设计反馈：理解并克服对话代理中的用户反馈障碍

Nikhil Sharma*
nsharm27@jhu.edu 约翰斯·霍普金斯大学 巴尔的摩，马里兰州，美国

Zheng Zhang
zhengzhang@adobe.com
Adobe 公司 圣何塞，加利福尼亚州，美国

Daniel Lee
dlee1@adobe.com
Adobe 公司 圣何塞，加利福尼亚州，美国

Namita Krishnan
namitak@adobe.com
Adobe 公司 圣何塞，加利福尼亚州，美国

Guang-Jie Ren
gren@adobe.com
Adobe 公司 圣何塞，加利福尼亚州，美国

Ziang Xiao
ziang.xiao@jhu.edu 约翰斯·霍普金斯大学 美国马里兰州巴尔的摩

Yunyao Li
yunyaol@adobe.com
Adobe 公司 美国加利福尼亚州圣何塞

arXiv:2602.01405v1 [cs.HC] 1 Feb 2026

摘要

高质量反馈对于有效的人机交互至关重要。它有助于弥合知识差距、纠正偏离，并在交互过程中及模型开发的整个过程中塑造系统行为。然而，尽管人类反馈至关重要，人类向人工智能提供反馈的频率往往很低，质量也不高。这一差距促使我们深入审视人类在与人工智能交互过程中提供的反馈。为了解并克服阻碍用户提供高质量反馈的挑战，我们开展了两项研究，考察人类与对话代理（CA）之间的反馈动态。我们的形成性研究以格赖斯准则为视角，识别出四种阻碍用户提供高质量反馈的反馈障碍——共同基础、可验证性、沟通和信息丰富度。在这些发现的基础上，我们提出了三项设计要求，并表明，融入与这些要求相契合的支架后，系统能够帮助用户提供更高质量的反馈。最后，我们呼吁更广泛的人工智能社区采取行动，推动大语言模型能力取得进展，以克服反馈障碍。

CCS 概念

• **以人为中心的计算 → 人机交互（HCI）**;**人机交互设计与评估方法**; **用户研究**;**人机交互实证研究**;**协作与社会计算**; •**信息系统**;

关键词

人机交互、人机协作、用户反馈、反馈障碍、反馈质量、人机协作、生成式人工智能、大语言模型、对话代理

ACM 参考文献格式：
Nikhil Sharma、Zheng Zhang、Daniel Lee、Namita Krishnan、Guang-Jie Ren、Ziang Xiao 和 Yunyao Li。2026。反馈设计：理解并克服对话代理中的用户反馈障碍。载于*2026*年 *CHI*人因计算系统大会论文集（*CHI '26*），*2026*年*4*月*13*日至*17*日，西班牙巴塞罗那。ACM，美国纽约州纽约市，23 页。https://doi.org/10.1145/3772318.3791600

1 引言

基础模型 [88] 和智能体人工智能 [28]的快速发展，使对话代理（CAs）成为人与人工智能交互的核心范式之一。这些系统如今能够基于复杂的知识库和动态上下文进行推理，以满足用户多元化的目标，同时利用人类反馈不断优化其在多轮对话中的回应。在多轮目标导向任务中，用户与模型拥有共同目标；高质量的人类反馈对于确保对话代理始终契合用户不断变化的意图和需求至关重要，从而帮助用户在长时间的对话中成功实现目标。

人类反馈在当今的机器学习（ML）研究中也发挥着关键作用 [7, 66, 126, 155]。直接偏好优化（DPO） [97] 或基于人类反馈的强化学习（RLHF） [90]等技术，会从人类反馈中提取偏好信号，以引导模型行为。随着更强大且嵌入于各种高风险领域的模型不断涌现，人类反馈质量将影响系统行为和输出 [31, 54, 81]。

然而，近期文献揭示了一个令人担忧的脱节现象：尽管人类反馈至关重要，但人们在现实场景中与对话代理互动时，很少提供高质量反馈 [26, 83, 112]。例如，*WildFeedback*[112], 这一数据集收录了超过 100 万次 ChatGPT 对话中真实场景里的人工反馈。数据显示，在 102 万次对话中，只有 38,992 次（3.89%）包含反馈。此外，人们给出的反馈

*本研究是在作者于 Adobe 实习期间完成的

本作品采用知识共享署名 4.0 国际许可协议。*CHI '26*，西班牙巴塞罗那© 2026 版权归所有者/作者所有。ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791600

往往含糊不清且不完整，例如“表达不够自然”“错了”和“错误，请再试一次”。即使是在复杂的多轮任务中——而高质量反馈在这类任务中最有助于实现用户目标——这一模式依然存在。这一结果揭示了一个关键差距：尽管人工智能系统旨在从人类输入中学习并进行调整，用户却始终未能提供清晰、具体的反馈来改善互动结果。尽管合成反馈已成为一种可扩展的替代方案[68]，，但合成反馈无法充分呈现人类互动时带来的情境细微差异和多元文化视角[108, 119]，因此，在当今的人工智能训练范式中，人类反馈仍不可替代。

与人类对会话代理的反馈形成对比的是，在许多日常情境中，人类会频繁提供高质量反馈；在教育和组织中，人们经常交换及时、具体的反馈。研究表明，人类善于提供有效反馈 [41, 121]。在协作工作环境中，人类经常提供结合上下文、可执行的反馈，帮助他人提升表现 [39]。即使在非正式场合，人们也会采取有效的反馈行为——从指导运动队到辅导同事——这表明提供高质量反馈的能力是人类社会互动的固有组成部分 [59]。这引出了一个关键问题：如果人类能够在其他领域有效地提供反馈，那么是什么阻碍了他们在与对话代理互动时提供高质量反馈?

人机交互研究探讨了如何在人类与人工智能的互动中引出有效反馈 [2, 25, 42, 89, 124]。提出的支架从能够与用户动态协商控制权的混合主动式系统 [3, 43, 44]，到降低用户提供纠正性输入门槛的交互式机器学习框架 [4, 29, 115]，不一而足。这些方法强调了一些设计策略，例如公开系统状态、组织用户输入，以及在对话中共享倡议，以提升用户反馈的质量和频率。然而，大多数技术是为基于规则或混合系统开发的，这些系统缺乏处理多轮复杂任务或在交互过程中纳入细致反馈的能力 [15, 152]。

大语言模型驱动的对话代理能够进行自然的多轮对话，这一新兴技术引出了新的问题：哪些障碍限制了用户对大语言模型驱动的对话代理提供反馈（**研究问题**一）？我们又该如何设计脚手架，以促进用户在更丰富、更有效的交互中持续提供高质量反馈（**研究问题2**）？

为探究对话代理中的反馈动态，我们开展了两项研究。为回答**研究问题**一，我们进行了形成性研究，招募了16名参与者，他们经常使用会话代理完成各种目标导向任务，例如创意写作、编程、图像编辑、研究和搜索。通过深入访谈，我们识别了人机交互中常见的失效情况、用户向模型提供反馈的方式，以及阻碍用户提供高质量反馈的因素。借助格赖斯准则[36], 这一视角——该准则描述了人类对话所遵循的合作原则——我们识别并定义了四种关键的**反馈障碍**。

在第二项研究中，我们设计并实现了*FeedbackGPT*，它整合了六种模型无关支架，每种支架都针对研究1中发现的一种特定反馈障碍。为回答**研究问题2**，我们招募了20名参与者开展被试内研究，让他们使用基线界面完成两项目标导向任务（ChatGPT）和 FeedbackGPT。通过混合方法分析，我们发现，提供能够最大限度减少反馈障碍的支架，可以让用户提供质量更高且更频繁的反馈。

总而言之，我们的工作主要有三项贡献：

- 首先，我们提供了关于**反馈障碍**的实证见解，这些障碍会阻碍用户在与对话代理交互时提供高质量反馈。
- 第二，我们介绍了FeedbackGPT的设计与实现，该系统整合了一套与模型和任务无关的支架，能够有效鼓励用户更频繁地提供更高质量的反馈。
- 第三，我们提出了一套实用的设计建议，用于在与对话代理交互的过程中引出高质量的用户反馈。

2 相关工作

2.1 人类–人工智能交互中的反馈

交互系统长期以来都需要人类反馈来完成任务并实现对齐。在人类–人工智能系统的设计与评估中，反馈通常被定义为系统采取行动之后所提供的信息，用以表明其行为在多大程度上达成了预定目标 [39, 114]。研究区分了两种主要的反馈形式：显式反馈，例如评分或纠正，以及隐式反馈，即从用户行为中被动获取的反馈（例如点击、跳过、停留时间）。早期的人在回路（人在回路）与交互式机器学习（交互式机器学习）研究将反馈视为模型和界面增量改进的关键 [82, 129]

在人机交互和机器学习文献中，如何 eliciting 有效的用户反馈长期以来一直是核心议题。混合主动式系统强调人类与代理之间的共同控制，使用户能够持续引导系统行为，而不是仅发出一次性指令 [3, 43] 。交互式机器学习进一步将用户视为协作者：用户反复标注、修正并重新调整示例的权重，使模型能够在实际情境中进行适应 [4, 27, 96]。与此相辅相成的是，关于解释性调试的研究表明，当系统呈现其推理过程时，人们能够更有效地理解和调试模型行为，并提供更具针对性的纠正性反馈 [62, 117]。

同样，推荐系统研究将反馈视为个性化的基础。近期关于交互式和对话式推荐系统的研究创建了相应界面，鼓励用户随着时间推移对建议进行批评和完善，而不是被动接受 [16, 40, 45]。例如，Xiao 等人 [138] 研究了音乐流媒体服务的语音助手如何 eliciting 显式反馈，并表明界面层面的选择会显著影响用户何时以及如何愿意提供修正。这些研究表明，经过周密设计的反馈渠道能够显著提升模型与用户偏好之间的对齐程度。

大语言模型驱动的对话代理承袭了混合主动式交互系统的发展脉络，但在几个关键方面有所不同。与传统的交互式机器学习或推荐系统不同，对话代理支持开放式多轮任务，其中目标、策略和评估标准会随时间不断演变。近期关于人类与大语言模型协作的研究记录了在复杂工作流程中维持共同目标、决定何时信任或纠正模型，以及协商倡议权方面的挑战 [50, 74, 149]。我们的工作在这一传统基础上，专门聚焦于协作的 反馈阶段： 我们

通过实证研究刻画 CA 交互中的反馈如何以及为何失效，并提出界面级支架，使反馈更容易表达、解读和纳入。

2.2 基于大语言模型的系统中的用户交互

用户通过提供两种类型的输入与由大语言模型驱动的系统交互：用户提示和用户反馈。二者在由大语言模型驱动的系统中发挥着不同的作用：提示词是前瞻性控制信号，旨在模型采取行动前防止错误；反馈则是在生成响应后进行的回顾性评估。在机器学习和对齐领域的文献中，二者都被用作引导和改进模型的信号。提示工程研究侧重于制定有效的初始指令，将提示词视为事前的任务特定模板，用于指定目标、约束和输出格式，而不改变模型参数 [100, 133]。人在回路的提示优化方法（如 CoTAL）将提示设计本身视为迭代学习问题，结合思维链提示与主动学习，使用户能够逐步完善评分提示 [21]。然而，Zamfirescu-Pereira 等人的实证研究 [143] 表明，即使是精通技术的非专家，也难以将自己的目标转化为有效的首轮提示，因此事后信息（反馈）对于引导模型至关重要。

机器学习领域的工作通常主要将反馈视为从大规模人类-大语言模型聊天日志中提取的训练信号。WildFeedback 会自动识别用户与 ChatGPT 在实际使用场景中的对话里以自然语言表达的批评意见和偏好，并利用这些内容构建偏好数据集；但研究发现，大多数可用的监督信号来自简短的重试和简洁的后续消息，而非丰富、结构化的批评意见 [112]。基于人类交互的强化学习（RLHI）同样将 WildChat-1M 语料库重新用作交互级监督的数据来源，区分修订先前回答的后续消息与开启新任务的消息，并强调只有少数轮次包含明确的纠正信号 [48]。关于反馈的相关研究还表明，结构松散的评论（例如点赞/点踩、简短回复）噪声大、稀疏，且倾向于集中在粗略的接受/拒绝判断上 [26]。

总的来说，这一以机器学习为中心的研究方向展现了从真实世界交互中学习的潜力，但也揭示了一个核心局限：高质量反馈数据十分稀缺、零散，而且很少能体现细致对齐所需的丰富结构。现有方法大多将反馈视为附加在轮次上的标签，却没有探究用户为何 一开始就很少提供详尽、规范的反馈。我们的工作从交互角度补充了这些研究：我们通过实证研究，考察日常使用会话代理时反馈如何失效，以及界面脚手架如何引导用户提供更结构化、更高质量的反馈，从而最终有望形成更优的训练信号。

2.3 与大语言模型驱动的对话代理进行人与人工智能交互

从人机交互的视角来看，大语言模型驱动的对话代理标志着一种转变：从狭窄、面向任务的基于规则的对话系统，转向支持日益复杂目标的通用混合主动界面。用户现在借助对话代理进行搜索、写作、编程、辅导、情感支持以及日常问题解决 [32, 46, 52, 67, 70, 107, 127, 142]。与早期系统不同，这些智能体被部署为始终在线的协作者，能够在单次会话中灵活转换角色（例如讲解者、共同作者、批评者）。近期的对齐综述认为，这一上下文要求一种双向的对齐观：模型必须适应人类目标，但人类也会随着时间在认知和行为上适应强大的助手 [47, 55, 73, 109]。在此类混合主动式协作中，反馈对于维持共同目标以及跨多个轮次协调复杂任务至关重要。

越来越多的人机交互和自然语言处理研究开始考察人们在真实场景中如何与对话代理实际互动。HALIE 等框架重新构想了评估方式，将重点放在完整的交互轨迹、第一人称体验，以及所有权和愉悦感等标准上，而非仅关注静态提示词下的任务准确率 [70]。对 WildChat 及相关语料库的大规模分析研究了“真实场景中的”信息获取对话，梳理用户提出的需求，以及在超过一百万段 Chat- GPT对话中包含值得核验的主张的语句 [49]。教育和社会文化研究表明，学生和语言学习者会将对话代理用作导师或文化信息提供者；除非明确鼓励他们采取更具批判性的互动实践，否则他们往往会过度信任流畅的回答 [22, 153]。近期关于典型人机协作行为的研究将多轮聊天轨迹归纳为委派、共同编辑和核验等模式，并发现用户经常将对话代理视为协作者，却出人意料地很少提供明确的评估性反馈 [83]。

多轮交互进一步凸显了反馈的重要性。人机大语言模型交互研究日益将跨轮次的任务分解、规划和完善视为核心设计问题，并指出，“纯提示”给终端用户带来了不切实际的负担，也使有效利用持续反馈变得困难 [142]。Laban 等人[65] 开展的多轮评估研究表明，当指令不够明确且信息逐步揭示时，模型往往会过早固守最初的猜测，难以随着用户补充澄清而及时调整，因此与单轮基线相比，表现会显著下降。然而，在这些评估中，用户侧通常被模拟或抽象处理：多轮现象的分析旨在了解模型鲁棒性和任务表现，而非刻画真实用户如何决定何时以及如何提供反馈。

总之，人机交互领域的研究已经开始将大语言模型驱动的对话代理作为混合主动式、多轮协作者进行评估，并分析真实使用轨迹，但对提供反馈这一微观层面的互动工作关注相对较少：人们如何组织批评意见、何时放弃修复，以及哪些障碍阻止他们提供更具信息量的指导。现有研究通常聚焦于任务结果、用户满意度或粗略偏好，将反馈视为未加区分的标量或简短文本标签。我们的工作通过将反馈研究从纯粹以结果或数据为中心的视角转向以互动为中心的视角，弥补了这一空白：我们系统地编码轮次级反馈行为和对话中断，识别阻碍用户提供高质量反馈的因素，并设计界面级支架，使此类反馈更易于表达、解读和纳入，从而补充近期关于双向、交互层面对齐的倡议 [53, 109, 142]。

2.4 有效反馈的科学

大语言模型驱动的对话代理在交互层面被明确设计为模仿人类对话伙伴。既有研究表明，计算机是社会行动者 [84]，并将对话代理描述为自然语言用户界面，能够通过文本、语音和其他模态“模仿人与人之间的交流”以及“模拟人类对话” [104, 125]。近期对基于大语言模型的对话代理的分析显示，即使缺乏真正的社会理解，它们也能令人信服地再现类人对话和亲社会行为 [8, 22, 56, 95]。因此，从用户的角度看，与对话代理互动类似于与另一个人互动：他们提出请求、解读回应，并通过与人类交流时所使用的相同对话资源，决定是否以及如何提供反馈。尽管底层机制存在根本差异，这种模仿仍使我们有理由将人类对人类反馈与沟通科学中的洞见引入，作为评估交互层面质量的规范性基线。

数十年来，教育学和心理学领域的研究已证实，反馈是推动学习和表现的重要力量。Kluger 和 DeNisi [59] 以及 Hattie 和 Timperley [39] 进行的元分析表明，设计良好的反馈干预能够在不同领域显著提升成就。有效的反馈通常及时、具体且以目标为参照：它能帮助学习者了解当前表现与期望标准之间的差距，并提供有关如何缩小差距的具体信息 [39, 114, 135]。相反，含糊不清、过于复杂或时机不当的反馈可能会令学习者不堪重负、被忽视，甚至损害表现 [59, 114]。这些研究结果表明，反馈并非任意回应，而是能够减少人们对如何改进的不确定性的结构化信息。

近期研究进一步强调，反馈还必须易于使用。Carless 和 Boud [12] 提出了“反馈素养”这一概念，即学习者解读、评估并依据反馈采取行动的能力。Nicol 和 Macfarlane-Dick [85] 同样指出，若要让反馈促进改进，学生需要有机会依据明确标准比较自己的作品、讨论反馈并自我调节修改。这些观点将反馈视为一种对话式、迭代式的过程，而非一次性传递：反馈提供者与接收者共同协商意义，双方都必须投入努力，反馈才能转化为更好的结果。对于对话代理而言，这意味着仅仅提供反馈渠道（例如点赞或点踩按钮，或自由文本框）并不足够；互动还必须为用户提供脚手架，帮助他们表达、解读并依据反馈采取行动。

