理解语音助手故障后用户信任的混合方法研究

Amanda Baughan*
baughan@cs.washington.edu
华盛顿大学，美国华盛顿州西雅图

Allison Mercurio
amercurio@google.com
Google Research 美国加利福尼亚州芒廷维尤

Ariel Liu
arielliu@google.com
Google Research 美国加利福尼亚州芒廷维尤

Xuezhi Wang
xuezhiw@google.com
Google Research 美国纽约州纽约市

Jilin Chen
jilinc@google.com
Google Research 美国加利福尼亚州芒廷维尤

Xiao Ma
xmaa@google.com
Google Research 美国纽约州纽约市

arXiv:2303.00164v2 [cs.HC] 3 Mar 2023

摘要

近年来，大语言模型在自然语言理解方面取得了巨大进步，但语音助手仍常常无法满足用户期望。本研究采用混合方法分析，探讨语音助手故障如何影响用户对语音助手的信任。为展示用户经历过的此类故障，我们构建了一个包含 199 起语音助手故障的众包数据集，并将其归纳为 12 种故障来源。基于访谈和调查数据，我们发现，某些故障（例如因过度捕获用户输入而导致的故障）比其他故障更容易损害用户信任。我们还考察了故障如何影响用户今后依赖语音助手完成任务的意愿。用户往往会在某项任务导致故障后，短暂停止使用语音助手完成该任务，之后再恢复类似的使用。我们证明了低风险任务（例如播放音乐）对于在故障后建立信任的重要性。

CCS 概念

• **以人为中心的计算** → **人机交互实证研究**。

关键词

语音助手、信任、问卷调查、访谈、数据集

ACM 引用格式：

Amanda Baughan、Allison Mercurio、Ariel Liu、Xuezhi Wang、Jilin Chen 和 Xiao Ma。2023年。混合方法研究：理解语音助手故障后的用户信任。载于《*2023*年计算系统中的人为因素大会论文集》（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约，16页。https://doi.org/10. 1145/3544548.3581152

1 引言

语音助手受到了业界和学术界的广泛关注，尤其是考虑到近期自然语言处理（NLP）领域的进展。在过去五年中，自然语言处理取得了长足进步，在标准数据集上的准确性大幅提升 [8, 18, 34, 56, 58, 60, 63], 自然语言处理模型在某些任务上的准确性已达到 99% 的最先进水平 [8, 34]。这让许多从业者和研究人员都开始设想，在不久的将来，语音助手可以用于越来越复杂的场景，包括支持医疗保健任务 [41, 55], 提供心理健康建议 [52, 66],以及辅助高风险决策 [17]。

然而，尽管自然语言处理模型的准确性不断提高、应用范围不断扩大，有证据表明，用户仍然不愿意使用语音助手，并且对其缺乏信任 [12, 29]。在美国，语音助手在家庭中十分普遍，据估计有72%的美国人使用过语音助手 [27]。然而，人们主要将其用于播放音乐、设置计时器和制作购物清单等基本任务 [12, 29, 35]。这是因为当语音助手出现故障时，例如错误回答问题，就会削弱用户信任 [27, 35]。用户信任对于各种技术的用户采用至关重要 [3], , 而在这种情况下，较低的用户信任会导致用户不愿尝试语音助手的新功能。

随着语音助手越来越依赖大型语言模型 [24, 47], , 我们认为，这些模型的高准确性与用户不愿将语音助手用于复杂任务之间的差距，或许可以用用户和自然语言处理从业者评估模型成功与否的方式不同来解释。标准的自然语言处理模型通常使用大量连贯的文本问答数据集 [48, 49] 或配对书面对话 [68]进行评估。与此同时，在实际使用中，用户的语音可能包含言语不流畅现象，例如重新开始说话和填充词；也可能涉及训练数据未涵盖的问题，或受到导致语音识别错误的背景噪声干扰。在问答任务中，自然语言处理模型的评估依据是：在训练数据集的一个子集上，模型能准确回答多少问题 [48, 49]。不难想见，人们与语音助手互动的方式多种多样，其中有许多超出训练数据的范围，从而可能造成使用上的摩擦。在用户看来，这些不准确的响应，也就是语音助手故障，可能令人沮丧。例如，只有百分之五的用户表示，他们使用语音搜索时从未感到沮丧 [13]。

我们认为，自然语言处理模型的评估方式与用户遭遇并感知故障的方式之间存在差距，这阻碍了语音助手技术进步的实际应用。因此，我们提出以下问题：用户目前会遇到哪些类型的语音助手故障?

*本研究作为 Google Research 实习项目的一部分开展。

特此授予许可：可免费出于个人或课堂使用目的，对本作品的部分或全部内容制作数字或纸质副本，但前提是不得为牟利或商业利益而制作或分发副本，并且副本必须在第一页附有本声明及完整引文。本作品中第三方组件的版权必须得到尊重。如有其他用途，请联系版权所有者/作者。*CHI '23*，*2023*年*4*月*23*日至*28*日，德国汉堡© 2023年版权归版权所有者/作者所有。ACM ISBN 978-1-4503-9421-5/23/04。https://doi.org/10.1145/3544548.3581152

这些故障又会如何影响用户信任？以人为本地理解自然语言处理故障的类型及其对用户信任的影响，将有助于技术人员确定关键故障的优先级并加以解决，从而推动语音助手在更多用例中的长期采用。

此外，尽管已有研究开始对用户与自然语言处理智能体之间的通信故障进行分类 [28, 46]，但鲜有研究探讨用户如何感知这些故障，以及这些故障如何进一步影响用户对语音助手的信任和使用。我们借鉴并拓展了以往的研究，做出以下贡献：

- **C1：** 在现有自然语言处理失败分类体系的基础上进行迭代，我们通过众包构建了一个数据集，其中包含用户在12种不同失败来源中经历的199次故障。
- **C2：** 通过定性和定量评估，研究这些不同故障如何影响用户信任，具体考察能力、善意和正直这几个维度。
- **C3：** 通过定性和定量分析，探讨信任如何影响未来的预期使用。

为此，我们开展了一项以人为中心、采用混合方法的语音助手故障研究。首先，我们对12名语音助手用户进行了访谈，以了解他们经历过哪些类型的故障，以及这些故障如何影响他们对语音助手的信任和后续使用。同时，我们在亚马逊土耳其机器人上向语音助手用户众包收集了一个故障数据集。最后，我们开展了一项问卷调查，以量化不同类型的故障如何影响用户对语音助手的信任，以及他们未来在各种任务中使用语音助手的意愿。

我们发现，不同类型的语音助手故障对信任的影响各不相同。我们的访谈和问卷调查显示，参与者更能谅解由误触发或自身请求存在歧义所导致的故障。在误触发的情况下，语音助手会因错误听到并识别激活短语而被激活，尽管用户并未说出该短语。用户更容易谅解这类故障，因为它并未妨碍他们实现目标。由歧义导致的故障发生在用户的请求存在多种合理解释，而语音助手的响应虽然准确回答了问题，却与用户的意图不一致时。用户往往会将这些故障归咎于自己。然而，过度捕获导致的故障会更严重地降低用户的信任；当语音助手在没有任何额外输入的情况下继续聆听时，用户会认为使用它是在浪费时间。

我们还发现，许多情况下，用户在语音助手发生故障后，会暂时停止用它执行某项特定任务；待信任重建后，他们便会再次使用。用户通常会通过让语音助手执行自己认为简单的任务（例如播放音乐）来重建信任，或者在不同的使用场景中让语音助手执行同一类任务。除了这些发现，我们还发布了一个包含199例语音助手故障的数据集，其中记录了用户输入、语音助手响应以及失败上下文，供研究人员用于未来研究用户对语音助手故障的反应。随着语音助手在各个行业中持续承担日益复杂且高风险的任务 [17, 41, 51, 55, 66]，我们希望这项研究能帮助技术人员理解、确定优先级并解决自然语言交互失败，从而提升并维持用户对语音助手的信任。

2 相关研究

许多领域的既有研究都探讨了用户与语音助手之间的交互，包括人机交互、以人为中心的人工智能、人机交互、科学技术研究（STS）、计算机媒介传播（CMC）和社会心理学。此外，自然语言处理（NLP）领域的一些研究，尤其是自然语言处理鲁棒性方面的研究，也关注了语音助手中的技术故障，并提出了相应的技术解决方案。本文将跨领域回顾与用户交互过程中语音助手故障相关的研究。文献综述安排如下：1）关于语音助手中的用户期望和信任的研究；2）人机交互（HCI）领域用于理解语音助手故障及其缓解策略的方法；3）自然语言处理（NLP）领域针对语音助手故障的研究，包括非流利现象和鲁棒性。

2.1 语音助手中的用户期望与信任

研究人员长期以来一直试图理解人们如何与自动化智能体互动，尤其是将这些体验与人际交流进行比较和对照。与他人交谈时，对话大体可分为功能性对话（也称为事务性对话或任务型对话）和社交对话（互动型对话），许多对话兼具两者 [11]。功能性对话旨在达成某个目标，参与其中的人通常各自承担有助于实现该目标的角色。相比之下，社交对话旨在与其中一位参与者建立、加强或维持积极关系。这类社交对话有助于建立信任、融洽关系和共同基础 [11]。

人们通常期望与语音助手进行功能性对话 [11]。缺少社交对话可能会降低用户对语音助手建立信任的能力。事实上，过往研究表明，具身对话代理在进行闲聊时更容易获得用户的信任 [5]，不过这一点会因用户人格类型和代理的具身程度而异 [6]。目前，人们表示自己不会使用语音助手完成广泛的任务，尽管从技术上讲，语音助手有能力完成这些任务 [27]。先前研究阐明了信任对于持续使用语音助手的重要性 [31, 35]，因为信任是用户采用语音助手的关键 [33, 45]，也会影响用户是否愿意扩大语音助手的任务范围 [27]。随着研究人员不断构想并开发语音助手的新能力，包括支持医疗保健任务 [41, 55]、提供心理健康建议 [52, 66]，以及辅助其他高风险决策 [17]，支持用户与语音助手之间建立信任尤为重要。

这便引出了一个问题：用户与语音助手之间的信任是如何建立的？随着自动化系统的广泛应用，机器信任正成为一个越来越重要的话题 [64]。具体而言，信任可被视为对系统有信心，以及愿意依据系统提供的建议采取行动的结合体 [37, 54]。此前的研究者已从以下角度考察了人们对机器的信任：

人们对机器按预期运行的能力 的信心、对其善意 （心怀好意）的判断，以及对其遵守伦理标准的正直 的信任 [36] 。总体而言，过往研究评估了准确性和错误等多种因素如何影响人们对算法的信任 [19, 20, 67]。针对语音助手，Nasirian 等人 [45] 和 Lee 等人 [33] 研究了质量如何影响人们对语音助手的信任及其采用情况，并发现信息质量和系统质量不会影响用户对语音助手的信任，但交互质量会产生影响。交互质量的测量依据 Ekinci 和 Dawes [22], 的一项研究，该研究通过李克特量表收集了受访者对语音助手能力、态度、服务态度和响应能力的评价。此外，根据用户定制语音助手的个性可以提升信任 [7], , 而性别不会影响用户对语音助手的信任 [59]。总体而言，既有研究表明，交互质量和社交对话对于建立用户与语音助手之间的信任至关重要，而这又会影响用户继续使用语音助手并扩大其任务范围的意愿。

2.2 人机交互领域应对语音助手故障的方法

然而，信任破裂偶尔也会出乎意料地发生，因为并非所有互动都能像人们预期的那样顺利。以往研究探讨了影响用户与语音助手互动及持续使用的各种问题，并表明，当用户对语音助手的期待超出其能力时，就会出现语音助手故障和用户挫折感 [31, 35]。

