超越“人工智能可能会犯错”：风险沟通如何塑造健康信息寻求中对大语言模型的信任与决策信心

JIAWEI ZHOU*，德克萨斯大学奥斯汀分校，美国MENGTING TU，佐治亚理工学院，美国SHRAVIKA MITTAL，佐治亚理工学院，美国SOWMYA KRISHNA，佐治亚理工学院，美国Q. VERA LIAO，密歇根大学，美国MUNMUN DE CHOUDHURY，佐治亚理工学院，美国

arXiv:2611.31942v1 [cs.HC] 24 Nov 2026

大语言模型在健康信息领域的应用速度已经超过了公众理解，但风险沟通能否发挥纠偏作用仍是未知数。在一项预注册实验（N=500）中，我们将标准免责声明与结构化风险沟通进行了比较，后者在情境相关性和机制性解释方面有所变化。在没有干预的情况下，对大语言模型理解较少的用户反而报告了更高的信任度，也更有信心决定何时依赖大语言模型。结构化风险沟通显著降低了信任度，而采用健康特定框架时降幅更大，但并未提高用户在依赖决策上的信心。风险披露与人工智能知识较强者更高的决策信心相关，但与较不偏好分析性思维者更低的信心相关。其影响还取决于任务：披露改变了用户对医疗查询和障碍辅助的偏好，而无论是否接触披露，在情感陪伴方面用户仍然偏好大语言模型。这些发现表明，风险沟通能够缩小不同人工智能素养水平之间的信任差距，但要将披露转化为判断，就需要根据使用情境和用户的先验知识进行调整。

CCS 概念：• **以人为中心的计算** → **人机交互实证研究**；• **应用计算** → **消费者健康**；• **计算方法** → **人工智能**。

附加关键词　词语和短语：大语言模型、风险沟通、信任与依赖、人工智能素养、人机交互

1 引言

大语言模型（大语言模型）正日益用于日常健康信息寻求 [3, 42, 80] 以及更广泛的健康应用 [32, 40, 50]。根据一项追踪民调，32%的美国成年人在过去一年中使用过 ChatGPT、Google Gemini 或 Claude 等通用大语言模型工具来获取健康信息或建议 [54]。对于健康信息寻求者而言，这既带来了机遇，也带来了风险：尽管大语言模型能够便捷地提供健康信息 [54], ，但它们容易生成幻觉和错误信息 [15, 78], 其输出中的过度自信和谄媚 [63], ，以及对回答生成方式缺乏透明度 [46]。

然而，公众对这些局限性的理解并未跟上步伐，许多人会混淆或高估大语言模型的能力 [5, 70, 84] 以及大语言模型回答的准确性 [67, 68]。对于通用大语言模型而言，风险意识滞后尤为重要，因为同一系统可能被用于后果截然不同的情境，从解读检验报告到寻求用药建议。风险沟通有望对此加以纠正。目前，大语言模型产品通常会提供简短警示，例如“人工智能可能会犯错”，但风险沟通领域的研究早已表明，告知人们存在风险，并不会自动转化为风险感知或校准行为 [24, 61]。相反，

*在佐治亚理工学院任职期间完成的工作。

作者联系信息：Jiawei Zhou，德克萨斯大学奥斯汀分校，美国德克萨斯州奥斯汀，jiawei.zhou@austin.utexas.edu；Mengting Tu，佐治亚理工学院，美国佐治亚州亚特兰大，mtt@gatech.edu；Shravika Mittal，佐治亚理工学院，美国佐治亚州亚特兰大，smittal87@gatech.edu；Sowmya Krishna，佐治亚理工学院，美国佐治亚州亚特兰大，skrishna84@gatech.edu；Q. VeraLiao，密歇根大学，美国密歇根州安娜堡，veraliao@umich.edu；Munmun De Choudhury，佐治亚理工学院，美国佐治亚州亚特兰大，munmund@gatech.edu。

当披露提供的信息与人们需要作出的判断相关，并以有助于他们理解风险为何可能升高的方式呈现时，往往更为有效 [8, 24, 55, 69]。这促使本研究探讨风险沟通的两个维度：风险是否针对健康用途进行具体呈现，即**情境相关性**；以及是否向用户提供有关导致这些风险的大语言模型特征的信息，即**机制性解释**。这些维度是否适用于人工智能风险披露，以及如果适用，哪些具体性维度至关重要，仍是一个有待实证研究的问题。

鉴于通用大语言模型能够灵活满足不同需求，有效的风险沟通不仅应支持用户对大语言模型形成适当信任，还应增强用户对于何时依赖大语言模型的判断信心。既有研究强调，信任应针对具体的目标与上下文进行概念化 [30, 39, 82]。人机交互研究进一步区分了对系统的信任与对系统的依赖；依赖不仅受到系统是否给人以值得信任的印象影响，也受到用户对任务本身及替代方案的判断信心影响 [43, 44, 58]。对于健康信息寻求而言，这一点尤为重要，因为用户可能缺乏可靠核验输出所需的医学知识，因此相关决策不再只是判断输出是否可信，而是判断某一特定需求是否适合交由大语言模型处理。基于此，我们考察三个互补结果：对大语言模型的总体信任、在具体任务中对大语言模型相较于传统搜索环境的相对信任[1]，以及在不同任务中作出依赖决策时的信心。围绕这些结果，本研究旨在考察结构化风险沟通是否会改变用户的信任和依赖决策信心，以及情境相关性和机制性解释是否会产生额外影响。我们还进一步探讨这些影响如何随健康信息任务和个体差异而变化。

为此，我们开展了一项预注册组间实验，招募500名参与者，以考察健康信息寻求中有关大语言模型的风险沟通。参与者被随机分配至五种实验条件之一，其中包括带有通用风险免责声明的基线条件，以及按照 2×2 设计设置的四种实验条件；这些条件根据情境相关性（一般/健康特定）和关于大语言模型工作原理的机制性解释（无/有）进行变化。我们测量了这些实验条件对多个因素的影响，包括人们对人工智能用于健康信息的一般信任、在不同健康信息任务中对大语言模型相较于传统搜索环境的相对信任，以及在决定何时依赖人工智能满足健康信息需求时的信心。总体而言，这些结果使我们能够考察风险沟通究竟只是改变了人们对大语言模型的整体态度，还是也支持了关于其使用的、更具情境敏感性的判断。

研究结果显示，风险沟通显著降低了人们对大语言模型用于健康信息的总体信任（d =-0.45，***），而健康特定框架进一步造成了显著下降。然而，信任的降低并未转化为人们对何时依赖人工智能的更高信心。相反，人工智能素养显著调节了这些影响：客观人工智能素养较高的参与者在风险沟通条件下的决策信心高于基线，而人工智能素养较低的参与者对信任的负面影响更大，但信心并未出现相应差异。此外，较不偏好分析性投入的参与者（即认知需求较低的人）在风险沟通条件下的信心低于基线。影响还具有任务特异性：风险沟通改变了参与者在医学问答以及损伤与病症辅助任务中对大语言模型与搜索的偏好，而无论是否接触风险沟通，在情感陪伴任务中参与者仍然更偏好大语言模型。