除有效性与可用性之外，人类沟通研究还解释了如何在对话中协调此类反馈。会话分析与语用学明确了使反馈交流成为可能的合作规范。格赖斯准则以及随后关于共同基础的研究描述了对话者如何使贡献保持真实、相关、信息充分且清晰，同时通过附和反馈与修正来建立相互理解 [19, 36]。在自然语言处理与人机交互中，这些准则已被采纳为聊天机器人和分析型代理的具体设计与评估启发式。例如，Xiao 等人 [140] 通过沿信息丰富度等格赖斯维度对回答进行编码，将基于聊天机器人的开放式调查中的回答质量加以操作化，相关性、具体性与清晰度，并使用这些编码将聊天机器人引出的回答与传统网页表单进行比较。同样，Panfili等人 [91] 通过格赖斯视角分析人类与Alexa等语音助手的互动，表明用户将离题或无意义的回复体验为对相关性与清晰度的违反，并认为此类违反对人与AI协作尤其具有破坏性。更广泛地说，近期研究明确将格赖斯准则实例化为判断系统回复是否信息适当、诚实、切题且可理解的标准，并用于为对话模型构建评估任务 [33, 61, 79, 92, 98, 105, 139]。

与此同时，社会与组织心理学阐明了为什么人们即使理解了规范，也不总是给予反馈。反馈寻求与给予被视为一种成本-收益权衡：只有当预期收益超过认知、时间和社交成本时，人们才会提供详细反馈 [6]。这一视角补充了会话规范，突显了塑造合作性反馈的动机障碍。

我们采用这些人人框架，作为分析人与会话代理反馈的规范性基线。鉴于对话代理旨在参与类人对话 [8, 56, 95,104, 125], ，即使其内部流程并非人类，仍会依照人类对话预期接受评判。在我们的分析中，格赖斯准则和反馈有效性标准为我们提供了一个视角，用于描述会话代理-用户交互在哪些方面未能达到合作且高质量的反馈交流。我们借助这一视角理解人与会话代理沟通中的反馈动态，并引出本文提出的反馈障碍和设计启示。

3 形成性研究

Shi等人 [112] 高亮了面向对话代理的真实场景中的人类反馈与人类自然向其他人类提供反馈之间的差异。为理解此类差异为何存在，我们开展了一项面对面的形成性研究，共有16名参与者。在这16人中，有11名从事复杂目标导向任务的有经验的会话代理用户，以及5名从事复杂目标导向任务的无经验的会话代理用户。；通过对N=11 名有经验用户访谈的分析，我们发现了反馈障碍（定义见下一节），这些障碍会阻碍用户在与对话代理互动期间提供“高质量”反馈（**研究问题一**）。我们还将研究发现推广至N=5 名缺乏经验的用户，即没有任务导向对话经验的会话代理用户。

3.1 方法与参与者

我们对16名参与者进行了时长30分钟的一对一访谈（有经验者，N = 11；无经验者，N = 5）。11名有经验的参与者来自大型国际科技公司和学术机构，这些机构正在积极探索将对话代理用于复杂目标导向任务。所有有经验的参与者目前或近期都曾在各种任务中使用对话代理，包括研究、创意写作、编程、图像编辑、决策和文章写作。无经验的参与者包括学校教师、退休人员、家务助理和高中毕业生。这些参与者几乎没有使用会话代理完成复杂目标导向任务的经验。相反，他们主要的

使用场景包括医疗帮助、服装推荐、语言学习、图像润饰和一般事实性问题。所有参与者人口统计特征和详细的使用场景均汇总于表1中。

访谈采用半结构化形式，围绕以下主题展开：

- **对话中断**：参与者在与对话代理对话时遇到的常见挑战有哪些？
- **中断修复（反馈）**：参与者使用了哪些策略来修复对话中断？
- **反馈有效性**：参与者在向对话代理提供反馈时面临哪些障碍？
- **反馈评估**：参与者采用哪些策略来评估其反馈的成效？

所有参与者均自愿参加本研究。本研究经公司内部审查委员会批准。征得参与者同意后，我们对访谈进行了转录和匿名化处理。

3.2 分析

我们采用 Braun 和 Clarke 的反思性主题分析方法 [11]。半结构化形式使我们能够灵活探索新出现的主题，并根据参与者的回答提出后续问题。第一作者对访谈进行了归纳式编码。随后，第一作者起草了编码手册，并与研究团队讨论和完善。之后，第一作者再次对文字稿进行编码，并在此基础上进一步开展多轮讨论和完善，最终形成更新后的编码手册。第一作者随后从编码后的文字稿中归纳出主题簇，以构建更广泛的研究发现和讨论。以下按此格式报告研究发现：引文_参与者α ($P\alpha$；Pi、Pj等支持。

3.3 立场声明

研究团队主要由具有不同立场性的作者组成。这些作者拥有两种截然不同的机构背景：学术和工业机构。除第三作者外，所有研究人员均已在学术机构接受过STEM背景的研究生培训，专业领域为人机交互（HCI）和机器学习（ML）。团队成员多元的机构与专业背景为该项目营造了丰富的跨学科环境，但也带来了对研究目标与实践的不同视角。来自学术机构的学者侧重于提供可推广的见解，以理解用户面临的反馈障碍；来自工业界的学者则侧重于提供可执行的见解，以克服反馈障碍。这一协作需要持续、开放的讨论与反思性，以弥合这些不同的优先事项。例如，早期研究问题的讨论旨在兼顾对反馈障碍的更广泛理解，以及供工业界克服反馈障碍的可执行见解。在用户研究的数据收集过程中，研究团队定期讨论，明确每位研究人员的个人立场性可能如何影响数据收集与解读。通过承认这些差异，我们旨在开展稳健的研究，充分利用学者们的多元背景。最终的研究综合了学术和工业机构双方的视角，并因有意识地调和多元背景与机构的过程而得到加强。

3.4 描述性发现

3.4.1 与对话代理互动时常见的失效情况有哪些？我们首先请参与者回忆与对话代理的对话令人不满意，或输出偏离预期的情形。参与者最常指出的互动失效是上下文丢失（16 位参与者中有 10 位），即代理逐渐偏离任务的原始目标。一位参与者解释说："最主要的问题是上下文漂移 .....它真的会忘记整场对话最初的目标是什么"（P2；P4、P5、P7、P9）。其他参与者则强调，时间较长、内容交错的会话会让代理难以维持正确的上下文："你可能得花好几个小时告诉它正确的上下文，否则就只能重新开一个聊天"（P9；P2、P4、P8）。

除了上下文漂移，参与者还经常报告质量问题。使用对话代理进行研究和编程时，他们遇到过捏造或不可靠的内容："它会编造一些我从未提过的事"（P8；P1、P9、P10、P11）。许多参与者也提到了指令遵循与谄媚之间的矛盾 [106, 107]。模型虽然试图遵从指令，却常常忽略细微之处，或未能质疑用户的假设："有时候，即使我错了，人工智能也会附和我——而一个人会说：'这说不通'"（P7；P2、P4、P10）；"Cursor 只会把你的命令当成铁律，也不会真的提出任何澄清问题"（P2；P4、P7、P9）；"[ChatGPT] 只会告诉我想听的话，这让我感到不自在"（P12；P13、P14、P16）。这些失效情况——上下文丢失、内容不可靠，以及过度或不足的遵从——共同反映了当前对话代理在与用户维持共同理解方面的不足。

3.4.2 用户如何提供反馈？当对话偏离轨道时，参与者会采用务实的修复策略，从不费力地重复输入，到彻底放弃智能体不等。最常见的第一步就是改述或重复请求："大多数时候，纯粹是因为懒得费事，我就直接重新输入提示词"（P2；P1、P3、P4、P7、P8、P10、P11）。还有一些人会尝试逐步完善输入，或分解任务："有时我会把任务拆分成更小的任务，然后手动一步步引导人工智能"（P4；P3、P5、P7、P9）；也有人会预先提供更多细节，以引导模型："对于编程任务，我会预先提供非常详细的提示词——输入越清晰，效果越好"（P11；P2、P4、P6）。

当上述策略失效时，参与者往往会重置交互或更换工具："有时情况会糟糕到你只能重新开始整段对话"（P2；P3、P4、P6、P9），或者"如果还是不行，我就换一个大语言模型——我会来回切换，因为总有一个大概能解决"（P6；P1、P5、P7）。在时间压力下，许多人转而接受不完美的输出并自己完成任务："自己做比一直提示更快，所以我会接受次优回复并手动修正"（P5；P1、P7、P9、P10、P11）。放弃在所有缺乏经验的用户中尤为常见，因为他们并不了解提供反馈的可供性："我不知道 [反馈] 是可能的"（P12；P13、P14、P15），以及"[ChatGPT] 从未要求我这样做"（P13；P12、P16）。

在各种策略中，参与者的反馈往往是反应式的、局限于最近一轮，并且转瞬即逝。他们只是对输出稍加推动，使其足以推进任务，但很少投入丰富的结构化反馈。这一模式引出一个自然的问题：这些日常反馈实践与教育和心理学 [39, 78,114], 中关于高质量 反馈的规范概念相比如何？这些概念强调目标参照、可执行、表达清晰和渐进式反馈。我们接下来将对此进行考察。

3.4.3 用户反馈是否符合高质量反馈的特性？对话代理需要高质量的人类反馈来完成复杂任务并弥合知识差距。我们根据教育和心理学文献中提炼出的四个维度来界定反馈质量——目标参照、可执行、表达清晰和渐进式反馈 [39, 78, 114]。下文将探讨参与者对对话代理的反馈实践与高质量反馈特性之间的契合与差异。

• **目标参照**：高质量反馈与当前目标紧密结合，并明确提及成功标准。教育文献强调，反馈应以明确的成功标准为依据，让学习者能够了解“ [他们] 相对于目标的表现如何”，而不是收到针对无关方面的评论 [39]。因此，目标参照反馈有助于弥合当前表现与预期结果之间的差距。尽管用户会积极努力维持上下文，例如再次强调指令：“不，我想让你按照我之前说的去做”（P1）；但参与者指出，当对话偏离方向时，用户往往无法提供目标参照的纠正性反馈，使对话回到正轨。他们最常见的策略是彻底放弃相关性，直接结束互动。一位参与者指出，当对话进展不顺时，“情况可能糟糕到你只能重新开始整个对话”（P2；P4、P7、P9）。这种重启放弃了形成性反馈，打断了教育文献所强调的、对改进至关重要的目标参照反馈回路。

• **可执行**：有效的反馈是具体且可执行的。在教育和形成性评价中，当反馈为学习者提供关于改什么以及如何改的具体信息，而非笼统评价（例如“做得好”）时，其效力最为显著 [114]。Hattie 和 Timperley [39] 强调，反馈不仅应描述当前表现，还应提供“下一步该往何处”的指导——即有助于缩小与目标差距的策略或修订。与文献一致，取得更好结果的参与者往往通过提供细化指令来实现这一点，例如“将任务拆[分]成更小的任务”（P4），或为编码任务预先提供“非常详细的提示词”（P11）。然而，用户反馈常常缺乏实现强可执行性所需的细节程度。参与者往往默认使用高层级、模糊的指令，例如要求人工智能“改善行文流畅度”或“修正词汇”（P1；P3、P4、P6、P9）。这种低可执行性倾向在低投入策略中也很明显。例如，少数参与者承认：“我只是重新提示它，这样更容易”（P8；P2、P7、P10、P11），而不是提供细化修正。从反馈研究的视角来看，这些行为类似于缺乏过程层面指导的评价性评论：它们表达了不满，但几乎没有提供关于具体下一步或策略的信息，从而限制了对话代理进行有意义调整的能力。

• **表述清晰**：高质量反馈应表达清晰、歧义尽可能少，从而支持共同基础。形成性反馈研究强调，评论应易于理解；当反馈含糊不清或晦涩难懂时，学习者就必须额外花费精力去解读对他们的要求，这会让反馈显得“毫无用处”或令人沮丧 [114]。在我们的研究中，用户的反馈往往表达得不够清晰，参与者也提到难以清楚表达问题：“有时我在语义上知道自己想要什么，但我无法措辞，让模型理解我的意思”（P3；P7、P8、P11）。这一困难反复出现，一位用户也证实，“解释为什么 [某个回复是错误的]很难”（P10；P3、P7、P8）。无法用语言清楚描述问题，导致反馈本身含糊不清，难以让对话代理据此采取行动；这与课堂研究的发现一致：未充分说明的评论会让学习者不确定该如何回应或改进 [114]。

• **渐进**：高质量反馈会随着时间推移，在渐进式形成性循环中逐步提供。教育研究表明，最具影响力的反馈并非在任务完成时给出的一次性评判，而是能够支持修改和再次尝试的信息 [39, 114]。然而，反馈的数量必须经过合理调节：一次给出太多评论可能会增加认知负荷，降低反馈效果 [114]。一位参与者直言：“我不太会提供人工智能反馈，也不会很频繁地提供。有时候完全不会提供”（P3；P5，P6）。用户放弃提供反馈的门槛很低；如果人工智能“基本的事情错了不止两次”，用户就会不再提供反馈（P8；P1，P5）；或者“如果任务在 15–20 分钟内仍然无法完成，就会彻底放弃任务”（P11；P6，P8）。这种模式导致反馈稀少，与文献所设想的渐进式形成性循环形成鲜明对比。

值得注意的是，先前的研究区分了任务层面的反馈（针对正确性）、过程层面的反馈（针对策略），以及自我调节反馈（针对对自身学习的监控）[39]。在我们的访谈中，参与者对对话代理的评论几乎完全属于任务层面——针对特定输出提出调整建议——而非有助于逐步校准他们与对话代理如何协作的过程导向型指导。综合来看，这些叙述表明，尽管用户有时会提供目标明确、可执行、表达清晰且渐进的反馈，但他们日常的做法往往至少在其中一个方面有所欠缺，而且很少超越局部的任务层面修正。这种由教育和心理学驱动的视角关注反馈本身的内容 和形式 ； 接下来，我们将结合基于格赖斯准则的对话视角，进一步理解阻碍用户达到这一理想状态的深层障碍。

3.5 研究发现——是什么阻碍用户提供“高质量”反馈?

上文的描述性发现表明，用户在实践中很少提供“高质量反馈”。为了了解是什么阻碍用户提供高质量反馈，我们将参与者的叙述归纳为用户提供反馈时反复面临的挑战。如相关工作（第 2.4 节）所述，对话代理明确是为人类沟通而设计的，并且常被视为社会行动者。因此，为了揭示用户在与对话代理互动时面临的障碍，我们采用格赖斯准则 [36] 作为分析视角，以剖析这些挑战。格赖斯准则是人类成功沟通所遵循的合作原则——关系 （切合主题）、质量准则 （真实且有证据支持）、方式准则 （清晰且有条理）和信息丰富度 （提供恰当数量的信息）。

关键在于，违规行为是相互的：模型失误（例如偏离、幻觉、冗长）会引发用户行为（例如放弃、规格不足、模糊修正），而这些行为本身也违背了合作原则。这种相互恶化解释了反馈循环为何会崩溃，也说明了为何需要支架 来同时应对模型局限和维持合作反馈循环的人类认知限制。第一作者使用四项准则对所有访谈进行了编码：每个访谈句子都分别标注一次用户侧准则违规，以及一次会话代理侧准则违规。详细的编码手册见补充材料。

3.5.1 关系准则：对话代理转移对话目标，用户则放弃互动。关系准则要求每个对话轮次都与互动的共同目标相关。在模型侧，参与者描述了即使是纠正性反馈，也常常无法让会话代理重新聚焦于共同目标。一位参与者这样概括这种偏离：“它只会让事情朝着你不想要的方向越陷越深”（P7；P2、P4、P8）。其他参与者指出，反复重述上下文或约束，也无法阻止模型围绕偏离目标的解读打转：“如果我给它反馈超过两三次，它就会开始反复迭代自己之前的答案，而不是寻找新的方向”（P8；P5、P6、P9）。在较长的互动中，用户觉得自己不得不不断重建共同基础：“你可能得花几个小时告诉它正确的上下文，或者重新开始一段聊天”（P9；P2、P4）。这些模式表明，会话代理未能利用反馈让对话重新贴合用户目标，因而违反了关系准则。

在用户侧，参与者面对这种错位时，选择退出互动，而不是提供更有针对性、以目标为导向的反馈。正如一位参与者所说：“有时候，说实话，重新开始……或者干脆自己写代码更容易。”（P4；P2、P5、P6、P7、P8）。另一位则说得更直接：“继续输入提示词不如自己动手快，所以我接受次优回复，然后手动修正。”（P5；P3、P9、P11）。反馈本应将对话拉回共同目标，但双方却都逐渐偏离：会话代理继续跑题，而用户放弃尝试修复，切断了有效反馈所依赖的关系纽带 [19]。

3.5.2 质的准则：对话代理产生幻觉，用户传播未经核实的输出。 质的准则要求说话者提供真实且有充分证据的贡献。在模型侧，参与者一致指出了捏造或幻觉输出。一位参与者指出，“五篇论文中，它编造了一或两篇”（P8；P1、P10、P11），另一位则解释道：“如果与研究相关或事实密集，我需要核查链接或所引论文是否真实。因此这需要更多精力”（P5；P6）。此类未能将信息建立在证据之上的情况会侵蚀信任。这些幻觉以及松散内容直接增加了评估反馈是否已被正确纳入的成本。在人际对话中，此类失误会削弱信任；而在对话代理中，它们给用户带来额外的核实负担，从而使其不愿提供更细致、以质量为导向的反馈。

从用户的角度看，尽管参与者没有亲自陈述错误说法，但他们表示，在时间和精力受限时，会放宽自己的证据标准。正如P6所说：“我不会逐行检查。我会看一下高层结构……这更像是一种直觉检查”（P6；P5、P7、P9）。另一位参与者坦言：“时间不够的时候……我就直接复制粘贴，不去考虑反馈。”（P9；P7、P10）。在截止日期压力较大或利害关系较低时，用户往往只做最低限度的核验便传播模型输出，而不是提供纠正性、基于证据的反馈。因此，模型侧的质量准则违反（幻觉）与用户侧捷径（直觉检查、跳过核验）相互强化，削弱了整个反馈回路的可靠性。