这不禁让人追问：以往研究是如何界定语音助手故障的? 一些研究在研究中采用了特定场景。例如，Lahoual 和 Frejus [31] 在家庭和驾驶场景中进行了评估。他们发现，语音识别效果不佳、对命令的理解有限以及连接性问题都会导致故障。Cuadra 等人 [16] 在特定任务中设置故障，例如尝试将路线指引发送至错误地点、给错误的人发送短信、播放错误类型的音乐，或添加详情有误的提醒 [16]。Mahmood 等人 [39] 模拟了网上购物场景，其中带有语音功能的人工智能助手会因使用与用户所要求商品同音异义的词而出错。例如，含义模糊的单词“bow”可以指蝴蝶结、射箭用的弓，或礼品包装用的缎带。Salem 等人 [53] 让参与者控制机器人的移动；在故障条件下，机器人会胡乱移动，无法正确响应用户输入。Candello 等人 [9] 将故障定义为有人提出的问题无法被理解，或超出语音助手的知识范围；在这种情况下，语音助手会转而提出一个无关的问题，使对话偏离原话题。

其他研究旨在借鉴人类之间交流的理论框架，对语音助手故障进行广泛分类 [10, 28, 46]。我们参考了 Herbert Clark 的人类交流共同基础模型，该模型依靠四个不同层面实现相互理解：通道、信号、意图和对话 [10]。Paek and Horvitz [46], 对此模型进行了扩展，将这四个层面应用于人机交互及其故障点。通道层面的错误包括人工智能未能注意到用户试图发起交流；信号层面的错误包括未能捕获用户输入（例如，由于转录错误）；意图层面的错误包括在理解转录输入的语义含义时出现错误；而对话层面的错误则发生在用户向人工智能请求执行未知操作时（例如，要求天气应用安排某项事务）。Hong 等人 [28] 以此模型为基础，专门将情境限定为自然语言处理失败，而非整个人工智能领域。根据对自然语言处理从业者的访谈，他们将类别重新命名为注意（通道）、感知（信号）、理解（意图）和响应（对话）。Hong 等人 [28] 聚焦于非常常见的故障，或虽罕见但代价高昂的故障，以涵盖用户与基于自然语言处理的系统交互时遇到的最重要且最频繁的故障。在本研究中，我们以现有的自然语言处理失败分类体系 [28],为基础，将应用场景缩小到仅限语音助手故障，并评估不同故障对用户信任及其未来使用意愿的影响。

目前，关于语音助手故障对用户信任影响的系统性评估还很少。Salem 等人 [53] 发现，如果机器人出现性能故障，这并不会影响参与者是否遵从其请求的决定，但会显著影响他们对机器人可靠性和可信度的感知。Mahmood等人 [39] 发现，与推卸责任的语音助手相比，会为错误承担责任并道歉的语音助手被认为更智能、更讨人喜欢，也更能有效地从故障中恢复。

有时，在发生故障后，用户会尝试改述、简化或夸张地清晰发音其命令，以继续使用设备 [31, 35, 43, 61]。如果用户反复无法通过语音助手修复故障，他们的信任就会减弱，并将命令范围缩小到失败风险较低的简单任务 [31, 35]。Lahoual 和 Frejus [31] 发现，在某些情况下，语音助手故障会逐渐侵蚀信任，严重到使用户完全放弃使用语音助手。然而，并非所有故障都需要自我修复。Cuadra 等人 [16] 的一项研究发现，当语音助手出错时，语音助手的自我修复会显著提升人们对智能语音助手的评价；但如果不需要进行纠正，则可能产生相反的影响。因此，理解哪些类型的故障最会削弱信任，也有助于我们判断何时应启动故障缓解策略。

2.3 面向语音助手故障的自然语言处理方法

自然语言处理领域的研究者也从略有不同的角度考察了语音助手故障，重点关注语音助手所依赖的不同自然语言处理组件的鲁棒性，例如用于自然语言推理 [44], [26, 问答和语音识别等任务的模型。自然语言处理鲁棒性可以定义为：当在一个与模型训练所用数据集分布不同的新数据集上进行测试时，模型性能会发生怎样的变化 [62]。在实践中，用户与语音助手的现实世界交互可能与开发过程中使用的数据有所不同，这模拟了自然语言处理鲁棒性研究中的数据分布偏移。 [42], [32]

研究表明，此类数据分布偏移会导致模型失效。在问答领域，最先进模型在基于 Wikipedia [49] 收集的标准基准数据集上进行阅读理解时，表现几乎达到人类水平。然而，Miller等人[42] 发现，当问答模型在不同主题领域（例如

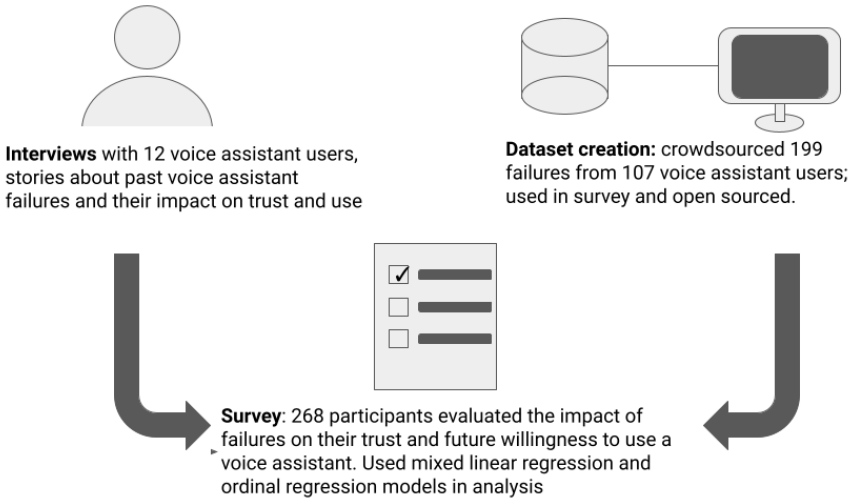

图 1：为分析语音助手故障对用户信任的影响，我们采用了混合方法研究，包括访谈和问卷调查。作为问卷调查材料的一部分，我们通过众包从107名语音助手用户那里收集了199个故障案例，并将该数据集作为我们的贡献之一。

《纽约时报》的文章、Reddit 帖子和亚马逊产品评论）上接受评估时，模型性能会下降。噪声输入也会损害模型性能。Lee 等人[32] 表明，语音识别错误会对机器阅读理解产生灾难性影响。Gupta等人 [26] 创建了问答数据集 Disflu-QA，其中人类引入了语境不流畅现象，这同样会导致模型性能下降。

尽管这些研究并未直接关注语音助手故障，但在用户与语音助手交互时，话题领域变化、语音识别错误和言语不流畅都十分常见。这些相似之处促使我们将自然语言处理鲁棒性文献与人机交互领域对系统故障的研究视角进行类比。通过理解不同类型的故障如何影响用户对语音助手的整体信任，我们便能尝试找出导致最严重、进而削弱信任的故障的根源所在的自然语言处理组件 [30]。随后，我们可以运用技术解决方案，提高系统中最关键部分的鲁棒性，从而以最高效率提升用户信任和长期使用参与度。

3 方法概述

既然我们已经确立了理解语音助手故障如何影响用户信任的重要性，接下来便开展一项混合方法研究。首先，为了进行定量评估，我们回顾了人机交互和自然语言处理领域的现有数据集，以寻找可用于问卷调查的故障案例。最终，现有数据集无法满足我们的需求。因此，我们通过众包方式收集了语音助手用户遭遇故障的数据集，并将其作为本研究贡献的一部分开放源代码。同时，我们访谈了 12 位语音助手用户，以了解他们经历过哪些类型的故障，以及这些故障如何影响他们对语音助手的信任和后续使用。

这些访谈旨在广泛了解用户对语音助手故障的想法、感受和行为，并为定量问卷调查的设计提供参考。最后，我们开展了一项问卷调查，以量化不同类型的故障如何影响用户对其语音助手的信任感知，以及他们未来使用语音助手完成各种任务的意愿。为报告这些发现，我们首先介绍众包故障数据集的收集过程，以及我们如何从中选取一部分用于问卷调查。接下来，我们介绍访谈和问卷调查，先描述数据收集与分析，再同时呈现研究结果。

4 语音助手故障数据集的众包构建

我们研究的首要目标是确定用户在使用语音助手时会遇到哪些类型的故障。我们首先评估了现有数据集在故障类型覆盖范围和广度方面是否适用。我们认为这些数据集不足以满足研究需求，因此着手通过众包构建故障数据集，并借鉴 Hong 等人的分类体系来指导数据收集。最后，我们对该数据集进行了清理并将其开源，作为本研究的一项贡献。

4.1 现有人机交互与自然语言处理数据集综述

我们首先考察了自然语言处理领域的基准数据集，这些数据集包含大量问答 [48–50],或对话 [25, 57, 68]。我们发现，现有的自然语言处理数据集无法涵盖各种可能的对话失败案例，因为它们侧重于用于训练的正确数据。此外，它们关注的是特定任务的表现，例如回答问题或对话生成，其范围比语音助手的多样化使用场景要窄得多。由于训练数据依赖于准确完成任务，

故障类型	序贯编码指南	故障来源	故障情境
注意	缺乏视觉或音频证据 语音助手已开始 聆听或视频或音频 语音助手的证据 已开始聆听 缺少提示	**触发遗漏** **误触发** **延迟触发**	用户说了某些话来触发语音助手，但 它没有响应。 用户没有说任何话来触发语音助手， 但它还是会被激活。 与系统延迟类似，用户说了些什么来 触发语音助手，但它回复得太迟，因而无法 派上用场。
感知	视觉或音频证据表明 语音被系统错误地 捕捉。例如， 被语音助手打断， 或看到它继续 在用户说完语音后继续聆听 完整内容，明显听错了一个 词，或有背景噪声的迹象 和串音。	**噪声信道** **过度捕获** **截断** **转录**	由于背景噪声，用户输入被错误捕获 。 语音助手捕获了超出预期的输入 可能是开始捕获输入过早，或结束捕获 过晚，并对与 用户的请求。 系统未能完整捕捉用户的语音，可能是因为 开始捕捉输入过晚，或结束得过早。 系统产生转录错误，通常表现为 发音相似的词语。
理解	怀疑音频已被 正确捕获，但未映射 到正确的操作。例如， 收到表明 无法完成某项操作的响应，而该操作 过去曾成功执行，或者 收到的响应 看似合理，但并不符合 输入意图的响应。	**歧义** **误解** **无法理解**	用户的 in- 可能有几种解释。 图可能有多种解释，而系统给出的响应看起来合理 准确，但不符合用户的意图。 系统将用户输入映射为错误动作， 或许对用户的意图做出了部分正确的推断， 但并不完全准确。 系统无法将用户输入映射到任何已知的 操作或响应。
响应	最后，假设输入 已被正确捕获并 理解，那么响应是否 生成的内容不正确、不清楚或没有 给出，或者存在其他错误?	**动作执行：无 操作** **动作执行：不 正确的操作**	如果系统听完了完整请求，但随后就关闭 ，既没有给出任何答复，也没有采取任何操作。 系统提供的信息有误。

表1：定性编码手册及所收集的各类故障描述。我们依次检查每个故障的故障类型：首先判断其是否属于注意力故障，再逐一检查其他类型，直到找到符合的类型。确定故障类型后，我们再评估适用的故障来源。

这些数据集不包含故障。虽然测试这些模型会产生少量错误（约 10%），但故障类型仅限于响应和理解类别，因为注意和感知故障被排除在训练此类模型的情境之外。这限制了这些数据集对于我们研究使用过程中发生的语音助手故障及其对用户信任的影响的用途。