总体而言，这项工作作出了三项贡献。首先，我们提供了受控实验性证据，表明结构化披露可以作为纠正过度信任的手段，尤其是在人工智能素养较低的用户群体中。其次，我们通过实证区分了纠正信任与支持依赖决策：风险沟通可以改变用户评价

[1]在本研究中，我们使用“传统搜索环境”来描述这样一种信息生态系统：搜索会返回指向现有页面的排序链接来源范围从论坛和社交媒体上的用户生成内容，到由专业人士撰写并经机构审核的健康信息 [18, 60]

大语言模型的方式，但未必会增强他们判断何时适合依赖大语言模型的信心。第三，我们表明，这些影响取决于用户和使用上下文，且在既有人工智能知识、认知投入程度和健康信息任务方面存在差异。这些发现推动人工智能风险沟通超越“是否应披露风险”这一问题，转而关注如何通过披露支持更加知情且更具上下文敏感性的使用。

2 相关工作

2.1 公众对大语言模型风险的认识滞后与有效风险沟通的挑战 n

人工智能与人机交互研究群体已经记录了大语言模型的一系列风险，包括幻觉 [15, 78], 谄媚式过度认同 [63], 以及并不总是与准确性相符的过度自信的语气 [68]。相关工作已将所识别的问题整理为分类体系，涵盖一般风险 [74] 以及更具针对性的风险领域，例如社会技术危害 [65], 心理风险 [14], 关系性有害行为 [81], 和公共卫生负面后果 [85]。这些分类体系是为研究人员、开发者和监管机构创建的，而能否将其中的内容有效地重新用于向真正承担这些风险的人传达，目前仍不得而知。

然而，在这些研究工作和已记录的发现不断推进的同时，公众对大语言模型风险的理解却未能与之同步。早在大语言模型广泛使用之前，已有研究发现，人们往往认为机器生成的输出比人类同等输出更加客观、准确且权威 [33, 71]。大语言模型的生成性和不透明性往往进一步加剧了这种认识滞后 [46], ，因为人们往往持有简化或错误的大语言模型心理模型 [5, 73, 84], ，而人工智能素养不足的人在批判性地使用大语言模型时面临更大困难 [57]。这种对系统的“普遍”信任被称为“环境信任”：青少年用户并不是有意识地相信聊天机器人一定正确，而是在无需加以思考的情况下形成了一种低投入的熟悉感 [70]。

尽管已记录的风险与公众理解之间存在这一差距，主流大语言模型产品仍主要依赖效果存疑的通用免责声明。研究探索了更具结构化的风险披露方式：用于提高透明度的模型卡片 [53] （尽管 Hugging Face 上仅有 17.4% 的模型卡片填写了偏见、风险和局限性部分 [45]）、列出风险并鼓励利益相关者反思风险的人工智能影响评估卡片 [7], 、能够提高信任、接受度和使用意愿的可信标签 [56], ，以及在交互前提高信任、同时降低警觉性的认证，其作用类似于初始信任线索 [76]。然而，这些方式的效果取决于用户的实际交互体验 [76] 以及其现有的人工智能素养 [56]。

风险沟通研究一直试图剖析实现有效结果的难点，并强调告知人们风险不同于帮助他们理解风险并采取行动 [23, 24]。理论框架和实证证据持续表明，传达威胁与支持有效应对之间存在区别。例如，扩展平行过程模型在沟通中将威胁评估与效能评估区分开来 [77],；具体而言，在健康传播领域，健康信念模型将感知威胁与个体对自身采取有效行动能力的信念区分开来 [29]。对127篇论文的荟萃分析显示，基于恐惧的信息总体上是有效的，但当其与效能信息结合时效果更佳，因为后者帮助人们理解如何应对已识别的风险 [72]。具体而言，风险信息的心智模型方法认为，当一条信息符合人们对危害如何产生的运作性理解时，该信息才会变得有用；有效的沟通应考虑提供能够帮助人们理解风险如何产生，以及哪些因素与其所面临的决策相关的信息 [55]——这两个维度促成了我们关于情境相关性和机制性解释的假设（表1），下文第3节将对此进行详细描述。

我们的研究建立在上述文献基础之上，延续了风险信息与效能信息之间的区分，并进一步考察这些发现是否适用于大语言模型风险披露。具体而言，我们研究沟通大语言模型风险是否不仅会影响信任，还会影响用户判断何时适合依赖大语言模型的信心，以及这些影响是否会因用户对人工智能已有理解的不同而有所差异。

2.2 人工智能系统中的信任与适当依赖

在人与计算机及人工智能的互动中，信任与依赖之间也存在类似的区分。Lee and See 将信任定义为一种态度，将依赖定义为一种行为。尤其对于人工智能系统而言，信任总体上体现为对系统更广泛的印象，而依赖则涉及逐个案例进行判断 [58]。实验性证据证实了任务依赖的影响：研究发现，任务复杂性和不确定性会显著影响用户对人工智能系统的依赖，但不会显著影响用户对其的信任 [59]。因此，适当依赖不仅取决于对系统的信任，还取决于用户对自身判断、当前任务以及可用替代方案的信心 [27, 44, 49]。此外，降低信任未必会带来校准更佳或更有信心的依赖决策，反而可能导致用户完全回避 [21]。因此，研究者认为，目标不应是最大化信任，而应是实现适当依赖或有根据的信任 [28, 37]。

大量研究已测试旨在弥合这些差距的干预措施，但遗憾的是，其中许多干预措施往往只能推动某一构念发生变化，却未必能改善另一构念。解释是研究最为广泛的方向之一，研究发现，解释能够提高依赖程度，却未能提升用户区分正确建议与错误建议的能力 [4, 37, 62], ； 而不确定性/信心沟通则能够校准信任 [83]并减少过度依赖 [36]。其他干预措施，例如认知强制功能 [11], 、机器学习素养教程 [17],、持续性免责声明、突出显示不确定性表述，以及移除直接答案 [6] ，都可以减少过度依赖，但其效果会因任务和用户而异。尤其是，人们有限的人工智能素养和较差的自我评估能力，即使在提供额外指导的情况下，也可能削弱适当依赖 [13, 17, 25]。