3.5.3 方式准则：对话代理很少跟进，用户也难以表达清晰。方式准则要求表达清晰、简洁且易于理解。在模型侧，参与者描述了一种流于表面礼貌，却缺乏真正澄清或调整的模式。正如 P2 所说：“Cursor……只是把你的指令当成法律，不会真正提出任何澄清问题”（P2；P3、P4、P6、P7）。即使反复输入，底层行为也往往没有改变：“即使我重新输入提示词，它道了歉，却还是重复同一个答案”（P6；P2、P4、P7）。其他人则觉得这只是风格冗余：“感觉过于客气了——几乎有点虚假。一个人只会说‘好吧，我试试看’，不会用不必要的废话来包装这句话。”（P1；P3、P6）。由于对话代理缺乏语气、面部表情和肢体语言等副语言线索，用户也难以判断系统是否真正理解了他们的反馈：“在人际对话中，你给别人反馈，有时能从他们的眼神中看出来。如果有人摸不着头脑，你也能看出他们很困惑。”（P2；P1、P7、P13、P12、P14、P16）。总的来说，这些问题违反了方式准则，因为它们让模型的回应更难理解和诊断。

在用户侧，参与者坦言很难清楚表达自己的需求。一位参与者解释说：“有时候我在语义上知道自己想要什么，但不知道该怎么措辞，才能让模型理解”（P3；P7、P8、P11）。另一位参与者回忆道：“一开始很难——我知道自己想要什么，但想不出合适的形容词”（P7；P3、P10）。P10也指出：“很难解释 [模型错在哪里] ”（P10；P3、P8）。在更多视觉任务中，表达障碍更加明显：“它无法准确把握可视化位置，而我也很难用语言描述这些位置。”（P8）。因此，即使用户发现了问题，他们提供的反馈往往仍然含糊或不够具体，降低了其对会话代理的实用性，也构成了用户侧对方式准则的违背。

3.5.4 数量准则：对话代理要么信息过多，要么信息不足，而用户则始终信息不足。数量准则要求相对于手头的任务，既不过多也不过少地提供信息。在模型侧，参与者既体验到了人工智能的信息过多，也体验到了其信息不足。在一

方面，会话代理会遗漏重要要素："它只挑两三样东西……我还得不断补上其余的指标，"（P7；P4、P5、P6、P9）。另一方面，它有时做得比要求的更多："它通常试图做比我要求的更多的事情，把我代码的其他部分搞乱了"（P4；P7、P9）。在这两种情况下，用户都需要先进行弥补——要么反复补充缺失的细节，要么撤销不想要的修改——然后才能开始对关键部分提供有针对性的反馈。预先提供详细上下文被视为一种负担，尤其是对新手或那些从人工智能的回复中学习的人而言："有时我想给出反馈，但又忍住了，因为我也在从它的回复中学习"（P9）。

在用户侧，数量准则违反主要表现为反馈过少或不足。许多参与者往往只做最少的调整："大多数时候，纯粹是因为懒得动，我就只重新输入提示词"（P2；P1、P3、P6、P7、P8、P9、P10、P11）。其他人则表示，他们会越来越沮丧，而不是提供更丰富的上下文："我通常会先对它吼三四次……如果还是不行，我就重新开一个聊天"（P6；P2、P3、P7、P9）。少数参与者在多次重试失败后确实会补充更多细节，例如："如果重试失败，我会补充更多上下文，或者让它检查整个代码库，然后再试一次"（P4、P3、P5、P9），但这并非常态。总体而言，用户很少根据会话代理为恢复所需的信息量来调整反馈，反而在极短输入与彻底放弃之间摇摆不定。模型侧的信息过多或不足与用户侧的信息不足相结合，使得难以维持形成性评估研究中所述的渐进式、范围恰当的反馈循环 [10, 39, 99, 114]。

3.5.5 反馈障碍。根据观察到的对格赖斯准则的相互违反，我们归纳出四个关键的**反馈障碍**，阻碍用户如今向对话代理提供高质量反馈。

共同基础障碍。

- 模型：模型未能维持既定的任务目标。这表现为上下文漂移、忽略明确约束，或悄然将用户目标替换为模型幻觉出的目标。
- 用户：用户脱离对话目标，从而切断协作关系。这表现为**明确放弃**（"我还是自己做吧"）或**重新开始聊天**，使所有共享上下文失效。
- 影响：这一障碍迫使用户不断重新调整对话代理，增加了交互的认知成本。它将传播和维持共同目标的全部负担都放在用户身上，使交互显得毫无成效，导致用户："花几个小时告诉它正确的上下文，或者重新开始一次聊天"（P9；P2、P5、P7、P8）。

可验证性障碍。

- 模型：模型提供的信息缺乏可验证证据的支撑。这包括生成幻觉、引用不存在的来源，或将过时信息当作当前事实呈现。
- 用户：用户未能在工作流程中遵守证据标准。这并非关乎撒谎，而是**跳过核验**，即接受可能有误的输出，或传播未经核实的信息。
- 影响：这一障碍迫使用户扮演事实核查者的角色，削弱了信任。由于缺乏可供性来评估反馈是否得到忠实采纳，核验工作会变得代价高昂，迫使用户采取捷径："我不会逐行检查，只会看整体结构……这更多是凭直觉检查"（P6；P5、P7、P9）。

沟通障碍。

- 模型：模型的沟通方式不清晰或没有帮助。这包括回避性礼貌（"空话"）、道歉循环，以及尤为重要的一点：当用户的请求含糊不清时，**未能提出澄清性问题**。
- 用户：用户未能提供清晰明确的反馈。这源于固有的**表达困难**，例如难以说明问题所在、给出含糊的指令（如"做得更好"），或无法用语言描述期望的修改。• 影响：这一障碍会导致令人沮丧的反复猜测。若没有人工智能提出有用的探询问题，要继续对话就需要付出很高的认知努力，使持续参与显得不合情理；当用户感知到的成本超过收益时，他们便会放弃任务。正如一位用户所解释的："如果我发现某项任务需要提示词七次才能得到满意的结果，我可能干脆就不再尝试了"（P3；P1、P5、P8、P11）。

信息充分性障碍。

- 模型：模型提供的信息量不恰当。这可能是信息不足（省略关键步骤；只处理输入的一部分），也可能是信息过载（进行未被要求的更改；提供无关细节）。
- 用户：用户提供的信息不足以让模型进行有意义的修正。这包括对话中敷衍的回复，例如**简单地重新提示**（"再试一次"），或缺少必要上下文的反馈（"这不对"）。
- 影响：这一障碍使用户陷入低效的猜谜游戏。由于缺少低门槛的方式来传达关于要改什么、在哪里改以及以何种粒度进行更改的结构化反馈，用户往往只提供最少的信息；这会使低质量输出的循环持续下去："除非我给出非常具体、细化到文件层面的指令，否则我会犹豫，因为不然它就会把一切都弄乱"（P11；P2、P7、P8、P10）。

3.5.6 综合与启示。我们的形成性研究揭示，人机交互中的反馈动态系统性地脆弱。这些失败并非单方面造成，而是反映了模型与用户之间合作原则的相互瓦解。模型的失败（偏离、捏造、歧义、信息过载和信息不足）会引发低质量的用户反馈，例如放弃和信息不足。反过来，这些行为又使模型无法获得实现对齐和任务成功所必需的目标参照、可执行、表述清晰且持续的反馈。

我们识别出的四种**反馈障碍**——共同基础障碍、可验证性障碍、沟通障碍和信息充分性障碍——为诊断会话代理交互核心阶段中的这些失调循环提供了一个简明框架，其中包括用户如何传播反馈、如何

模型和用户如何与之互动，以及双方如何评估输出结果。总体而言，这些障碍表明，如果缺乏适当的支持，提供高质量反馈所需的认知成本和交互成本[5, 64] 很快就会超过用户感知到的收益。

要让人机协作真正有效，我们必须设计能够明确消除这些反馈障碍的系统。在下一节中，我们将把这项分析转化为设计要求，旨在减轻反馈障碍，并构建支架，以恢复与会话代理之间的合作沟通原则。

4 设计用于解决反馈障碍的系统

前文的分析将这四种反馈障碍视为具体障碍，它们阻碍用户为有效的人机协作提供高质量反馈。

克服反馈障碍，需要解决这些障碍在人机交互三个基本方面造成的干扰：第一，它们妨碍用户传达反馈，也就是用户表达目标、修正和约束的难易程度。第二，它们削弱用户与模型对反馈的参与，包括系统解读、澄清和回应用户提供的输入的能力。第三，它们损害用户评估反馈结果的能力，使用户难以确认系统是否正确采纳了他们的修正。反馈障碍共同阻碍了反馈的传达、参与和评估，形成一种循环：用户无法提供改善交互所必需的反馈。

基于这三个交互领域以及反馈障碍对它们造成的扭曲，我们现在将研究结果转化为三项设计要求。任何对话代理都必须满足这些要求，才能打破这些障碍并支持高质量反馈。

4.1 设计要求

每项期望目标对应一个或多个反馈障碍，并阐明系统如何降低当前阻碍用户提供高质量反馈的交互成本。

4.1.1 建立持久且清晰易懂的共享参照框架。当用户与会话代理无法对任务保持稳定的共同理解时，就会产生共同基础障碍[19]。我们的研究表明，当前对话代理具有即时性、逐轮交互的特征，常常导致上下文偏离，由此引发共同基础障碍。这种失误迫使用户不断投入大量精力，让智能体重新了解情况，成为交互成本的重要来源[5, 64]。因此，首要的期望目标是建立一个能够**维持并呈现对任务的持久共同理解**的系统。这与外部认知的基本原则相符：可见且持久的人工制品可以减轻用户记忆和追踪任务状态的认知负担[57, 87, 101, 147]。系统不应只依赖短暂的聊天记录，而应将目标和约束外化为共享且可编辑的产物。当反馈能够以这个稳定的参照框架为依托时，便可支持传达这一支柱，让用户能够准确而自信地表达意图。

4.1.2 面向主动低摩擦交互进行设计。当用户必须独自承担表达修正意见、确定细致程度以及修复误解的全部负担时，就会出现沟通和信息丰富度障碍，进而导致用户过早放弃。第二个期望目标是**设计支持相互协作的混合主动交互，以降低表达和修正的成本**。系统不应要求用户不断编写更加具体的输入，或自行诊断模型故障，而应主动参与维持沟通清晰度。这包括询问用户补充缺失的约束、指出歧义、提出可能的解释，以及提供结构化操作（例如“仅将更改应用于第 2 节”）。这与交互式机器学习的理念一致：将人机协作视为一种相互协作的混合主动过程[3, 4, 43]。当用户输入含糊不清时，系统主动寻求澄清，并提供结构化且结合上下文的操作，就能弥合“执行鸿沟”，减少用户表达意图所需的认知负荷[132]。

4.1.3 提供透明可验证推理。当用户无法检查会话代理如何解读其反馈，或无法验证反馈是否得到忠实纳入时，便会出现可验证性障碍。这种不透明性会使用户陷入盲目接受或穷尽式人工检查的两难境地，而这两者都会妨碍持续反馈。为应对这一问题，第三项期望目标要求系统<style id='3'>提供透明且可验证的推理，使反馈回路清晰可见并可供审计。 这是可解释人工智能（XAI）的核心原则之一：让系统的内部逻辑变得易于理解，有助于用户建立准确的心智模型并校准其信任[37, 69, 80, 122]。通过对交互进行设计，使结果清晰易懂——例如使用可视化差异呈现变化——可以弥合“评估鸿沟”，让用户反馈产生的影响立即显现。通过使反馈与结果之间的联系直接且“易于浏览”，系统能够实现低成本评估，使用户既能评估模型输出，也能评估其反馈的有效性。

4.2 将设计要求操作化以分解反馈障碍

为将设计要求操作化，我们构建了一个轻量级、基于 Web 的会话代理系统（FeedbackGPT），采用了用户熟悉的类 ChatGPT 界面（图 1）。我们引入这些支架时，没有修改模型权重或模型的核心行为，从而确保改进源于交互设计，而非模型调优。下文将介绍并说明这些支架和系统的设计依据，并详细阐述实现选择，以确保系统可复现、可解释。由于形成性研究发现了许多潜在的支架，我们优先选择了符合以下条件的支架：(i) 与现有的基于聊天的工作流相辅相成；(ii) 在当前会话代理的可供性范围内可行；(iii) 与模型及领域无关；(iv) 直接针对已识别的反馈障碍。这些支架只是众多可能实现方式中的一种；我们的目标是展示，如何通过将设计要求操作化的支架消除反馈障碍，从而获得质量更高的反馈。

下面，我们通过以下方式描述每个脚手架：(1) 确定其满足的设计期望，(2) 说明其缓解了哪项反馈障碍，以及 (3) 详述该功能如何在实践中实现这一点。

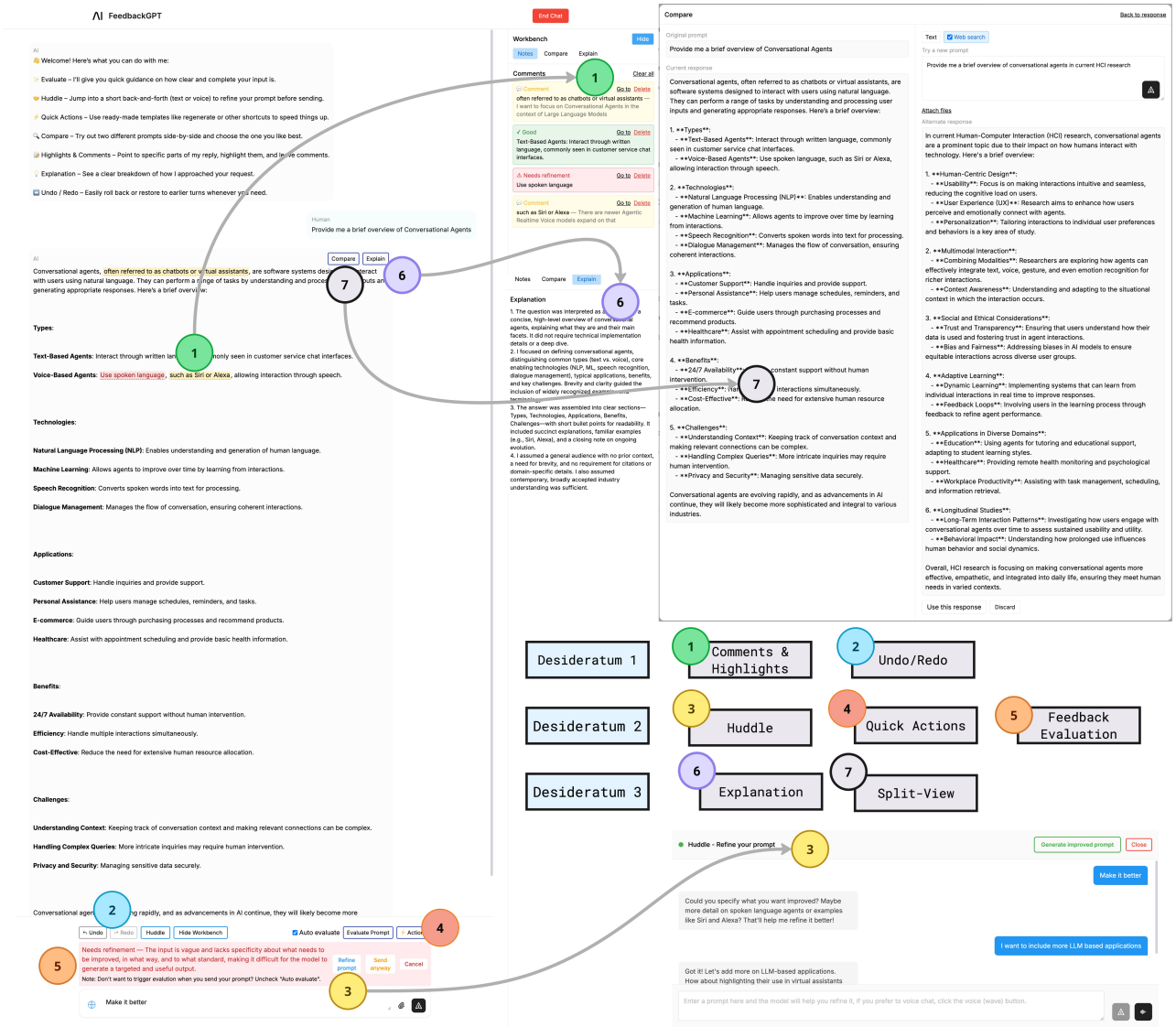


图 1：FeedbackGPT 概览。该系统由三个全新的界面组件组成：侧边栏、面板和弹出窗口。界面为用户提供七种不同的支架，以实现第 4.2 节所述的三个期望目标。

界面组件。 FeedbackGPT 保留了熟悉的聊天流和编辑器，同时引入内联控件、持久的右侧侧边栏以及辅助底部面板。

4.2.1 期望目标 ***1*** 的支架：构建持久且清晰易懂的共享参照框架。 这些功能旨在通过展示任务状态，并将反馈锚定到特定上下文中，克服**共同基础障碍**，从而稳定用户与会话代理之间的共享参照框架。

内联评论与高亮：受协作文档工具的启发，内联评论与高亮允许用户通过高亮和原位评论，将反馈锚定到模型输出中的特定片段[9]。与其使用模糊的指代（例如“第二段”），不如将反馈关联到具体文本；这种脚手架

将持久共享框架这一期望目标付诸实践：反馈会作为对话状态中明确且可检查的一部分持续保留。这些锚点会汇集在侧边栏中，并作为结构化约束传入下一次输入，促使模型逐一确认并处理每个对应位置。这样一来，内联评论与高亮便能直接缓解共同基础障碍，减少上下文漂移，并确保反馈始终基于对任务共同且稳定的表征。

