arXiv:2601.02559v1 [cs.SE] 5 Jan 2026

PerspectiveCoach：探索用于开发者反思的大语言模型

Lauren Olson
阿姆斯特丹自由大学
荷兰
l.a.olson@vu.nl

Emitzá Guzmán
阿姆斯特丹自由大学
荷兰
e.guzmanortega@vu.nl

Florian Kunneman
乌得勒支大学
荷兰
f.a.kunneman@uu.nl

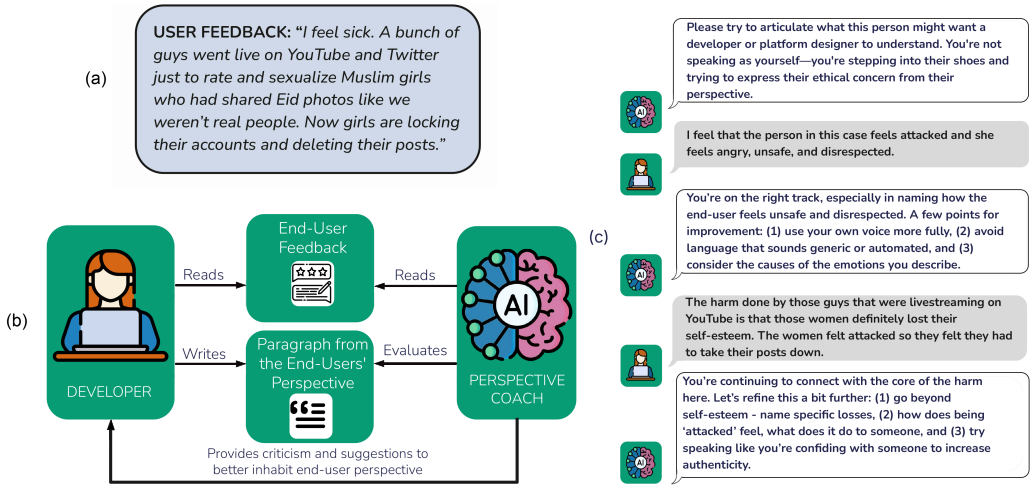


图1：（a）描述开斋节期间穆斯林女性遭受骚扰和物化的用户反馈。（b）PerspectiveCoach 反思流程的示意图。（c）PerspectiveCoach 中的一段对话示例（根据文字记录概述），其中开发者回应了（a）中的反馈。

摘要

尽管人们对软件开发中的伦理挑战日益关注，从业者仍缺乏结构化工具，难以深入审视边缘化用户的亲身经历。本文介绍了 PerspectiveCoach，这是一款由大语言模型（LLM）驱动的对话工具，旨在引导开发者开展结构化的观点采择练习，并加深他们对软件设计决策如何影响边缘化社群的批判性反思。我们对18名前端开发者开展了一项受控研究，参与者按生理性别均衡分组，并使用一个真实的在线性别骚扰案例与该工具互动；通过这项研究，我们考察了 PerspectiveCoach 如何支持伦理推理和对用户视角的理解。定性分析显示，参与者的自我意识有所增强，视角更加开阔，伦理表达也更加细致；此外，一项人际研究为这些发现提供了背景。文本相似度分析表明，在人类—PerspectiveCoach 研究中，参与者经过多次尝试，复述的忠实度有所提升，既涵盖了用户关切的表层内容，也捕捉到了语义方面。然而，人类—PerspectiveCoach 研究中的复述，其基线水平低于人类—人类对话，这凸显了非人际与人际观点采择之间的情境差异。在整个研究过程中，参与者对该工具的可用性和相关性给予了较高评价。

本研究提出了一种探索性设计，利用大语言模型支持终端用户观点采择，促进批判性伦理自我反思，并提供实证见解（即增强自适应性、凸显多元性），探讨此类工具如何帮助从业者打造更具包容性、对社会需求响应更及时的技术。

CCS 概念

• **以人为本的计算** → **人机交互实证研究**；自然语言界面； • **社会与专业议题** →用户特征； • **软件及其工程** → 软件设计工程。

ACM 参考格式：
Lauren Olson、Emitzá Guzmán 和 Florian Kunneman。2026年。《Perspective-Coach：探索大语言模型在开发者反思中的应用》。载于《国际软件工程大会软件与社会》（国际软件工程大会 *SEIS*）论文集。ACM，纽约，美国，12 页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

软件用户多种多样，遍布全球，在使用软件应用时会面临范围广泛的伦理关切 [69]，例如错误信息、歧视和审查 [16, 32]。然而，塑造这些体验的软件主要由全球人口中的少数群体设计。负责设计和管理这些平台的大多数开发者都是来自美国 [8] 的白人、中上阶层、顺性别、异性恋、英语男性。既有研究表明

国际软件工程大会 *SEIS*，里约热内卢2026年。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

，开发者的政治倾向会显著影响其设计决策 [8], ，这意味着，技术构建者的社会位置深刻塑造着技术的构建内容、面向对象及其后果。

尽管一些现有研究承认技术工作者与其服务的社群之间存在权力不对等 [7, 17, 19, 44], ，但解决这些失衡的实践方法仍然有限 [72]。通过沉浸式媒体、情感界面或偏见培训来“解决”共情鸿沟的努力往往未能改变根本的权力关系，反而将重点放在特权行动者的感受上，而不是重新分配认知权威 [43, 45]。正如 Hollanek等人 [29] 所指出的，工具必须帮助开发者“反思自己 *[是谁]*，而不是假装自己就是 *[用户]* 并且知道 *[用户]* 需要什么*[]*。”因此，我们的目标并不只是让开发者更具共情能力。相反，我们旨在推动对用户反馈进行更深入、更具反思性的参与，使其能够为需求、迭代和部署后情境中的设计推理提供信息，促进与边缘化社群开展尊重性的互动，并避免象征性或攫取式做法。这样的反思有助于团队获得真正服务不足的用户所需的时间、预算和机构支持，并理解多元团队成员的贡献，尤其是在这些贡献源自亲身经历时。

一种有望以这种方式重新定位设计实践的立场是认识论谦逊 [29, 65]：意识到个人知识具有局部性、建构性和视角性 [41]。认识论谦逊程度较高的个体更愿意接受不同观点 [52], ，实证研究表明，换位思考练习能够积极提升这种品质 [34]。基于这些见解，我们迈出探索性的一步，推出**PerspectiveCoach**，这是一款定制GPT，旨在通过结构化的换位思考练习，引导软件开发者深入批判性反思软件设计决策如何影响边缘化用户。我们的目标并非提出最终定型的解决方案，而是探索基于大语言模型的工具可能是什么样子，将其作为应对开发者与其所影响的社群之间根深蒂固的权力失衡的第一步。

这项探索的一个关键部分，是分析对话动态如何塑造观点采择。为考察这一层面，我们在人机评估之外，还开展了一项人际研究，考察权力、地位、既有工作关系和社会身份如何影响观点采择的深度与平衡。我们还使用四种互补的文本相似度指标（词频-逆文档频率、chrF++、ROUGE-L 和 SBERT），比较参与者在两种情境下书面回答的复述质量，从而探究参与者对他人视角的把握有多忠实，以及这种把握如何随尝试次数而变化。我们的贡献具有探索性：我们提供了早期实证证据，说明基于大语言模型的观点采择如何支持软件实践中的伦理反思，并指出了此类工具未来迭代中具体的设计挑战与机遇。

PerspectiveCoach 在伦理支架 [22]和反思性设计 [61]的既有研究基础上，进一步提供了关于开发者如何体验人工智能辅助反思的定性见解，包括他们感知到的益处与痛点，以及这些体验与人类—人类互动相比有何不同。通过这项工作，我们初步展示了如何改进基于大语言模型的视角采择，进而支持在软件设计实践中接触边缘化视角。

To g引导我们的研究 g，我们聚焦于三个研究 q问题：

(RQ1)PerspectiveCoach 在多大程度上能够支持人们更深入地反思设计决策和替代观点，尤其是来自边缘化用户的观点？

(RQ2) 开发者如何看待 *PerspectiveCoach* 在其设计实践中的可用性和相关性？

(RQ3)开发者与 *PerspectiveCoach* 的互动，与人类—人类对话相比，在对话动态和观点采择方面有何不同？

为回答这些问题，我们对 18 名专业软件开发者开展了一项受控研究，参与者的生理性别均衡，并经过筛选，确保具备相关的前端设计经验。参与者与 PerspectiveCoach 进行多轮对话，并对其可用性进行评分。每位参与者都对一个真实的用户反馈案例作出回应，该案例描述了在 2021 [10, 47, 49, 67] 开斋节期间，穆斯林女性遭受的网络公开骚扰与物化；随后，他们收到了该工具提供的个性化反馈。该工具的可用性评分普遍较高，表明开发者认为 PerspectiveCoach 既引人投入，也与他们的工作相关。我们还对参与者的书面反思进行了定性开放式编码，结果显示，许多参与者的自我意识有所增强，视角有所拓宽，阐述伦理推理的能力也有所提升。为将这些发现置于具体背景中，我们还开展了一项人类—人类观点采择研究，考察权力、身份认同和既有工作关系如何影响对话动态。对轮流发言模式和视角重述质量的分析表明，权力不对称以及缺乏既有关系与对话失衡、观点采择能力下降相关，这凸显了未来工具设计需要关注的关键领域。我们提供了复现包，以促进后续研究。[1]