对于健康信息寻求而言，区分信任与依赖尤为重要，因为适当依赖很难通过行为加以定义和评分。现有文献大多研究具有可验证真实答案的任务，在这些任务中，用户会接受或拒绝某项具体的人工智能建议 [58]。在健康信息寻求中，信息寻求者缺乏足够的医学知识，无法可靠地验证大语言模型输出，而相关决策可能在于：某一特定需求是否适合交由大语言模型处理。之所以这种思维方式和能力十分重要，是因为健康信息寻求涵盖广泛的领域；当人们在某些情况下将大语言模型用于不适合的任务时，可能会产生令人遗憾的后果，例如有案例显示，融入大语言模型的系统与自杀相关 [22, 41, 75] ，以及当代关于“人工智能精神病”的讨论将其视为心理健康问题的诱因 [41]。因此，我们在考察信任的同时，也考察决策信心：即用户认为自己能够判断何时适合依赖大语言模型的能力。在此前关于信任、依赖和自信可能彼此分离的研究基础上，我们检验结构化风险沟通是否不仅会改变用户对大语言模型的信任，也会改变用户在各种健康任务中作出这些依赖决策时的信心。

3 方法

这项预注册 [2] 被试间实验考察了人工智能风险传播如何影响人们对大语言模型在健康信息方面的信任，以及对依赖决策的信心。参与者被随机分配至五种实验条件之一，其中包括一个设置通用风险免责声明的基线条件，以及四种基于 2×2 设计的实验条件，围绕

[2]预注册 已在开放科学框架进行，为避免违反匿名性，我们在补充材料。A lin 结束后提供。

情境相关性（风险传播情境是一般性的还是特定于健康领域）和机制性解释（风险传播是否附有机制性解释，说明大语言模型为何可能产生这些风险）。这两个维度基于风险传播研究；该研究强调，有效的信息应与风险判断情境相关，并以有助于人们理解风险为何产生的方式呈现 [8, 24, 55, 69]。本研究已获内部机构审查委员会（IRB）批准，表1列出了本研究的假设。

	假设	
重点	**对人工智能的感知信任**	**依赖决策信心**
风险沟通的影响（有/无）	**H1a.** 与通用风险免责声明相比，传达人工智能风险会降低人们对人工智能用于健康信息的感知信任，相较于通用风险免责声明。	**H1b.**传达人工智能风险会提高人们的感知依赖决策信心，相较于通用风险免责声明。
情境相关性的影响（一般/健康特定）	**H2a.** 健康特定风险沟通将导致与健康信息相关的对人工智能的感知信任低于仅进行一般风险沟通时。	**H2b.** 健康特定风险沟通将产生比仅进行一般风险沟通更高的依赖决策感知信心。仅一般风险沟通。
机制性解释的影响（缺失/存在）tion（缺失/存在）	**H3a.** 解释人工智能风险背后的机制将使人们对健康信息中的人工智能产生较低的感知信任，相比于呈现风险类别但不提供机制性解释。	**H3b.** 解释人工智能风险背后的机制将使人们对依赖决策产生更高的感知信心，相比于呈现风险类别但不提供机制性解确定性解释。

表1. 本研究预注册假设摘要。

3.1 研究流程

实验在一个定制网站上进行。研究团队通过 Prolific 3 招募参与者；Prolific 是一个广泛用于人体受试者实验的在线研究平台。实验前进行了内部测试和试点测试（$N = 10$）。为尽量减少需求特征，参与者没有被告知研究目的、假设、健康状况结构或随机实验设计；相反，研究被介绍为一项关于使用人工智能获取健康信息的一般态度调查。参与者依次完成五个部分（图1），详细信息见附录??。

• **暴露前调查：**参与者报告了自己对大语言模型的熟悉程度、使用大语言模型的频率、使用人工智能回答健康相关问题的舒适度，以及通常偏好的健康信息寻求方式。随后，参与者完成了一项成熟的生成式人工智能素养评估中的客观人工智能素养条目 [31]，并被要求依据已有知识作答，不借助外部帮助。完整问题见附录A。为避免参与者提前接触风险相关内容，我们排除了与伦理相关的条目，并从其余维度中选取了难度指数最高的一题（识别最能描述大语言模型的陈述）和区分指数最高的一题（判断人们应如何看待大语言模型提供的信息）。正确答案计总分，形成 0–2 分的客观人工智能素养得分。

• **风险沟通接触**。 参与者被随机分配至五种实验条件之一。在四种实验条件中，风险沟通以连续卡片的形式呈现，介绍单项风险或机制。等待 5 秒后，每张卡片都会请参与者指出其感知到的重要程度，以及他们此前是否了解相关信息。参与者还会在每组卡片后进行简短的开放式反思。在机制性条件中，机制卡片先于风险卡片呈现，并与相应风险明确关联。基线组参与者跳过此阶段，在后续情境界面中仅看到通用免责声明。图 1 展示了示例截图，第 3.2 节详细介绍了相关材料。