撤销与重做：当今对话代理缺少撤销功能，这会阻碍用户尝试并限制用户控制[9]。撤销与重做通过让用户回溯对话快照而不丢弃整个交互，扩展了持久的参照框架。通过捕获对话状态并将其展示

通过简单的按钮，系统让用户探索替代的反馈路径——例如，尝试不同的批评意见或不同的约束——而不会有失去来之不易的上下文的风险。这减少了在反馈出错时放弃并重新开始聊天的需要，从而削弱共同基础障碍，并使迭代优化反馈更加安全。

4.2.2 期望目标二的支架：面向主动低摩擦交互的设计。 这些功能旨在克服**沟通与信息性障碍**，通过降低提供高质量反馈所需的付出，并将反馈完善转变为协作式、来回交流的过程，而非单方面的独白。

反馈协作区： 当用户的反馈含糊不清或难以表达清晰（这是沟通障碍的一个关键方面）时，反馈协作区会打开一个面板，其中同一模型会遵循新的指令，主动提出有针对性的澄清性问题；在此空间中，用户可以逐步将宽泛的反应（例如“这听起来不太对”）完善为具体、可执行的指令，并遵循高质量反馈的特征（3.4.3）。通过共同承担澄清责任，并以具体的后续问题提示用户，反馈协作区将互惠、低摩擦交互这一期望目标落到实处，帮助用户克服表达困难，从而同时应对沟通障碍和信息丰富度障碍。

快捷操作： 快捷操作通过将常见且价值较高的反馈模式转化为一键操作，解决信息量不足这一障碍。用户可以调用可重复使用且可编辑的模板，例如“仅根据突出显示的更改重新生成”，无需反复重新输入冗长的指令。即使是少量的重复输入也会让用户不愿提供详细反馈， [132],这些快捷方式降低了添加丰富约束的边际成本，使用户更有可能提供足够丰富的信息来指导会话代理。

反馈评估： 反馈评估通过在消息发送前提供实时指导，帮助用户克服自身反馈中的歧义。系统依据提示启发式规则和交际原则评估用户的输入，然后提出简短建议，以提升具体性、结构和清晰度。这直接针对用户未遵循方式准则和数量准则的情况：界面不再期待用户自然而然地产生高质量反馈，而是以尽可能少的额外操作教会他们如何完善消息，从而推动实现低摩擦、互惠式反馈交流这一期望目标。

4.2.3 期望目标三的支架：提供透明可验证推理。这些功能旨在通过使模型的推理过程以及用户反馈的影响变得透明且可审计，克服**可验证性障碍**。

解释： 不透明的输出迫使用户充当事实核查者。「解释」脚手架提供了一键式控件，可提示模型详细说明其所使用的上下文、优先遵循的指令，以及它如何理解用户的反馈并据此生成回复。受解释性调试启发[62, 63], , 此功能将透明、可验证推理这一期望目标付诸实践：它帮助用户建立对系统更准确的心智模型，诊断错误的根本原因，并判断糟糕的结果究竟源于他们的反馈、模型的推理，还是两者兼有。这种做法不同于某些模型生成的“思考”或中间推理轨迹；后者主要面向模型内部，通常对用户隐藏。解释面向用户，并在回复生成后产生，因此能够以最终输出为依据，也就更忠实地反映模型实际生成的内容。

分屏比较：为了避免用户盲目接受第一个看似合理的输出，分屏比较功能允许用户并排评估根据不同反馈生成的不同版本回复[120]。通过降低比较评估的成本并使其在视觉上更加醒目，这一脚手架支持可验证推理这一期望目标：用户可以直接看到不同的反馈策略如何改变模型的行为，并受到鼓励，批判性地评估反馈结果，而不是传播可能有缺陷的回复。

4.2.4 实现。开发系统时，我们使用*Next.js* 框架构建前端，使用 *Node.js* 和 *Express* 构建后端，最后使用*MongoDB* 作为数据库。我们使用 TypeScript 作为系统开发语言。

我们利用 OpenAI 的 Responses API 与模型交互，支持使用工具、上传文件和网页搜索。主要使用的模型是*GPT-4o*，用于基于文本的任务；以及*gpt-4o-realtime- preview-2024-12-17*，用于语音模式。

对于主界面中单流聊天窗口里的对话，我们没有修改默认的系统提示词，以确保不改变模型的底层行为。当用户提供高亮和评论时，我们会将其作为（上下文、评论）对添加到用户输入中。绿色高亮会被转换为“良好”评论，红色高亮则会被转换为“有待改进”评论。

在 Huddle 模式下，我们会更改系统提示词，使模型能够依据 OpenAI 提示指南提供反馈。我们还指示模型以简短的 1–2 行回复作答，让互动更具对话感。同样，在评估用户输入时，我们使用自定义系统提示词，让模型依据 OpenAI 提示指南评估输入（提示词详情见补充材料）。

5 研究 2：解决反馈障碍的可供性如何改变用户反馈?

为了评估我们旨在提供可供性、尽量减少反馈障碍的支架如何影响用户反馈质量，我们开展了一项为期一小时的被试内研究，采用面对面形式（无法到场的参与者则以虚拟形式参与），共有 20 名参与者。参与者与 ChatGPT 和 FeedbackGPT 互动，完成两项写作任务。下文将介绍我们的研究设计和测量方法。

5.1 研究流程

如图 2 所示，研究流程包括五个部分：任务前调查、第一次协同写作任务、第一次任务调查、第二次协同写作任务、第二次任务调查、任务后调查，最后是一次开放式访谈，了解参与者使用所互动的两个系统的体验。

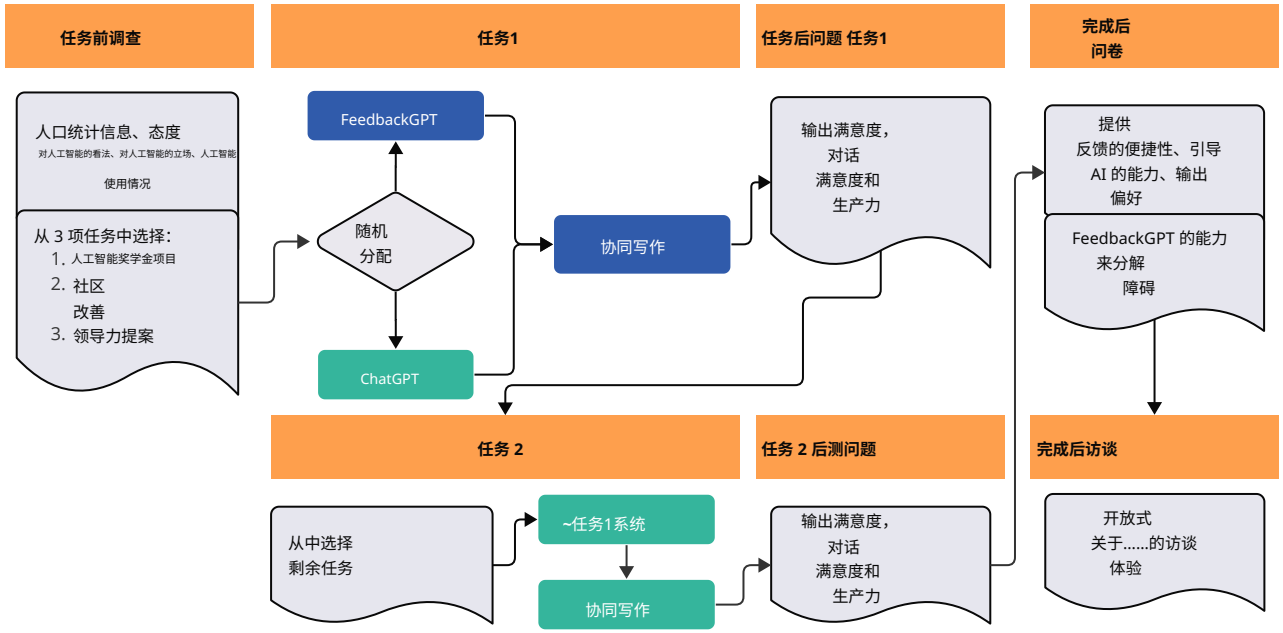


图 2：研究 2 的整体研究流程。在任务前调查中，参与者回答了有关其以往对对话式人工智能的使用经验、既往态度以及人口统计问题。随后，他们从三个可选主题中选择一项任务。之后，他们被随机分配至一个系统，并使用该系统完成任务。完成任务后，他们需要回答三个任务后问题，内容涉及其结果、对话和生产力。接着，他们从剩余的两个选项中选择第二项任务，并被分配至在任务1中未使用的系统，使用该系统完成任务并回答任务后问题。之后，参与者需要回答三个关于其使用 FeedbackGPT 体验的一般性问题，随后回答八个旨在了解 FeedbackGPT 中反馈障碍影响的问题。

在任务前调查中，参与者需要评价自己以往使用人工智能的经验，以及对对话代理的态度。之后，参与者分别完成两个任务，每个任务的主题和系统都不同。在每一轮中，参与者会被随机分配到一个系统，该系统要么是基线（Chat-GPT），要么是 FeedbackGPT。他们必须至少花 20 分钟与系统互动，并完成自己选择的任务。完成任务后，他们需要使用五点李克特量表评价输出质量、对话质量和生产力。

完成两项任务后，参与者完成了任务后调查。其中，他们需要评价与 ChatGPT 相比，为 FeedbackGPT 提供反馈的难易程度、感知到的对话质量和最终输出质量。随后，针对每个反馈障碍，他们需要回答两个问题，以评价自己使用 FeedbackGPT 的体验。所有任务后调查均采用五点李克特量表进行测量。详细问卷见补充材料。

5.2 协同写作任务

我们希望参与者与系统进行互动，以完成一项需要多轮次交互的任务，并且参与者必须提供反馈才能成功完成任务。我们发现，以用户个人经历为主题的协同写作任务最符合我们的目的。在我们的设置中，用户与模型之间存在固有的知识差距，只有用户向代理提供反馈才能弥合这一差距。此外，与需要特殊界面的编程和图像编辑不同，协同写作任务不需要特殊的界面元素，并且在 Gemini 和 ChatGPT 等不同平台上具有一致性。我们选择的三个主题如下（详细的任务描述见附录 A.2）：

- 变革性人工智能应用场景的奖学金申请。
- 社区改善计划提案
- STEM领导力发展招募宣传稿

5.3 测量指标

我们采用了三类测量指标，分别用于评估：(1) 用户反馈的客观质量，(2) 参与者对各系统的主观体验，以及 (3) 他们认为 FeedbackGPT 的可供性克服四种反馈障碍的能力。此外，我们还收集了任务开始前的态度和人口统计信息，以提供上下文。

5.3.1 反馈质量与障碍（基于日志）。

分析单位。 我们将反馈轮次 定义为首次系统回复之后、包含旨在塑造系统后续输出的评价性、纠正性或指导性内容的任何用户消息（例如，“这一节过于笼统，请补充具体示例”

或“保留引言，但缩短方法部分”）。对于每位参与者，我们从 ChatGPT 和 FeedbackGPT 两个会话片段中提取了所有反馈轮次及其之前的模型回复。

编码方案—反馈质量。 根据高质量反馈的定义（第 3.4.3 节），我们从目标参照、可执行和表达清晰这三个二元维度，对每个反馈轮次进行人工标注；同时统计反馈轮次总数，以评估渐进性。对于每个反馈轮次，每个维度均标记为存在（1）或不存在（0）。[1]。详细的编码手册见补充材料。第一作者和第二位编码员各自独立标注了随机抽取的 20% 反馈轮次，以评估可靠性；目标参照、可执行和表达清晰的 Cohen's κ 分别为 0.895、0.878 和 1。双方讨论并解决了分歧。随后，第一作者对其余数据进行了编码。

编码方案—反馈障碍。 与此同时，第一作者依据反馈障碍的定义（3.5.5），对每个模型回复中存在的反馈障碍进行编码。按照上述相同步骤，第二位编码员独立编码了随机抽取的 20% 样本。Cohen's κ：共同基础=1、可验证性=0.95、沟通=1、信息丰富度=1。

衍生指标。 对于每位参与者 p 和每种系统条件$s \in$ {ChatGPT, FeedbackGPT}，我们计算了各维度 d（目标参照、可执行和表达清晰）的比例：

- **维度比例：**

$$d_{p,s} = \frac{\#\{\text{feedback turns coded as } d \text{ for } p,s\}}{\#\{\text{feedback turns for } p,s\}} \times 100.$$

这些标准化比率使我们能够在控制轮次数量差异的同时，比较不同系统的反馈质量。

渐进式反馈的计算方式为：

$$\text{Vol}_{p,s} = \frac{\sum_{\text{feedback turns } t \text{ for } p,s} \text{characters}(t)}{\#\{\text{feedback turns for } p,s\}},$$

即该情节中每个反馈轮次的平均字符数。

5.3.2 反馈障碍的影响。 为评估 FeedbackGPT 的支架是否减少了四种反馈障碍（3.5.5），我们结合任务后调查量表，对开放式回答和访谈进行了定性分析。在任务后调查中，参与者针对其使用 *FeedbackGPT*的体验，回答了聚焦反馈障碍的问题（五点李克特量表）。

针对每项障碍，我们构建了一个包含两个条目的分量表。对于每个分量表，我们计算了两个条目的参与者层面均值。每个双条目分量表的内部一致性均可接受（$0.701 \leq \alpha \geq 0.852$）。此外，我们将三个全局条目作为独立指标，用于衡量反馈的整体便捷程度以及对不同系统的输出偏好。详情请参阅补充材料。

开放式访谈。 为帮助理解问卷评分，参与者参加了简短的研究后访谈，分享他们使用这两个系统及其支架的体验。

[1]轮次可以同时体现多个维度，例如，一条消息既可能高度相关，也可能十分具体。

每个任务后的满意度和生产力。 每项任务结束后，参与者还会针对刚刚使用的系统（FeedbackGPT 或 ChatGPT），使用五点李克特量表完成三个条目（输出满意度、对话满意度和感知生产力），评分范围从非常不同意 （1）到非常同意 （5）。

5.4 参与者

我们通过多家大型国际科技公司和大学的邮件列表及内部渠道招募参与者。纳入标准为：（1）英语流利；（2）此前熟悉对话代理（例如 ChatGPT ）。

共有 22 名参与者开始了研究，其中 2 人因任务未完成及未通过注意力检测题而被排除。我们的分析基于其余 20 名完成全部研究流程（两项任务、两种系统、任务后调查和访谈）的参与者。

该研究通过面对面或虚拟方式（通过视频通话共享屏幕）进行，每次时长约一小时。所有参与者均自愿参加该研究。所有参与者均提供了知情同意，且该研究已获公司内部审查委员会批准。

在20名参与者中，6人认定为女性，13人认定为男性，1人认定为非二元性别或第三性别。教育程度的中位数为学士学位。年家庭收入中位数介于50,000美元和100,000美元之间，年龄中位数为25–34岁。

5.5 分析

研究 2 的主要目标是比较基线 ChatGPT 界面和脚手架式 FeedbackGPT 界面对以下两方面的影响：(1) 用户反馈的客观质量准则；(2) 参与者提供反馈并引导人工智能时的主观体验。每位参与者均以随机顺序使用这两个系统；所有推断分析均将系统 （ChatGPT 与 FeedbackGPT）视为被试内因素。

用户层面分析。对于反馈质量指标（相关性、具体性、清晰度比例和数量）以及参与者层面的调查综合得分（感知交互质量、障碍子量表，以及提供反馈的便捷性等全局条目），我们首先按上文所述计算了每位参与者在各系统上的得分。随后，我们使用威尔科克森符号秩检验对各系统进行比较；该检验是配对 t检验的非参数替代方法，鉴于样本量较小（$N = 20$）且若干测量呈非正态分布，采用该方法是合适的。

我们以均值和标准误报告描述性统计（例如，$M \pm SE$），并为威尔科克森检验提供精确的 p值。遵循人机交互与心理学领域的常见做法 [20],，我们将 $p < .05$ 的结果视为具有统计显著性，并将 $0.05 \leq p < 0.10$ 视为边缘显著，表明这些是需谨慎解读的提示性趋势。

轮次层面分析。 对于在轮次或片段层面定义的结果（例如，某反馈轮次是否表现出给定属性的二元指标），我们使用广义估计方程（GEE）来处理嵌套于参与者之内的观测之间的非独立性。我们将参与者 ID 建模为聚类变量，采用可交换工作相关结构，并使用：

- 用于二元结果的logit连接（例如，某一轮次是否被编码为相关），以及

- 使用恒等链接函数处理连续型结果（例如，每个片段的字符数）。

定性分析。对开放式问卷回答和任务后访谈的定性分析遵循第3.2节所述的反思性主题分析程序。我们使用这些主题来三角验证并解释定量研究结果，尤其是关于参与者如何体验反馈障碍以及为应对这些障碍而设计的可供性。

6 结果

6.1 操纵检验

为确保比较公平，我们检验了参与者在两种条件下接触反馈障碍的程度是否相近。对于每种障碍，我们计算了参与者个体比率（存在该障碍的轮次所占比例），并使用描述性统计和广义估计方程逻辑回归模型进行系统间比较，其中系统（FeedbackGPT 与 ChatGPT）作为预测变量。

关系和消歧的障碍率较低，可验证性的障碍率中等。平均而言，参与者在使用 ChatGPT 时，1.45% 的轮次中遇到关系违规；使用 FeedbackGPT 时，这一比例为 3.50% ± 2.64；可验证性失败的比例分别为 54.33% ± 8.83（ChatGPT）和 47.55% ± 7.84（FeedbackGPT）；消歧问题的比例分别为 1.94%±1.17 和 5.85%±2.75；信息量不足的比例分别为 4.38% ± 2.18 和 2.25% ± 1.27。这些用户层面的差异均未达到统计显著性（所有威尔科克森检验 $p \geq .173$）。

与这些描述性模式一致，广义估计方程模型发现，任何一种障碍的存在都不存在显著的系统效应：关系（优势比（OR）= 1.17，$p = .878$）、可验证性（OR = 1.14，$p = .715$ = 3 85 $p = .071$ ）、消歧（OR .，）以及信息丰富度（OR = 0.56，$p = .399$）。综上，这些结果表明，参与者在两种条件下经历的模型侧故障程度相近，因此能够公平检验支架的效果。

