了7点李克特量表。（1—强烈不同意，7—强烈同意）

第1部分

先前经验

- 请列出你过去使用生成文本进行写作的所有经历（无——使用 Word 或句子建议进行写作——使用自动补全进行写作——使用自动纠错进行写作——使用智能回复功能——其他（+ 文本字段））

- 有时会有其他人代您撰写文本吗？例如，工作中的助理为您准备文本、演示文稿或演讲稿？（是 / 否）

对人工智能的态度

- 对人工智能的相关性和实用性的感知 [54]– 人工智能是推动积极变化的力量。– 应增加对人工智能研究的资助。– 人工智能总体上很有帮助。– 有必要使用人工智能。

- 感知到的人工智能能力 [54]。您如何评价自己在以下方面的信心：– 解释什么是人工智能。– 围绕人工智能展开对话。– 我对人工智能的了解。

- 对 [55]中人工智能创造力的评估– 我认为人工智能可以独立发挥创造力。– 我相信人工智能可以独立创造新事物。– 人工智能开发的产品可以被视为创意作品。

第2a部分：人类生成文本和AI生成文本各迭代一次（平衡化）

所有权感

与研究1相同

领导感

与研究1相同

控制感

与研究1相同

文本匹配

- 研究1中的问题
- 我觉得这段文本是专门为我写的。

第2b部分：阅读两段文本后

对人工智能的态度

- 关于人工智能的相关性、有用性和创造力，以及参与者人工智能能力的问题，与研究2第1部分中的问题相同
- 用于评估拟人化 和智能 [7], 的 Godspeed问卷，使用于 [73]

关于作者身份的开放问题

- 您是如何决定为 <人工智能/Sasha>撰写的明信片标注作者归属的?
- 在您看来</style}，是否应该强制要求对借助 <人工智能/人类> 创作的文本作出标注?

- 你认为**其他人**期待你在文本中注明 <人工智能/Sasha> 吗？为什么?

偏好

- 哪 你更喜欢哪张明信片？（人工智能写的那张 - Sasha 写的那张 - 我无所谓）

C 研究1：博客式网站



图11。我们创建的博客式网站截图。图片来源：网络背景由 Freepik 上的 kjpargeter 提供。顶部明信片由 Pexels 上的 Chris Schippers 提供。底部明信片由 Freepik 上的 gstudioimagen 提供。页眉图片为作者自有。

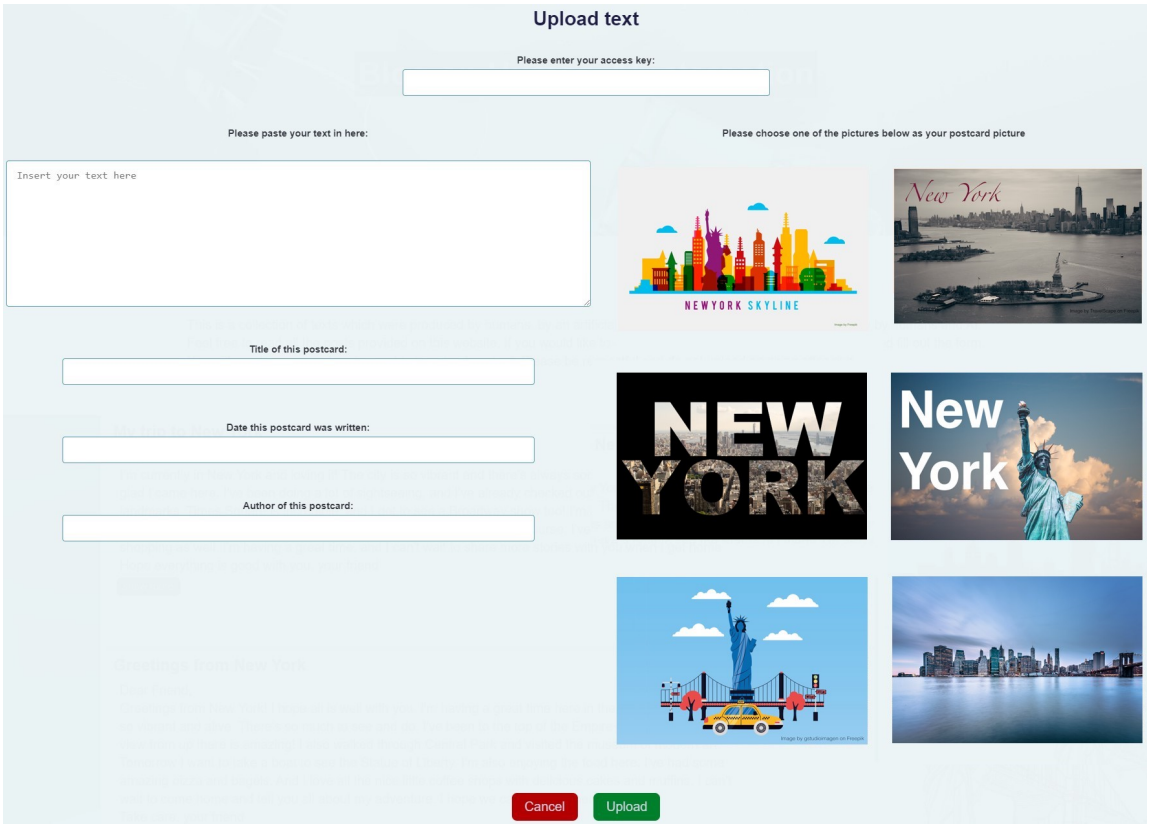


图12。我们为研究1创建的博客式网站中上传表单的截图。原始图片来源：左上：Freepik。右上：TravelScape 在 Freepik 上发布。中左：Pixabay（文字下方的照片）。中右：ParentRap 在 Pixabay 上发布。左下：gstudioimagen 在 Freepik 上发布。右下：Chris Schippers 在 Pexels 上发布。