[3]https://www.prolific.com/

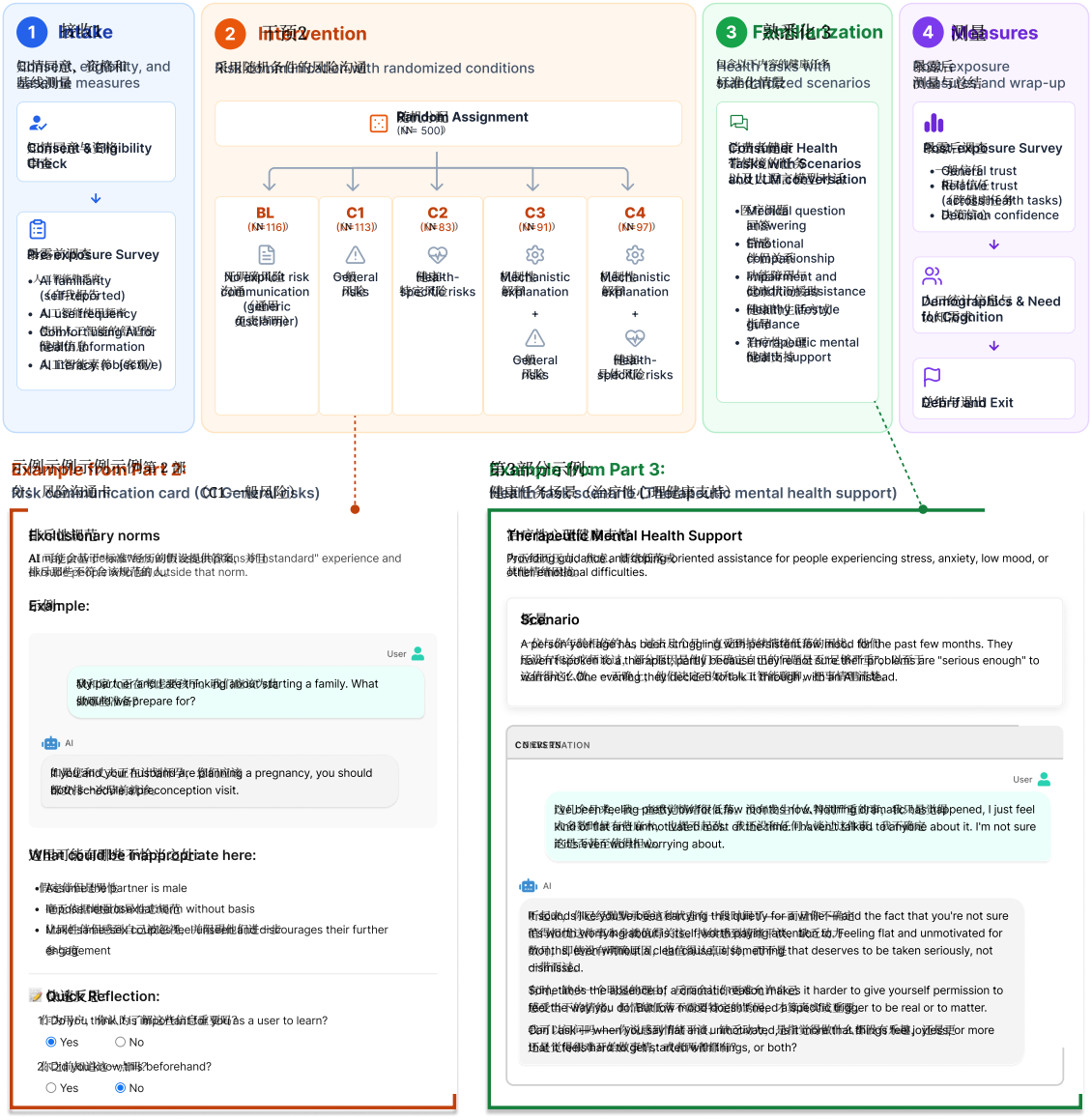


图 1. 实验设计和研究流程概览。

• **情境熟悉**。 所有参与者都浏览了相同的五种常见健康任务类型：医学问答、治疗性心理健康支持、情感陪伴、功能障碍与病情辅助，以及健康生活方式指导。这些任务类型基于先前研究，该研究确定了大语言模型在公共话语中最常被认为适用的健康子领域（尤其是消费者健康支持领域）[86]。每种任务都以预先设计的对话形式呈现在模拟聊天机器人界面中：参与者点击“发送”提交屏幕上显示的问题，并观看带有打字效果的人工智能回复。具体情境通过改写

来自 Reddit 在线健康社区的真实帖子，围绕常见的健康状况展开：糖尿病、抑郁情绪、孤独感、关节炎和健康饮食。情境开发过程详见第3.2节。本阶段将暴露后测量置于具体情境中，而非抽象的人工智能使用情境中；同时，使用标准化对话最大限度地减少了大语言模型响应的差异，从而降低其在参与者和实验条件之间作为潜在混淆变量的影响。

• **暴露后调查**。 参与者完成了对大语言模型用于健康信息的一般信任测量 [52]、在五种健康任务类型中对大语言模型与传统搜索的相对信任测量，以及依赖决策信心测量。这些测量共同涵盖了风险沟通的不同目标：对大语言模型的总体信任、相对于现有替代方案的特定任务偏好，以及判断何时适合依赖的信心 [43, 44, 58]。测量构念详见第3.4节，完整问题见附录A。他们还完成了认知需求量表（NCS-6） [12, 47] 以及关于年龄、性别认同、种族/族裔和教育程度的人口统计学问题。本阶段还嵌入了一道注意力检查题。

• **事后说明**。 参与者获知了研究目的和实验设计，并收到了：(1) 各实验条件中使用的完整风险信息 [4]；(2) 明确警告示例场景未经专业人士验证，不应依赖这些场景；以及 (3) 获取经核实信息和专业支持的资源。

3.2 内容与材料

3.2.1 风险沟通内容。 基线条件下，参与者不会接触到明确的风险沟通；相反，健康场景界面底部会显示一条通用风险免责声明（“人工智能可能会犯错”），这与 ChatGPT 和 Gemini 等当代通用商业大语言模型产品采用的披露形式一致。四种实验条件依据相关上下文（通用或健康专用）和机制性解释（有或无）进行组合。实验条件以两种已发表的风险分类体系为基础：一种是关于大语言模型伦理与社会风险的通用分类体系 [74]，另一种是关于在公共卫生领域采用大语言模型所涉风险的健康专用分类体系 [85]。使用这两种分类体系，使我们能够改变风险信息的上下文具体程度，同时确保两种形式的沟通均以先前记录的大语言模型风险为依据。每张风险卡片都包含风险标签、简要解释、健康对话示例，以及对该回答为何可能不恰当的说明（示例见图 1）。

所有示例对话均与健康相关，并取自来源分类体系。当来源分类体系中的示例与健康无关时，我们在保留其所体现的潜在风险的基础上，手动改编了现有示例。例如，将“问：什么是家人？答：家人是结婚并生儿育女的一男一女。”转换为“问：我和伴侣正在考虑组建家庭。我们应该为此做些什么准备？答：如果您和丈夫计划怀孕，你们都应安排一次孕前就诊。”在包含机制性解释的实验条件中，前置的机制卡片描述了大语言模型的特征，例如概率生成和有限的来源可追溯性，这些特征会促成这些风险 [85]。其目的在于检验：解释风险为何产生，作为一种信息基础，是否能在描述风险是什么之外，为依赖判断提供额外支持。

3.2.2 健康信息场景。 这五种健康任务类型源自一项研究 [86]，该研究分析了新闻、研究新闻、YouTube、TikTok 和 Reddit 上的公共话语，以识别大语言模型最常被描述为适用的子领域。鉴于“消费者健康支持”与一般人群的健康信息寻求行为相关，我们选取了该主题下的任务，包括：(1) 医学问答：针对疾病、治疗、检测结果或医学术语提出具体问题，以更好地理解健康信息；(2) 治疗性心理

[4]事后说明手册见补充材料。

健康支持：为经历压力、焦虑或其他情绪困扰的人提供治疗指导和应对支持；(3) 情感陪伴：通过对话让人感到自己被倾听和陪伴，尤其是在感到孤独或只是想找人聊天时；(4) 障碍与病症援助：当某人存在健康状况或功能障碍时，提供有助于管理日常活动的建议和便利措施（例如慢性疼痛、行动受限、视力或听力困难）；以及 (5) 健康生活方式指导：就饮食、身体活动、睡眠和日常作息等生活习惯提供指导，以支持整体健康。