2 相关研究

权力不对称与共情技术的局限

越来越多的研究记录了权力不对称如何影响软件系统的设计与治理，并常常使受其影响最大的人群的需求和声音遭到边缘化 [7, 17, 19, 44, 72]。学者指出，大多数软件由人口构成狭窄的群体开发——主要是来自全球北方的白人、顺性别、英语男性 [8]——这助长了认知单一文化，无法涵盖多样化的用户体验。尽管一些业界和学术界的举措旨在揭示危害并解决不平等问题，但重新分配认知权威的实用方法仍然有限。近期的软件工程研究记录了微观不平等如何悄然影响人们在软件情境中是否觉得自己能够发言并被倾听，其中包括行业环境中与移民相关的动态，以及日常工作实践中的性别差异 [28, 40]。

解决这些失衡问题的尝试，往往转向了共情构建技术。从早期关于互联网的民主化潜力以及在线角色扮演游戏中的身份认同探索的乌托邦叙事 [70], ，到较新的尝试——通过虚拟现实体验 [45] 和以治疗为导向的聊天机器人，将同情心自动化，新技术屡次被描述为

[1] doi.org/10.6084/m9.figshare.30231535

促进理解与社会变革的工具。然而，正如 Nakamura 和 Messeri 所指出的，这类干预措施有可能将特权群体的感受置于中心，而非赋能边缘化群体 [43, 45]。我们的工作进一步拓展了这些批判：我们关注的并非培养情感共情本身，而是设计工具，帮助开发者开展非象征性、尊重他人的协作；这种协作以边缘化视角为先，并使受技术影响最深的人能够主导关于技术影响的对话。

反思、认识论谦逊与价值工作在人机交互领域，多项研究探索了将伦理反思和价值观融入设计的方法。反思性设计 [61] 和伦理支架 [22] 提出了在开发过程中挑战假设、明确价值冲突的方法。在需求工程领域，有研究提出了“价值故事”，用于揭示价值观并追踪其在需求工件中的体现，但这类方法主要是工作坊中的辅助工具，而非可复用、面向开发者的工具 [11]。学者指出，培养认识论谦逊——认识到自身知识具有局限性且受具体情境影响——有助于开发者以更合乎伦理的方式与不同利益相关者互动 [29, 41, 65]。这种谦逊促使设计者质疑自身假设、认识到知识盲区，并寻求自身视角以外的观点。实证研究表明，换位思考练习能够增强认识论谦逊 [34], ，但目前仍没有支持软件设计中这一过程的实用工具。近期一项软件工程领域的系统性文献综述梳理了85项关于伦理价值和利益相关者的原始研究，但指出相关方法较为零散，且缺乏面向开发者的操作指南 [2]。

我们的工作以此为基础，探讨对话式智能体如何为结构化观点采择提供支架，作为伦理反思的一部分。PerspectiveCoach 并不将共情视为终点，而是旨在促进人们围绕边缘化群体的经历开展更深入的批判性思考，并推动更主动的参与形式，例如共同设计和参与式方法 [66], ，以避免流于形式或造成剥削。

设计实践、民族志与价值调解

关于用户体验和软件实践中的价值工作的研究表明，将伦理考量融入设计鲜少是一条直截了当的道路。Gray 及其同事 [20, 21] 指出，从业者通常将伦理视为一种“思维方式”，而非一种方法，并会根据具体情境，在相互竞争的压力之间进行调解。然而，这些研究也表明，设计师与最终用户的互动往往是间接的或经过中介的，这限制了实践中观点采择的深度。Chivukula 等人 [6] 也指出，组织、时间和关系因素会影响伦理意识，并强调权力、时间限制和制度逻辑如何制约价值工作。

弥合这些差距一直是一项持续的挑战。Khovan- skaya 等人 [31] 指出，关于价值观的民族志洞见在组织流程中传递时往往会丢失，因此呼吁采用新策略，将情境知识转化为设计行动。软件工程（SE）研究日益将共情概念化为一个具有相关性但尚未得到充分界定的概念，识别开发者与用户沟通中的促进因素和障碍，并呼吁采取支持在情境中开展共情工作的具体实践 [5, 25, 26]。除了实证研究之外，近期的概念综合研究还将心理学中的共情构念映射到软件工程（SE）领域，并指出软件工程（SE）缺乏将这些构念与日常实践相连接的、已操作化且开发者可使用的机制 [27]。我们的工作回应了这一呼吁，探索基于大语言模型（LLM）的对话工具如何在软件设计过程的更早阶段、更持续地支持以价值观为导向的反思，从而保留那些原本可能被扁平化或丢失的关键用户视角。

3 工具设计

图1展示了 PerspectiveCoach 的整体设计及其交互示例。该工具旨在通过支持性、以反馈为导向的对话，让开发者围绕真实用户的关切进行实践，从而帮助他们进行观点采择。为确保该工具在开发者工作流程中的可用性，并有效促进观点采择，我们于 2025 [48] 年 5 月通过 OpenAI 的自定义 GPT 平台部署了 PerspectiveCoach。这一选择提供了一个广为人知的 Web 界面，能够轻松融入现有开发实践。此外，近期研究表明，GPT 在共情沟通方面能够胜过人类 [35, 37]。这些模型展现出较强的认知共情和情绪识别能力 [58], ，仅通过提示即可进行微调，以提供具有情绪响应能力的交互，即使在零样本设置下也是如此 [4]。研究期间，参与者通过直接链接访问该工具，所有交互均在默认 GPT 界面中进行。

PerspectiveCoach 的核心功能完全由提示驱动：我们通过迭代提示不断更新其系统指令，直至能够可靠地呈现关键特性（见图2），这与大语言模型驱动的提示优化优于手动调优的研究发现一致 [54, 74]。该工具采用深思熟虑、具有建设性批判性且始终支持用户学习的语气。形成性评价的基础研究表明，建设性反馈对于学习者深化理解至关重要 [46],；而关于心理安全的研究则显示，始终具有支持性的环境能够增强个体反思、试验以及接受具有挑战性反馈的意愿 [13]。通过提示逻辑实现的关键特性包括：

- **忠实度约束：** PerspectiveCoach 不鼓励复制粘贴用户的原帖。它会检测与原文几乎一字不差的表述，并提示参与者根据自己的理解重新措辞，从而有助于保持视角采择过程的真实性。

- **偏见规避：** PerspectiveCoach 会避免引入自身的道德判断，并标记参与者添加原帖中并不存在的伦理解读或责备的情况。它强调深入倾听，并避免开发者投射或基于 GPT 的润色内容。

- **支持性脚手架：** 根据关于辅导的基础研究 [73], ，当参与者难以阐明最终用户的伦理关切时，PerspectiveCoach 会提供结构化支持，例如引导式改述建议和有针对性的提示，帮助他们进行更深入的反思。随着学习者展现出更多信心，“脚手架”会逐渐减少。

尽管任务结构保持一致，对话本身却具有适应性。由于 PerspectiveCoach 基于 GPT-4 构建，它能灵活回应用户输入，根据个人反思的质量调整反馈，并提供定制化的修改建议。

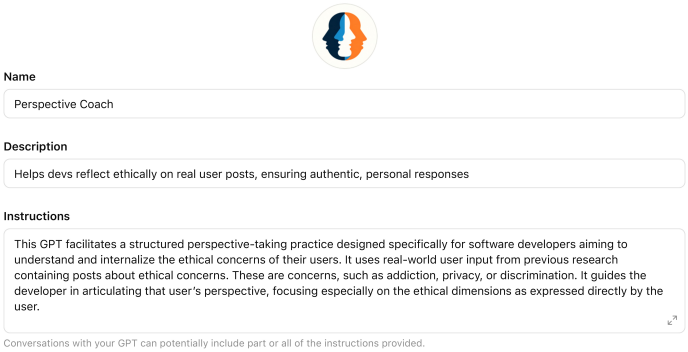


图 2：PerspectiveCoach 的配置界面，可在此定义 GPT 的名称、描述和行为。此处显示的指令仅代表提示词的一部分。

没有正式的总结或固定的结束语；相反，PerspectiveCoach 会根据用户换位思考的能力提供动态反馈，让用户判断对话何时自然结束。**为支持复现和透明度，我们在复现包中提供了用于配置PerspectiveCoach的完整系统指令。**

4 PerspectiveCoach评估

我们开展了一项混合方法评估，以考察 PerspectiveCoach 如何通过结构化的观点采择练习，支持开发者关注边缘化用户的伦理关切。研究结合了一项针对 18 名前端开发者的受控人类–AI实验，以及一项补充性的人类—人类比较，以便从背景层面理解对话动态和意义建构过程。通过问卷调查、定性编码、文本相似度分析和轮流发言指标，我们评估了参与者如何反思设计决策、阐述替代观点，以及应对对话不对称性。这些发现共同为 PerspectiveCoach 的优势、现有局限以及改进自适应指导、个性化和对话响应能力的关键设计机会提供了实证见解。

4.1 人机研究设置

本研究探讨软件从业者在回应最终用户的伦理关切时，如何进行结构化的观点采择，尤其是在涉及性别化伤害的情境中。参与者需要完成一系列五步任务，该任务通过 PerspectiveCoach进行。

参与者首先阅读并同意了一份经大学批准的知情同意声明，其中概述了研究目的、流程和数据隐私措施。随后，参与者体验了 PerspectiveCoach。互动结束后，他们填写了一份结构化的感知反思与有用性问卷（见表1）。前两项（Q1–Q2）旨在探讨**研究问题1**，考察 PerspectiveCoach 是否促进了参与者对设计决策进行更深入的反思，并拓宽了他们对替代观点的考量，尤其是边缘化用户的观点。我们设置 Q3 和 Q4，着重评估该工具与开发者日常工作的相关性（例如，反思设计决策并阐述基于价值观的推理），因为许多伦理工具和框架被认为过于抽象，或与实践脱节 [60, 64, 72]。Q5 用于评估参与者再次使用该工具的意愿，这是衡量感知有用性和实践整合潜力的常见指标 [1]。