6.2 降低反馈障碍、提高反馈质量的脚手架

接下来，我们将从四个维度考察 FeedbackGPT 是否提高了用户反馈质量：目标参照性、可执行性、表达清晰度和渐进性。对于前三个维度，我们分析其占比；对于渐进性，则按照第 5.3 节所述，分析每个反馈轮次的平均字符数。

目标参照——脚手架帮助参与者始终围绕目标。与 ChatGPT（CG）相比，FeedbackGPT（FG）使参与者的反馈更具目标参照性。平均而言，参与者的相关反馈比例从 32.49% ± 7.07 （CG）上升至 58.43% ± 6.56 （FG），提升了 +25.94%（SE=10.08）；威尔科克森符号秩检验：n=20，W=40.5，$p < .05$ 。这一客观提升与任务后问卷数据相符，参与者强烈认同 FeedbackGPT 能够提供**目标参照反馈（将目标/标准纳入考量）**（$\bar{x} = 3.60, p < .01$）。此外，我们发现有 2 名参与者重新开始了与 ChatGPT 的对话，但没有证据表明有人重新开始与 FeedbackGPT 的对话。

参与者将这些收获归因于那些使任务目标“始终可见”的功能。例如，锚定式评论和对比帮助参与者在各轮次中维持意图：“使用 FeedbackGPT，我更长时间地维持了对话状态”（P9），以及“我比较了几个提示词，看看我更喜欢哪个回答”（P11）。行内控制还增强了人们对更改会被局部化的信任：“标记好/坏让我确信模型不会改动我想保留的部分”（P15）；“行内评论功能……让我能够指出具体出错的部分并要求修正”（P18）。

6.2.1 可执行性——支架促成了精准、详细的反馈。 归一化具体性率从 65.86% ± 4.58 (CG) 上升至 90.62% ± 2.60 (FG)， +24.76 增幅为（SE=4.34）；威尔科克森检验：n=20，W=1.0, $p < 0.01$。任务后评分也印证了这一结果，参与者认同 FeedbackGPT 有助于提升具体性（明确指出要修改的内容）（$\bar{x} = 3.48, p < .01$）。

访谈揭示了推动这一变化的两种机制。首先，针对特定位置添加注释，使用户能够精准修改，而无需冗长解释：“我可以在特定区域添加多条评论，而不用……手动复制内容并把所有事情都解释一遍”（P8）。其次，鼓励输入优化和简短来回交流的支架促使用户在生成前提供更详细的输入。正如 P6 所说：“Huddle 基本上是在提示我补充细节……我就会提供这些信息。”系统这种积极主动的参与促使用户提供具体细节，与他们通常的做法形成对比：模型往往会“自行填补空白”（P11）。

6.2.2 清晰表达——尽管参与者认为有所改进，模糊的反馈仍然存在。我们未观察到清晰度有显著差异：77.59% ± 6.59 84.85 ± 4.77 Δ = +7.26 =8 21 （CG）与 %（FG）；（SE .）；威尔科克森检验：n=20，W=35.0，$p = 0.463$。这一结果揭示了一个值得注意的矛盾。从定性反馈来看，一些参与者认为 FeedbackGPT 帮助他们提升了输入质量：“自动评估和提示词优化提升了我的提示词质量”（P6；P15）。任务后评分有力地支持了这一看法，用户认为 FeedbackGPT 显著改善了沟通（表达反馈）（$\bar{x} = 3.69, p < 0.01$）。然而在实际操作中，参与者仍然会提供模糊、含混的输入，例如“再来一个版本”和“删减到 950 个词”。这种脱节——参与者感觉表达更轻松，但这一感受并未转化为客观上更清晰的反馈——与自我报告的对话质量评分相吻合，后者也未显示出显著改善（FG=3.70 vs. CG=3.63）。

6.2.3 渐进式——支架以增加付出为代价，促进了更丰富的反馈。 使用 FeedbackGPT 时，参与者提供的反馈量显著增加。按每轮字符数计算，使用 FG 时参与者写了 585.06 ± 68.55 个字符，而使用 CG 时为 242.33 ± 35.24 个字符；威尔科克森检验：n=20，W=32.0, $p < .01$。这与任务后调查数据一致，调查显示用户认为 FeedbackGPT 在信息丰富度（提供所需信息）方面更优（$\bar{x} = 3.93$， $p < 0.01$）。

访谈揭示，这种参与度的提升是以增加投入为代价的。参与者觉得自己更有掌控感，生产力也更高——“我最喜欢的交互方式是发表评论并重新生成”（P2）——但也注意到了相应的代价：“因为我需要与评论互动，所以认知负荷更高了，但这让我更有生产力，也更有掌控感”（P2）。这种更有成效但也更费力的工作流程 解释了为何反馈质量准则和数量的客观提升，并未转化为感知生产力 的主观评价显著提高（FG=3.65 对比 CG=3.63），也未转化为输出质量 的主观评价显著提高（FG=3.91 对比 CG=3.87）。许多参与者认为这种权衡可以接受，因为这种互动更像是

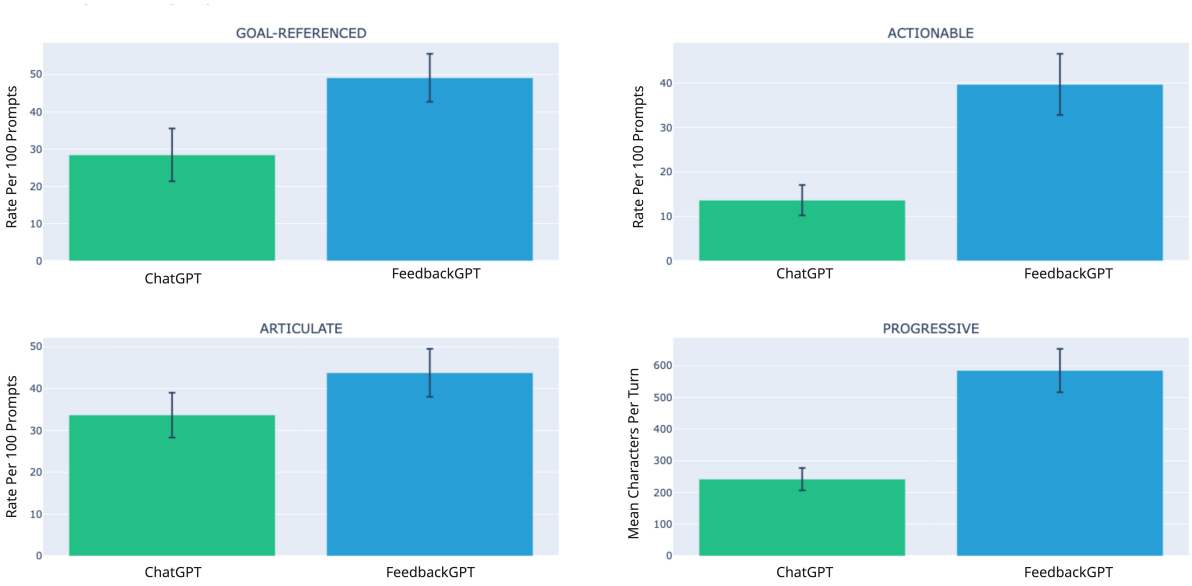


图 3：用户反馈分布散点图。x 轴表示 ChatGPT 给出高质量反馈的比例，y 轴表示 FeedbackGPT 每 100 条反馈中给出高质量反馈的比例。可以看出，FeedbackGPT 让反馈更具目标参照性、可执行性和渐进性。然而，用户在表达清晰的反馈时仍然面临困难。

类人协作。正如 P3 所解释的："这感觉更像是在和人互动，因为你可以指向某个地方，而对方知道你在说什么。"最终，尽管这些支架成功降低了提供高质量反馈的门槛，但认知负荷的增加拉低了用户对整体体验的主观评价。尽管如此，参与者确认"在 FeedbackGPT 中提供反馈更容易"（$\bar{x} = 3.69, p < 0.01$），而且它"提升了我引导人工智能的能力"（$\bar{x} = 3.63, p < 0.01$）。

7 讨论

我们的两项研究表明，现代人机协作中的反馈回路存在系统性失灵，这并非因为人类或模型单方面"失败"，而是由于合作原则相互崩塌。形成性研究发现了四种**反馈障碍**——共同基础、可验证性、沟通和信息丰富度——这些障碍源于模型局限与用户认知限制的共同作用。第二项研究表明，配备轻量级、与模型和任务无关的支架的系统**FeedbackGPT**能够在实践中缓解这些障碍，使用户能够提供明显更多具有目标参照性、可执行性和渐进性的反馈。

与此同时，我们的研究结果也揭示了一个关键的权衡。虽然这些支架让用户能够更有效地参与协作，但也增加了他们感知到的认知负荷：20 名用户中有 9 名认为，"提供恰当反馈的负担"仍完全落在自己身上。这表明，仅仅提供工具还不够；要实现有效的人机协作，就必须更深入地重新思考交互范式本身。此外，尽管个别支架可能会融入未来的商业对话代理界面，我们的持久贡献在于表达清晰地阐明设计要求，以及其所针对的根本反馈障碍——这些内容应当指导可供性的构建及其组合方式，而不是临时拼凑。下文将阐释我们的研究结果，为下一代会话代理提出设计启示，并呼吁更广泛的人工智能领域采取行动，应对模型层面的挑战。

7.1 从独白到对话：解读支架式支持的影响

我们的结果表明，FeedbackGPT 中的支架开始将互动从交易式、基于命令的独白转变为更具协作性的对话。在基线条件（ChatGPT）下，用户将会话代理视为需要接受指令的工具；当指令未能奏效时，他们便采用重新输入提示词或放弃等低成本修复策略。这种互动是单向且脆弱的。

这些支架营造了共享对话空间的感觉。行内评论等功能让用户能够将反馈锚定在特定位置，从而更准确地表达意图，并减少冗长且脱离上下文的解释的必要性。正如 P8 所说，这意味着他们可以"在特定区域添加多条评论，而不必……手动复制内容并逐一解释"。这与外部认知的原则相符 [101],：可见且持久的人工制品能够分担认知负担，并支持共享参照框架。然而，参与度的提升也伴随着代价。参与者表示，尽管他们觉得效率更高、掌控力更强，但认知负荷也更高——"提供恰当反馈的负担落在了我身上"（P19；P17、P2）。这种"富有成效的摩擦"是一项关键发现：有效协作需要付出心智努力。虽然我们的支架降低了提供反馈的门槛，但组织优质反馈这一行为仍然需要耗费大量认知

任务。这解释了为何反馈质量的客观提升并不总能转化为更高的轻松程度或效率主观评价。这表明，未来的系统不仅必须提供反馈工具，还应更积极地分担协作过程本身的认知负荷 [132]。综合来看，这些发现表明，支架的价值不只在于改善互动体验本身，还可能影响用户后续如何参与并取得对 AI输出的所有权。

在我们的访谈中，脚手架对后续表现的影响十分明显。参与者始终描述了更强的作者身份感和更高的感知质量："我觉得自己更拥有 FeedbackGPT 的输出了，因为我有了更好的控制力。"（P16；P3、P5、P19）；以及"……我看到自己的提示词质量不断提高，这也提升了输出质量"（P20；P5、P7）。虽然我们没有观察到自我报告的输出质量存在显著差异，但我们预计，随着系统更善于纳入用户反馈，以及用户越来越适应提供反馈，这些交互层面的变化将转化为可衡量的后续收益。未来的研究可以通过过程导向指标来量化这些收益，例如最终产物中用户生成内容所占的比例、达到稳定输出所需的轮次，或在存在真实值的领域中衡量客观任务表现。

7.2 真正协作式智能体的设计启示

我们的研究发现，尤其是参与者提供的定性反馈，指向了三项关键设计原则，助力系统超越单纯的辅助，实现真正的协作。

- **面向主动性与混合主动式进行设计**。 参与者明确表示，希望人工智能不只是被动回应；他们期待它成为能够主动提出自身完善建议（P6）、在生成完整回复前提供总体计划（P5、P20），甚至"说服"他们采纳更好想法（P15）的协作者。这要求我们回归人机交互中的混合主动交互原则，让系统能够主动引导用户、提出澄清性问题，并共同承担任务成功的责任。目标不仅是回应反馈，而是共同创造解决方案。

- **促进互惠性与相互 grounding**。 许多参与者仍觉得这种互动是"单向的"。正如 P3 所说："我给模型留评论，但模型好像不会给我留评论。"用户期待一种互惠对话，让人工智能也能提供自己的反馈和见解。P19 对此做了很好的总结："对话是双向的，你知道，我不仅想提供反馈，也想收到反馈。"未来的系统应纳入显式 grounding 机制，让模型确认自己对内容的理解。正如 P8 所设想的，评论应该变成"在那条评论里与朋友交谈"，从而实现来回交流和澄清。

- **让反馈的价值更易于理解**。 如果用户了解反馈的用途并看到其益处，就更有可能提供高质量反馈。当前带有简单赞成/反对按钮的系统缺乏透明度。组织科学表明，当反馈的益处清晰明确时，人们会提供更多反馈 [1, 6]。向用户展示他们的反馈如何在当前轮次之外发挥作用，可以帮助他们成为人工智能开发的合作伙伴，从而提升信任、对齐和安全性。

7.3 面向更广泛人工智能社区的行动呼吁

尽管我们的工作展示了界面与交互设计的力量，但从根本上克服反馈障碍，需要大语言模型核心能力同步取得进展。我们敦促人工智能社区解决导致这些障碍的模型侧因素：

- **要克服共同基础障碍，**模型需要的不只是更长的上下文窗口 [76]；它们还需要**上下文鲁棒注意力和记忆**。"中间丢失"效应 [77] 是共享指代的一项关键失效，使用户不得不持续重新构建上下文。研究如何缓解这一效应 [18, 150], , 同时开发更忠实的思维链推理 [71, 123] 和明确的规划能力 [130], , 对于维持一个会随对话更新的稳定共同目标至关重要。

- **要克服可验证性障碍，**事实核查的负担必须从用户身上转移开来 [94]。幻觉仍然是信任的核心障碍 [51, 118], ；尽管检索增强生成（RAG） [72] 有所帮助 [86, 113], , 但其有效性对检索质量 [30] 较为敏感。我们需要在设计上可验证的模型。这要求在三个关键领域取得进展： **（i）校准** [34, 116],, 使模型能够表达可靠的置信度分数，从而帮助处理副语言特征； **（ii）有原则的弃答** [128, 131, 136],, 使模型在不确定时能够寻求帮助；以及 **（iii）可靠的内联引用** [14, 145, 146], , 以证据为依据支撑论断。

- **为克服沟通和信息传递方面的障碍，**模型必须接受**多轮协作对话** [65, 144], 的训练，而不仅仅是遵循单轮指令 [90, 110]。这需要制定相应的基准和训练方案，以奖励主动性、寻求澄清和自我批评，从而反映参与者的期望 [13, 17, 24, 35, 75, 111]。此外，模型还必须学会根据用户线索调整其沟通风格，借助个性化 [151, 154] 和副语言 [38, 58, 60, 102, 103]方面的研究，以适当的详细程度提供信息，就像人类所做的那样 [39, 114]。

7.4 利用脚手架式界面中的用户反馈进行基于人类反馈的强化学习

如前文（第 2.3 节）所述，目前基于人类反馈的强化学习和偏好优化流程主要从粗粒度的轮次级信号（例如点赞/点踩、完整回答比较）中学习，这限制了奖励模型的粒度以及其能够支持的协作行为类型 [90, 97, 134, 141]。诸如 FeedbackGPT 这样的脚手架式界面则能够产生结构化、高质量的偏好数据：片段级高亮和评论、协作讨论记录以及提示词评估分数，可以自然地构成丰富的偏好元组（例如反馈前后回答、能够更好满足标记的约束的竞争性修订），并可输入基于人类反馈的强化学习

和 DPO 式训练流程，包括在线偏好学习和持续偏好学习 [93, 97, 141]。此外，用户何时打开协作讨论、请求解释或反复标记同一错误的日志，会暴露出模型询问澄清性问题或进行自我批评的“错失机会”；这些正是近期关于主动式和协作式大语言模型的研究所针对的状态 [23, 83, 137, 148]。

8 局限性

作为这项实证研究的一部分，我们发现了若干局限性。首先，我们的样本并不能代表总体。尽管我们努力招募背景和使用场景各异的参与者，但样本中仍存在不均衡，尤其体现在性别和教育水平方面。此外，由于第二项研究的性质及其较高的认知负荷，愿意参与并完成任务的参与者可能存在选择偏差。

其次，我们的任务聚焦于协同写作，以便使研究发现具有普遍适用性；然而，其他具有特定任务界面的交互界面可能会引发用户的不同行为。

第三，我们的支架并不全面，而只是对设计要求的一种具体实现。不同的功能组合可能会引发用户的不同行为。此外，我们尚未研究不同操作之间，或这些操作与用户之间，如何相互作用，从而引发用户和模型的不同反馈及行为。我们将这一探索留待未来工作。

最后，我们也承认，随着用户逐渐熟悉界面和模型行为，他们的反馈行为会随时间发生变化。鉴于我们的研究时间较短，我们无法预测用户行为将如何演变。我们将对用户反馈行为开展纵向研究留作未来工作。

9 结论

随着对话代理成为复杂多轮任务中不可或缺的协作者，其效能在很大程度上取决于高质量的用户反馈——而这种资源往往稀缺且难以获取。通过一项形成性研究，以及对采用脚手架的系统（FeedbackGPT）与基线系统（ChatGPT）进行的对照比较，我们以实证方式表明：（1）自然发生的反馈受到四种反馈障碍（共同基础、可验证性、沟通和信息丰富度）的阻碍，这些障碍源于用户与模型之间的互惠性失灵；（2）将持久共享参照、互惠互动和可验证推理付诸实践的轻量级、模型无关支架，能够显著提升用户反馈的相关性、具体性和数量。这些发现凸显了一项关键洞见：实现高质量反馈，不仅要依靠更强大的模型，还要通过交互设计让提供反馈在认知上负担得起、在社交层面具有互惠性，并且能带来看得见的影响。