除了这些基准数据集，我们还考察了纳入口语表达模式的数据集，例如 Spoken SQuAD 数据集 [32] 和 Disflu-QA 数据集 [26],，以及人机交互数据集，例如 ACE 数据集 [1],、Niki and Julie 语料库 [2], 和一个语音助手故障视频数据集 [15]。在这些情况下，我们发现这些数据集仍仅限于理解和响应层面的故障 [26, 32]，或者故障上下文十分具体，未必能涵盖用户可能遇到的各种故障 [1, 2]。Cuadra 等人 [15]的视频数据集最符合我们的需求，但我们仍认为，实验室内问答这一应用场景对于我们的研究目的而言过于狭窄。因此，我们决定从用户处众包收集语音助手故障数据集，并在开展有关用户信任的定量调查时使用这些故障。

4.2 数据集收集

4.2.1 流程。 我们要求众包工作者提交他们在使用语音助手时经历过的三次故障。我们从一个包含 12 种故障的分类体系中选取三种特定类型，随机分配并展示给所有工作者，且每种类型展示的次数相同。我们用于询问特定故障类型的分类体系改编自 Hong 等人此前的研究 [28],，将故障归为注意、感知、理解和响应问题，如表1所示。每个问题首先根据我们的分类体系中的定义，询问用户是否能回忆起语音助手发生故障的情形。例如，为了了解触发遗漏故障，我们询问：“你是否曾想要唤醒语音助手，但它没有响应？”如果有，我们会请这些工作者填写：1. 他们对语音助手说了什么；2. 语音助手如何响应；3. 故障上下文，包括周围环境中发生了什么；以及 4. 故障发生的频率，从 1（使用时很少发生）到 5（每次使用时都会发生）。除频率问题采用多项选择形式外，其余问题均以文本输入框的形式呈现。此外，我们还请众包工作者酌情分享一个他们尚未有机会分享的额外故障

。这样做是为了捕捉不属于已向他们展示的三种类别中任何一种的故障，随后我们根据分类体系对这些故障进行分类。

收到这些故障案例后，我们将提交示例中的语音助手类型匿名化，并统一将激活词替换为“语音助手”。随后，我们修正了语法和拼写错误，以提升清晰度。对于偏离任务要求、含义不清或与其他提交内容完全重复的故障案例，我们也予以删除。最后，我们注意到，用户提交故障案例时选择的某些类别有误，因此根据我们制定的编码手册重新归类，具体如表1所示。两位评审者反复对提交的101个故障案例进行编码，最终一次编码会议的评审者间一致性达到70%。随后，一位研究人员重新对整个数据集进行了编码。最终的数据集共包含107名不同众包工作者提交的199个故障案例，分属12个类别。

4.2.2 众包工作者特征。 我们通过亚马逊土耳其机器人招募众包工作者。共有107名众包工作者为我们的数据集作出贡献。我们要求工作者满足以下条件：HIT 批准率超过98%、获批HIT超过1000个、拥有 AMT 大师资格、来自美国、年满18岁，并且至少每周使用一次语音助手。人数最多的年龄范围是35至44岁（$n = 46$），其次是25至34岁（$n = 32$）和45至54岁（$n = 16$）；其余人群分别为55至64岁（$n = 8$）、18至24岁（$n = 1$），另有1人选择不回答。58名众包工作者为男性，44名为女性，1人选择不回答，另有1人自我认同同时为男性和女性。他们使用亚马逊 Alexa（$n = 59$）、GoogleAssistant（$n = 62$）和苹果 Siri（$n = 40$）等商用语音助手，许多人会组合使用其中几种（$n = 47$）。91名众包工作者是英语母语者，13名不是。人数最多的族裔为白人（$n = 58$），另有39人认同为亚裔。3名众包工作者未提供任何人口统计信息。任务平均耗时15至20分钟，参与者获得5.00美元的报酬。

4.2.3 最终数据集。 总的来说，我们最终确定的数据集包含来自107名用户的199个故障案例，涵盖根据Hong 等人提出并在表1中更新的分类体系划分的12种不同故障类型。我们最常收到的故障是由误解（$n = 38$）、触发遗漏（$n = 25$）和噪声信道（$n = 22$）造成的。用户提交最少的故障类型是截断（$n = 7$）、过度捕获（$n = 7$）和延迟触发（$n = 8$）。大多数众包工作者提交的故障表示，这些故障“在我使用时很少发生”（$n = 87$）或“在我使用时有时发生”（$n = 84$）。表2列出了12个类别中的故障示例。

平均而言，发生频率最高的故障是无法理解（$m = 2.15$，在我使用时有时发生，$sd = 0.67$）和动作执行：错误（$m = 2.00$，在我使用时有时发生，$sd = 0.88$）。其余故障来源的平均报告频率介于1.0（在我使用时很少发生）和2.0（在我使用时有时发生）之间。发生频率最低的故障是由延迟触发（$m = 1.25$，$sd = 0.46$）和歧义（$m = 1.39$，$sd = 0.78$）造成的。

随后，我们从数据集中选取了 60 个故障纳入问卷调查，以量化不同故障对用户信任的影响。下一节将对此进行更详细的阐述。该数据集已开放源代码[1]，供研究人员未来用于回答与语音助手故障相关的研究问题。

5 访谈和调查方法

收集并整理语音助手故障数据集后，我们便准备好回答研究问题：语音助手故障会如何影响用户信任？为此，我们首先对 12 人进行了探索性访谈，了解他们在经历语音助手故障后的想法、感受和行为。随后，我们利用这些发现以及数据集中收集的故障来设计并开展问卷调查。问卷调查量化了各种语音助手故障对用户信任的影响，衡量指标包括用户对语音助手能力、善意和正直的感知，以及他们未来任务中使用语音助手的意愿。在此，我们介绍访谈和问卷调查的方法，并在此之后综合呈现两项研究的结果。

5.1 访谈方法

5.1.1 访谈流程。访谈首先询问参与者为何开始使用语音助手，以及他们通常会向语音助手提出哪些类型的问题。我们还询问他们通常会在什么时间、什么地点使用语音助手，以了解他们对语音助手的总体体验。

完成这些准备后，我们请参与者尽可能详细地讲述一次使用语音助手时遇到错误的经历。我们询问他们当时想做什么、为什么想做、是否有其他人在场，以及周围环境中是否还发生了其他事情。我们进一步询问故障发生后用户的感受，以及他们对语音助手理解自己并提供准确信息的能力有何看法。我们询问参与者当时如何应对故障。最后，我们询问他们在故障发生后如何使用语音助手，包括对它的信任程度，以及是否改变了某些行为以减少未来发生故障的可能性。所有访谈均以远程方式进行。

5.1.2 访谈参与者。在招募过程中，我们请参与者提交人口统计信息、使用语音助手的频率，以及使用的设备类型。此外，我们还要求参与者用简短文字（1-3 句话）概述一次使用语音助手时遇到故障的经历。我们根据人口统计分布，以及他们对故障经历描述的详细程度来选取参与者。

我们的 12 位参与者均居住在美国。他们每周至少使用语音助手 1-3 次（$n = 2$）；大多数人表示每天都会使用语音助手（$n = 8$），其余参与者（$n = 2$）每周使用 4-6 次。大多数参与者在移动设备上使用语音助手（$n = 11$），其中五位参与者也使用智能家居设备上的语音助手。一位参与者只使用智能家居设备上的语音助手。参与者表示，他们使用常见的商用语音助手，例如亚马逊 Alexa（$n = 2$）、Google 助理（$n = 7$）和苹果 Siri（$n = 8$）。参与者的年龄介于 18 至 50 岁之间，其中人数最多的一组（$n = 5$）年龄范围为 18-23 岁。3 位参与者年龄为 41-50 岁，2 位为 31-40 岁，2 位为 24-30 岁。我们的 6 位参与者

[1]https://www.kaggle.com/datasets/googleai/voice-assistant-failures

故障来源	情境	用户说的话	语音助手的反应
触发遗漏	我让她设置一个十分钟的计时器，当时我独自一人，身边没有其他人。	语音助手，设置一个 10 分钟。	[没有响应。]
误触发	我当时正在和团队进行电话会议，并且正在给同事 Sherry 打电话。语音助手却误开启了。	[与同事交谈时] “你能共享屏幕吗？”	[回应与同事的对话——同事。] “请稍等，我来帮您处理。”
延迟触发	这发生在我开车的时候。	语音助手，给我显示路线到国家公园。	[语音助手花了太长时间才作出回应，以至于它还没来得及回应，你就又一次询问路线。]
噪声信道	我的孩子们在背景中玩耍狗也在叫，我不得不提高我的语音，并尝试多次让我的手机听见，尽管它就在我面前几英寸处。	语音助手，什么是 the 天气?	[它没有意识到我的请求已经结束了]，还一直转圈。]“抱歉，我没太听明白你的意思。”
过度捕获	我当时在告诉它关灯。那里只有我一个人。	语音助手，关灯。	[它一直聆听了很久，久到你转身自己把灯关掉。]
截断	我让语音助手计算一道数学题，但它打断了我。	语音助手，你能帮我计算乘积吗？54、39、33 和 22?	“54乘以39乘以33等于69,498。”
转录	我询问了我所在城市的天气状况。当时除了我，没有其他人在场。	语音助手，今天的温度是多少?	“佐治亚州玛丽埃塔今天的温度是 65 华氏度。”
歧义	我独自在家看终极格斗冠军赛时，问了一架战斗机有多老。	语音助手，Johnny Walker?	“Johnny Walker 于 1865 年成立。” [它指的是威士忌公司，而不是战斗机。]
误解	我让它播放《万圣节》的主题曲。我当时正和母亲坐在一起。	语音助手，播放主题曲，《万圣节》电影的主题曲。	[它播放了一段恐怖音效原声带，而不是那首歌。]
无法理解	我当时正试图运行一个例行程序，叫醒我的孩子们。	语音助手，叫醒双胞胎。	“抱歉，我不知道。” [不过，我已经设置了一个“叫醒双胞胎”的例行程序，之前一直运行正常过去。]
操作执行：无操作	我问一部电影什么时候会在影-院上映，它却一直转着灯光，来来回回地over.	语音助手，《尚-气》什么时候上映?	[停顿了很长时间，然后熄灭指示灯，没有回应。]
操作执行：错误动作	我独自在家里的客厅。当时我想查一下 Taco Bell 营业到几点。	语音助手，Glenwood 上的 Taco Bell 几点关门?	“Taco Bell 营业至凌晨 1 点。” [开车前往 Taco Bell，我这才发现它晚上11:30就关门了。]

表2：用户提交的语音助手故障汇总表，包括失败上下文、用户说的话以及语音助手说的话。

认同为女性，5 位参与者认同为男性，1 位参与者认同为非二元性别。3 位参与者认同为亚裔，3 位认同为白人，3 位认同为黑人或非裔美国人，2 位认同为西班牙裔、拉丁裔或西班牙裔出身，1 位同时认同为白人和黑人或非裔美国人。所有参与者的母语均为英语。参与者获得了价值 50 美元的礼品卡作为报酬，每次访谈时长约为 30 分钟。