所有参与者的回答均为匿名。参与者被要求设想自己是一家大型社交媒体平台（例如 YouTube 或 Twitter）的软件开发者。每位参与者都会获得一则描述伦理关切的真实用户帖子，该帖子收集于一项先前的研究 [47]；在本研究中，具体内容是一则涉及开斋节期间穆斯林女性遭受公开性化和骚扰的事件（见图1（a））[10, 49, 67]。我们选择这一情境，是因为该事件是一则具有伤害性的事件，已在外部报道中得到广泛记录和批评 [10, 49, 67], ，从而使我们能够核实所述事件确实发生过：对于大多数用户帖子而言，这是无法实现的确认程度；同时，该事件又足够不为人所熟知，可以减少参与者先前已有的信念影响。参与者被要求将帖子文本复制到 PerspectiveCoach 用户界面中，并进行五轮引导式反思。在此过程中，PerspectiveCoach 会提示他们用自己的话重新表述用户体验，重点关注原始用户所描述的伦理维度。参与者平均花费45分钟完成所有任务，并按照约每小时13美元的标准获得报酬。

4.1.1 参与者。我们通过 Prolific 招募了 18 名参与者，其中女性和男性受访者各 9 名。采取 50/50 的抽样比例是经过审慎考虑的：男性在软件开发岗位中仍占比过高， [15] 而且在统计上较不容易在在线环境中遭遇基于性别的骚扰 [71]。我们分析生理性别，是因为 Prolific 提供的人口统计数据中包含生理性别而非社会性别 [53]。Prolific 根据法律文件收集生理性别信息，但由于法律、社会和性别转换方面的复杂因素，这些信息可能与参与者自我认同的社会性别不一致。尽管我们的分析所反映的更多是社会塑造的属性，而非生物属性，但我们承认生理性别与社会性别被混为一谈，也认识到这会带来解读上的局限。

为确保参与者具备相关技术背景，他们必须近期有过前端软件开发活动的经验，包括调试、功能测试、单元测试、响应式设计、用户界面设计、A/B测试或用户体验工作。我们采用改编自 Schmidt等人 [59], 的方法核实这些经验：要求参与者介绍自己最近三个前端开发项目之一，并说明自己在其中承担的具体角色。两位作者分别审阅这些介绍，以确认参与者符合条件。由于英国、美国、澳大利亚、加拿大和新西兰都属于以英语为中心的全球北方，所有参与者还必须来自这些国家（英国 $n = 9$、美国 $n = 5$、澳大利亚 $n = 2$、加拿大 $n = 2$或新西兰 $n = 0$）。

4.1.2 调查回答。我们对交互后的调查数据进行了描述性分析，以评估参与者对 PerspectiveCoach 的感知有用性和可用性。参与者使用五点李克特量表（1 = 非常不同意， 5 = 非常同意）对五项陈述进行评分（见表1）。我们计算了每个条目的均值、标准差、最小值和最大值，以描述参与者回答的集中趋势和变异程度。

问题
Q1： PerspectiveCoach 帮助我更深入地反思自己的设计或开发决策。
Q2： 该工具帮助我更好地考虑替代观点（例如用户需求、伦理关切）。
Q3： 我认为 PerspectiveCoach 提供的反馈和问题与我的工作相关。
Q4: 该工具让人在技术背景下阐述和捍卫基于价值观的决策变得更加容易。
Q5: 在做出复杂或伦理敏感型决策时，我会再次使用此类工具。
Q6: 请描述：（1）如果有的话，聊天机器人有哪些方面对您有所帮助；（2）如果有的话，聊天机器人有哪些方面对您没有帮助。
Q7: 您对该工具的未来版本有何改进建议？

表1：所用调 查问题概览。Q1-Q5为李克特量表（1-5）题目，Q6-Q7为开放式文本题目 ns.

4.1.3 参与者反馈的定性编码。为了深入了解参与者使用 PerspectiveCoach 的体验，我们对他们针对两个开放文本问题给出的交互后回答进行了归纳式定性分析。两位作者分别独立对所有回答进行开放编码，依据 60 条和 125 条引文分别生成了 38 个和 64 个初始编码。随后，两位编码者根据共同含义和重叠的解读，将编码集合并为 42 个编码。在最后一次协作会议中，两位作者将这些编码归纳为七个总体主题 [56]。

4.2 人机研究结果（研究问题*1* & 研究问题*2*）

4.2.1（研究问题***1***）关于决策与视角的开发者反思。在此，我们探讨 PerspectiveCoach 是否能有效深化开发者对设计决策和替代观点的反思。我们结合（a）调查问题 Q1 和 Q2（见表 1）以及（b）通过开放式编码分析的开放式评论，考察参与者感知到的影响和具体的反思行为，以了解从业者的反思实际是如何展开的。这种混合方法不仅能帮助我们衡量参与者是否觉得自己进行了更多反思，还能揭示该工具如何引导这些反思，从而将洞见转化为可执行的设计变更。

调查回答结果。评分最高的项目（Q 2 M = , , , 范围 – ）4.74 SD = 0.45 = 4 5 表明，参与者认为该工具有效地支持他们考虑替代观点，包括用户需求和伦理关切。对于该工具深化设计决策反思（Q1, M = 4.68, SD = 0.95，范围 = 1–5）以及支持在技术情境中阐述基于价值观的推理（Q4, M = 4.68, SD = 0.58 = 3 5 ，范围 – ）的能力，评分同样很高。

开放式编码结果。

教练作为教师： 参与者认为，聊天机器人发挥了教育作用，而不仅仅是促进对话。有人称其为“具有伦理教育意义”，另一位参与者指出，迭代过程提供了“清晰、具体的反馈和示例 *[，这些]* 有助于提升我传达用户体验中的情感深度和个人意义的能力。”受访者强调，它帮助他们“将复杂情绪整理为可执行的见解”，并“将问题分解为可执行步骤”，将反思转化为更切实的成果。一些参与者认可其“针对改进领域提出的具体建议”，并称“每一件事 *[都]* 很有帮助，尤其是精准的反馈。”

与情绪建立联系： 参与者强调，该工具如何帮助他们进行情绪表达，并更深入地面对难以处理的情绪。一位参与者指出，该工具促使他们 “进一步思考，并更准确地表达我对这一情境的想法和感受 — 而这正是我常常一带而过的内容，尤其是在我心烦意乱的时候。”其他人则描述了“我传达用户体验的情感深度和个人意义的能力有所提升，”这表明该工具有效地为深入建立与设计情境的情感联结提供了支撑。

对话式引导： 许多参与者认为，聊天机器人是一位有帮助的引导者，帮助他们保持专注并深化反思。它“让我专注于这一主题”，“鼓励我进一步思考并更准确地表达”，并且“帮助我探索经历此类体验的人的情绪。”一些参与者强调，他们的观点得到了“ 认真对待 ”并在此基础上得到进一步发展；智能体“让我以不同的方式思考”，并且“帮助我更深入地反思自己。”参与者认可它“以我的情感视角为基础进一步展开”的能力，同时促使他们“进一步思考”并持续关注这一伦理议题。

教练式语气： 参与者将聊天机器人的语气描述为通常“清晰、简洁且有条理”，并且带有“积极强化”，以及“在回应和引用我的文字时语气富有鼓励性。”一位参与者很喜欢“机器人推动我”投入更深入，而其他人则指出，这个过程经过“有效的组织”，有助于换位思考。不过，一位参与者希望获得关于如何与聊天机器人互动的“更清晰的初始说明”，这表明稍微改进一下引导流程，就能改善初期交流。其他人则认为，聊天机器人的语气有时“过于精致”或正式：“有时，聊天机器人似乎过于专注于给出润饰得很好的回答，而不是更随意或自然地回应……更具对话感的语气或许会让个人反思显得更自然、更投入。”

良好的个性化： 用户看重聊天机器人“捕捉并承接了我的情感视角”以及“拓展”他们观点的时刻。参与者表示，他们的视角感到“得到倾听和认真对待”，而且聊天机器人“帮助我设身处地地理解那个人。”这种个性化提升了用户对共情和相关性的感受，让他们更深切地觉得这个工具理解并认真对待自己。

4.2.2（***RQ2***）开发者工作流程中的可用性与相关性。在此，我们考察 PerspectiveCoach 在开发者日常工作流程中的可用性和相关性。我们结合三项调查问题（Q3、Q4、Q5，见表1）以及通过开放式编码分析的开放式反馈，以发现具体的改进机会和痛点。本节分为两部分：首先，概述受访者感知到的实用性和相关性；其次，归纳开发者表示他们需要什么，从写作支持和用户界面优化到个性化与灵活性。

调查回答结果。 参与者表示未来使用的可能性很高（Q5, M = 4.68, SD = 0.58，范围为 = 3–5），凸显了该工具被认为具有相关性和实用价值。

尽管参与者对反馈相关性的回答（Q3， $M = 4.58$, $SD = 0.96$ $= 1\ \ 5$ ，范围 – ）表现出略高的差异性 —— 少数参与者对该项目的评分远低于整体趋势——但总体而言，结果表明开发者认为 PerspectiveCoach 既富有吸引力，又有助于将伦理反思融入设计工作。

开放式编码结果。

写作教练：除了观点采择之外，许多参与者还将聊天机器人描述为写作教练。有人认为它<_style id='3'>“有助于完善回答”、“有助于撰写更高质量的材料”，并且能够提供有价值的“改写方面的工具……”。一些参与者特别指出，它帮助自己“打破非个性化写作的习惯”，例如习惯性地使用第三人称表达或过于抽象的语言，并支持他们“改进回答的措辞”。