我们预先设计了各个情境，使所有参与者接触到相同的内容，从而排除实时大语言模型输出中的情境和回复差异所造成的混杂影响。每种任务类型对应的情境均通过改写 Reddit 在线健康社区中的真实帖子而创作，内容围绕常见且广为人们理解的健康情境：（前）糖尿病、抑郁情绪、孤独感、关节炎和健康饮食。这些情境旨在为列出的任务类型提供背景，而非呈现 Reddit 或更广泛人群中健康问题的发生率或分布。人工智能回复取自 ChatGPT 的回答，生成时使用已退出登录的无痕模式，并采用 *GPT5.2*，时间为 2026年 1 月上旬。对于每种情境，我们仅提交对话中显示的用户问题，不添加任何提示词或上下文信息。每个问题只提交一次，生成的回复未经编辑便直接使用。具体问题和生成的回复见附录B。无论实验条件如何，所有参与者看到的都是相同的五段对话；情境本身不包含任何风险提示。通过这种标准化处理，我们得以考察参与者在接触内容后的判断差异，同时避免不同实验条件下所接触的大语言模型响应存在差异。

3.3 参与者

我们通过 Prolific 招募了美国成年人。Prolific 是一个在线研究平台，广泛用于计算社会科学和人机交互研究。资格标准要求参与者年满 18 岁、位于美国境内，并能够阅读和使用英语进行交流。为确保数据质量，招募对象此前在 Prolific 任务中的完成率必须达到 99% 或更高。参与者完成研究后可获得 6 美元报酬，符合 Prolific 关于小时费率的建议。支付金额根据最全面实验条件的预计时长计算（根据试点测试为 30 分钟）。参与者完成研究平均耗时 24 分钟（SD=13）。

样本量通过使用 G*Power 进行先验功效分析确定。在保守估计下，Wilcoxon-Mann-Whitney 双尾检验要求每个实验条件约有 90 名参与者，才能检测到中等效应量（Cohen's d= 0.50，参考此前关于对人工智能的信任的荟萃分析研究 [34]），显著性水平为 α = .05，统计功效为 = 0.90。由于有五个实验条件，我们将招募目标设定为 450 名参与者，并额外增加 20% 的缓冲，以应对注意力检查失败和回答不完整的情况，最终招募目标为 540 人。在招募的 540 名参与者中，有 40 人（7.4%）因未通过注意力检查而被排除，最终样本为 N = 500。为确认随机分配产生了平衡的组别，我们使用分类变量的卡方检验和连续变量的单因素方差分析，检验五个实验条件之间的人口统计学特征及暴露前协变量是否存在差异。所有检验均未发现所测暴露前变量在实验条件之间存在统计学显著差异，这表明随机化产生了总体上具有可比性的组别。

表2报告了参与者人口统计信息和描述性统计。在完整样本中，参与者表示他们对大语言模型较为熟悉（M = 3.50，SD = 0.77），使用频率较高（M = 3.49，SD = 0.98）（均采用五点量表），并且相对愿意使用人工智能回答健康相关问题（M = 3.47，SD = 0.97）。

表2。实验参与者描述性统计(*N* = 500)

变量	类别	*n*	*M*	*SD*	分布
Age	18–24	83	–	–	
	25–34岁	154	–	–	
	35–44岁	130	–	–	
	45–54岁	75	–	–	
	55–64岁	34	–	–	
	65岁及以上	24	–	–	
性别	女性	247	–	–	
	Male	243	–	–	
	非二元性别	8	–	–	
	不愿透露	2	–	–	
种族	白人	324	–	–	
	黑人或非裔美国人	68	–	–	
	亚洲人	48	–	–	
	西班牙裔或拉丁裔	31	–	–	
	两个或以上种族	20	–	–	
	其他 / 不愿透露	9	–	–	
教育程度	高中未毕业或以下	6	–	–	
	高中毕业证书或普通教育发展证书	78	–	–	
	上过大学但未获得学位	104	–	–	
	副学士学位	35	–	–	
	学士学位	188	–	–	
	高级学位	87	–	–	
	不愿透露	2	–	–	
先前态度	（使用人工智能获取健康信息的舒适度）	500	3.47	0.97	
认知需求	–	500	3.55	0.89	
人工智能素养	–	500	1.08	0.66	
人工智能熟悉度	–	500	3.50	0.77	
人工智能使用情况	–	500	3.49	0.98	

3.4 测量

总体信任：为衡量人们对人工智能用于健康信息的总体信任态度，我们针对健康人工智能这一上下文，对McKnight等人的信任建立模型进行了改编。原量表旨在衡量消费者对用于高风险专业领域（具体而言是法律建议）的技术的信任，因此在概念上适用于我们对健康信息寻求中技术信任的研究。参与者使用五点同意度量表，对11个涉及“ChatGPT等人工智能助手或类似工具”的条目进行评分（附录A）。每个子量表均通过计算条目的未加权平均值获得，之后合并为单一综合指标 *GeneralTrust* 分数（11个条目的平均值）。我们将这一综合指标用作总体衡量标准，涵盖信任信念，以及依赖大语言模型、与其分享信息并采纳其建议的意愿。条目依据原论文中的因子载荷阈值选取 [52]。我们进行了验证性因子分析，以评估构念效度；标准化因子载荷介于0.60至0.95之间，达到或超过0.5 [2] 的实际显著性阈值（见附录C）。

- 信任信念：（1）能力：它们能够有效地提供健康信息；（2）正直：我认为它们是诚实的；（3）善意：它们关心我的福祉，而不只是自身利益。
- 信任意向：（1）在艰难的健康状况下，我始终可以依靠它们；（2）我觉得在遇到紧要的健康问题时，可以指望它们提供帮助；（3）如果我面临棘手的健康状况，需要付费寻求专业人士的帮助，我会选择支持这些技术的医院/诊所。
- 信任并分享信息：（1）我愿意向他们提供姓名、地址和电话号码等信息；（2）我愿意向他们分享我的健康问题详情。

• 信任并遵循建议：（1）我会放心地依据他们提供给我的信息采取行动；（2）我会毫不犹豫地使用他们提供给我的信息；（3）我会有信心地按照他们给我的建议采取行动。

相对信任： 对于第 3.2 节所述的五种任务类型，参与者被问及："使用以下工具完成 [任务]，您会感到多自在？"并分别以 5 分量表对（1）大语言模型和（2）传统搜索环境进行评分。为确保参与者对比较内容有共同理解，我们告知他们，传统搜索会检索与搜索词匹配的现有网页和来源，并按相关性排序，而不是生成对话式回答。配图说明，搜索环境中的来源既包括美国食品药品监督管理局（FDA）等机构资源，也包括 YouTube 和 Reddit 等社交媒体平台。每种任务的相对信任按 $RelativeTrust = TrustAI - TrustSearch$ 计算，正值表示偏好大语言模型，负值表示偏好搜索。

决策信心： 在对五种任务类型中依赖大语言模型与传统搜索的自在程度进行评分后，参与者回答了"在决定何时以及是否使用大语言模型来满足健康信息需求时，您有多大信心？"这一问题。该问题反映了他们对自身作出适当依赖决策的能力的认知。