展望未来，我们的研究为该领域提出了双重议程。在设计方面，从业者必须整合支架，帮助用户传播、参与反馈并对其进行评估，从而维持合作规范。在模型侧，记忆、校准和多轮训练方面的进展，对于减轻用户负担并实现真正的互惠对话至关重要。更广泛地说，让丰富的人类反馈重新融入流程，不仅仅是可用性问题，更是构建可信度更高、适应性更强且与社会更加协调的人工智能系统的根本要求；这能使用户从被动的消费者转变为安全、有效的人工智能开发过程中的积极伙伴。

致谢

本研究部分由 Schmidt Sciences 颁发的可信人工智能科学奖资助。

参考文献

[1] Nicole Abi-Esber、Jennifer E Abel、Juliana Schroeder 和 Francesca Gino。2022。“只是想让你知道……”低估了他人对建设性反馈的需求。《人格与社会心理学杂志》123，6（2022），1362。[2] Daehwan Ahn、Abdullah Almaatouq、Monisha Gulabani 和 Kartik Hosanagar。2024。模型可解释性和结果反馈对人工智能信任的影响。载于*2024*年*CHI*人因计算系统大会论文集（美国夏威夷州檀香山）（*CHI '24*）。计算机协会，美国纽约州纽约市，文章 27, 25 10 1145 3613904 3642780 页码。doi: . / . J.E. Allen、[3] 1999 C.I. Guinn 和 E. Horvtz。. 混合主动交互。*IEEE* 智能系统及其应用14，5（1999），14–23。doi:10.1109/5254. 796083[4] Saleema Amershi、Maya Cakmak、William Bradley Knox 和 Todd Kulesza。2014。赋权于民：人类在交互式机器学习中的作用。《人工智能杂志》 35, 4 2014 105 120 10 1609 35 4 2513（12月），–。[5] 2003 doi: . /aimag.v i . Christopher J. Anderson。. 无所作为的心理学：决策回避的形式源于理性与情绪。《心理学通报》 129 2003 139 167 =895727（），–。https://ssrn.com/abstract [6] F. Anseel、A. S. Beatty、W. Shen、F. Lievens 和 P. R. Sackett。2015。30年后情况如何？对寻求反馈行为的前因与结果进行元分析综述。《管理学杂志》41，1（2015），318–348。doi:10.1177/0149206313484521[7] Neeraj Arora、Ishita Chakraborty 和 Yohei Nishimura。2025。用于营销研究的人工智能—人类混合体：将大语言模型（LLM）用作协作者。《营销学杂志》89，2 2025 43 70 10 1177 00222429241276529（），–。arXiv:https://doi.org/ . / doi:10.1177/00222429241276529[8] Ariel Flint Ashery、Luca Maria Aiello 和 Andrea Baronchelli。2025。大型语言模型群体中涌现的社会惯例与集体偏见。《科学进展》11，20 2025 9368 10 1126（），eadu [9] 9368 doi: . /sciadv.adu 2013 Alberto Bacchelli 和 Christian Bird。. 现代代码评审的预期、成果与挑战。载于 *2013* 年第 *35* 届国际软件工程大会（*ICSE*）。 712 721 10 1109 2013 6606617 –。doi: . /ICSE. . Paul [10] 1998 Black 和 Dylan Wiliam。. 评估与课堂学习。《教育评估：原则、政策与实践》5，1（1998），7–74。arXiv:https://doi.org/ 10 1080 0969595980050102 10 1080 0969595980050102 . / doi: . / [11] 2023 Virginia Braun 和 Victoria Clarke。. 迈向良好的主题分析实践：避免常见问题，成为有见识的研究者。《国际跨性别健康杂志》24，1（2023），1–6。arXiv:https://doi.org/10.1080/26895269.2022.2129597 doi:10.1080/26895269.2022.2129597[12] David Carless 和 David Boud。2018。学生反馈素养的发展：促进反馈的采纳。《高等教育评估与评价》43，8（2018），1315–1325。arXiv:https://doi.org/10.1080/02602938.2018.1463354 doi:10.1080/02602938.2018.1463354[13] Amartya Chakraborty、Paresh Dashore、Nadia Bathaee、Anmol Jain、Anirban Das、Shi-Xiong Zhang、Sambit Sahu、Milind Naphade 和 Genta Indra Winata。2025。T1：面向多轮智能体规划的面向工具对话数据集。*arXiv*预印本 *arXiv:2505.16986*（2025）。[14] Joseph Chee Chang、Amy X. Zhang、Jonathan Bragg、Andrew Head、Kyle Lo、Doug Downey 和 Daniel S. Weld。2023。CiteSee：为科学论文中的引文补充持久且个性化的历史背景。载于*2023*年*CHI*人因计算系统大会论文集（德国汉堡）（*CHI '23*）。计算机协会，美国纽约州纽约市，文章 737, 15 10 1145 3544548 3580847 [15] 2017 页码。doi: . / . Hongshen Chen、Xiaorui Liu、Dawei Yin 和 Jiliang Tang。. 对话系统综述：最新进展与新前沿。《*ACM SIGKDD* 探索通讯》19，2（2017），25–35。[16] Li Chen 和 Pearl Pu。2007。基于偏好的推荐组织中混合式评价系统的评估。载于*2007*年*ACM* 推荐系统大会论文集（美国明尼苏达州明尼阿波利斯）（*RecSys '07*）。计算机协会，美国纽约州纽约市，169–172。doi:10. 1145/1297231.1297263

[17] Maximillian Chen、Ruoxi Sun、Tomas Pfister 和 Sercan Arik。2025。澄清学习：通过基于行动的对比式自训练进行多轮对话。载于国际学习表征会议，Y. Yue、A. Garg、N. Peng、F. Sha 和 R. Yu（编），第2025卷。32244–32279。https://proceedings.iclr.cc/paper_files/paper/2025/file/4ffd05ca3cf3985f4572af015b4cfc1e-Paper-Conference.pdf [18] Yeounoh Chung、Gaurav T. Kakkar、Yu Gan、Brenton Milne 和 Fatma Özcan。2025。长上下文就是你所需要的一切吗？利用大语言模型的扩展上下文实现自然语言到SQL。*VLDB*基金会会刊18，8（2025年4月），2735–2747。doi:10.14778/3742728.3742761[19] Herbert H Clark 和 Susan E Brennan。1991。交际中的奠基。载于社会共享认知视角。美国心理学会，127–149。[20] Jacob Cohen。1992。统计功效入门。《心理学通报》112，1（1992），155–159。doi:10.1037//0033-2909.112.1.155[21] Clayton Cohn、Ashwin T S、Naveeduddin Mohammed 和 Gautam Biswas。2025。CoTAL：用于可泛化形成性评估评分的人在回路中的提示工程。arXiv: 2504 02323 [ ] 2504 02323 . cs.CL [22] 2025 https://arxiv.org/abs/ . David Wei Dai、Hua Zhu 和 Guanliang Chen。. 与大语言模型驱动的聊天机器人交互如何塑造人们对文化的理解？批判性交互能力（CritIC）的必要性。《应用语言学年度评论》 45 2025 28 49 10 1017 0267190525000054（），–。doi: . /S Yang Deng、Lizi [23] Liao、Wenqiang Lei、Grace Hui Yang、Wai Lam 和 Tat-Seng Chua。2025。主动式对话人工智能：进展与机遇综述。《*ACM*信息系统汇刊》43，3，文章67（2025年3月），45页。doi:10.1145/3715097 [24] Kaustubh Deshpande、Ved Sirdeshmukh、Johannes Baptist Mols、Lifeng Jin、Ed-Yeremai Hernandez-Cardona、Dean Lee、Jeremy Kritz、Willow E. Primack、Summer Yue 和 Chen Xing。2025。MultiChallenge：挑战前沿大语言模型的真实多轮对话评测基准。载于计算语言学协会论文集：*ACL 2025*，Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，18632–18702。doi:10.18653/v1/ 2025 958 .findings-acl. Berkeley J Dietvorst、Joseph P Simmons 和 [25] Cade Massey。2015。算法厌恶：人们看到算法出错后会错误地避开算法。《实验心理学杂志：总论》 144, 1 2015 114 126 10 1037 0000033（），–。[26] 2025 doi: . /xge Shachar Don-Yehiya、Leshem Choshen 和 Omri Abend。. 自然发生的反馈十分普遍、可提取且有用。arXiv:2407.10944 [cs.CL] https://arxiv.org/abs/ 2407 10944 . John J. Dudley 和 Per Ola Kristensson。2018。[27] 交互式机器学习用户界面设计综述。《*ACM*交互式智能系统汇刊》8，2，文章8（2018年6月），37页。doi:10.1145/3185517[28] Zane Durante、Qiuyuan Huang、Naoki Wake、Ran Gong、Jae Sung Park、Bidipta Sarkar、Rohan Taori、Yusuke Noda、Demetri Terzopoulos、Yejin Choi、Katsushi Ikeuchi、Hoi Vo、Li Fei-Fei 和 Jianfeng Gao。2024。智能体人工智能：多模态交互前沿综述。arXiv:2401.03568 [cs.AI] https://arxiv.org/ abs/2401.03568[29] Jerry Alan Fails 和 Dan R Olsen Jr. 2003。交互式机器学习。载于第八届智能用户界面国际会议论文集。ACM，39–45。doi:10.1145/604045.604056[30] Wenqi Fan、Yujuan Ding、Liangbo Ning、Shijie Wang、Hengyun Li、Dawei Yin、Tat-Seng Chua 和 Qing Li。2024。关于RAG与大语言模型相结合的综述：迈向检索增强大语言模型。载于第*30*届*ACM*知识发现与数据挖掘特别兴趣小组会议论文集（西班牙巴塞罗那）（*KDD '24*）。计算机协会，美国纽约州纽约市，6491–6501。doi:10.1145/3637528.3671470[31] Patrick Fernandes、Aman Madaan、Emmy Liu、António Farinhas、Pedro Henrique Martins、Amanda Bertsch、José G. C. de Souza、Shuyan Zhou、Tongshuang Wu、Graham Neubig 和 André F. T. Martins。2023。弥合差距：整合（人类）反馈以实现自然语言生成的综述。《计算语言学协会汇刊》11（2023），1643–1668。doi: 10 1162 _ _00626 . /tacl a [32] Jie Gao、Simret Araya Gebreegziabher、Kenny Tsu Wei Choo、Toby Jia-Jun Li、Simon Tangi Perrault 和 Thomas W Malone。2024。人类与大语言模型交互模式分类体系：初步探索。载于*CHI*人因计算系统大会扩展摘要（*CHI '24*）。ACM，1–11。doi:10.1145/3613905.3650786[33] Yubin Ge、Ziang Xiao、Jana Diesner、Heng Ji、Karrie Karahalios 和 Hari Sundaram。2023。我该问什么：一种用于对话式调查后续问题生成的知识驱动方法。载于第*37*届亚太语言、信息与计算会议论文集。113–124。[34] Jiahui Geng、Fengyu Cai、Yuxia Wang、Heinz Koeppl、Preslav Nakov 和 Iryna Gurevych。2024。大语言模型中的置信度估计与校准综述。载于*2024*年北美计算语言学协会分会年会：人类语言技术论文集（第一卷：长篇论文），Kevin Duh、Helena Gomez 和 Steven Bethard（编）。计算语言学协会，墨西哥城，6577–6595。doi:10.18653/v1/2024.naacl-long.366[35] Zhibin Gou、Zhihong Shao、Yeyun Gong、yelong shen、Yujiu Yang、Nan Duan 和 Weizhu Chen。2024年。CRITIC：大语言模型可以通过工具交互式批评进行自我纠正。载于《第十二届国际学习表征会议》。https://openreview.net/forum?id=Sx038qxjek [36] H. P. Grice。1975年。《逻辑与会话》。Brill，荷兰莱顿，41–58。doi:10.1163/9789004368811 003[37] David Gunning 和 David W Aha。2019年。美国国防高级研究计划局的可解释人工智能（XAI）项目。《人工智能杂志》40，2（2019年），44–58。[38] J. J. Guyer、P. Briñol、T. I. Vaughan-Johnston、L. R. Fabrigar、L. Moreno 和 R. E. Petty。2021年。通过语音传达的副语言特征会影响信心评价和评价性判断。《非语言行为杂志》 45, 4 2021 479 504 2021 07 06 10 1007 10919 021 00374 2（），–。[39] 2007 arXiv: - - doi: . /s - - - John Hattie 和 Helen Timperley。. 反馈的力量。《教育研究评论》77，1（2007年），81–112。[40] Chen He、Denis Parra 和 Katrien Verbert。2016年。交互式推荐系统：现状调查及未来研究面临的挑战与机遇。《专家系统与应用》56（2016年），9–27。doi:10.1016/j. eswa.2016.02.013[41] M. Henderson、M. Bearman、J. Chung、T. Fawns、S. Buckingham Shum、K. E. Matthews 和 J. de Mello Heredia。2025年。生成式人工智能与教师反馈的比较：学生对其有用性和可信度的看法。《高等教育评估与评价》（ 2025 1 16 10 1080 02602938 2025 2502582），–。[42] 2020 doi: . / . . Donald Honeycutt、Mahsan Nourani 和 Eric Ragan。. 征求人在回路中的用户反馈用于交互式机器学习，会降低用户对模型的信任及其对模型准确性的印象。《AAAI人类计算与众包会议论文集》8，1（2020年10月），63–72。doi:10.1609/hcomp.v 8 1 7464 i . Eric Horvitz。. 混合主动用户界面的原则。载于《SIGCHI计算机系统中的人为因素会议论文集》。ACM，159–166。[43] 1999 doi:10.1145/302979.303030[44] Eric Horvitz、Tim Paek 和 Mani Subramani。1999年。面向效用的偏好评估中的持续计算策略。载于《第十五届人工智能中的不确定性会议论文集》。Morgan Kaufmann Publishers Inc.，269–278。[45] Dietmar Jannach、Ahtsham Manzoor、Wanling Cai 和 Li Chen。2021年。对话推荐系统综述。《ACM计算调查》54，5，文章 105 2021 36 10 1145 3453154（5月），页码。doi: . / Juyong Jiang、[46] Fan Wang、Jiasi Shen、Sungju Kim 和 Sunghun Kim。2025年。代码生成大语言模型综述。《ACM软件工程与方法学汇刊》（2025年7月）。doi:10.1145/3747588。已接收待刊。[47] Yuyang Jiang、Longjie Guo、Yuchen Wu、Aylin Caliskan、Tanu Mitra 和 Hua Shen。2025年。超越单向影响：多轮人与大语言模型交互中的双向观点动态。arXiv:2510.20039 [cs.HC] https://arxiv. org/abs/2510.20039[48] Chuanyang Jin、Jing Xu、Bo Liu、Leitian Tao、Olga Golovneva、Tianmin Shu、Wenting Zhao、Xian Li 和 Jason Weston。2025年。真实世界人机交互时代：从用户对话中进行强化学习。arXiv:2509.25137 [cs.AI] https://arxiv. org/abs/ 2509 25137 . Hideaki Joko、Shakiba Amirshahi、Charles L. A. Clarke [49] 和 Faegheh Ha- sibi。2025年。WildClaims：野外信息获取对话（聊天）。arXiv: 2509 17442 [ ] 2509 17442 。cs.IR [50] 2025 https://arxiv.org/abs/ 。Wendy Ju 等。. 与人工智能体在实地协作：一项关于情境化人机交互的研究。载于《2025年CHI人因计算系统大会论文集》。ACM。[51] Adam Tauman Kalai、Ofir Nachum、Santosh S. Vempala 和 Edwin Zhang。2025年。语言模型为何产生幻觉。arXiv:2509.04664 [cs.CL] https: // arxiv.org/abs/2509.04664[52] Dongjin Kang、Sunghwan Kim、Taeyoon Kwon、Seungjun Moon、Hyunsouk Cho、Youngjae Yu、Dongha Lee 和 Jinyoung Yeo。2024年。大语言模型能成为优秀的情感支持者吗？缓解情感支持对话中的偏好偏差。载于《第62届计算语言学协会年会论文集》（第1卷：长篇论文），Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷， 15232 15261 10 18653 1 2024 813 –。doi: . /v / .acl-long。[53] Yoonsu Kim、Brandon Chin、Kihoon Son、Seoyoung Kim 和 Juho Kim。2025年。将格赖斯准则应用于人与大语言模型交互循环：参与式方法带来的设计启示。CoRR abs/2503.00858（2025年3月）。https://doi.org/ 10 48550 2503 00858 . /arXiv. . Hannah Rose Kirk、Andrew [54] M. Bean、Bertie Vidgen、Paul Röttger 和 Scott A. Hale。2023年。大语言模型反馈学习的过去、现在与更美好的未来：面向主观人类偏好与价值观。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， 2409 2430 10 18653 1 2023 148 –。doi: . /v / . [55] emnlp-main。Hannah Rose Kirk、Iason Gabriel、Chris Summerfield、Bertie Vidgen 和 Scott A. Hale。2025年。人类与人工智能的关系为何需要社会情感对齐。《人文与社会科学传播》12（2025年），728。doi:10.1057/s41599-