5.1.3 访谈分析。访谈内容由自动转录服务完整转录，并通过演绎与归纳相结合的过程进行分析 [14]。我们采用演绎分析，评估这些参与者经历了哪些类型的故障。为使演绎分析有据可依，我们使用了与数据集分析相同的编码手册，如表1所示。我们首先找出参与者讨论不同故障的实例，然后将编码手册应用于这些实例。我们使用的线索包括参与者所处环境中的情况，以及在适当情况下，用户对故障发生原因的自身感知。首先，我们确定故障属于四种故障类型中的哪一种：注意、感知、理解或响应。首先，为了判定是否发生了注意力故障，我们考察了是否有证据表明语音助手准确响应了唤醒短语，例如视觉或听觉线索所示，或参与者叙述中的其他证据。其次，我们根据参与者是否认为语音助手准确解析了其输入、我们对其叙述的评估，或其他音频／视觉线索，来判断是否存在感知错误。接下来，在假设输入已被正确解析的前提下，我们采用相同的方法，探究语音助手是否准确理解了输入的语义（理解故障）。最后，在假设其他方面均已正确理解的前提下，我们将响应故障归为语音助手未采取操作，或未能针对已准确理解的命令采取正确操作。一旦确定故障类型，我们便进一步按照表1所述明确故障来源。在数周时间里，我们通过讨论和比较，在线下非同步交流及会议中解决了分歧。

在进行这项分析的过程中，我们还通过归纳方式识别出与这些故障对未来任务和恢复策略的影响相关的主题。为开展这项分析，两位研究人员通读了十二份访谈记录，另有一位研究人员审阅了其中五份，以进一步拓展和丰富主题。研究人员在数周内多次会面，交流笔记和主题，最终通过归纳分析形成了四个不同的主题。在这些主题中，我们报告了其中两个

因为它们具有新颖性，尤其是与未来任务导向和恢复策略相关。

5.2 调查方法

为了量化访谈中获得的发现，我们设计了一份问卷调查，探究用户在分类体系中的十二种不同故障发生后对语音助手的信任，以及他们在故障发生后使用语音助手完成各种任务的意愿。

5.2.1 流程。 问卷调查包括筛选题、核心任务和人口统计信息部分。参与者必须年满 18 岁、使用英语与语音助手互动，并且经常使用语音助手，才能参与调查。通过筛选的参与者必须先阅读并同意一份数字同意书，才能继续。

核心任务的说明如下：“接下来的问题将询问您对语音助手能力的看法，前提是该语音助手犯了一个错误。请想象这些错误是由您以前使用过的语音助手所犯的。请将每种情境视为独立事件，不受其前后情境的影响。本问卷调查大约需要 *20* 分钟。”随后，参与者会看到 12 种不同的失败情景，并需要通过两个独立的问题对自己的信任程度进行评分。

第一个问题通过信心评分来衡量对语音助手的信任，涵盖三个维度：能力、善意和正直。之所以选择这些维度，是因为以往关于信任的研究已经确定，这些要素可以解释信任度的很大一部分 [36, 40]。在语音助手的情境下，能力是指语音助手准确响应用户输入的能力。善意是指产品的意图有多善。最后，正直代表其会遵守伦理标准。

我们要求参与者以百分比的形式，在 0—100 的范围内以 10 为步长，对自己对语音助手的能力、善意和正直的信心进行评分，以复现以往研究对信任的概念化方式 [36]。具体评分依据以下陈述：

- （能力）该语音助手通常能够准确地响应命令。
- （善意）该语音助手旨在满足用户提出的命令。
- (正直) 该语音助手不会伤害其用户。

第二个问题评估了用户对语音助手完成不同信任程度任务的信任：高信任、中等信任和低信任。为选择这些任务，我们在 Mechanical Turk 上对 88 名语音助手用户开展了一项小型问卷调查。我们提出了 12 个不同的问题；每个问题首先给出一个语音助手故障的示例（每种故障来源各一个），然后询问：“您会多大程度上信任该语音助手执行以下任务：”提供天气预报、播放音乐、编辑购物清单、给同事发短信以及转账。用户可以选择完全信任、在一定程度上信任，或完全不信任。

在人们对语音助手播放音乐和预测天气的信任程度方面，两者没有显著差异（$Z = 2.06$，$p = 0.078$）。根据成对比较，使用 -检验，并在有序混合模型的方差分析中采用 Holm 顺序 Bonferroni 校正后发现，人们对语音助手编辑购物车或给同事发短信的信任程度也没有显著差异（$Z = 1.39$，$p = 0.21$）。我们发现，在播放音乐、给同事发短信和转账这几项任务之间存在显著差异：故障发生后，用户对语音助手播放音乐的信任度最高，对给同事发短信的信任度较低，对转账的信任度则更低。因此，我们选择播放音乐、给同事发短信和转账，分别代表所需信任度低、中、高的任务。因此，在询问参与者对语音助手能力、善意和正直的看法后，我们还询问他们对语音助手执行以下任务的信任程度：播放音乐、给同事发短信和转账。这些问题采用线性量表呈现，范围从 1（“我完全不信任它”）到 5（“我完全信任它”），每级递增 1。

我们在问卷调查中加入了一个开放式可选问题，供参与者分享他们希望补充的任何其他内容。问卷调查最后包含人口统计学问题，涉及性别、种族、族裔、是否为英语母语者、所使用的语音助手类型，以及作为控制变量的一般信任倾向。一般信任倾向根据以下问题的回答进行测量：“一般来说，您认为大多数人值得信任，还是认为与人打交道时需要格外小心？”选项范围从 1（与人打交道时需要格外小心）到 5（大多数人值得信任）。用于该问卷调查的问卷已作为补充材料提交。

5.2.2 来自我们数据集的材料。为了在问卷调查中呈现十二种故障来源，我们从创建的数据集中选取了材料。我们从十二个类别中各选取了五个故障。我们要求这些故障必须由团队中两名意见一致的成员进行编码（参见表2中的数据集示例）。对于每种故障来源，我们通过随机选择，决定向每位用户呈现五个可能故障中的哪一个。这些故障在数据集的“Survey”列中标记。

5.2.3 参与者。我们从亚马逊 Mechanical Turk 招募参与者。我们首先开展了一项小规模试点研究（$n = 27$），据此确定参与者平均约需 20 分钟完成问卷调查，并将报酬设定为 9 美元。剔除未通过注意力检查或作答始终选择同一答案的参与者后，我们最终共有 268 名参与者。这些参与者必须符合以下条件：拥有 AMT 大师资格；已有超过 1000 个 HIT 获批；年满 18 岁；居住在美国；获批率高于 97%；且未参与过我们此前的任何研究。

参与者中人数最多的是年龄范围为 35-44 岁的人（$n = 106$），其次是 25-34 岁（$n = 68$）、45-54 岁（$n = 52$）和 55-64 岁（$n = 33$）；在 18-24 岁、65-74 岁和 75+ 这几个年龄段中，每个年龄段有 2-4 名参与者。134 名参与者认同自己是男性，132 名认同自己是女性，2 名认同自己属于非二元性别。大多数参与者是白人（$n = 210$），21 名参与者是黑人，15 名是亚裔。其余参与者认同自己是混血，或选择不作答。

	能力				善意				正直			
	F	df	残差	*p*	*F*	df	残差	*p*	*F*	df	残差	*p*
(截距)	9844.95	1	422.35	<0.001	8436.32	1	355.86	<0.001	10153.41	1	325.54	<0.001
一般信任	3.07	4	286.67	0.017	1.78	4	299.1	0.133	4.69	4	311.17	0.001
故障类型	17.17	3	2656.78	<.001	8.87	3	2711.23	<.001	20.56	3	2772.09	<.001

表3：在控制用户的基线信任倾向后，基于对三个线性混合模型进行方差分析的结果，语音助手故障类型显著影响了用户对语音助手的信任，包括对其能力、善意和正直的信任。故障类型被编码为分类变量，一般信任被编码为有序值。参与者 ID 是一个随机分类变量。

6 结果：语音助手发生故障后的信任

在访谈中，我们发现参与者报告了涉及全部四种故障类型以及十二种故障来源中的十种的故障。访谈中唯一未提及的两种故障来源，是注意力故障类型中的触发遗漏和触发延迟。为了了解哪些类型的故障对用户信任的影响最为显著，我们分析了各种故障如何影响用户对语音助手能力、善意和正直的信心。我们使用了六个混合线性回归模型，将能力、善意或正直的对数归一化信心作为数值因变量。请注意，我们从两个不同层面进行分析。首先是四种宽泛的“故障类型”层面（注意、感知、理解和响应），然后进一步考察各故障类型下的十二种具体“故障来源”。因此，对于信任的每个维度，我们都将故障类型或故障来源以及一般信任倾向编码为自变量，即每个信任维度对应两个回归模型。故障类型和故障来源被编码为分类变量，一般信任倾向则被编码为有序值。在所有模型中，PID 均被编码为随机分类变量。

对回归模型进行方差分析后发现，故障类型（注意、感知、理解、响应）确实会显著影响能力感知（$F(3,\ 2656。78) = 17.17,\ p < .001$）、善意（$F(3,\ 2711。23) = 8.87, p < 0.001$），在控制一般信任倾向后，正直方面也呈现了这一结果（$F(3, 2772.09) = 20.56, p < .001$）（见图2和表3）。我们发现，与由“注意”导致的故障（包括触发遗漏、误触发和延迟触发）相比，“响应”这一故障类型（包括动作执行：不执行动作和动作执行：错误动作）在能力（$m = 43.6, \beta = -0.155, p < .001$ $m = 52.5\ \beta = -0.072\ p = 0.013$）、善意（, , ）和正直（$m = 57.3, \beta = -0.124, p < .001$）方面，对语音助手用户信任的损害更为显著。注意力失败在0至100%的量表上，对能力、善意和正直的平均信任评分分别为49.9、56.0和67.7。我们还发现，与注意力失败相比，感知故障显著降低了用户对语音助手能力（$m = 44.8, \beta = -0.122, p < .001$）和善意（$m = 53.6, \beta = -0.047, p = .05$）的信心，但对正直（$m = 61.4, \beta = -0.014, p = 0.484$）没有可测量的影响。理解故障使用户对善意（$m = 57.9, \beta = 0.054, p = 0.031$）和正直（$m = 65.0, \beta = 0.063, p = 0.003$）保持了较高的信心，但对能力没有可测量的影响（$m = 50.2,\ \beta = 0.015,\ p = 0.592$），与注意力失败相比。

总体而言，响应故障在能力、感知和正直方面的平均得分最低。这些类别之间最显著的差异体现在理解故障上：如图2所示，这类故障通常最能维持用户对能力、善意和正直的信任。因此，在下文评估不同故障来源对信任的影响时，响应：错误动作 被选作参照变量，所有报告的贝塔系数均以此类别为基准。下面，我们将详细探讨用户在访谈和问卷调查中面对不同故障来源时的反应。

6.1 注意故障

注意力失败是指语音助手未能准确响应激活尝试的任何故障。这类故障是在访谈中最少被提及的。在问卷调查中，由触发遗漏导致的故障尤其损害用户对语音助手

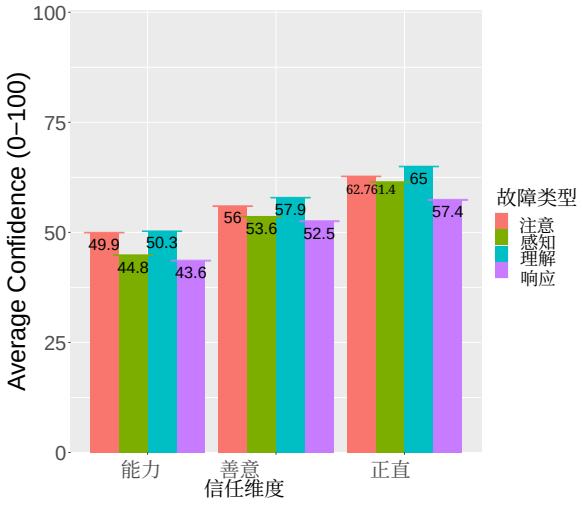

图2：按故障类型划分的信任三个维度（能力、善意和正直）的平均得分。与感知故障和响应故障相比，参与者在经历注意故障和理解故障后，在三个信任维度上的信心更高。误差线表示置信区间。