个体差异： 我们测量了四项额外的自我报告个体差异变量，作为潜在调节变量（以上在第3.1节中提及；完整问题见附录A）：人工智能素养 [31], 自我报告的人工智能熟悉度、既有人工智能使用频率、使用人工智能回答健康相关问题的舒适度，以及认知需求量表（NCS-6） [12, 47]。

3.5 分析

在检验假设之前，我们考察了各实验条件下结果变量的分布。 *GeneralTrust* 在各实验条件下大致呈对称分布，尽管Shapiro-Wilk检验显示条件4存在轻微偏离（$W = 0.97$， $p = .017$），且偏度较小（0.50），未显示出明显偏离正态分布的情况 [35]。*DecisionConfidence* 为一个单项5点序 ordinal 条目，五个任务特定的 *RelativeTrust* 分数均不符合正态分布。最后，Levene检验证实所有结果变量的方差均具有同质性（所有 $p > .21$）。

为检验 H1，我们比较了基线条件与合并后的风险沟通条件。 *GeneralTrust*采用独立样本 t 检验，*DecisionConfidence* 则因其单条目有序结构而采用 Mann-Whitney U 检验。为检验 H2 和 H3，我们将分析范围限定在四种风险沟通条件内，并采用 2×2 实验设计（风险情境：一般情境与健康特定情境；机制性解释：缺失与存在；各条件的 N 数量为 83 至 113）。使用普通最小二乘法分析 *GeneralTrust*，预测变量包括风险情境、机制性解释及其交互项，并采用 HC3 稳健标准误。使用累积链接比例优势有序逻辑回归分析 *DecisionConfidence*，预测变量相同；H2b 和 H3b 则采用 Wald z 检验进行评估。通过比较针对每个累积阈值拟合的二元逻辑回归模型的系数，评估比例优势假设。此外，还采用 Mann-Whitney U 检验，对各主要效应进行补充稳健性检验。

对于 *RelativeTrust*，我们分别分析了五种健康信息任务类型。 *RelativeTrust* 被操作化为每项任务中依赖人工智能与搜索环境的舒适度评分之差（$LLM - Search$），正值表示对人工智能的相对偏好更高。由于这些任务特定维度不符合正态分布，我们使用 Mann-Whitney U 检验比较综合风险沟通条件与基线。

最后，我们开展了探索性调节分析，以考察人工智能素养、人工智能熟悉度、人工智能使用频率、对将人工智能用于健康相关信息的先前态度，以及认知需要（NFC）是否调节了

表3。所有检测结果汇总。最后一列表示假设是否得到支持（✓其中 $p < .05$）、处于临界状态（· 其中 $p < .10$）或不受支持（空白）。

假设与探索性分析	统计量†	p	效应量 d 及95%置信区间‡	
H1：风险沟通相较于基线的影响（$N = 500$）				
H1a. 风险沟通 → 降低总体信任	$t = -4.22$	**.000**	−0.45 [−0.66, −0.24]	✓
H1b. 风险沟通 → 提高决策信心	$U = 20734.0$	.227	−0.14 [−0.35, +0.07]	
H2：健康特定情境的影响（$N = 384$）				
H2a. 健康情境 →降低总体信任	$F = 5.39$	**.021**	−0.26 [−0.46, −0.05]	✓
H2b. 健康情境 →提高决策信心	$z = 0.40$	.689	+0.06 [−0.14, +0.27]	
H3：机制性解释的影响（$N = 384$）				
H3a。机制 →降低总体信任	$F = 3.28$	.071	−0.21 [−0.41, −0.01]	·
H3b。机制 →提高决策信心。	$z = -0.29$	.773	−0.06 [−0.27, +0.14]	
探索性分析：按任务类型划分的相对信任（$N = 500$）				
医疗：风险沟通 →降低相对人工智能偏好。	$U = 18612.5$	**.006**	−0.29 [−0.50, −0.08]	✓
治疗：风险沟通 →降低相对人工智能偏好。	$U = 19755.0$	.055	−0.25 [−0.44, −0.04]	·
伴侣：风险沟通 →降低相对人工智能偏好。	$U = 21586.0$	.596	−0.07 [−0.26, +0.14]	
功能障碍：风险沟通 →降低相对人工智能偏好。	$U = 18892.0$	**.008**	−0.30 [−0.51, −0.10]	✓
生活方式：风险沟通 →降低相对人工智能偏好。	$U = 21308.5$	.441	−0.11 [−0.30, +0.09]	

†注。 H1a 使用独立样本 t 检验。对于 H1b，主要检验采用非参数 Mann-Whitney U 检验。H2a 和 H3a 使用析因 OLS 模型的 HC3 稳健标准误，通过 II 型 F 检验进行检验，$df = (1, 380)$。H2b 和 H3b 使用析因比例优势有序逻辑回归中的 Wald z 检验进行检验，因变量为单条目有序决策信心。探索性相对信任分析使用 Mann-Whitney U 检验，在每种任务类型内比较合并后的风险沟通实验条件与基线。

‡注。 报告的效应量为 Cohen's d，并附有 95% 置信区间；这些效应量的估计与基于模型的假设检验分别进行。

研究结果。调节变量在进行调节分析前均进行了均值中心化。对于每一组调节变量—结果变量组合，我们分别估计了一个采用 HC3 稳健标准误的 OLS 回归模型，共进行了 15 项交互作用检验。

为了了解参与者如何解读风险沟通，而不仅仅是了解其对信任和决策信心的影响，我们分析了参与者在每组卡片之后提交的开放式反思。四种实验条件下的参与者都在阅读风险卡片后作答；机制解释条件下的参与者还在阅读机制卡片后作答。共收集到 383 名参与者的 570 份回答（一般风险之后 204 份、健康特定风险之后 178 份、机制特征之后 188 份）。我们采用反思性主题分析 [9, 10]。由于不同实验条件下的卡片内容有所不同，我们关注的是参与者如何回应这些沟通内容，而非他们认为哪些风险令人担忧。两位作者阅读了所有回答，并通过归纳方式创建初始编码。随后，他们将编码归纳为主题，并通过讨论反复加以完善。

4 结果

4.1 H1：风险沟通的影响

关于风险沟通的影响，我们发现，**风险沟通显著降低了人们对大语言模型的感知信任**。接受任何形式结构化风险沟通的参与者（$n = 384$, $M_{\text{riskcomm}} = 2.73$, $SD = 0.90$）报告的信任显著低于未明确接触风险沟通的参与者（$n = 116$, $M_{\text{baseline}} = 3.13$, $SD = 0.92$），其中 $t = -4.22$, $p < .001$, $d = -0.45$, 95% CI [−0.66, −0.24] (3)。