025-04532-5[56] Asger Kirkeby-Hinrup 和 Jakob Stenseke。2025年。大语言模型交互心理学：恐怖谷与其他心智。《心理学与人工智能杂志》 1, 1 2025 10 1080 29974100 2025 2457627 ( )。doi: . / . . David Kirsh 和 Paul Maglio。 . 论认识性行动与实用性行动的区分。《认知科学》18，4（1994年），513–549。[58] S. H. Klein。2025年。类人社交线索对人们针对文本对话代理的社交反应的影响——一项荟萃分析。《人文与社会科学传播》12（2025年），1322。doi:10.1057/s41599-025-05618-w [59] Avraham N Kluger 和 Angelo DeNisi。1996年。反馈干预对表现的影响：历史回顾、荟萃分析及初步反馈干预理论。《心理学通报》119，2（1996年），254–284。[60] Dimosthenis Kontogiorgos 和 Julie Shah。2025年。询问机器人：利用人类非语言线索估计对解释的需求。载于《2025年ACM/IEEE人机交互国际会议论文集》（澳大利亚墨尔本）（HRI '25）。IEEE出版社，717–728。[61] Lea Krause 和 Piek T.J.M. Vossen。2024年。自然语言处理中的格赖斯准则：综述。载于《第17届国际自然语言生成会议论文集》，Saad Mahamood、Nguyen Le Minh 和 Daphne Ippolito（编）。计算语言学协会，日本东京，470–485。doi:10.18653/v1/2024.inlg- main.39[62] Todd Kulesza、Margaret Burnett、Weng-Keen Wong 和 Simone Stumpf。2015年。交互式机器学习个性化的解释性调试原则。载于《第20届智能用户界面国际会议论文集》（美国佐治亚州亚特兰大）（IUI '15）。计算机协会，美国纽约州纽约市，126–137。doi:10.1145/2678025.2701399[63] Todd Kulesza、Simone Stumpf、Margaret Burnett、Sherry Yang、Irwin Kwan 和 Weng-Keen Wong。2013年。过多、不足还是恰到好处？解释影响终端用户心智模型的方式。载于《2013年IEEE可视化语言与以人为中心的计算研讨会》。 3 10 10 1109 2013 6645235 –。doi: . / [64] VLHCC. . Indah Kurniawan、Alexander Soutschek 和 Matthew Apps。 . 选择现在阻力最小、而非以后：将认知努力推迟到未来会降低努力折扣。《心理学公报与评论》30，1（2023年），1–15。doi:10.3758/s13423-022-02198-7[65] Philippe Laban、Hiroaki Hayashi、Yingbo Zhou 和 Jennifer Neville。2025年。大语言模型在多轮对话中迷失方向。arXiv:2505.06120 [cs.CL] https: //arxiv.org/abs/ 2505 06120 。Vivian Lai、Yiming Zhang、Chacha Chen、Q. [66] Vera Liao 和 Chenhao Tan。2023年。选择性解释：利用人类输入使可解释人工智能实现对齐。《ACM人机交互汇刊》7，CSCW2，文章357（2023年10月），35页。doi:10. 1145/3610206[67] Daniel Lee、Nikhil Sharma、Donghoon Shin、DaEun Choi、Harsh Sharma、Jeongh- wan Kim 和 Heng Ji。2025年。ThematicPlane：连接隐性用户意图与图像生成的潜在空间。载于《第38届ACM用户界面软件与技术研讨会（UIST Adjunct '25）附属论文集》。计算机协会，美国纽约州纽约市，文章120，3页。doi:10.1145/3746058.3758376[68] Harrison Lee、Samrat Phatale、Hassan Mansoor、Thomas Mesnard、Johan Ferret、Kellie Lu、Colton Bishop、Ethan Hall、Victor Carbune、Abhinav Rastogi 和 Sushant Prakash。2024年。基于人工智能反馈的强化学习与基于人类反馈的强化学习：利用人工智能反馈扩展基于人类反馈的强化学习。载于《第41届国际机器学习会议论文集》（奥地利维也纳）（ICML'24 1071，28）。JMLR.org，文章页码。[69] John D Lee 和 Kelly A See。2004年。自动化信任：为适当依赖而设计。《人为因素》46，1（2004年），50–80。[70] Mina Lee、Megha Srivastava、Amelia Hardy、John Thickstun、Esin Durmus、Ashwin Paranjape、Ines Gerard-Ursin、Xiang Lisa Li、Faisal Ladhak、Frieda Rong、Rose E Wang、Minae Kwon、Joon Sung Park、Hancheng Cao、Tony Lee、Rishi Bommasani、Michael S. Bernstein 和 Percy Liang。2023年。评估人类与语言模型的交互。《机器学习研究汇刊》（2023年）。https://openreview.net/forum?id = 9 1 2025 [71] hjDYJUn l S. Lee、X. Wang、A. Zhang 等。 . 深度模型型的忠实且与人类对齐的自我解释。npj《人工智能》1（2025年），21。doi:10. 1038/s44387-025-00023-9[72] Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel、Sebastian Riedel 和 Douwe Kiela。2020年。面向知识密集型自然语言处理任务的检索增强生成。载于《第34届神经信息处理系统国际会议论文集》（加拿大不列颠哥伦比亚省温哥华）（NIPS '20 793，16）。Curran Associates I nc.，美国纽约州雷德胡克，文章页码。[73] Jingshu Li、Tianqi Song、Beichen Xue 和 Yi-Chieh Lee。2025年。我们塑造人工智能，继而人工智能塑造我们：人类通过社会影响与人工智能对齐。载于《2025年国际学习表征会议双向人机对齐研讨会》。https://openreview. net/forum?id=64rCWVC78p[74] Zhiyu Lin、Upol Ehsan、Rohan Agarwal、Samihan Dani、Vidushi Vashishth 和 Mark O. Riedl。 [n. d.]。超越提示词：探索混合主动式共创系统的设计空间。《计算语言学国际会议》 创造力（ [ ] 10434407 无日期）。 [75] https://par.nsf.gov/biblio/ Zicheng Lin、Zhibin Gou、Tian Liang、Ruilin Luo、Haowei Liu 和 Yujiu Yang。2024。CriticBench：用于批评-纠正推理的大语言模型基准测试。载于计算语言学协会研究成果：*ACL 2024*，Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，1552–1587。doi:10.18653/v1/2024.findings-acl.91[76] Jiaheng Liu、Dawei Zhu、Zhiqi Bai、Yancheng He、Huanxuan Liao、Haoran Que、Zekun Wang、Chenchen Z hang、Ge Zhang、Jiebin Zhang、Yuanxing Zhang、Zhuo Chen、Hangyu Guo、Shilong Li、Ziqiang Liu、Yong Shan、Yifan Song、Jiayi Tian、Wenhao Wu、Zhejian Zhou、Ruijie Zhu、Junlan Feng、Yang Gao、Shizhu He、Zhoujun Li、Tianyu Liu、Fanyu Meng、Wenbo Su、Yingshui Tan、Zili Wang、Jian Yang、Wei Ye、Bo Zheng、Wangchunshu Zhou、Wenhao Huang、Sujian Li 和 Zhaoxiang Zhang。2025。长上下文语言建模综合综述。arXiv: 2503 17407 [ ] 2503 17407 。cs.CL [77] https://arxiv.org/abs/。Nelson F. Liu、Kevin Lin、John Hewitt、Ashwin Paranjape、Michele Bevilacqua、Fabio Petroni 和 Percy Liang。2024。迷失在中间：语言模型如何使用长上下文。《计算语言学协会汇刊》 12 2024 157 173 10 1162 _ _00638 （），–。doi:。/tacl a Edwin A [78] Locke 和 Gary P Latham。。目标设定与任务表现的新进展。第24卷。Routledge，纽约。[79] Erik Miehling、Manish Nagireddy、Prasanna Sattigeri、Elizabeth M. Daly、David Piorkowski 和 John T. Richards。2024。对话中的语言模型：人机交互的会话准则。载于计算语言学协会研究成果：*EMNLP 2024*，Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编）。计算语言学协会，美国佛罗里达州迈阿密，14420 14437 10 18653 1 2024 843 –。doi:。/v / . [80] findings-emnlp。Tim Miller。。人工智能中的解释：来自社会科学的见解。人工智能267（2019），1–38。[81] Taywon Min、Haeone Lee、Yongchan Kwon 和 Kimin Lee。2025。通过影响函数理解人类反馈的影响。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文）。计算语言学协会，27471–27500。doi:10.18653/v 1 2025 1333 / .acl-long。Eduardo Mosqueira-Rey、Elena [82] Hernández-Pereira、David Alonso-Ríos、José Bobes-Bascarán 和 Ángel Fernández-Leal。2023。人在回路中的机器学习：现状。人工智能评论56（2023），3005–3054。doi:10.1007/s10462-022-10246-w[83] Sheshera Mysore、Debarati Das、Hancheng Cao 和 Bahareh Sarrafzadeh。2025。真实场景中大语言模型辅助写作所呈现的典型人机协作行为。载于*2025*年自然语言处理实证方法会议论文集，Christos Christodoulopoulos、Tanmoy Chakraborty、Car- olyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，16819 16846 10 18653 1 2025 852 –。doi:。/v / . [84] emnlp-main。Clifford Nass、Jonathan Steuer 和 Ellen R Tauber。。计算机是社会行动者。载于*SIGCHI*计算机系统中的人为因素会议论文集。ACM，72–78。[85] David J. Nicol 和 Debra Macfarlane-Dick。2006。形成性评价与自我调节学习：教学系统模型及优质反馈实践的七项原则。高等教育研究31，2（2006），199–218。doi:10.1080/03075070600572090[86] Cheng Niu、Yuanhao Wu、Juno Zhu、Siliang Xu、KaShun Shum、Randy Zhong、Juntong Song 和 Tong Zhang。2024。RAGTruth：用于开发可信检索增强语言模型的幻觉语料库。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，10862–10878。doi: 10 18653 1 2024 585 。/v / .acl-long。Donald A Norman。。让我们 [87] 变聪明的事物：机器时代对人类特质的捍卫。Basic Books。[88] OpenAI。2022。ChatGPT。https://openai.com/blog/chatgpt[89] Jianzhe Ou、Krzysztof Z Gajos、Wojciech Matusik、Wilmot Li 和 Maneesh Agrawala。2022。ARTEMIS：用于创意三维建模的人机循环。载于第*35*届*ACM*用户界面软件与技术年会论文集（*UIST '22*）。75–88。doi:10.1145/3526113.3545685[90] Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。2022。训练语言模型遵循人类反馈指令。载于神经信息处理系统进展 *NeurIPS*。https://arxiv.org/ abs/2203.02155 [91] Laura Panfili、Steve Duman、Andrew Nave、Katherine Phelps Ridgeway、Nathan Eversole 和 Ruhi Sarikaya。2021。以格赖斯视角看人机交互。美国语言学会会议论文集6，1（2021年3月），288。doi: 10 3765 6 1 4971 。/plsa.v [92] i。Dojun Park、Jiwoo Lee、Hyeyun Jeong、Seohyun Park 和 Sungeun Lee。2024。面向韩语的大语言模型语用能力评估。载于第*38*届太平洋亚洲语言、信息与计算会议论文集，Nathaniel Oco、Shirley N. Dita、Ariane Macalinga Borlongan 和 Jong-Bok Kim（编）。东京外国语大学，日本东京，

日本，256–266。https://aclanthology.org/2024.paclic-1.25/ [93] Ryan Park、Rafael Rafailov、Stefano Ermon 和 Chelsea Finn。2024。将长度与质量从直接偏好优化中分离。载于计算语言学协会研究成果：*ACL 2024*，Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，4998–5017。doi:10.18653/v1/2024.findings-acl.297[94] Petr Parshakov、Iuliia Naidenova、Sofia Paklina、Nikita Matkin 和 Cornel Nesseler。2025。用户喜欢大语言模型生成的内容——直到他们知道这是人工智能生成的。论文2503.16458。arXiv.org。doi:None [95] Sandra Peter、Kai Riemer 和 Jevin D. West。2025。拟人化对话代理的益处与危险。美国国家科学院院刊122，22（2025），e2415898122。doi:10.1073 2415898122 /pnas。Nicolas Pfeuffer、Lorenz Baum、Wolfgang [96] Stammer、Benjamin M. Abdel-Karim、Patrick Schramowski、Andreas M. Bucher、Christian Hügel、Gernot Rohde、Kris- tian Kersting 和 Oliver Hinz。2023。解释型交互式机器学习——为机器学习项目建立行动设计研究流程。商业与信息系统工程65，6（2023），677–701。https://aisel.aisnet.org/bise/vol 65 6 4 /iss / Rafael Rafailov、Archit Sharma、[97] Eric Mitchell、Stefano Ermon、Christopher D. Manning 和 Chelsea Finn。2023。直接偏好优化：你的语言模型其实暗中充当奖励模型。载于第*37*届神经信息处理系统国际会议论文集（美国路易斯安那州新奥尔良）（*NIPS '23*）。Curran Associates Inc.，美国纽约州雷德胡克，第2338篇文章，共14页。[98] Fardin Saad、Pradeep K. Murukannaiah 和 Munindar P. Singh。2025。以格赖斯准则为基础实现有效协作。载于第*24*届自主智能体与多智能体系统国际会议论文集（美国密歇根州底特律）（*AAMAS '25*）。自主智能体与多智能体系统国际基金会，美国南卡罗来纳州里奇兰，1812–1820。[99] D. Royce Sadler。1989。形成性评价与教学系统的设计。《教学科学》18，2（1989），119–144。doi:10.1007/BF00117714[100] Pranab Sahoo、Ayush Kumar Singh、Sriparna Saha、Vinija Jain、Samrat Mondal 和 Aman Chadha。2025。大语言模型提示工程的系统综述：技术与应用。arXiv:2402.07927 [cs.AI] https://arxiv.org/abs/ 2402 07927 。Mike Scaife 和 Yvonne [101] Rogers。1996。外部认知：图形表征如何发挥作用？《国际人机研究期刊》45，2（1996），185–213。[102] Klaus R. Scherer、Harvey London 和 Jared J. Wolf。1973。自信的声音：副语言线索与受众评价。《人格研究杂志》 7, 1 1973 31 44 10 1016 0092 6566 73 90030 5 （），–。doi:。/ - ( ) - Juliana [103] Schroeder、Michael Kardas 和 Nicholas Epley。2017。人性化声音：在意见分歧中，语音揭示而文本掩盖了更善于思考的心智。《心理科学》28，12（2017），1745–1762。[104] Anna-Maria Seeger、Jella Pfeiffer 和 Armin Heinzl。2021。与类人对话代理发短信：为拟人化而设计。《信息系统协会期刊》22，4（2021），8。[105] Vidya Setlur 和 Melanie Tory。2022。如何与分析型聊天机器人对话？重新审视格赖斯准则，以设计分析型对话行为。载于*2022*年*CHI*人因计算系统大会论文集（美国路易斯安那州新奥尔良）（*CHI '22*）。计算机协会，美国纽约州纽约市，第29篇文章，共17页。doi:10.1145/3491102.3501972[106] Mrinank Sharma、Meg Tong、Tomek Korbak、David Duvenaud、Amanda Askell、Sam Bowman、Esin DURMUS、Zac Hatfield-Dodds、Scott Johnston、Shauna Kravec、Timothy Maxwell、Sam McCandlish、Kamal Ndousse、Oliver Rausch、Nicholas Schiefer、Da Yan、Miranda Zhang 和 Ethan Perez。2024。迈向对语言模型谄媚现象的理解。载于国际学习表征会议，B. Kim、Y. Yue、S. Chaudhuri、K. Fragkiadaki、M. Khan 和 Y. Sun（编），第2024卷，110–144。https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf [107] Nikhil Sharma、Q. Vera Liao 和 Ziang Xiao。2024。生成式回音室？大语言模型驱动的搜索系统对多元信息寻求的影响。载于*2024*年*CHI*人因计算系统大会论文集（美国夏威夷州檀香山）（*CHI '24*）。计算机协会，美国纽约州纽约市，第 1033, 17 10 1145 3613904 3642459 页。doi:。/。Nikhil [108] Sharma、Kenton Murray 和 Ziang Xiao。。伪多语者：多语言大语言模型中的信息差异研究。载于*2025*年计算语言学协会美洲分会会议：人类语言技术论文集（第*1*卷：长篇论文）。计算语言学协会，8090–8107。doi:10.18653/v1/ 2025.naacl-long.411 [109] Hua Shen、Tiffany Knearem、Reshmi Ghosh、Michael Xieyang Liu、Andrés Monroy-Hernández、Tongshuang Wu、Diyi Yang、Yun Huang、Tanushree Mitra、Yang Li 和 Marti Hearst。2025。双向人机对齐：新兴挑战与机遇。载于*CHI*人因计算系统大会扩展摘要论文集（*CHI EA '25*）。计算机协会，美国纽约州纽约市，第857篇文章，共6页。doi:10. 1145/3706599.3716291 [110] Lingfeng Shen、Sihao Chen、Linfeng Song、Lifeng Jin、Baolin Peng、Haitao Mi、Daniel Khashabi 和 Dong Yu。2024年。《奖励不一致对RLHF的层层影响》。载于《第十二届国际学习表征会议》。https://openreview.net/forum?id=MeHmwCDifc[111] Li Shi、Houjiang Liu、Yian Wong、Utkarsh Mujumdar、Dan Zhang、Jacek Gwizdka 和 Matthew Lease。2025年。《论辩体验：通过大语言模型生成的多角色辩论减少争议性议题上的确认偏误》。arXiv: 2412 04629 [ ] 2412 04629 。cs.HC https://arxiv.org/abs/ 。[112] Taiwei Shi、Zhuoer Wang、Longqi Yang、Ying-Chun Lin、Zexue He、Mengt- ing Wan、Pei Zhou、Sujay Kumar Jauhar、Xiaofeng Xu、Xia Song 和 Jennifer Neville。2024年。《WildFeedback：通过现场用户交互与反馈对齐大语言模型》。载于NeurIPS 2024行为机器学习研讨会。https://openreview.net/forum?id=07QCozT1pi [113] Kurt Shuster、Spencer Poff、Moya Chen、Douwe Kiela 和 Jason Weston。2021年。《检索增强可减少对话中的幻觉》。载于《计算语言学协会论文集：EMNLP 2021》，Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih（编）。计算语言学协会，多米尼加共和国蓬塔卡纳，第3784–3803页。doi: 10 18653 1 2021 320 。/v / .findings-emnlp。Valerie J. Shute。.[114] 《聚焦形成性反馈》。《教育研究评论》78卷，1期（2008年），153–189页。arXiv:https://doi.org/10.3102/0034654307313795 doi:10.3102/0034654307313795[115] Patrice Y. Simard、Saleema Amershi、David Maxwell Chickering、Alicia Edelman Pelton、Soroush Ghorashi、Christopher Meek、Gonzalo A. Ramos、Jina Suh、J ohan Verwey、Mo Wang 和 John Wernsing。2017年。《机器教学：构建机器学习系统的新范式》。CoRR abs/1707.06742（2017年）。arXiv: 1707 06742 1707 06742 。http://arxiv.org/abs/ 。M. Steyvers、[116] H. Tejeda、A. Kumar 等。.《大语言模型知道什么，以及人们认为它们知道什么》。《自然·机器智能》7卷（2025年），221–231页。doi:10.1038/s42256-024-00976-7[117] Simone Stumpf、Vidya Rajaram、Lida Li、Weng-Keen Wong、Margaret Burnett、Thomas Dietterich、Erin Sullivan 和 Jonathan Herlocker。2009年。《与机器学习系统进行有意义的交互：三项实验》。《国际人机研究期刊》67卷，8期（2009年），639–662页。doi:10.1016/j.ijhcs. 2009.03.004[118] Y. Sun、D. Sheng、Z. Zhou等。2024年。《人工智能幻觉：对人工智能生成内容中失真信息的全面分类》。《人文与社会科学传播》11卷（2024年），1278。doi:10.1057/ s41599-024-03811-x[119] Yan Tao、Olga Viberg、Ryan S Baker 和 René F Kizilcec。2024年。《大语言模型的文化偏见与文化对齐》。《PNAS Nexus》3卷，9期（2024年9月）。doi: 10 1093 346 。/ pnasnexus/pgae Maryam Tohidi、William Buxton、Ronald Baecker 和 Abigail Selle[120] n。2006年。《获得恰当的设计，并把设计做好》。载于《SIGCHI计算机系统中的人为因素会议论文集》。1243–1252页。[121] Keith J. Topping。2009年。《同伴评价》。《理论与实践》48卷，1期（2009年），20–27页。arXiv:https://doi.org/10.1080/00405840802577569 doi:10.1080/ 00405840802577569 [122] Amy Turner、Meena Kaushik、Mu-Ti Huang 和 Srikar Varanasi。2022年。《校准对人工智能辅助决策的信任》。谷歌学术 谷歌学术 导航至（2022年）。[123] Miles Turpin、Julian Michael、Ethan Perez 和 Samuel R. Bowman。2023年。《语言模型并不总是说出它们所想的内容：思维链提示中的不忠实解释》。载于第三十七届神经信息处理系统大会。https://openreview.net/forum?id=bzs4uPLXvi [124] Andrea Vaccaro 等。2024年。《人与人工智能协作时的表现往往不如独自工作：一项荟萃分析》。《自然·人类行为》（ 2024 10 1038 41562 024 02024 1 ）。doi: . /s - - - Michelle M.E. Van [125] Pinxteren、Mark Pluymaekers 和 Jos G.A.M. Lemmink。2020年。《对话代理中的类人沟通：文献综述与研究议程》。《服务管理期刊》31卷，2期（2020年），203–225页。doi:10.1108/JOSM-06-2019-0175[126] Vanshika Vats、Marzia Binta Nizam、Minghao Liu、Ziyuan Wang、Richard Ho、Mohnish Sai Prasad、Vincent Titterton、Sai Venkat Malreddy、Riya Ag- garwal、Yanwen Xu、Lei Ding、Jay Mehta、Nathan Grinnell、Li Liu、Sijia Zhong、Devanathan Nallur Gandamani、Xinyi Tang、Rohan Ghosalkar、Celeste Shen、Rachel Shen、Nafisa Hussain、Kesav Ravichandran 和 James Davis。2025年。《大型基础模型人机协作调查》。arXiv:2403.04931 [cs.AI] https://arxiv.org/abs/2403.04931[127] Shen Wang、Tianlong Xu、Hang Li、Chaoli Zhang、Joleen Liang、Jiliang Tang、Philip S. Yu 和 Qingsong Wen。2024年。《用于教育的大语言模型：调查与展望》。CoRR abs/2403.18105（2024年）。arXiv:2403.18105 doi:10. 48550/ARXIV.2403.18105[128] Wenxuan Wang、Shi Juluan、Zixuan Ling、Yuk-Kit Chan、Chaozheng Wang、Cheryl Lee、Youliang Yuan、Jen-tse Huang、Wenxiang Jiao 和 Michael R. Lyu。2025年。《学会提问：当大语言模型智能体遇上模糊指令》。载于《2025年自然语言处理实证方法会议论文集》，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet