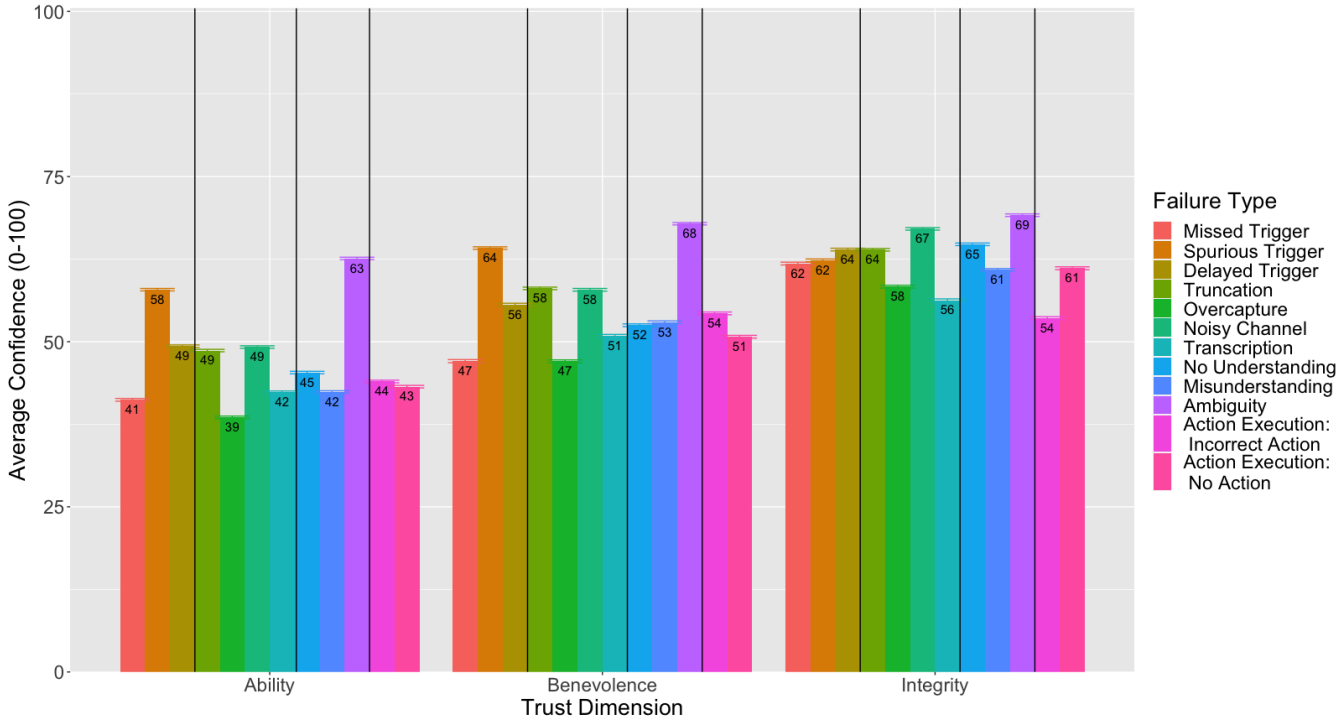

图3：按故障类型划分的信任的三个维度（能力、善意和正直）的平均得分。与其他故障类型相比，参与者在经历由歧义和误触发导致的故障后，在三个信任维度上的信心更高，尤其是相较于触发遗漏和过度捕获失败。

能力的信心（$m = 41.2$，$\beta = -0.10$，$p = 0.05$）和善意感知（$m = 47.1$，$\beta = -0.228$，$p < 0.001$）。然而，与参考值（动作执行：错误动作）相比，这类故障对正直的影响是积极的（$m = 61.8$，$\beta = 0.194$，$p < 0.001$）。没有访谈参与者报告因触发遗漏而导致的故障。由于所有故障对正直的影响都优于参考值，因此在其余结果中我们不再报告这一指标。更多详情请参见图3和表4。

只有P7和P12在访谈中报告经历过注意力失败，而且两者都是误触发。如表1所示，这些故障是指语音助手在没有激活短语的情况下被激活。P7报告说，

> “我觉得在对话时，如果我把它插着电，而且房间里有好几个人，他们在聊天之类的，我觉得有时它可能会听到一个 *[激活短语]*，但其实并没有人说。要是发生过这种情况，比如它完全莫名其妙地激活了一两次，那倒也没让我感到不高兴或怎样，但我当时就是觉得，我并没有说 *[激活短语]*。你为什么要激活？发生什么事了？你为什么要这样做？”

P7还说，他们当时正在工作，“它肯定是在什么地方听到了*[激活短语]*。然后，在我正试图做 *[工作]*，时，它开始说话，我只好停下来对它说，嘿，停下。”他们说：“这真的会让我 *[很生气]*。”同样，P12报告称，这类故障“既令人恼火又同时很有趣”。他们说，这些故障之所以有趣，是“因为有时候，比如说，当你通常在呼叫 *[语音助手]* 时，它需要更长时间才能回应；但当你没有和它说话时，它却会自动弹出来……就好像我根本没在和你说话，但你在我跟你说话时却可以回应我。”正如图3和表4所示，由误触发导致的故障对用户对语音助手能力（$m = 57.9$，$\beta = 0.36$ $p < 0.001$ $m = 64.2$ $\beta = 0.236$ $p < 0.001$，）和善意（，，）的信任印象产生了相对更有利的影响。总体而言，这些故障似乎是对用户信任危害最小的故障之一。

与参照变量（响应：错误动作）相比，由延迟触发导致的故障同样有利于用户的能力感知（$m = 49.3$，$\beta = 0.16$，$p = 0.001$）。延迟触发故障是指语音助手在启动时出现延迟，可能因此未能及时提供正确响应，以至于响应虽正确却已无用，但也不一定会出现这种情况。此类故障对善意没有可测量的影响（$m = 55.6$，$\beta = 0.043$，$p = 0.306$）。没有参与者在访谈中报告过由延迟触发导致的故障。

6.2 感知失败

用户报告了表1中所列全部四种故障类型的故障，包括截断、过度捕获、噪声信道和转录。感知故障表明，语音助手未能准确捕获用户输入。转录是最常见的故障来源，远远

	能力				善意				正直			
	betas	se	*Z*	*p*	*betas*	se	*Z*	*p*	*betas*	se	*Z*	*p*
(截距)	3.56	0.046	77.413	<0.001	3.796	0.048	79.083	<0.001	3.754	0.045	83.422	<0.001
一般信任	0.23	0.096	2.396	0.017	0.111	0.111	1	0.317	0.339	0.106	3.198	0.001
触发遗漏	-0.10	0.048	-1.958	0.05	-0.228	0.043	-5.302	<0.001	0.194	0.037	5.243	<0.001
误触发	0.36	0.046	7.761	<0.001	0.236	0.041	5.756	<0.001	0.212	0.037	5.73	<0.001
延迟触发	0.16	0.046	3.391	0.001	0.043	0.042	1.024	0.306	0.249	0.036	6.917	<0.001
截断	0.16	0.046	3.522	<0.001	0.126	0.042	3	0.003	0.263	0.037	7.108	<0.001
过度捕获	-0.13	0.047	-2.809	0.005	-0.193	0.042	-4.595	<0.001	0.133	0.037	3.595	<0.001
噪声信道	0.14	0.046	2.978	0.003	0.076	0.042	1.81	0.07	0.326	0.036	9.056	<0.001
转录	-0.06	0.047	-1.213	0.225	-0.101	0.042	-2.405	0.016	0.093	0.037	2.514	0.012
无法理解	0.06	0.047	1.17	0.242	-0.043	0.042	-1.024	0.306	0.28	0.037	7.568	<0.001
误解	-0.017	0.046	-0.37	0.711	-0.023	0.042	-0.548	0.584	0.202	0.037	5.459	<0.001
歧义	0.46	0.046	9.913	<0.001	0.302	0.042	7.19	<0.001	0.361	0.036	10.028	<0.001
无操作	-0.003	0.047	-0.064	0.949	-0.09	0.042	-2.143	0.032	0.186	0.037	5.027	<0.001

表4：三个混合线性回归模型的结果，展示了语音助手故障如何影响用户在能力、善意和正直方面对语音助手的信任。参照故障来源：错误动作。

多于截断、过度捕获和噪声信道这三类故障各仅记录到的一次故障。

截断故障表明，语音助手过早停止聆听输入，只根据用户部分预期输入进行了操作。P12报告称："我使用 *[语音助手]*发送消息之类的，有时它会把一些词转写成文本，但不是所有词都会转写。所以我发送一条消息所花的时间比预期更长，因为它只会接收一小部分词，没有完整地聆听。"他们说："这很令人恼火，非常烦人。"截断故障对能力感知（$m = 48.7$，$\beta = 0.16$，$p < 0.001$）和善意感知（$m = 58.1$ $\beta = 0.126$ $p = 0.003$ 3, , , ）具有较为有利的相对影响。如图所示，与其他感知故障相比，这些方面维持了更高的相对信任。

过度捕获失败表明，语音助手在用户输入结束后仍继续聆听。正如 P8 所说，有时，"它不知道什么时候该开始搜索我说的话，只是不停地听，却不采取操作，尽管它显示自己正在聆听。"他们试图理解这种故障，说："我发现，在不同的设备上，它的反应时间 *[各不相同]*。"他们说："这在浪费我的时间。按理说，这只会花两到三分钟，"但他们也说："如果你一直摆弄它，情况只会更糟。"与所有其他故障类型相比，过度捕获导致的故障尤其损害用户对语音助手能力的信心（$m = 38.5$，$\beta = -0.10$，$p = .05$）和对其善意的信任（$m = 47.0$，$\beta = -0.228$，$p < 0.001$），其总体均值也最低。

有一例中，用户认为其经历的故障是由背景噪声造成的，这表明存在噪声信道故障。P9 说道："有时候……我会尝试使用一项能够识别歌曲之类内容的功能……但它就是无法识别出来，只会给我一条消息，比如'抱歉，我无法理解您的意思。'"他们还说："我知道当时声音很大……但我觉得它应该能够听懂。所以我觉得这有点令人恼火。"然而，他们表示，这次故障并未影响他们对语音助手准确性或能力的看法，并说道："总体来说，它在其他事情上的准确性还是相当不错的。"用户认为噪声信道故障对能力感知的影响更为正面（$m = 49.2$，$\beta = 0.14$，$p = 0.003$），而对善意没有可测量的影响（$m = 57.9$，$\beta = 0.076$，$p = 0.07$）。

如图3所示，他们获得的信任水平与因截断导致的故障所获得的信任水平相近。

九位参与者提到了与其输入转录有关的故障，表明他们认为语音助手未能准确捕捉自己说的话。这些故障包括无法理解音乐团体的名称（P7）、错误转录短信（P2）、错误转录数字序列（P4）、无法理解愤怒、含混或含糊不清的语音（P3、P5、P9），以及无法理解口音（P8）或其他语言（P6、P9）。P7 说，当语音助手无法理解他们想点播的音乐人时，会让他们感到"有一点点沮丧"。不过，他们"并不会因此特别责怪 *[语音助手]*，因为它在其他方面做得都很好。"然而，说到用语音助手处理西班牙语或法语等其他语言时，"它从来没有成功过，没能用另一种语言播放那首不同的歌曲"（P9）。这让该参与者觉得，"它就是没有能力理解我说的另一种语言"（P9）。与参照变量相比，转录导致的故障对问卷调查中的能力感知没有造成可测量的影响（$m = 42.3$，$\beta = -0.06$，$p = 0.225$）；然而，如图3所示，在感知层面，这类故障对信任的影响比其他故障来源更大。转录失败确实对善意感知产生了负面影响（$m = 51.0$，$\beta = -0.101$，$p = 0.016$）。