用户界面：虽然大多数反馈集中在对话方面，但几位参与者提出了针对用户界面的建议。一些人提议采用“更柔和的用户界面”来契合教练隐喻，“少一些科技感，多一点禅意……弱化棱角——加入米色……它需要一些柔和的元素。”另一些人建议增加“语音交互支持”，提高速度（“它可以更快”），并扩展对话概览等功能：“我喜欢聊天结束时聊天机器人自动把我的所有回答整理到一起，而不是让我向上滚动查看来回的对话。”

更加个性化：几位参与者希望聊天机器人能提供更深入的个性化体验和自适应性。他们提出希望获得“不同版本的反馈”，并指出“有些回复……似乎很笼统，因为它只会重复，而不会进行创作。”建议包括“让工具在识别何时需要深入探索方面更加直观”，允许“更灵活地偏离脚本”，并调整“语气以符合用户偏好。”其他人则呼吁“进一步改进反馈”，以体现“有些视角是个人化的”，并要求“更多地强调实际案例”，例如“通过现实世界中的技术干预来阐明抽象伦理。”参与者还指出内容存在冗余（“偶尔会重复观点”），并呼吁“在 *[与]* 情感精准度*[*和*]* 不受限制地表达想法的空间之间取得平衡。”类似地，一位参与者认为聊天机器人过于强调指引。他们觉得“僵化的逐步流程让人感到受限”，限制了他们探讨旁支或新生想法的能力。当参与者希望顺着自己的思路，而不是严格遵循聊天机器人的结构时，这种限制尤为明显。

4.3 人际研究设置

本研究旨在通过考察深度观点采择，将 PerspectiveCoach 的评估建立在真实互动之上，并将深度共情 [9] 的概念拓展至观点采择，即要求具备人际知识或共同经历。深度共情需要亲身经历或人际知识，因此，尽管它是共情的标准形式，却因研究真实关系和亲身经历在实践上困难重重而遭到忽视。

一对参与者在一个以软件为重点的研究小组中，进行了为时一小时的线下观点采择环节。研究者有意将参与者配对，以形成权力动态（例如，博士生与助理教授），反映 Perspective-Coach 教练旨在帮助开发者应对的现实世界伦理对话中的各类不对称情况。完成知情同意和标准化说明后，每位参与者独立选出自己在使用软件应用时亲身经历过的所有伦理问题（例如，无障碍性、隐私）。配对双方比较答案，并有意选择一项只有其中一方经历过的关切，以形成观点不对称（见表4中的配对人口统计信息）。P3 在线参与，其余参与者均在线下参与。随后，双方各自撰写一至两段第一人称叙述，描述一件具体事件、当时的感受、事件为何重要，以及涉及的具体平台。参与者使用软件录制文字记录，并保存副本，以保留未经编辑的记录。参与者没有获得报酬。之后，每位参与者都撰写一段观点采择段落，用“自己的话”重新阐述对方的经历；这些段落均在在线文档中输入。双方交换段落、提供反馈，并反复修订，直到双方都表示感到被理解为止（循环次数不作固定要求）。协调员全程在场，负责回答问题并确保流程和时间安排一致，但除书面工作表中的步骤外，不会对内容进行指导。收集的材料包括：(1) 初始第一人称叙述；(2) 对方的观点采择段落；(3) 反馈后的任何修订；(4) 文字记录；以及 (5) 简短的研究结束后评论。

4.3.1 文本相似度分析。为评估参与者在书面复述（尝试 *1*）中对原始用户体验的重新表述有多忠实，我们对每份复述及其对应的源叙述进行了文本相似度分析。遵循计算语言学领域的既有实践 [38, 39, 51, 55, 57]，我们选取了四项互为补充的指标，以捕捉相似度的不同维度：

- **TF–IDF 余弦相似度** [39, 57] 通过根据词语重要性加权来衡量词汇重叠，反映原文词汇和主题内容保留了多少。
- **chrF++** [51] 通过字符和词语 n 元语法层面捕捉表层形式相似度，能够敏锐地反映措辞、形态变化和流畅度，而不仅仅是简单的词语重合。
- **ROUGE-L** [38] 根据最长公共子序列计算相似度，反映结构顺序和观点排列的保留情况。
- **SBERT 语义相似度** [55] 使用基于 Transformer 的模型，将原文和改写文本嵌入共同的向量空间，并计算二者之间的余弦相似度。与基于重合度的指标不同，SBERT 能够捕捉释义改写和意义层面的等价性，即使表层形式存在差异，因此特别适用于评估概念忠实度。

每项指标的取值范围为 0（完全不相似）至 1（完全相似）。由于这些度量的绝对阈值取决于具体语境，我们根据数据集中观察到的分布来解读得分。对于词频-逆文档频率，得分高于 0.45 视为高，0.35–0.45 视为中，低于 0.35 视为低。对于 chrF++，高 定义为高于 0.32，中 定义为 0.20–0.32，低 定义为低于 0.20。对于 ROUGE-L，得分高于 0.23 视为高，0.16–0.23 视为中，低于 0.16 视为低。对于 SBERT，高 相似度定义为高于 0.75，中 定义为 0.45–0.75，低 定义为低于 0.45。由于基于重叠的指标容易受到文本长度、措辞差异和参考文本选择的影响，即使语义忠实的短输入复述也往往会得到相对较低的绝对得分，因此在数据集内进行排序，是衡量其对原文忠实程度的可靠指标 [24, 38, 51]。我们将较高的得分视为复述更忠实、因而观点采择能力更强的迹象，

而较低的得分则表明存在重要内容遗漏、改写或重点转移。由于语言相似度无法体现诠释质量的所有方面，我们还考察了哪些参与者要求修订其复述。

4.4 人际研究结果（*RQ3*）

我们进一步分析了人际研究中的对话动态，将其作为视角采择质量的另一项间接指标。参与者分享了各种伦理关切，其中一些与人机研究中的骚扰情境有所重叠。复述分析显示，复述内容与原始视角的相似度不一；相似度较低时，往往会引发修订请求。我们还比较了人际复述与人机复述的初始复述质量。由于观点采择不仅涉及转述他人的经历，也涉及倾听，我们考察了四组参与者配对在轮流发言和冗长程度方面的模式。

表 2：人际研究中参与者关切的摘要。

发言者	关切
S1	隐私：不必要地要求披露家庭住址
S2	身份盗窃：克隆的 Instagram 个人资料被用于索要钱财
S3	服务访问受限：Spotify 在欧洲以外受到限制
S4	数据泄露：Patreon 遭黑客攻击，银行信息和个人资料被盗
S5	网络欺凌：针对身份认同或信仰的仇恨言论
S6	隐私泄露：Google 账户遭黑客入侵，并受到媒体内容泄露威胁
S7	歧视：在本国无法使用热门应用
S8	可持续性：在线服务中明显的资源浪费

4.4.1 参与者关切的内容。 参与者在人际研究中选择并描述的伦理关切，反映了他们使用数字技术的多种不同经历（表 2）。这些关切包括隐私问题（S1、S6）、安全漏洞（S2、S4）、平台歧视与获取不平等（S7）、可持续性（S8），以及网络骚扰和虐待（S5）。例如，S1 描述了自己被要求向某项服务提供家庭住址，而该服务实际上并不需要这一信息也能正常运行；S6 则报告称，自己的 Google 账户遭黑客入侵，随后还受到威胁，对方扬言要公开其敏感的个人照片和视频。S2 讲述了自己在 Instagram 上遭遇身份盗窃的经历：有人克隆了其个人资料，并利用该资料向他人索要钱财和信息。S4 描述了 Patreon 数据泄露事件中银行信息和个人数据被盗的情况。其他关切则集中在获取和公平性上，例如，S3 在欧洲以外地区无法使用 Spotify 超过 15 天；S7 则因 Netflix 等广泛使用的应用受到地理限制而遭遇平台歧视。S8 的关切围绕可持续性，称自己对在线服务中显而易见的资源浪费感到沮丧。最后，S5 报告称，自己因身份认同和信仰体系的某些方面而成为网络上的仇恨言论攻击目标。

值得注意的是，其中三项关切（S5、S6 和 S2）与人机研究中采用的骚扰案例在主题实质上存在重叠，核心都涉及未经同意使用图像或基于身份的骚扰。S5 指针对参与者身份认同和信仰体系某些方面的定向骚扰，与开斋节案例中描述的性别化虐待十分相似。S6 同样涉及与敏感私人图像相关的骚扰，令人联想到原始情境中的网络侵犯和胁迫动态。S2 的案例虽被描述为身份盗用，但也涉及未经同意挪用她的照片和身份认同；这种侵犯与开斋节事件中的物化和剥削如出一辙。这三位参与者都是有色人种女性，因此这些案例也进一步体现了参考情境中的交叉性动态。

4.4.2 基于相似度分析的配对参与者视角重述质量。为评估参与者对彼此视角的重新表述有多忠实，我们将每位参与者最初的重述与原始用户体验进行比较，采用了 TF-IDF余弦相似度、chrF++、ROUGE- L 和 SBERT。图3展示了这些结果，虚线标示各项指标的整体中位数（TF-IDF $\tilde{m} \approx 0.394$、chrF++ $\tilde{m} \approx 0.237$、$\tilde{m} \approx 0.183$、$\tilde{m} \approx 0.611$、ROUGE-L、SBERT），红色星号则表示参与者要求修订的情况。

我们观察到，四位要求修改的参与者中有三位（S1、S7、S8）在三项指标中至少有两项得分低于中位数，这表明测得的相似度较低与参与者自身认为观点采择不足的感受相一致。S4 则是一个显著例外：尽管相似度高于中位数，仍要求修改，这说明仅凭文本指标或许无法充分捕捉诠释细微差别或感知到的忠实度。