Peng（编）。计算语言学协会，中国苏州，第21773–21784 10 18653 1 2025 1104 页。doi: . /v / .emnlp-main。Zijie J. Wang、Dongjin Choi、Shenyu Xu 和 Diyi Yang。2021年。《让人类参与自然语言处理：一项调查》。载于首届人机交互与自然语言处理桥接研讨会论文集，Su Lin Blodgett、Michael Madaio、Brendan O'Connor、Hanna Wallach 和 Qian Yang（编）。计算语言学协会，线上，第47–52页。https://aclanthology.org/2021.hcinlp-1.8/ [130] Hui Wei、Zihao Zhang、Shenghua He、Tian Xia、Shijia Pan 和 Fei Liu。2025年。《PlanGenLLMs：大语言模型规划能力的现代综述》。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，第19497–19521页。doi:10.18653/v1/2025.acl-long.958 [131] Bingbing Wen、Jihan Yao、Shangbin Feng、Chenjun Xu、Yulia Tsvetkov、Bill Howe 和 Lucy Lu Wang。2025年。《了解自身局限：大语言模型弃答研究综述》。《计算语言学协会汇刊》13 2025 529 556 10 1162 _ _00754 ( ), –。doi: . /tacl a [132] Andrew Westbrook 和 Todd S. Braver。2015.《认知努力：一种神经经济学方法》。认知、情感与行为神经科学15卷（2015年），395–415页。doi:10.3758/s13415-015-0334-y [133] Jules White、Quchen Fu、Sam Hays、Michael Sandborn、Carlos Olea、Henry Gilbert、Ashraf Elnashar、Jesse Spencer-Smith 和 Douglas C. Schmidt。2023年。《提示词模式目录：借助ChatGPT改进提示工程》。载于第*30*届程序模式语言大会论文集（美国伊利诺伊州蒙蒂塞洛）*(PLoP '23)*。The Hillside Group，美国，第5篇文章，共31页。[134] Genta Indra Winata、Hanyang Zhao、Anirban Das、Wenpin Tang、David D Yao、Shi-Xiong Zhang 和 Sambit Sahu。2025年。《语言、语音和视觉任务中的人类反馈偏好调优：一项综述》。人工智能研究杂志82卷（2025年），2595–2661页。[135] Benedikt Wisniewski、Klaus Zierer 和 John Hattie。2020年。《重探反馈的力量：教育反馈研究的荟萃分析》。《心理学前沿》10 2020 3087 10 3389 2019 03087 ( ), 。doi: . /fpsyg. . [136] Cheng-Kuang Wu、Zhi Rui Tam、Chao-Chung Wu、Chieh-Yen Lin、Hung-yi Lee 和 Yun-Nung Chen。2024年。《我需要帮助！评估大语言模型寻求用户支持的能力：文本到SQL生成案例研究》。载于*2024*年自然语言处理实证方法会议论文集，Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编）。计算语言学协会，美国佛罗里达州迈阿密，第2191–2199页。doi:10.18653/v1/ 2024.emnlp-main.131 [137] Shirley Wu、Michel Galley、Baolin Peng、Hao Cheng、Gavin Li、Yao Dou、Weixin Cai、James Zou、Jure Leskovec 和 Jianfeng Gao。2025年。《CollabLLM：从被动应答者到主动协作者》。载于第四十二届国际机器学习大会。https://openreview.net/forum?id=DmH4HHVb3y [138] Ziang Xiao、Sarah Mennicken、Bernd Huber、Adam Shonkoff 和 Jennifer Thom。2021年。《让我问你一个问题：语音助手如何引出明确的用户反馈？》*ACM*人机交互汇刊论文集5卷，CSCW2（2021年），1–24页。[139] Ziang Xiao、Michelle X Zhou、Wenxi Chen、Huahai Yang 和 Changyan Chi。2020年。《如果我没听错：运用积极倾听技能构建和评估访谈聊天机器人》。载于*2020*年*CHI*人因计算系统大会论文集。1–14页。[140] Ziang Xiao、Michelle X Zhou、Q Vera Liao、Gloria Mark、Changyan Chi、Wenxi Chen 和 Huahai Yang。2020年。《介绍一下你自己：使用人工智能驱动的聊天机器人开展包含开放式问题的对话式调查》。*ACM*计算机与人类交互汇刊（*TOCHI*）27卷，3期（2020年），1–37页。[141] Fengli Xu、Qianyue Hao、Chenyang Shao、Zefang Zong、Yu Li、Jingwei Wang、Yunke Zhang、Jingyi Wang、Xiaochong Lan、Jiahui Gong、Tianjian Ouyang、Fanjin Meng、Yuwei Yan、Qinglong Yang、Yiwen Song、Sijian Ren、Xinyuan Hu、Jie Feng、Chen Gao 和 Yong Li。2025年。《迈向大型推理模型：大语言模型强化推理综述》。*Patterns* 6卷，10期（2025 101370 10 1016 2025 101370 ），。doi: . /j.patter. . Diyi Yang、Sherry Tongshuang Wu 和 Marti A Hearst。2024年。《大语言模型时代的人机交互》。载于*2024*年北美计算语言学协会分会年会论文集：人类语言技术（第*5*卷：教程摘要）。34–38页。[143] J.D. Zamfirescu-Pereira、Richmond Y. Wong、Bjoern Hartmann 和 Qian Yang。2023年。《为什么约翰尼不会编写提示词：非人工智能专家如何尝试（并未能）设计大语言模型提示词》。载于*2023*年*CHI*人因计算系统大会论文集（德国汉堡）*(CHI '23)*。计算机协会，美国纽约州纽约市，第437篇文章，共21页。doi:10.1145/3544548. 3581388 [144] Chen Zhang、Xinyi Dai、Yaxiong Wu、Qu Yang、Yasheng Wang、Ruiming Tang 和 Yong Liu。2025年。《大语言模型多轮交互能力调查》。arXiv:2501.09959 [cs.CL] https://arxiv.org/abs/2501.09959 [145] Jiajie Zhang、Yushi Bai、Xin Lv、Wanjun Gu、Danqing Liu、Minhao Zou、Shulin Cao、Lei Hou、Yuxiao Dong、Ling Feng 和 Juanzi Li。2025年。《LongCite：使大语言模型能够在长上下文问答中生成细粒度引文》。载于计算语言学协会论文集：*ACL 2025*，Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，5098–5122。doi:10.18653/v1/ 2025.findings-acl.264 [146] Jingyu Zhang、Marc Marone、Tianjian Li、Benjamin Van Durme 和 Daniel Khashabi。2025年。《设计上可验证：让语言模型引用预训练数据》。载于*2025*年计算语言学协会美洲分会会议：人类语言技术论文集（第*1*卷：长篇论文），Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会，新墨西哥州阿尔伯克基，3748 3768 10 18653 1 2025 191 –。doi: . /v / .naacl-long。[147] Jiajie Zhang 和 Donald A Norman。1994。分布式认知任务中的表征。认知科学 18, 1 (1994), 87–122。[148] Michael JQ Zhang、W. Bradley Knox 和 Eunsol Choi。2025年。《对未来对话轮次建模，以教会大语言模型提出澄清性问题》。载于第十三届国际学习表征会议。https: //openreview.net/forum?id=cwuSAR7EKd [149] Shuning Zhang、Hui Wang 和 Xin Yi。2025年。《从能动性视角探索人机共创中的协作模式与策略：顶级人机交互文献的范围综述》。*ACM* 人机交互论文集 9, 7, Article CSCW413 (Oct. 2025), 43 pages。doi:10.1145/3757594 [150] Zhenyu Zhang、Runjin Chen、Shiwei Liu、Zhewei Yao、Olatunji Ruwase、Beidi Chen、Xiaoxia Wu 和 Zhangyang Wang。2025年。《发现于中间：语言模型如何通过即插即用位置编码更好地利用长上下文》。载于第*38*届神经信息处理系统国际会议论文集（加拿大不列颠哥伦比亚省温哥华）（*NIPS '24*）。Curran Associates Inc.，美国纽约州雷德胡克，Article 1943, 21 pages。[151] Zhehao Zhang、Ryan A. Rossi、Branislav Kveton、Yijia Shao、Diyi Yang、Hamed Zamani、Franck Dernoncourt、Joe Barrow、Tong Yu、Sungchul Kim、Ruiyi Zhang、Jiuxiang Gu、Tyler Derr、Hongjie Chen、Junda Wu、Xiang Chen、Zichao Wang、Subrata Mitra、Nedim Lipka、Nesreen K. Ahmed 和 Yu Wang。2025年。《大语言模型的个性化：综述》。机器学习研究汇刊（2025）。https://openreview.net/forum?id=tf6A9EYMo6 Survey Certification。[152] Zheng Zhang、Ryuichi Takanobu、Qi Zhu、MinLie Huang 和 XiaoYan Zhu。2020年。面向任务的对话系统的最新进展与挑战。中国科学：技术科学 63, 10 (2020), 2011–2027。[153] Yukun Zhao、Zhen Huang、Martin Seligman 和 Kaiping Peng。2024年。风险和亲社会行为线索会引发人工智能聊天机器人类似人类的响应模式。《科学报告》14 (2024), 7095。doi:10.1038/s41598-024-55949-y [154] Zheng Zhao、Clara Vania、Subhradeep Kayal、Naila Khan、Shay B Cohen 和 Emine Yilmaz。2025年。《PersonaLens：对话式人工智能助手个性化评估的基准测试》。载于计算语言学协会论文集：*ACL 2025*，Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编）。计算语言学协会，奥地利维也纳，18023 18055 10 18653 1 2025 927 –。doi: . /v / .findings-acl。[155] Henry Peng Zou、Wei-Chieh Huang、Yaozu Wu、Chunyu Miao、Dongyuan Li、Aiwei Liu、Yue Zhou、Yankai Chen、Weizhi Zhang、Yangning Li、Liancheng Fang、Renhe Jiang 和 Philip S. Yu。2025年。《呼吁协作智能：为何人类-智能体系统应先于人工智能自主性》。arXiv:2506.09420 [cs.AI] https://arxiv.org/abs/2506.09420

A 附录

A.1 研究1：参与者表

见表1

A.2 详细任务描述

A.2.1 任务*1*：*25*万美元 *STEM*奖学金申请。**参与者指令：** 你正在申请一项25万美元的 STEM奖学金，该奖学金面向那些怀揣大胆而富有创意的愿景、希望以变革性的方式在各自领域运用人工智能的人士。

要求（500–1000字）：

- **你是谁：**分享你的背景、价值观以及个人动力。
- **你为何具备独特资质：**突出你的个人优势、经历与视角。
- **你的人工智能愿景：**你将开展怎样的创新人工智能项目或倡议?
- **奖学金的影响：**说明它将如何推进你的个人历程，并创造更广泛的积极社会影响。

重要：

- 必须为 500–1000 个单词（超出此范围的申请将被取消资格）。
- 力求创作一篇真实、富有想象力且主题鲜明的文章，将清晰度与个人叙事融为一体。
- 你需要在所撰写文章的第一分钟进行演讲。

A.2.2 任务 *2*：*50,000* 美元社区改善补助金。**参与者须知：**你将通过 50,000 美元的社区改善补助金，为所在社区提出一项本地解决方案。

要求（500–1000 个单词）：

- **社区与问题陈述：**描述你的社区及其面临的一个紧迫问题。为什么这个问题亟待解决?
- **拟议方案：** 你提议采取什么具体行动或开展什么项目（例如公共空间、安全改善或社区福利）？
- **你的角色与资历：**你与社区改善的个人联系、相关经验或过往参与经历。
- **基金用途与可持续性：** 资金将如何使用？资助结束后，该项目将如何继续?
- **社区影响：** 谁将从中受益？这将如何提升社区福祉?

重要提示：

- 必须为 500–1000 字，少于或超过此范围将被直接取消资格。
- 你需要为你已经起草的文章的第一分钟准备一段演讲。

A.2.3 任务*3*：*1*万美元竞赛——*5*分钟个人陈述。**参与者须知：**你将参加一场竞赛，争取赢得10000美元。参赛者需要展现出迎接未知挑战的准备，并有能力成为各自领域的领导者。该基金旨在培养能够推动变革、促进经济发展的领导者。

要求：

- 5分钟演讲（总计~500–1000字）。
- 请同时提交：– 文字稿。– 你朗读其中一部分的 1 分钟音频录音。

在这段陈述中，请着重表达：

- 你从事什么工作，以及这项工作为何重要。- 你能带来怎样的独特视角或专业知识。- 明确的请求或下一步行动（例如，与你合作或支持你的工作）。
- 结合过往实例，说明你为何具备成为所在领域领导者的条件。
- 如果当选，你将带来的影响，以及你为什么想成为领导者。

重要事项：

- 时间/字数限制严格——5分钟（~500–1000字）。
- 你需要在演讲的第一分钟进行陈述。

参与者	年龄范围	性别	人工智能使用情况	教育背景
P1	18–24岁	Male	编程，家庭作业	高中毕业生
P2	25–34岁	Male	编程、信息寻求、研究	博士学位
P3	18–24岁	Male	决策、写作、规划	学士学位
P4	25–34岁	Male	图像生成、编程	博士学位
P5	25–34岁	女性	文章写作、编程	学士学位
P6	25–34岁	女性	自动化、编程	学士学位
P7	18–24岁	女性	烹饪、自动化、研究	高中毕业生
P8	25–34岁	Male	学习、图像生成	博士学位
P9	18–24岁	Male	图像生成、信息寻求	学士学位
P10	18–24岁	Male	写作	学士学位
P11	18–24岁	非二元性别	学习、编程、决策	高中毕业生
P12	35–44	女性	消息传递、信息寻求、烹饪、决策	学士学位
P13	18–24岁	女性	信息寻求，决策	高中毕业生
P14	75–84岁	Male	信息寻求，电子邮件，决策	高中毕业生
P15	45–54岁	Male	信息寻求，决策	学士学位
P16	45–54岁	女性	信息寻求、决策	学士学位

表1：参与者人口统计特征及报告的人工智能用例专业知识。（P12-P16为缺乏经验的用户）