6.3 理解失败

我们发现，参与者提交的故障涵盖以下所述的所有理解失败类别。

由于无法理解而导致的故障，使语音助手完全无法将输入映射为操作或响应。P6 说："我当时正计划去度假……那是我朋友的单身派对……我就问，*[语音助手]*,哈瓦苏湖在哪里？有多远？……她却说：'抱歉，我没听懂你在说什么。'"这让 P6 不禁质疑："我为什么还要用你？"不过，他们也表示："用计时器的话，它表现得很好。"与参照变量相比，未能理解的失败并未显著影响用户对能力（$m = 45.3$，$\beta = 0.06$，$p = 0.24$）或善意（52.5 $\beta = -0.043$ $p = 0.306$, , ）的信任。

当语音助手将用户输入映射到某项操作，但映射结果只部分符合用户意图、并不完全准确时，就会发生误解失败。例如，P4 解释说，当他们对语音助手说“‘带我回家。’它通常会为我导航回家，但有时会显示‘带我回家’这句话的搜索结果。”同样，P1 解释了使用语音助手网上购物时，有时它会“调出错误的商品，或者错误的地点”。他们说自己感到“失望和沮丧。”与参照变量相比，误解失败并未对能力感知（$m = 42.4$，$\beta = -0.017$，$p = 0.711$ $m = 52.9$ $\beta = -0.023$ $p = 0.584$ ）或善意（， ，）产生可测量的影响。

由歧义导致的故障，是指从捕获的输入中可以看出用户意图存在多种合理解释，但系统未能妥善处理这些歧义的情况。例如，P10表示：“我想去必胜客……但它一直告诉我附近城市有一家，而不是我觉得离我大概*10*分钟路程的那家。所以我问了几次，结果还是不行，于是我就拿出手机，自己查了一下，然后出发了。”他们说自己“有点困惑，因为通常来说，当我向*[*语音助手*]*询问某件事时，我会得到符合预期的响应。”如图3和表4所示，由歧义导致的故障对用户关于语音助手能力（$m = 62.6$、$\beta = 0.46$、$p < 0.001$）和善意（$m = 67.9$、$\beta = 0.302$、$p < 0.001$）的印象影响较为有利。总体而言，这类故障维持了最高水平的用户信任。

6.4 响应失败

响应故障有两种可能类型：错误动作，即系统提供了错误的信息；无操作，即语音助手完全没有响应。

错误动作故障指的是命令似乎被准确理解，但响应中提供的信息有误。例如，P1说，有时会使用“语音助手为我提供前往该地点的最佳路线。”虽然语音助手通常能准确响应这一命令，但有时“它会给我一条非常绕远、非常耗时的路线。”如图3所示，错误动作导致的故障使用户对能力（$m = 44.0$）和善意（{v6）的感知处于相对平均的水平，而对正直性的感知最低（$m = 53.6$）。

多位用户经历了由无操作导致的故障，即语音助手完全没有回应输入。P2 说：“我确实遇到过几次这种情况，也挺让人沮丧的……我会说‘回复’*[*短信*]*，然后说完了，却什么也没发出去。那时候我的手上真的全是东西，因为我正在擀这些饼干面团，我只好停下手头的事，回到手机旁，自己手动发短信。”另一位参与者也经历了由无操作导致的故障，他说：“今天早上我刚醒来时，我说：*[*语音助手*]*,外面天气怎么样？它最初几秒钟加载了一下……几秒后，它说：‘发生错误。请几分钟后重试。’我等一两秒，然后再问一次，它就会把信息告诉我”（P10）。这位参与者说，由于信息一直“准确”，他们“仍然会非常信任它。无操作导致的故障与错误动作相比，无操作导致的故障对能力没有可衡量的相对影响（43.2 $\beta = -0.003$ $p = 0.949$ ， ，），但对善意有轻微但显著的负面影响（$m = 50.7$，$\beta = -0.09$，$p = 0.032$）。

7 响应失败与语音助手的未来使用

用户描述了在故障确实发生时用来缓解故障的各种策略。有些情况下，用户表示会完全停止使用语音助手来完成某项特定任务。例如，P12在使用语音助手发送短信时遇到截断故障后说，他们要么“得重来一次，要么干脆就不做了”。后来，P12说，他们不再遇到这种故障，是因为现在“几乎不用它”来完成同一项任务了。因此，有些用户觉得语音助手在执行任务时出故障，“也没什么大不了”（P5）；而另一些用户则表示，在这些相同的任务中，他们会“少用一些”（P2）。

我们发现，在许多不同类型的故障中，包括转录、误解和歧义，用户总体上继续使用语音助手，但至少在一段较短的时间内避开导致故障的任务，这一模式是一致的。例如，P2 表示，他们使用语音助手时需要更加小心，因为有时他们说出一个姓名，而“它却会显示 *[*出*]* 另一个姓名。”他们说，在发生此类事件后，

> “我仍然会使用 *[*语音助手*]*。不过，我想接下来会发生的是，你会逐渐建立起那种信任……所以接下来的几次，我会自己进入联系人并按下按钮，你知道，如果我正走向汽车，一只手拿着钥匙，而且已经过了一段时间。于是我会想，好吧，再试一次。我认为，这种情况下，你必须逐渐重新建立信任，再给它一次机会。至少我是这么做的。”

P12 也表达了类似的看法，他说：“比如说你要打开 *Spotify* 或者做类似的事。我觉得它大概会执行命令，而不是发送消息……不同的任务，你懂的，信任程度也不同。” P5 也有类似的感受，他说：“我觉得大多数语音助手的问题在于，如果我试着输入复杂的搜索查询，它就不太理解我的意思，或者会让人很沮丧，于是我干脆直接输入自己要找的内容。”即使故障当下得到了解决，用户仍然对使用语音助手处理同类任务心存疑虑。

有趣的是，有时用户在语音助手发生故障后，仍会继续用它完成同一类任务，只是稍微调整了使用方式。例如，P1 在线上购买毛衣时遭遇了理解故障，于是开始“不那么依赖它，多自己搜索。”他们说：“以后我会记住，有些任务就不再用它来做，而是自己完成，*[*尤其是*]* 找难以找到的商品时。”不过与此同时，“我还是会让它帮忙做其他任务。”对 P1 来说，这包括“寻找毛衣以外的其他商品。我会让它搜索杂货之类的商品，并在线比价购物。”购买不同的商品这一任务足够不同，因此仍能维持这位用户的信任。P7 也经历了类似的情况，在

	播放歌曲				给同事发短信				转账			
	β	se	Z	p	β	se	Z	p	β	se	Z	p
能力	0.048	0.003	16.000	<0.001	0.064	0.003	21.333	<0.001	0.076	0.005	15.2	<0.001
善意	0.043	0.003	14.333	<0.001	0.028	0.003	9.333	<0.001	0.017	0.005	3.4	0.001
正直	0.019	0.002	9.500	<0.001	0.024	0.003	8	<0.001	0.03	0.004	7.5	<0.001
一般信任	-0.146	0.108	-1.352	0.176	-0.094	0.134	-0.701	0.483	-0.192	0.218	-0.881	0.378

表5：三个混合有序回归的结果。这些回归根据用户对语音助手的能力、善意和正直的感知，对用户信任语音助手能够执行任务的程度进行建模。为便于解读，表中未列出切点计算和 state1 计算。

他们遇到了转录错误，后来通过拼写出“*hyerpop duo 100 gecs*”的名字“*G-E-C-S*”来解决。他们说，这样纠正很有帮助，因为“*[助手确实]* 理解了我说的话。”尽管他们在点播这位艺人的音乐时遇到了故障，但直到今天，他们仍“继续 *[用它播放歌曲]*。它是个非常好用的音乐播放器，”但对于某些音乐人，他们还是“有点担心，我觉得语音助手在理解他们时可能会遇到困难……*[语音助手]*。”

用户通常会根据感知到的任务复杂度来理解故障。P12 认为，他们最信任的任务是“打开应用程序”，其次是“拨打电话”。他们解释说：“我想把它排在第一位，但有时候，联系人姓名没有被注册。你知道，自己说出姓名的方式和语音 *[助手]* 识别的方式不一样。”P2 对语音助手的评价也类似，他们说：“最擅长的是获取网站信息。”不过，他们也表示：“要获取更多个性化消息、联系人信息之类的内容，你必须非常注意自己说什么以及怎么说。”

基于这些发现，我们假设，用户在语音助手发生故障后的信任，会以不同程度影响他们使用语音助手执行不同任务的意愿。如表5所示，我们使用三个混合有序回归模型来分析用户对这三项任务的信任，并将用户对语音助手能力、善意和正直的信心评分作为自变量。用户对语音助手播放歌曲、给同事发短信和转账的信任以有序值编码。用户对语音助手能力、善意和正直的信心以数值编码。一般信任倾向以数值编码，PID 则以随机类别值编码。我们发现，用户对语音助手的能力、善意和正直的感知与其使用语音助手执行未来任务的意愿呈正相关。总体而言，人们对语音助手播放歌曲抱有中等程度的信任（$m = 3.29,\ sd = 1.30$），对语音助手给同事发短信的信任较低（$m = 2.34,\ sd = 1.18$），对语音助手转账的信任最低（$m = 1.56,\ sd = 0.91$）。

具体而言，与善意相比，能力感知对人们使用语音助手播放歌曲的意愿影响更大（$\beta = 0.048,\ p < .001$）；善意也显著影响人们使用语音助手播放歌曲的意愿，但影响程度略低（$\beta = 0.043,\ p < .001$）。正直的影响更小，不过仍与人们对语音助手播放歌曲的信任程度显著正相关（$\beta = 0.019,\ p < .001$）。给同事发短信和转账时也呈现出同样的模式：能力与人们信任语音助手的意愿呈最强的正相关，其次是善意，然后是正直，即人们愿意信任语音助手执行这些任务的程度。

8 讨论

通过访谈、问卷调查和众包的语音助手故障数据集，我们开展了一项混合方法研究，探讨语音助手故障及其对用户信任和未来使用意愿的影响。随着语音助手的底层技术在准确性和能力方面不断提升，其应用对人类健康与福祉的影响也日益重大 [17, 41, 55, 66], 我们将讨论研究发现，以期提升用户对语音助手的信任和长期使用参与度。

用户一贯依赖语音助手来查找信息并执行不同复杂程度的任务。与先前研究相似 [35], , 那些希望持续使用语音助手执行可能导致故障的任务的人，已经形成了复杂的心智模型，了解哪些任务可以信赖语音助手来完成。与先前研究不同 [35], , 即使语音助手在复杂任务上反复出现故障，人们通常也不一定会因此完全放弃使用。许多用户认为语音助手的准确性始终很高，因此能够原谅故障，并在短暂间隔后继续使用语音助手执行这些任务。在修复对复杂任务的信任期间，许多参与者仍会使用语音助手完成他们认为较简单的任务，例如信息检索和播放音乐。

我们发现，那些让用户觉得自己浪费了时间的故障，例如由触发遗漏和过度捕获导致的故障，往往会进一步削弱用户对语音助手能力和善意的感知。相比之下，在用户更了解语音助手为何发生故障的情形下，例如因歧义和转录导致的故障，用户通常觉得自己能够设法应对或提前预料。不过，如果用户认为转录故障是因为用另一种语言使用设备所致，就会因此放弃使用该语言的语音助手。同样，发生误触发故障时，用户并不觉得自己失去了使用语音助手的优势，因此这类故障对能力感知的损害较小。对语音助手正直影响最大的故障来源是动作执行：错误动作，因为参与者对语音助手在发生此类故障后不会造成伤害这一说法更加怀疑。