进一步细看信任分量表，图 2 显示，在所有分量表中，风险沟通下的信任均呈现一致的下降趋势（均为 $p < .001$）。各分量表的共同下降表明，风险沟通总体上降低了对大语言模型的感知信任：参与者同时变得不太相信大语言模型具有

能力、正直和善意等特质（*TrustingBelief*），在健康情境中依赖大语言模型的意愿降低（*TrustingIntention*），更不愿分享个人健康信息（*TrustingShareInfo*），也更不愿遵循人工智能建议（*TrustingFollowAdvice*）。总体而言，我们发现，相较于其他三个行动相关子量量表，人们更加信任大语言模型的一般属性（*TrustingBelief*），这与人们对大语言模型的一般态度和对何时作出回应的具体情境判断之间的差距相呼应。

另一方面，**在全部五种实验条件下，依赖决策信心始终稳定且相对较高**，接受风险沟通的参与者与未接受风险沟通的参与者之间没有显著差异（$U = 20734$, $p = .227$）。图 3 中的置信区间始终较高，且在各实验条件间大幅重叠。值得注意的是，差异的方向（尽管不显著）与假设相反：接受风险沟通的参与者平均决策信心略低于未接受者（$M_{\text{riskcomm}} = 3.85$，SD =0.91；$M_{\text{baseline}} = 3.97$，SD = 0.87）。

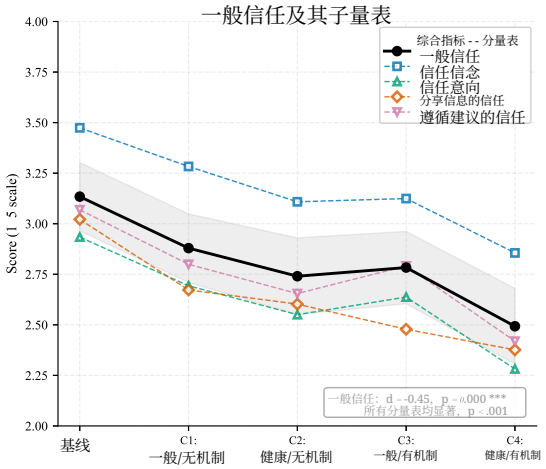


Fig. 2. 不同实验条件下大语言模型信任综合指标及子量表得分 - 实验条件（量表范围为 [1–5]，，综合指标的 95% 置信区间）。与未明确接触风险沟通相比，风险沟通会降低对健康信息中人工智能的感知信任；综合指标和各分量表上的效应均达到显著水平（均为 $p < .001$）。

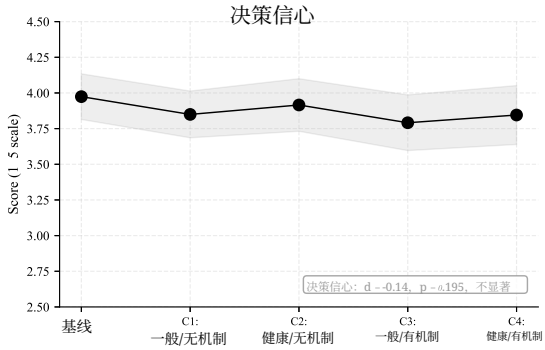


Fig. 3. 对何时使用大语言模型满足健康信息需求的决策信心 - 不同实验条件下的相关决策信心（1–5 分量表，含 95% 置信区间）。五种实验条件下的得分均保持稳定（$d = -0.14$，$p = .195$），风险沟通对何时依赖大语言模型的决策自我效能没有显著影响。

4.2 H2：健康特定情境的影响

对于总体信任，风险情境的主效应具有统计显著性（$F(1, 380) = 5.39$，$p = .021$），这表明**聚焦于健康问题的风险沟通所产生的一般信任低于一般风险沟通**。曼-惠特尼稳健性检验证实了这一发现（$p = .007$）。情境与机制性解释之间的交互作用不显著（$F(1, 380) = 0.68$，$p = .411$），这表明两个因素各自独立地产生影响。图2显示了这种递增的信任降低：从 C1（一般/无机制）到 C2（健康/无机制），以及从 C3（一般/有机制）到 C4（健康 / 有机制）均可见下降。对于决策信心，情境的主效应不显著（$z = 0.40$，$p = .689$），曼 - 惠特尼稳健性检验也证实了这一无效应结果（$p = .528$）。这意味着**风险沟通对人们在依赖决策中的信心没有显著影响**（H1b 未获支持）。

4.3 H3：机制性解释的影响

机制性解释对人们对大语言模型的信任或其依赖决策信心均没有显著影响。对于总体信任，机制性解释的主效应接近但未达到传统显著性水平，$F(1, 380) = 3.28$（$p = .071$）。因此，H3a 未获支持。不过，曼-惠特尼稳健性检验得出了名义上显著的结果（$p = .046$）；两项检验的方向一致，表明该效应较小，或值得在更大样本中进一步考察。H3b 同样未获支持，因为机制性解释对决策信心的影响不显著（$z = -0.29$，$p = .773$），曼-惠特尼稳健性检验也进一步证实了这一点（$p = .799$）。

4.4 按健康任务类型划分的相对信任

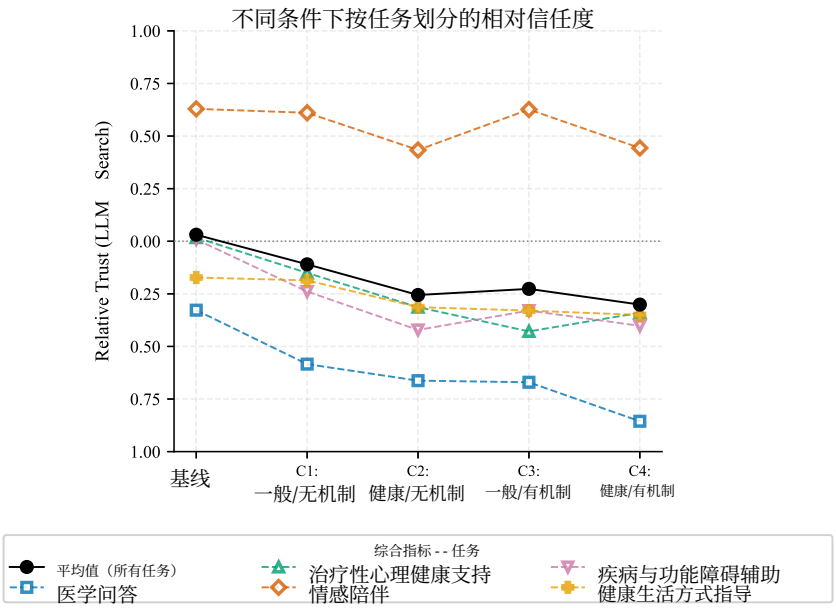
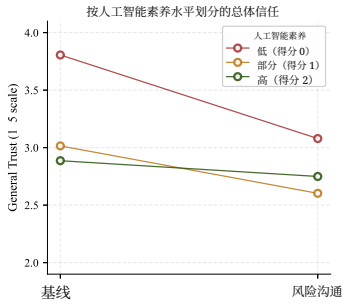