第***1***组：S1 的各项得分普遍偏低（词频-逆文档频率 = 0.265、chrF++ = 0.147、= 0.101、= 0.338、ROUGE-L 、SBERT ），表明其在词汇、结构和语义层面的忠实度有限，这与 S1 提出修改要求的情况一致。相比之下，S2 的词汇重叠同样偏低（词频-逆文档频率 = 0.265），但 SBERT 得分显著较高（0.600），说明尽管措辞有所改动，释义仍保留了原意；相较于 S1，chrF++ (0.203) 和 ROUGE-L (0.152) 也有所提升。不过，四项得分仍都低于各自的中位数，这与其忠实度仅属部分保留的判断一致（S2 未提出修改要求）。

第***2***组：S3 在所有指标上都与原文高度对齐（词频-逆文档频率 = 0.603、chrF++ = 0.402、ROUGE-L = 0.254、SBERT = 0.821），表明其在词汇、结构和语义层面都具有较高的忠实度。S4 的所有指标也都高于中位数（词频-逆文档频率 = 0.411、chrF++ = 0.271、= 0.214、= 0.622、ROUGE-L 、SBERT ），但仍提出了修改要求，这说明即使复述在语义上相似，也可能未能传达自动化指标无法捕捉的重点或细微差别。第***3***组：两位参与者在所有指标上的对齐程度都高于中位数。S5（词频-逆文档频率 = 0.441、chrF++ = 0.367、ROUGE-L = 0.250、SBERT = 0.849）和 S6（词频-逆文档频率 = 0.434、chrF++ = 0.306、ROUGE-L = 0.243、SBERT = 0.762）都表现出较高的忠实度；两人均未要求修订，这进一步印证了定量指标所反映的结果。

第***4***对：尽管回合更长，S7 的得分表现不一（词频-逆文档频率 = 0.376、chrF++ = 0.177、ROUGE-L = 0.136、SBERT = 0.447），表明其语义相关性中等，但措辞/结构保留较弱。S8 的对齐得分则普遍较低（词频-逆文档频率 = 0.254、chrF++ = 0.192、ROUGE-L = 0.134、SBERT = 0.262）。两者都提出了修改要求，这说明冗长程度并不能保证忠实度；即便主题重叠，语义相似度仍可能较低。

4.4.3 对比人类—人类与人类–*AI*的观点采择。总体而言，人机交互条件下的相似度分数呈现出明显的

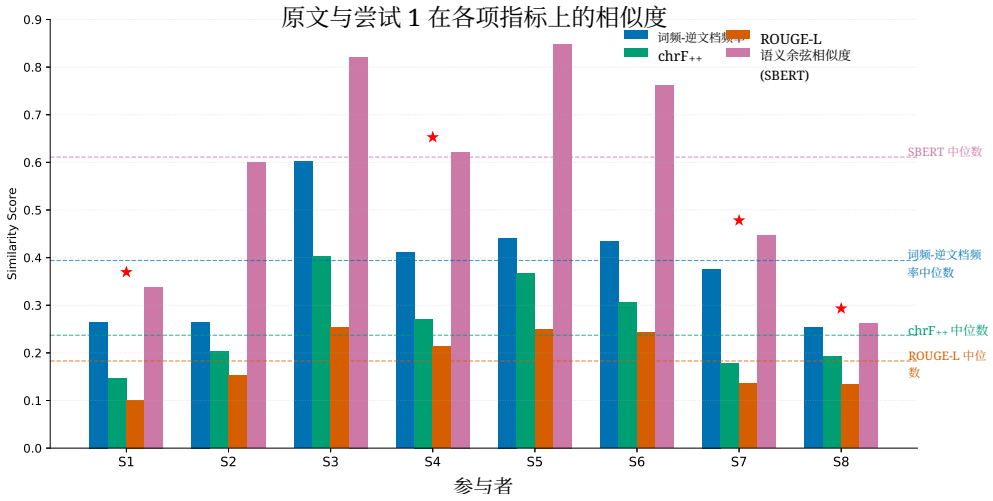


图3：参与者的改述（尝试 1）与原始用户体验在四项指标（TF-IDF余弦相似度、chrF++、ROUGE-L 和基于 SBERT 的语义余弦相似度）上的相似度。虚线表示各项指标的总体中位数。红色星号表示要求修订其尝试 1 的参与者。

表3：人类–AI 在各次尝试中相似度的总体提升。上：从尝试 1 到最佳尝试的变化。下：从尝试 1 到最终尝试的变化。DoM =均值之差；MoD = 差值均值。

指标	趋势	斜率	DoM Δ%	MoD Δ	MoD Δ%	分数↑
		尝试 *1* →最佳尝试				
词频-逆文档频率	↑	0.0386	27.2%	0.1254	55.5%	0.9474
chrF++	↑	0.0182	30.0%	0.0647	68.6%	0.8947
ROUGE-L	↑	0.0085	18.3%	0.0379	54.6%	0.9474
SBERT	↑	0.0352	6.4%	0.0715	14.9%	0.6842
		尝试 *1* →最终尝试				
词频-逆文档频率	↑	0.0153	27.1%	0.0921	43.2%	0.7368
chrF++	↑	0.0027	7.8%	0.0434	56.1%	0.6842
ROUGE-L	↑	0.0033	17.9%	0.0206	39.4%	0.7895
SBERT	↓	-0.0180	-10.8%	-0.0066	1.1%	0.4737

提升，尤其是在比较尝试 1 与表现最佳的尝试时（表3）。四项指标均随时间推移而上升，其中 chrF++ (+30.0% DoM， +68.6% MoD) 和词频-逆文档频率（+27.2% DoM，+55.5% MoD）的相对增幅最大。语义相似度也提升了 6.4%（DoM）和 14.9%（MoD）；与首次尝试相比，近 69% 的参与者在最佳尝试中表现出更高的语义对齐。相比之下，将尝试 1 与最终尝试进行比较时，增幅较小且不那么稳定。词频-逆文档频率和 ROUGE-L 总体上继续上升（分别为 +27.1% 和 +17.9%），但语义相似度总体略有下降（–10.8% DoM），这表明最佳尝试往往出现在最终尝试之前。综上，这些结果表明，参与者通常会在多次尝试中不断完善回答，大多数人在最终轮次之前就达到了相似度峰值。

对尝试 1 中人类–AI与人类—人类条件下的平均相似度得分进行比较（图 4）可以发现，后者在所有指标上与原始叙事的对齐程度始终更高

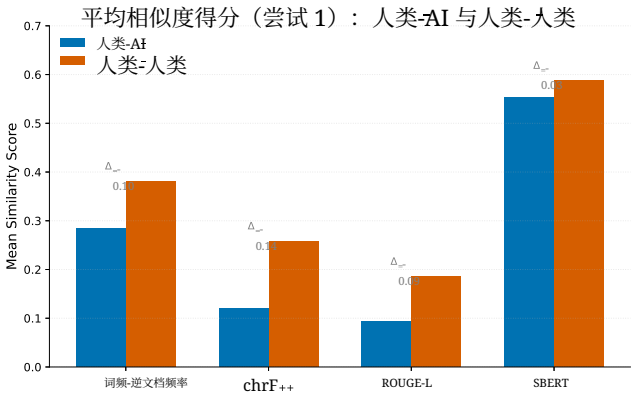


图4：人类–AI与人类—人类在四项指标上的平均相似度得分（尝试1）。每对上方的 Δ 显示 $HA - HH$。

在人与人对话中，词频-逆文档频率（0.3810 vs. 0.2845）、chrF++(0.2581 vs. 0.1210)、ROUGE-L（0.1854 vs. 0.0936）和语义相似度（0.5875 vs. 0.5529）的平均得分均较高。这些差异（从语义相似度的 –0.0346 到 chrF++的 –0.1371）表明，与 AI 互动的参与者相比，与另一位人类合作的参与者起步时与原始视角的基线对齐程度更高。

4.4.4 通过轮流发言分析对话质量。我们的目标是对比人类—人类的观点采择与使用 Perspective Coach 的人类–AI互动。一个关键的情境差异在于轮流发言：在人类–AI对话中，机器人总会作出回应，而且回应通常较长、信息量较大；在人类—人类配对中，发言长度和阐述则取决于角色、熟悉度和权力。

表4：基于轮次的参与度指标（轮次、每轮词数的中位数/平均数、总词数）。带星号的配对*之间存在正式工作关系。

Pair	发言者人口统计特征	轮次	Med.	Mean	总计
P1	S1（白人男性助理教授）	58	6.0	20.8	1204
	S2（有色人种女性博士）	53	6.0	8.5	449
P2*	S3（有色人种男性博士）	46	13.5	20.0	919
	S4（白人男性助理教授）	47	9.0	14.8	696
P3	S5（有色人种女性高年级博士生）	53	11.0	14.5	766
	S6（有色人种女性低年级博士生）	52	7.0	10.1	523
P4*	S7（有色人种男性博士生）	54	14.5	27.2	1469
	S8（白人男性副教授）	56	9.5	17.5	978

第1–4组（人类—人类）的发言机会（发言轮次数）总体上较为均衡，但发言长度会随关系和角色而变化，详情见表4。

第***1***组：发言轮次较为均衡，但阐述不对称：地位较高的参与者发言长得多（平均词数为 20.8 对 8.5；总词数为 1204 对 449，见表4）。这反映了实地情境中的风险：开发者一方往往设定议程并展开阐述，而社区成员则作出较简短的回应。第***2***组：发言轮次数几乎相同，但资历较浅者的发言更长（平均词数为20.0对14.8；总词数为919对696）。简短而明确的工作关系似乎促成了资历较浅者的阐述，这与双方工作关系良好的情况相符。