先前研究指出，发生故障时可以通过某些方式修复信任。Cuadra 等人 [16] 发现，当语音助手主动承认故障并尝试修复信任时，

人们对其智能程度的感知会随之提升。此外，Mahmood 等人 [39] 发现，故障发生后，道歉等故障缓解策略有助于恢复人们对语音助手亲和力和智能程度的感知。Xiao等人[65]表明，将语音助手定位为学习者，并帮助用户了解何时向语音助手提供反馈，可以改善用户对语音助手的感知。Fischer等人 [23] 建议，语音助手的响应应支持对话的顺畅推进，尤其是在响应无法帮助用户时。我们的研究表明，故障发生后，用户会自然而然地通过让语音助手处理不同任务，或在同一任务中改换主题来修复信任，例如网上购物时购买不同商品，或播放由其他艺人演唱、而非导致故障的艺人创作的音乐。

通过定量分析，我们发现某些类型的故障比其他故障更为关键。这一发现有助于在人机交互和自然语言处理领域中，优先采用最能恢复信任的故障恢复策略。例如，Cuadra 等人 [16] 采用的语音助手自我修复方法，在语音助手因触发遗漏或过度捕获用户输入而发生故障时，可能最为有效。此外，我们还可以尝试识别语音助手技术栈中导致关键故障的具体组件，并利用自然语言处理鲁棒性技术，改善这些模型在用户交互过程中的表现。例如，噪声信道和转录失败可以建模为对输入的微小扰动，这方面已有充分研究 [4, 21]。语音识别模块也可能需要着力提升转录的可靠性，尤其是对于低资源语言 [38]。

我们的开源数据集还为未来的研究人员提供了具体且全面的故障示例（199 个来自真实世界的示例，包含情境、查询和响应），供其重复使用，以开发故障缓解策略；同时，该数据集还提供了一个经过完善的分类体系，用于对语音助手故障进行分类，这一点也得到了既有研究的支持。虽然既有研究 [28] 有助于自然语言处理从业者预测并规划各种自然语言处理技术中的故障，但我们的数据集专门针对语音助手所发生的故障。我们预计，未来的研究人员将能够在开展与语音助手故障、信任和缓解策略相关的研究时，使用以人为中心的故障示例。

局限性

我们的研究存在一些方法学局限，现详述如下。首先，我们的数据集收集和访谈依赖于回顾性报告以及对故障的回忆，而不是在原情境中观察这些故障。这使我们的数据受到回忆偏差的影响，因此应从这一角度解读我们的结果。例如，在访谈中，没有任何参与者回忆起由触发遗漏或延迟触发导致的故障，尽管在问卷调查中，触发遗漏被认为会相对严重地损害人们对能力和善意的感知。此外，我们的问卷调查依赖于收集用户对假设情境的反馈。未来的研究可以利用我们的数据集，系统地引入故障，并通过经验抽样法或日记研究捕捉由此对用户信任产生的影响，从而在我们的研究发现基础上进一步展开。我们的参与者样本也都是经常使用语音助手的用户，这表明与其他人群相比，他们可能更容易原谅错误 [35]。此外，我们没有探讨通过语音以外的界面使用对话代理的情况，例如具身对话代理或基于文本的对话代理。由于具身界面和文本界面具有更多可供用户据以判断和与系统互动的潜在可供性， [5, 6], 故障的影响可能无法完全推广到这些使用场景。

9 结论

总之，通过一项混合方法研究，我们发现语音助手用户会遇到各种各样的故障，从语音助手误触发，到它给出的回应无法满足用户需求。这些不同类型的故障对用户信任的影响各不相同，而用户信任又会影响他们今后使用语音助手执行任务的意图。具体而言，我们发现，与错误的动作执行、触发遗漏或过度捕获导致的故障相比，误触发和歧义导致的故障对用户信任的损害较小。我们还发现，在故障发生后，人们会先通过播放歌曲等简单任务重建对语音助手的信任，然后才重新使用语音助手的全部功能。我们还贡献了一个包含199个故障案例的数据集，以帮助未来的研究人员和从业者在我们的工作基础上继续推进。我们希望，通过进一步致力于理解、预防和修复语音助手故障，能够建立语音助手用户对这些设备的信任，并使他们受益于语音助手日益丰富多样的功能。

参考文献

[1] Deepali Aneja、Daniel McDuff 和 Mary Czerwinski。2020年。《人机交互中的对话错误分析》。载于《第20届ACM智能虚拟代理国际会议论文集》。计算机协会，美国纽约州纽约市，论文3，共8页。https://doi.org/10.1145/ 3383652.3423901[2] Ron Artstein、Jill Boberg、Alesia Gainer、Jonathan Gratch、Emmanuel Johnson、Anton Leuski、Gale Lucas 和 David Traum。2018年。《Niki和Julie语料库：人类、机器人与虚拟代理之间的协作式多模态对话》。载于《第十一届国际语言资源与评测会议论文集》（LREC 2018）。欧洲语言资源协会（ELRA），日本宫崎。https://aclanthology.org/L 18 1463 -

[3] Tammy Bahmanziari、J Michael Pearson 和 Leon Crosby。2003年。信任对技术采纳重要吗？一种政策捕获方法。《计算机信息系统杂志》43卷4期（2003年），46–54页。[4] Yonatan Belinkov 和 Yonatan Bisk。2017年。合成噪声和自然噪声都会破坏神经机器翻译。arXiv预印本 arXiv:1711.02173（2017年）。[5] Timothy Bickmore 和 Justine Cassell。2001 年。关系型代理：建立用户信任的模型与实现。载于《 SIGCHI计算系统中人的因素会议论文集》（美国华盛顿州西雅图）（CHI '01）。计算机协会，美国纽约州纽约市，396–403页。https: //doi.org/ 10 1145 365024 365304 . / . Timothy Bickmore 和 Justine Cassell。2005年。

[6] 与具身对话代理进行社交对话。Springer Netherlands，荷兰多德雷赫特，23–54页。https://doi.org/ 10.1007/1-4020-3933-6 2_[7] Michael Braun、Anja Mainz、Ronee Chadowitz、Bastian Pfleging 和 Florian Alt。2019年。随时为您服务：设计语音助手个性以改进汽车用户界面。载于《2019年CHI计算机系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约市，1–11页。https://doi.org/10.1145/ 3290605.3300270[8] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习器。《神经信息处理系统进展》33卷（2020年），1877–1901页。[9] Heloisa Candello、Claudio Pinhanez、Mauro Pichiliani、Paulo Cavalin、Flavio Figueiredo、Marisa Vasconcelos 和 Haylla Do Carmo。2019年。受众对实体空间中对话界面用户体验的影响。载于《2019年CHI计算机系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约市，

[10] 1 13 10 1145 3290605 3300320 – 。 1996 https://doi.org/ 。 / 。Herbert H. Clark。《运用语言》。剑桥大学出版社，英国剑桥。https://doi.org/10.1017/CBO9780511620539[11] Leigh Clark、Nadia Pantidi、Orla Cooney、Philip Doyle、Diego Garaialde、Justin Edwards、Brendan Spillane、Emer Gilmartin、Christine Murad、Cosmin Munteanu、Vincent Wade 和 Benjamin R. Cowan。2019年。怎样才算一次好的对话？设计真正具备对话能力的代理所面临的挑战。载于《2019年CHI计算机系统中的人为因素会议论文集》（英国苏格兰格拉斯哥）（CHI '19）。计算机协会，美国纽约州纽约市，1–12页。https://doi.org/ 10 1145 3290605 3300705 . / 。Jamie Condliffe。人工智能语音助手应用日益普及，但人们并不使用它们。《麻省理工科技评论》。https: //www.technologyreview.com/2017/01/23/154449/ai-voice-assistant-apps-are-proliferating-but-people-dont-use-them/.

[12] 2017 [13] Toby A. Cox。2020年。Siri和Alexa故障：语音搜索令人沮丧之处。https://themanifest.com/digital-marketing/resources/siri-alexa-fails-frustrations-with-voice-search[14] John W Creswell 和 Cheryl N Poth。2016年。《质性探究与研究设计：五种方法的选择》。Sage出版社，千橡市。[15] Andrea Cuadra、Hansol Lee、Jason Cho 和 Wendy Ju。2021年。看着我跟你说话：帮助语音助手识别错误的视频数据集。arXiv预印本 arXiv:2104.07153（2021年）。[16] Andrea Cuadra、Shuran Li、Hansol Lee、Jason Cho 和 Wendy Ju。2021年。我的错！修复智能语音助手错误可改善交互。《ACM人机交互论文集》5卷 CSCW1（2021年），1–24页。[17] Celso M De Melo、Kangsoo Kim、Nahal Norouzi、Gerd Bruder 和 Gregory Welch。2020年。利用智能虚拟助手降低认知负荷并改善作战人员的问题解决能力。《心理学前沿》11卷（2020年），554706。[18] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向变换器预训练。arXiv预印本 arXiv:1810.04805（2018年）。[19] Berkeley J Dietvorst、Joseph P Simmons 和 Cade Massey。2015年。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：综合》144卷1期（2015年），114页。[20] Mary T Dzindolet、Linda G Pierce、Hall P Beck 和 Lloyd A Dawe。2002年。人类和自动化辅助工具在视觉检测任务中的感知效用。《人的因素》44卷1期（2002年），79–94页。

[21] Javid Ebrahimi、Anyi Rao、Daniel Lowd 和 Dejing Dou。2018年。HotFlip：用于文本分类的白盒对抗样本。载于计算语言学协会第*56*届年会论文集（第*2*卷：短论文）。计算语言学协会，澳大利亚墨尔本，31–36。https://doi.org/ 10 18653 1 18 2006 。/v /P - Yuksel Ekinci 和 Philip L Dawes。。

[22] 2009 一线服务员工人格特质、交互质量与消费者满意度的消费者感知。《服务业期刊》29，4（2009），503–521。[23] Joel E. Fischer、Stuart Reeves、Martin Porcheron 和 Rein Ove Sikveland。2019年。语音界面设计的进展性。载于第一届国际对话用户界面会议论文集（爱尔兰都柏林）（*CUI '19*）。计算机协会，纽约州纽约市，美国，文章26，8页。https://doi.org/10.1145/3342775.3342788[24] Jack FitzGerald、Shankar Ananthakrishnan、Konstantine Arkoudas、Davide Bernardi、Abhishek Bhagia、Claudio Delli Bovi、Jin Cao、Rakesh Chada、Amit Chauhan、Luoxin Chen、Anurag Dwarakanath、Satyam Dwivedi、Turan Gojayev、Karthik Gopalakrishnan、Thomas Gueudre、Dilek Hakkani-Tur、Wael Hamza、Jonathan J. Hüser、Kevin Martin Jose、Haidar Khan、Beiye Liu、Jianhua Lu、Alessandro Manzotti、Pradeep Natarajan、Karolina Owczarzak、Gokmen Oz、Enrico Palumbo、Charith Peris、Chandana Satya Prakash、Stephen Rawls、Andy Rosenbaum、Anjali Shenoy、Saleh Soltan、Mukund Harakere Sridhar、Lizhen Tan、Fabian Triefenbach、Pan Wei、Haiyang Yu、Shuai Zheng、Gokhan Tur 和 Prem Natarajan。2022年。Alexa教师模型：用于自然语言理解系统的数十亿参数编码器的预训练与蒸馏。载于第*28*届*ACM*知识发现与数据挖掘会议论文集（美国华盛顿特区）（*KDD '22*）。计算机协会，纽约州纽约市，美国，2893–2902。https://doi.org/10.1145/3534678.3539173[25] Karthik Gopalakrishnan、Behnam Hedayatnia、Qinlang Chen、Anna Gottardi、Sanjeev Kwatra、Anu Venkatesh、Raefer Gabriel 和 Dilek Hakkani-Tür。2019年。Topical-Chat：迈向知识支撑的开放域对话。载于*Interspeech*论文集

2019 1891 1895 10 21437 2019 3079 。–。

[26] https://doi.org/。/Interspeech。- Aditya Gupta、Jiacheng Xu、Shyam Upadhyay、Diyi Yang 和 Manaal Faruqui。2021年。Disfl-QA：用于理解问答中言语不流畅现象的基准数据集。载于*ACL*论文集。[27] Paige Hayes、Jason Wagner、Mark McCaffrey 和 Matt Hobbs。2018年。消费者洞察系列：为语音革命做好准备。https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/voice-assistants.html [28] Matthew K. Hong、Adam Fourney、Derek DeBellis 和 Saleema Amershi。2021年。借助AI实战手册规划自然语言交互失败。载于*2021*年*CHI*计算系统中的人为因素会议论文集（日本横滨）（*CHI '21*）。计算机协会，纽约州纽约市，美国，文章

386, 11 10 1145 3411764 3445735 页。https://doi.org/。/。

[29] 2022 Tatum Hunter。。Siri和Alexa快把用户惹恼了。《华盛顿邮报》。https://www.washingtonpost.com/technology/2022/03/07/voice-assistants-wrong-answers/。（访问日期：2022年06月08日）。[30] Rinat Khaziev、Usman Shahid、Tobias Röding、Rakesh Chada、Emir Kapanci 和 Pradeep Natarajan。2022年。FPI：大规模对话助手中的故障点隔离。载于*2022*年计算语言学协会北美分会会议：人类语言技术：产业赛道论文集。计算语言学协会，混合形式：西雅图，华盛顿州

+ 141 148 10 18653 1 2022 17线上，–。

[31] 2019 https://doi.org/。/v / .naacl-industry。Dounia Lahoual 和 Myriam Frejus。。当用户帮助语音助手：从监督到故障解决。载于*2019*年*CHI*计算系统中的人为因素会议扩展摘要集（英国苏格兰格拉斯哥）（*CHI EA '19*）。计算机协会，纽约州纽约市，美国，1–8。https://doi.org/ 10 1145 3290607 3299053 。/。Chia-Hsuan Lee、

[32] Szu-Lin Wu、Chi-Liang Liu 和 Hung-yi Lee。2018年。Spoken SQuAD：一项减轻语音识别错误对听力理解影响的研究。*Interspeech 2018*论文集（2018年），3459–3463。[33] One-Ki Daniel Lee、Ramakrishna Ayyagari、Farzaneh Nasirian 和 Mohsen Ahmadian。2021年。交互质量与信任在人工智能语音助手系统使用中的作用。《系统与信息技术期刊》23，2（2021），154–170。[34] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*（2019年）。[35] Ewa Luger 和 Abigail Sellen。2016年。“就像有一个非常糟糕的私人助理”：对话代理的用户期望与体验之间的鸿沟。载于*2016*年*CHI*计算系统中的人为因素会议论文集（美国加利福尼亚州圣何塞）（*CHI '16*）。计算机协会，纽约州纽约市，美国，

5286 5297 10 1145 2858036 2858288 –。https://doi.org/。/。

[36] Xiao Ma、Jeffrey T. Hancock、Kenneth Lim Mingjie 和 Mor Naaman。2017年。Airbnb房东个人资料的自我披露与感知可信度。载于*2017*年*ACM*计算机支持的协同工作与社会计算会议论文集（美国俄勒冈州波特兰）（*CSCW '17*）。计算机协会，纽约州纽约市，美国，2397–2409。https://doi.org/10.1145/2998181。2998269

[37] Maria Madsen 和 Shirley Gregor。2000年。衡量人机信任。载于第11届澳大拉西亚信息系统会议，第53卷。6–8。[38] Alexandre Magueresse、Vincent Carles 和 Evan Heetderks。2020年。低资源语言：过去工作与未来挑战综述。arXiv预印本 arXiv:2006.07264（2020年）。[39] Amama Mahmood、Jeanie W Fung、Isabel Won 和 Chien-Ming Huang。2022年。真诚地承认错误：缓解人工智能错误的策略。载于2022年CHI计算系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，纽约州纽约市，美国，文章 578, 11 10 1145 3491102 3517565 页。[40] https://doi.org/。/。Roger C Mayer、James H Davis 和 F David Schoorman。1995。组织信任的整合模型。《管理学评论》20，3（1995），709–734。[41] Nikita Mehandru、Samantha Robertson 和 Niloufar Salehi。2022年。在医疗环境中可靠且安全地使用机器翻译。载于2022年ACM公平性、问责制与透明度会议（韩国首尔）（FAccT '22）。计算机协会，纽约州纽约市，美国，2016–2025。https://doi.org/ 10 1145 3531146 3533244 。/。John Miller、Karl Krauth、Benjamin Recht 和 Ludwig Schmidt。2020年。自然分布偏移对问答模型的影响。载于第37届国际机器学习会议（ICML'20）论文集。JMLR.org，文章641，12页。[42] Chelsea Myers、Anushay Furqan、Jessica Nebolsky、Karina Caro 和 Jichen Zhu。2018年。用户克服语音用户界面障碍的模式。载于2018年CHI计算系统中的人为因素会议论文集（加拿大魁北克省蒙特利尔）（CHI '18）。计算机协会，纽约州纽约市，美国，1–7。https://doi.org/10.1145/3173574.3173580[43] Aakanksha Naik、Abhilasha Ravichander、Norman Sadeh、Carolyn Rose 和 Graham Neubig。2018年。自然语言推理的压力测试评估。arXiv预印本 arXiv:1806.00692（2018年）。[44] Farzaneh Nasirian、Mohsen Ahmadian 和 One-Ki Daniel Lee。2017年。人工智能语音助手系统：从交互与信任视角进行评估。载于美洲信息系统会议。[45] Tim Paek 和 Eric J Horvitz。2013年。不确定性下作为操作的对话。arXiv预印本 arXiv:1301.3883（2013年）。[46] Yury Pinsky。2021年。响亮而清晰：人工智能正在改善助手对话。https://blog.google/products/assistant/loud-and-clear-ai-improving-assistant-conversations/。（访问日期：2022年12月12日）。[47] Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。了解自己不知道什么：SQuAD中的不可回答问题。arXiv预印本 arXiv:1806.03822（2018年）。[48] Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。2016年。SQuAD：用于文本机器阅读理解的100，000+ 个问题。arXiv预印本 arXiv:1606.05250（2016年）。[49] Siva Reddy、Danqi Chen 和 Christopher D Manning。2019年。CoQA：对话式问答挑战。计算语言学协会学报7（2019），249–266。[50] Samantha Robertson 和 Mark Díaz。2022年。理解与被理解：用户在机器翻译辅助聊天中识别并修复误译的策略。载于2022年ACM公平性、问责制与透明度会议（韩国首尔）（FAccT '22）。计算机协会，纽约州纽约市，美国，2223–2238。https: //doi.org/ 10 1145 3531146 3534638 。/。Tulika Saha、Saichethan Reddy、Anindya Das、Sriparna Saha 和 Pushpak Bhattacharyya。2022年。可以倾诉的对象：迈向缓解心理痛苦的激励型虚拟助手。载于2022年计算语言学协会北美分会会议：人类语言技术论文集。计算语言学协会，美国西雅图， 2436 2449 10 18653 1 2022 174 –。https://doi.org/。/v / .naacl-main。Maha Salem、Gabriella Lakatos、Farshid Amirabdollahian 和 Kerstin Dautenhahn。2015年。你会信任一个（有故障的）机器人吗？错误、任务类型和个性对人机协作与信任的影响。载于第十届ACM/IEEE人机交互国际会议论文集（美国俄勒冈州波特兰）（HRI '15）。计算机协会，纽约州纽约市，美国， 141 148 10 1145 2696454 2696497 –。https://doi.org/。/。William Seymour 和 Max Van Kleek。2021。探索语音助手中信任、拟人化与关系发展的交互。ACM人机交互学报5，CSCW2（2021），1–16。[55] Emre Sezgin、Yungui Huang、Ujjwal Ramtekkar 和 Simon Lin。2020年。在健康危机与疫情期间，语音助手支持医疗服务交付的准备度。《NPJ数字医学》3，1（2020），1–4。[56] Shaden Smith、Mostofa Patwary、Brandon Norick、Patrick LeGresley、Samyam Rajbhandari、Jared Casper、Zhun Liu、Shrimai Prabhumoye、George Zerveas、Vijay Korthikanti 等。2022年。使用DeepSpeed和Megatron训练Megatron-Turing NLG 530B大型生成式语言模型。arXiv预印本 arXiv:2201.11990（2022年）。[57] Yueming Sun 和 Yi Zhang。2018年。对话推荐系统。载于第41届ACM信息检索研究与开发国际会议（美国密歇根州安娜堡）（SIGIR '18）。计算机协会，美国纽约州纽约市， 235 244 10 1145 3209978 3210002 –。[58] https://doi.org/。/。Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kul-shreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du等。2022年。LaMDA：对话应用语言模型。*arXiv*预印本 *arXiv:2201.08239*（2022年）。https://arxiv.org/abs/2201.08239[59] Suzanne Tolmeijer、Naim Zierau、Andreas Janson、Jalil Sebastian Wahdatehagh、Jan Marco Marco Leimeister和Abraham Bernstein。2021年。默认女性化？——探索语音助手性别和音高对特质归因与信任归因的影响。载于*2021*年*CHI*计算机系统中的人为因素会议扩展摘要（日本横滨）（*CHI EA '21*）。计算机协会，美国纽约州纽约市，文章455，7页。https://doi.org/10.1145/ 3411763.3451623[60] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser和Illia Polosukhin。2017年。注意力机制是你所需的一切。载于神经信息处理系统进展，I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan和R. Garnett（编），第30卷。Curran Associates公司。https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf[61] Julia Velkovska和Moustafa Zouinar。2018年。自然对话的幻觉：在家庭环境中与智能助手互动。载于*2018*年*CHI*计算机系统中的人为因素会议论文集。[62] Xuezhi Wang、Haohan Wang和Diyi Yang。2021年。衡量并提升自然语言处理模型的鲁棒性：一项问卷调查。*arXiv*预印本 *arXiv:2112.08313*（2021年）。[63] Jason Wei、Maarten Bosma、Vincent Y Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M Dai和Quoc V Le。2021年。微调语言模型是零样本学习器。*arXiv*预印本 *arXiv:2109.01652*（2021年）。[64] Peter Winter和Annamaria Carusi。2022年。"如果你要信任机器，那么这种信任就必须建立在某种基础之上"：开发用于肺动脉高压（PH）早期诊断的人工智能（AI）中的验证与信任共同建构。科学与技术研究 35，4（2022年），58–77。[65] 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 eliciting 显式用户反馈？*ACM*计算机与人机交互论文集5，CSCW2，文章388（2021年10月）， 24 10 1145 3479532 页。https://doi.org/。/。Samuel Yang、Jennifer Lee、Emre Sezgin、Jeffrey Bridge、Simon Lin等。2021年。语音助手提供的产后抑郁症临床建议：使用苹果Siri、亚马逊 Alexa、Google 助理和微软Cortana进行的横断面调查。*JMIR移动健康与电子健康* 9，1（2021年），e24045。[67] Ming Yin、Jennifer Wortman Vaughan和Hanna Wallach。2019年。理解准确性对机器学习模型信任的影响。载于*2019*年*CHI*计算机系统中的人为因素会议论文集（英国苏格兰格拉斯哥）（*CHI '19*）。计算机协会，美国纽约州纽约市，1–12。https://doi.org/10.1145/3290605.3300509[68] Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela和Jason Weston。2018年。对话智能体个性化：我有一只狗，你也养宠物吗？*arXiv*预印本 *arXiv:1801.07243*（2018年）。