第***3***组：发言轮次较为均衡；资历较深的参与者发言更长（平均词数为14.5对10.1；总词数为766对523），这表明在身份认同保持不变时，资历与阐述程度相关。第***4***对：发言轮次较为均衡；资历较浅者的发言明显更长（平均词数为27.2对17.5；总词数为1469对978），这再次印证了协作关系中“资深者担任协调员／资历较浅者负责产出”的模式。

5 讨论

(研究问题*1*) 我们的研究结果表明，PerspectiveCoach 能够切实帮助开发者反思其设计决策并考虑替代观点，尤其是边缘化用户的观点。参与者高度评价了该工具深化反思的能力（Q1, $M = 4.68$）以及拓展观点采择的能力（Q2, $M = 4.74$），定性反馈也展现了这一过程在实践中是如何展开的。参与者称这款聊天机器人“具有伦理教育意义”，并赞赏它能够将复杂的伦理问题拆解为可执行步骤，帮助他们“梳理复杂情绪”，并深化情绪表达。这些发现表明，在软件设计情境中，结构化的对话提示可以为批判性反思和共情工作提供支架；传统伦理工具往往因过于抽象或缺乏相关性而难以胜任这些工作 [60, 64, 72]。

然而，结果也揭示了重要的局限性。参与者提到，有时会出现内容泛泛的重复和结构过度僵化的情况。这些发现表明，若工具能根据用户需求动态调整，并避免僵化、规定性过强的流程，就更有利于深入反思。参与者呼吁提供更具体、以实例为基础的指导，并减少泛泛的重复，这指向一种**多元而非单**一的理念：我们将这种诉求视为一个契机，即打造一款能够汇集多种视角，而非仅仅提供一个“中立”的替代代表。我们希望强调，在设计 PerspectiveCoach 时，应让边缘化用户的声音居于核心，同时避免将他们简化为静态形象。因此，我们为 PerspectiveCoach 的未来版本提出两项**设计承诺**：**（1）**为每个边缘化群体提供轮换呈现、情境多元的叙述，而非单一情境的叙述，以回应参与者对内容泛泛、重复冗余的反馈，同时展现边缘化群体需求的多维性；**（2）**提示用户记录具体的设计变更，并在之后的环节中重新审视，以回应参与者对实用、以实例为导向的建议的呼吁，将反思与行动联系起来。

(研究问题*2*) 参与者普遍认为 PerspectiveCoach 易于使用，且与日常设计工作相关。参与者表示未来很可能会继续使用该工具（Q5, $M = 4.68$），并认为其反馈切合实际（Q3, $M = 4.58$），这凸显了该工具将伦理推理融入现有工作流程的潜力，使其不至于显得像是外加环节。参与者还将这款聊天机器人描述为“写作教练”，认为它提升了语言表达的清晰度和准确性，打破了非人格化写作的习惯，并帮助他们阐述基于价值观的推理——这些技能对于记录软件项目中的伦理依据至关重要。

与此同时，参与者指出了可用性方面仍有改进空间。一些人希望引导入门更加清晰，语气更接近日常对话，并且能够偏离预设脚本流程。另一些人则希望获得更加个性化的反馈，并能更灵活地探索旁支话题。这些见解指向了几项可付诸实践的设计策略：（1）在流程中提供一段前言，说明如何进行互动，以及何时应追求精确、何时可以自由联想；（2）加入一个语气选择器（例如，对话式、临床式、反思式），以便定制体验；（3）通过变化策略减少冗余：追踪对话历史，并引入新的角度（来自同一社群的其他视角），而不是重复先前的观点。参与者还建议采用更柔和的视觉界面、提供可选的语音交互，并加快响应节奏，让体验更像协作式辅导，而非评估。

总的来说，这些发现表明，PerspectiveCoach 已被视为一种易于使用且有价值的工具，可将伦理反思融入开发实践。然而，增强其适应性、个性化程度和用户能动性，是提升其相关性并在多样化的工作流程和用户偏好中持续保持用户参与的关键。

(*RQ3*) 我们对文本相似度和对话动态的分析揭示了 PerspectiveCoach 相较于人类—人类对话的潜力与当前局限。积极的一面是，人机交互条件下的参与者在多次尝试中显著改善了复述（表3）。从尝试 1 到之后的最佳尝试，四项指标均有所提升，包括词频-逆文档频率（+27.2%）、chrF++（+30.0%）、ROUGE-L（+18.3%）和 SBERT（+6.4%），这表明结构化互动帮助参与者随着时间推移捕捉用户体验的表层和语义方面。然而，将尝试 1 与最终尝试进行比较时，语义相似度略有下降（–10.8%），这表明语义对齐的峰值往往更早出现。我们要求参与者完成五轮互动，但当 PerspectiveCoach 评估认为参与者已经充分理解该视角后，后续交流便从复述转向了对开发者自身位置性和技术干预的反思。

自身位置性和技术干预。例如，在充分理解该视角后，一位参与者对PerspectiveCoach说："这让我开始质疑一切：我们有多少所谓'中立'的设计选择 *[.....]* 实际上促成了这一点？"另一位参与者反思道："这让我重新思考，作为一名开发者，我是否真的在做对社会有益的事情 *[.....]* 它将我的思考引向了为所有人，尤其是女性，创建更安全的平台。"

与人际对话环节的比较进一步为这些结果提供了背景。人类对话中的参与者在所有指标上的初始基线都更高，这表明人际情境（即社会利害关系、认知要求、任务结构）可能有助于人们从一开始就更准确地捕捉含义。此外，对话动态也存在显著差异：人类—人类配对中的发言轮次数总体上较为均衡，但发言长度往往因地位、角色和既有关系而异。研究表明，权力不对称会限制弱势发言者的叙事自由，导致自我噤声（提供的信息不足）和掠夺性表达（提供的信息过多）[12, 33, 36, 42]。例如，在第1组中，地位较高的参与者发言明显更长，这与混合性别及权力不对称情境中记录的模式相呼应 [3, 30, 75]。这些发现表明，尽管 AI 能确保边缘化视角持续得到表达，却无法复现人类对话中的相互调整、协商与情境共享。然而，Messeri 的《虚幻之地》警示我们，情感技术让边缘化表征近在咫尺，却让真实的边缘化群体成员远离设计流程 [43], ；而 Nakamura 则提醒我们，共情工作可能满足用户的道德自我形象，却无法改变权力格局 [45]。因此，我们为 PerspectiveCoach 的下一版本提出另外两项**设计承诺**：**(3)** 引导式练习模式，让开发者通过演练，在真实对话中实践对话平等与空间再分配，以应对人际研究中观察到的轮流发言不对称问题；在该研究中，地位较高的参与者主导了阐述。**(4)** 邀请原帖发布者 或其他边缘化群体的维护者创建可编辑、可撤销的情境，以回应我们的发现：参与者提出的修订要求并不总是与文本相似度分数一致。

6 局限性与未来工作

我们的工作属于探索性研究，应结合若干局限性来解读。首先，PerspectiveCoach 是早期设计探针，而非可投入生产使用的系统。其对话脚手架、反馈策略和交互流程均经过有意简化，目的是测试可行性和用户反应，而非最大化性能。未来版本应纳入更多自适应反馈、语气控制机制，以及支持倾听导向交互的机制。

其次，两项研究都采用了规模较小且具有特定特征的样本。人类–AI研究纳入了来自WEIRD英语国家的18名前端开发者。这种有意聚焦于全球北方开发者视角的做法限制了可推广性；未来的研究应探讨这些效应是否也适用于规模更大、人口统计特征更多元的样本。此外，人际研究仅涉及从我们自己的研究社群中招募的四对参与者。尽管这一规模较小且彼此熟悉的样本限制了我们能够捕捉到的视角广度，但它也带来了独特的优势：由于我们对参与者的职业背景、人际关系和沟通风格都有第一手了解，因此能够以更强的情境敏感性解读他们的互动和权力动态。鉴于已有大量证据论述边缘化对话者与主导地位对话者之间的对话动态，我们的目的并不是再次证明这些观点，而是在人类—人类版本的任务中考察：当人类–AI设计由对话工具介导时，可能获得或放弃什么。我们的研究是开展这一任务的起点；通过对两项研究的完整文字记录进行分析，将获得更深入的洞见。

第三，我们的实验任务聚焦于单一情境：开斋节期间针对穆斯林女性的骚扰与物化。选择这一情境是为了凸显一个显著的关切，但它并未涵盖开发者会遇到的全部伦理问题。应用于隐私、无障碍性、劳动或其他领域相关的关切时，参与者的回应和效果可能有所不同。此外，由于使用了 Prolific，我们必须考虑社会赞许偏差的可能性。在 Prolific 上，参与者的报酬与任务审核通过与否挂钩，这可能会促使他们给出被认为较好的回答，而非真实的回答。PerspectiveCoach 依赖 OpenAI 的 GPT 平台，因此具有非确定性：相同的提示在不同会话中可能产生略有差异的输出。这有助于个性化反思，但也使复现和一致性更难保证。其次，我们的文本相似度指标虽有参考价值，但并不能全面衡量视角采择质量。同样，我们的对话指标（发言轮次数、冗长程度）只是倾听和投入程度的间接指标。未来工作可通过对对话内容和互动动态进行定性分析，补充这些衡量方式。一场人际会话在线上进行，这可能会微妙地改变对话流程（由于延迟，轮次间隔更长、重叠发言更少）[14, 63, 68]；不过，对治疗会谈的比较研究常发现，视频与面对面场景在投入程度和关系结果方面总体相似 [18, 23, 50, 62]。鉴于我们的任务强调重新表述和相互肯定，且仅有一对参与者在线上进行会谈，我们预计偏差有限。

最后，作为一项探索性研究，我们考察了受控的单次会议情境中的工具使用情况。我们并未评估换位思考实践或认识论谦逊如何随时间演变，也未探究 PerspectiveCoach 如何融入现实世界的软件开发工作流程。要了解此类工具在专业环境中的持续影响和实际效用，还需要进行纵向部署、采用更多样化的数据集，并开展实地评估。

7 结论

通过将多元性、问责以及对话空间的再分配置于开发者实践的核心，未来版本的 PerspectiveCoach 可以超越"理解边缘化用户"，转而促进"与他们共同理解"。如此一来，它为重新构想对话式人工智能如何支持更公平的设计实践提供了起点：它并非替代交流，而是一种帮助开发者做好准备、以更谦逊、更专注、更关怀的态度参与对话的工具。这项探索性研究展现了基于大语言模型的工具在弥合技术设计者与受技术影响最深的人群之间长期存在的鸿沟方面的潜力。

参考文献

[1] Icek Ajzen。1991。计划行为理论。《组织行为与人类决策过程》50，2（1991），179–211。doi:10.1016/0749-5978(91)90020- T 认知自我调节理论。[2] Razieh Alidoosti、Patricia Lago、Maryam Razavian 和 Antony Tang。2025年。探索软件系统中的伦理价值：系统性文献综述。《系统与软件杂志》 226 2025 112430 10 1016 2025 112430 ( )，。doi: . /j.jss. 。Kristin J [3] Anderson 和 Campbell Leaper。。对对话打断中性别效应的荟萃分析：谁、什么、何时、何地以及如何。《性别角色》39，3 1998 225 252（），–。[4] Joost Broekens、Bernhard Hilpert、Suzan Verberne、Kim Baraka、Patrick Gebhard 和 Aske Plaat。2023年。大语言模型中诵现的细粒度情感处理能力。载于《2023年情感计算与智能交互国际会议（ACII）论文集》。IEEE，马萨诸塞州剑桥，美国，1–8。doi:10.1109/ACII59096.2023.10388177[5] Lidiany Cerqueira、João Pedro Bastos、Danilo Neves、Glauco Carneiro、Rodrigo Spínola、Sávio Freire、José Amancio Macedo Santos 和 Manoel Mendonça。2025年。探索软件工程中的共情：从从业者视角的灰色文献分析中获得的见解——RCR报告。《ACM软件工程与方法论汇刊》（2025年10月）。doi:10.1145/3771771 已接收。[6] Shruthi Sai Chivukula、Chris Rhys Watkins、Rhea Manocha、Jingle Chen 和 Colin M. Gray。2020年。塑造伦理意识的用户体验实践维度。载于《2020年CHI计算系统中的人为因素大会论文集》（美国夏威夷州檀香山）（CHI '20）。计算机协会，纽约，美国， 1 13 10 1145 3313831 3376459 –。doi: . / . 内容来源与真实性联盟（C2PA）。2023年。C2PA危害 [7] 建模——规范1.0版。https://c2pa.org/specifications/specifications/1.0/security/Harms_Modelling.html。访问日期：2025年5月14日。[8] Sasha Costanza-Chock。2020年。《设计正义：建设我们所需世界的社区主导实践》。麻省理工学院出版社，马萨诸塞州剑桥。[9] Molly J Crockett。2025年。共情：深与浅。可见于SSRN 5862422（2025年）。[10] Shoaib Daniyal。2021年。Ritesh Jha（又名 Liberal Doge）：在直播中肆意攻击巴基斯坦女性的人。https: // w w w. n e w s l a u n d r y. c o m/ 2 0 2 1/ 0 5/15/rites h-jha-aka-liberal-doge-the-man-behind-the-livestream-spewing-hate-against-pakistani-women 访问日期：2025年5月14日。[11] Christian Detweiler 和 Maaike Harbers。2014年。价值故事：将人类价值观融入需求工程。。载于 REFSQ研讨会，第1138卷。Springer Nature，德国埃森，2–11。[12] Kristie Dotson。2011年。追踪认识论暴力，追踪噤声实践。《希帕提娅》26，2（2011），236–257。doi:10.1111/j.1527-2001.2011.01177.x [13] Amy Edmondson。1999年。工作团队中的心理安全与学习行为。《行政科学季刊》44，2（1999），350–383。[14] David W Edwards。2025年。电信延迟对说话者转换时机的影响。《语音通信》171（2025），103226。[15] 美国平等就业机会委员会（EEOC）。2024年。高科技劳动力与行业的多样性：2014—2022年。https: //www.eeoc.gov/sites/default/files/2024-09/20240910_Diversity% 20in%20the%20High%20Tech%20Workforce%20and%20Sector%202014-2022.pdf 访问日期：2024-12-27。[16] Yagil Elias、Tom P. Humbert、Lauren Olson 和 Emitza Guzman。2025年。软件有哪些不道德之处？荷兰的用户感知。载于《2025年IEEE/ACM第47届国际软件工程大会：社会中的软件工程（ ICSE-SEIS）》。IEEE计算机学会，洛斯阿拉米托斯，美国，118–129。doi:10.1109/ICSE-SEIS66351.2025.00017[17] 伦理工具包。2024年。负责任人工智能伦理工具包。https://ethicstoolkit.ai。访问日期：2025年5月14日。[18] Ephrem Fernandez、Yilma Woldgabreal、Andrew Day、Tuan Pham、Bianca Gleich 和 Elias Aboujaoude。2021年。视频心理治疗与面对面心理治疗：疗效及其与治疗类型和目标关系的荟萃分析。《临床心理学与心理治疗》28，6（2021），1535–1549。[19] Google Research。2019年。用于模型报告的模型卡。https://modelcards.withgoogle.com。访问日期：2025年5月14日。[20] Colin M. Gray。2016年。“这更像是一种思维方式，而非一种方法”：用户体验从业者对设计方法的构想。载于《2016年CHI计算系统中的人为因素大会论文集》（美国加利福尼亚州圣何塞）（CHI '16）。计算机协会，纽约，美国，4044–4055。doi:10.1145/2858036.2858410[21] Colin M. Gray 和 Shruthi Sai Chivukula。2019年。用户体验实践中的伦理调解。载于《2019年CHI计算系统中的人为因素大会论文集》（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，纽约，美国， 1 11 10 1145 3290605 3300408 –。doi: . / . Colin M. Gray、Shruthi Sai Chivukula、 [22] Thomas V Carlock、Ziqing Li 和 Ja-Nae Duane。2023年。为实践共鸣搭建以伦理为重点的方法框架。载于《2023年ACM交互系统设计大会论文集》（美国宾夕法尼亚州匹兹堡）（DIS '23）。计算机协会，纽约，美国，2375–2391. doi:10.1145/3563657.3596111[23] Hannah Greenwood、Natalia Krzyzaniak、Ruwani Peiris、Justin Clark、Anna Mae Scott、Magnolia Cardona、Rebecca Griffith 和 Paul Glasziou。2022年。远程医疗与面对面心理治疗对较少见心理健康状况的疗效比较：随机对照试验的系统综述与荟萃分析。JMIR Mental Health 9, 3 (2022年), e31780. [24] Max Grusky。2023年。Rogue Scores。载于第61届计算语言学协会年会论文集（第1卷：长篇论文），Anna Rogers、Jordan Boyd-Graber 和 Naoaki Okazaki（编）。计算语言学协会，加拿大多伦多， 1914 1934 10 18653 1 2023 107 – 。 [25] 2024 doi: 。/v / .acl-long。Hashini Gunatilake、John Grundy、Rashina Hoda 和 Ingo Mueller。。软件开发者与用户互动中的共情促进因素与障碍：一项混合方法案例研究。ACM软件工程与方法学汇刊 33, 4 (2024年), 1–41. [26] Hashini Gunatilake、John Grundy、Rashina Hoda 和 Ingo Mueller。2025年。软件工程中共情的作用——社会技术扎根理论。ACM软件工程与方法学汇刊 (2025年)。投稿中。[27] Hashini Gunatilake、John Grundy、Ingo Mueller 和 Rashina Hoda。2023年。共情模型与软件工程——初步分析与分类体系。《系统与软件杂志》203 (2023年), 111747. [28] Emitza Guzmán、Ricarda Anna-Lena Fischer 和 Janey Kok。2024年。关注差距：软件开发中的性别、微观不平等与障碍。《实证软件工程》29, 1 (2024年), 17. [29] Tomasz Hollanek。2024年。设计工具包的伦理政治：负责任的人工智能工具，从大型科技公司指南到女性主义构思卡片。《人工智能与伦理》5 (2024年), 2165–2174。[30] Christopher F Karpowitz、Tali Mendelberg 和 Lee Shaker。2012年。审议式参与中的性别不平等。《美国政治学评论》106, 3 (2012年), 533–547. [31] Vera Khovanskaya、Phoebe Sengers、Melissa Mazmanian 和 Charles Darrah。2017年。重新弥合设计与民族志之间的鸿沟。载于2017年CHI计算系统中的人为因素大会论文集（美国科罗拉多州丹佛）（CHI '17）。计算机协会，美国纽约州纽约市， 5373 5385 10 1145 3025453 3026051 – 。doi: 。/ 。Daan Kieft、 [32] Laura Duits 和 Emitza Guzmán。2025年。用户的底线在哪里？关于软件的伦理关切。载于2025年IEEE第33届国际需求工程会议（RE）。 155 166 10 1109 63999 2025 00024 – 。 [33] 2025 doi: 。/RE 。 。Judy Kim 和 Molly J Crockett。 。低权力限制了发言者可使用的叙事空间。载于认知科学学会年会论文集，第47卷。[34] Foivos Kotsogiannis、Ioanna Spentza 和 Chrysanthi Nega。2024年。观点采择对智识谦逊的影响及其与确认偏误的关系。载于认知科学学会年会论文集，第46卷。认知科学学会，荷兰鹿特丹，5377–5384. [35] Timo Kühbacher、Tim Schlippe 和 Kristina Schaaff。2025年。哪个聊天机器人是最有共情力的教师？载于《教育人工智能技术：新进展与创新实践》，Tim Schlippe、Eric C. K. Cheng 和 Tianchong Wang（编）。施普林格·自然新加坡，新加坡，56–73. [36] Rebecca Kukla。2014年。施为力量、惯例与话语不公。《希帕提娅》29, 2 (2014年), 440–457. [37] Yoon Kyung Lee、Jina Suh、Hongli Zhan、Junyi Jessy Li 和 Desmond C. Ong。2024年。大语言模型会生成被认为具有共情力的回应。（ 2024 63 71 10 1109 63134 2024 00012 ），– 。doi: 。/ACII 。 。 [38] 2004 Chin-Yew Lin。 。Rouge：用于自动评估摘要的软件包。载于《文本摘要的新领域》。ACL Anthology，西班牙巴塞罗那，74–81. [39] Christopher D Manning。2008年。《信息检索导论》。Syngress Publishing，马萨诸塞州。[40] Valentin Markulj、Kousar Aslam 和 Emitzá Guzmán。2024年。软件行业中的微观不平等与移民背景。载于第46届国际软件工程大会：社会中的软件工程论文集。IEEE，葡萄牙里斯本，23–33. [41] D Matthews。2006年。认识论谦逊：科学哲学视角。施普林格，105–137. [42] Rachel Ann McKinney。2016年。摘录话语。《社会理论与实践》(2016年), 258–284。[43] Lisa Messeri。2024年。《虚幻之地：洛杉矶的虚拟与其他现实》。杜克大学出版社，美国北卡罗来纳州达勒姆。[44] Microsoft Azure 架构中心。2024年。负责任的创新——设计指南。https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/。访问日期：2025年5月14日。[45] Lisa Nakamura。2020年。为自己的负面感受而感到愉悦：美德虚拟现实与种族共情的自动化。《视觉文化杂志》19, 1 (2020年), 47–64. [46] David J Nicol 和 Debra Macfarlane-Dick。2006年。形成性评价与自我调节学习：一种模型与优质反馈实践的七项原则。《高等教育研究》31, 2 (2006年), 199–218. [47] Lauren Olson、Emitza Guzmán 和 Florian Kunneman。2023年。沿着边缘：边缘化社群对社交平台的伦理关切。载于

2023年IEEE/ACM第45届国际软件工程大会：社会中的软件工程（ICSE-SEIS）。IEEE，IEEE，澳大利亚墨尔本，71–82. [48] OpenAI。2023年。推出 GPT。https://openai.com/index/introducing-gpts/。访问日期：2025年12月30日。[49] Geeta Pandey。2021年。Sulli Deals：印度女性在一款“病态”应用上被在线拍卖。https://www.bbc.com/news/world-asia-india-57764271 访问日期：2025年5月14日。[50] Tessa Peasgood、Mackenzie Bourke、Nancy Devlin、Donna Rowen、Yaling Yang 和 Kim Dalziel。2023年。在线访谈与面对面访谈对健康状态估值的随机对照比较。《社会科学与医学》323 (2023年), 115818. [51] Maja Popović。2015年。chrF：用于自动机器翻译评估的字符 n 元语法 F 分数。载于第十届统计机器翻译研讨会论文集。ACL Anthology，丹麦哥本哈根，392–395. [52] Tenelle Porter 和 Karina Schumann。2018年。智识谦逊与对立观点的开放态度。《自我与身份认同》17, 2 (2018年), 139–162. [53] Prolific。2025年。导出人口统计数据。https://researcher-help.prolific.com/en/article/b2943f。访问日期：2025年12月30日。[54] Reid Pryzant、Dan Iter、Jerry Li、Yin Lee、Chenguang Zhu 和 Michael Zeng。2023年。借助“梯度下降”和束搜索进行自动提示优化。载于2023年自然语言处理实证方法会议论文集，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，新加坡，7957–7968. doi:10.18653/v1/2023.emnlp-main.494[55] Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。（2019年11月），3982–3992. doi:10.18653/v1/D19-1410[56] Johnny Saldana。2021年。《定性研究者编码手册》。SAGE Publications Ltd，伦敦：。http://digital.casalini.it/9781529755992 1975 [57] Gerard Salton、Anita Wong 和 Chung-Shu Yang。。用于自动索引的向量空间模型。Commun. ACM 18, 11 (1975年), 613–620. [58] Kristina Schaaff、Caroline Reinig 和 Tim Schlippe。2023年。探索 ChatGPT 的共情能力。载于2023年第11届国际情感计算与智能交互会议（ACII）。 1 8 10 1109 59096 2023 10388208 – 。doi: 。/ACII 。 。[59] Jan-Hendrik Schmidt、Sebastian Clemens Bartsch、Martin Adam 和 Alexander Benlian。2023年。问责不一致及其对人工智能开发者工作满意度的影响。载于ECIS。AIS，挪威克里斯蒂安桑，1–19. [60] Andrew D. Selbst、Danah Boyd、Sorelle A. Friedler、Suresh Venkatasubramanian 和 Janet Vertesi。2019年。社会技术系统中的公平性与抽象。载于公平、问责与透明度会议论文集（美国佐治亚州亚特兰大）（FAT* ’19）。计算机协会，美国纽约州纽约市， 59 68 10 1145 3287560 3287598 – 。doi: 。/ 。Phoebe Sengers、Kirsten Boehner、[61] Shay David 和 Joseph ’Jofish’ Kaye。2005年。反思性设计。载于第四届批判性计算十年会议：在理性与感性之间（丹麦奥胡斯）论文集（CC ’05）。计算机协会，美国纽约州纽约市，49–58. doi:10.1145/1094562.1094569[62] Patrik D Seuling、Johannes C Fendel、Lukas Spille、Anja S Göritz 和 Stefan Schmidt。2024年。视频会议心理治疗与面对面心理治疗中的治疗联盟：系统综述与荟萃分析。《远程医学与远程护理杂志》30, 10 (2024年), 1521–1531. [63] Lucas M Seuren、Joseph Wherton、Trisha Greenhalgh 和 Sara E Shaw。2021年。到底轮到谁了？视频媒介互动中的延迟与轮流发言组织。《语用学杂志》172 (2021年), 63–78. [64] Sean Sirur、Jason R.C. Nurse 和 Helena Webb。2018年。我们做到了吗？理解遵守《通用数据保护条例》（GDPR）所面临的挑战。载于第二届多媒体隐私与安全国际研讨会论文集（加拿大多伦多）（MPS ’18）。计算机协会，美国纽约州纽约市， 88 95 10 1145 3267357 3267368 – 。doi: 。/ 。Rachel Charlotte [65] Smith、Heike Winschiers-Theophilus、Daria Loi、Rogério Abreu de Paula、Asnath Paula Kambunga、Marly Muudeni Samuel 和 Tariq Zaman。2021年。设计实践去殖民化：迈向多元宇宙性。载于2021年CHI计算系统中的人为因素大会扩展摘要集（日本横滨）（CHI EA ’21）。计算机协会，美国纽约州纽约市，文章 83, 5 10 1145 3411763 3441334 页。doi: 。/ 。Clay [66] Spinuzzi。。参与式设计方法论。《技术传播》52, 2 (2005年), 163–174. [67] Bismme Taskin。2021年。“什么是同意”，因性别歧视和种姓歧视性辱骂、在线拍卖女性而被立案的“Liberal Doge”发问。 h t t p s:// t h e p r i n t. i n/ i n d i a/ w h at-is-c onsent-asks-liberal-doge-booked-for-sexist-casteist-slurs-auctioning-women-online/809379/ 访问日期：2025年5月14日。[68] Ying Tian、Siyun Liu 和 Jianying Wang。2024年。在线音频、在线视频与面对面对话中轮流发言差异的语料库研究。《语言与言语》67, 3 (2024年), 593–616. [69] Neelam Tjikhoeri、Lauren Olson 和 Emitzá Guzmán。2024年。以最佳方式达成最佳结果：应用程序评论中的伦理关切。《实证软件工程》29, 6 (2024年), 138. [70] Sherry Turkle。2011年。《屏幕上的生活》。Simon and Schuster，美国纽约州纽约市。

[71] Emily A Vogels。2021年。在线骚扰现状。第13卷。皮尤研究中心，华盛顿特区。[72] Richmond Y Wong、Michael A Madaio 和 Nick Merrill。2023年。像工具包一样看待：工具包如何构想人工智能伦理工作。《*ACM* 人机交互汇刊》7，计算机支持的协同工作1（2023年），1–27。[73] David Wood、Jerome S Bruner 和 Gail Ross。1976年。辅导在问题解决中的作用。《儿童心理学与精神病学杂志》17，2（1976年），89–100。[74] Yongchao Zhou、Andrei Ioan Muresanu、Ziwen Han、Keiran Paster、Silviu Pitis、Harris Chan 和 Jimmy Ba。2022年。大语言模型已达到人类提示工程师的水平。载于第十一届国际学习表征会议。国际学习表征会议，线上。[75] Dean H Zimmerman 和 Candace West。2008年。对话中的生理性别角色、打断与沉默。载于迈向批判社会语言学。约翰·本杰明出版公司，211–236。