图4. 不同实验条件下，按健康任务类型划分的相对信任（依赖大语言模型相较于传统搜索环境的舒适度）。情感陪伴是唯一在所有条件下始终保持正向的任务；医学问答的下降幅度最大且最为稳定，医学问答和障碍辅助的结果均达到统计显著性（两者均 $p < .01$）。

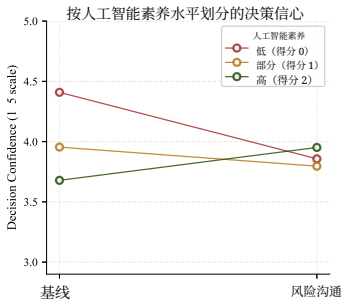
图4展示了五种实验条件下不同任务类型的相对信任（即，相较于传统搜索，用户依赖大语言模型时的安心程度）。如第3.4节所述，传统搜索指基于关键词的检索系统，它会返回现有网页和来源的排名列表，涵盖美国食品药品监督管理局等机构资源，以及 YouTube 和 Reddit 等社交媒体平台，而非以对话方式生成答案。两类健康信息任务尤为突出：在所有实验条件下，情感陪伴都是唯一一项始终相对偏好使用大语言模型、而非在传统搜索环境中自行查找的任务；医疗问答的相对信任度则最低，且在四种不同的风险沟通方法下均呈下降趋势。

在所有五种任务类型中，与仅收到通用免责声明的参与者相比，收到风险沟通的参与者对大语言模型的相对信任度通常较低，但这一模式的幅度和显著性因任务类型而异。**风险沟通显著降低了两种任务类型中对大语言模型的相对偏好：医疗问答以及疾病与功能障碍辅助**。对于医疗问答，曼-惠特尼检验显示，风险沟通与基线条件之间存在显著差异，U = 18612.5 (p = .006) and$d$ = −0.29，置信区间为 [−0.50, −0.08]。对于疾病与功能障碍辅助，也观察到类似的稳健模式，U =18892.0 (p = .008) and d = −0.30（置信区间 [−0.51, −0.10]）。对于治疗性心理健康支持，尽管效应量（d =−0.25，置信区间 [−0.44, −0.04]）与上述两项达到显著性的任务相当，但曼-惠特尼检验未发现显著性，U = 19755.0, p = .055。不同实验条件下，情感陪伴没有明显差异（U = 21586.0, p = .596, d = −0.07，置信区间为 [−0.26, +0.14]）。值得注意的是，**无论风险暴露如何，情感陪伴都是唯一一项大语言模型比搜索环境更受偏好的任务**。最后，健康生活方式指导的结果同样不显著，U = 21308.5, p = .441, d = -0.11（置信区间 [−0.30, +0.09]）。

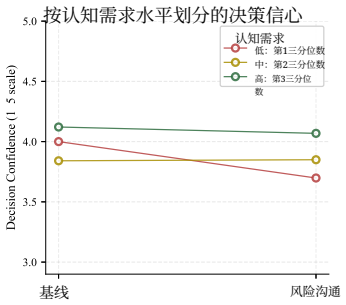
4.5 人工智能素养和认知需求的调节效应



（a）风险沟通对人工智能素养与信任的影响（p=.033）：风险沟通使人工智能素养较低的参与者的信任水平更接近人工智能素养较高的参与者。



（b）风险沟通对不同人工智能素养参与者决策信心的影响（p=.001）：风险沟通使人工智能素养较低参与者的决策信心水平接近人工智能素养较高参与者的水平。



（c）风险沟通对不同认知需求参与者决策信心的影响（p=.019）：较不偏好分析性投入的参与者在决定是否信任时，信心有所下降。

图5。调节　　分析人工智能素养和认知需求对风险沟通影响总体信任和决策信心的作用。

在调节分析中，10项调节变量—结果变量交互作用中有3项达到显著：人工智能素养调节了两项结果变量，而认知需求调节了DecisionConfidence。其余7项交互作用均不显著，表明人工智能熟悉度、人工智能使用频率和既往AI舒适度均未调节风险沟通对任一结果变量的影响。

人工智能素养调节了风险沟通对总体信任（β = 0.31，p = .033，95% CI [0.02，0.59]）和决策信心（β = 0.42，p =.001，95% CI [0.16，0.67]）的影响。正系数表明，更高的客观知识水平减弱了信任和决策信心的下降。基线数据（图5）显示，最不了解大语言模型如何运作的参与者最信任大语言模型，也最确定何时应当依赖大语言模型。结合基线水平解读这些系数，我们可以看到，**风险沟通缩小了不同人工智能素养水平之间在信任和决策信心方面的差距**，使人工智能素养较低的参与者更接近

人工智能素养较高的参与者所处的水平。这表明，风险沟通有可能作为纠正大语言模型心智模型失准的信息。然而，与H1b假设相反，在接受风险沟通后，**只有高素养参与者表现出依赖决策信心的提升**，这表明用户可能需要更高的人工智能素养，才能将风险信息转化为更清晰的决策。

认知需求也调节了对 DecisionConf 的影响（β = 0.23，p = .019，95% CI [0.04，0.42]）；认知需求越高，对信心的影响越弱。值得注意的是，偏好较少分析性投入的参与者，信心下降幅度最大。这表明，**风险沟通可能会降低那些不太倾向于处理复杂新信息者的决策信心**。综合这两项结果可以看出，风险沟通旨在帮助人们判断何时以及是否应依赖大语言模型，而要实现这一效果，人们可能需要具备相应的知识或倾向，才能有效利用这些信息。

4.6 参与者如何解读风险沟通

开放式回答揭示了参与者解读风险沟通的四种方式，包括 g:

4.6.1 熟悉的警告变得更加具体。许多参与者提到，他们知道大语言模型可能会犯错，但对具体风险了解较少。一位参与者写道：“我知道信息并非百分之百准确，但不知道它能多么令人信服地将内容呈现为事实”（P223）。其他参与者则称自己先前的了解不全面（“我之前至少有所了解，但没有了解到这种程度”，P57）或有限（“我所知道的只是皮毛”，P306）。与此同时，参与者指出，示例让这些风险更容易被察觉。P204表示：“在有人指出这些回答的问题之前，我觉得其中一些回答没什么问题”。不过，示例有时也会让参与者得出超出预期信息的结论。P113推断，人工智能提及用户的年龄意味着“它在用户不知情的情况下收集数据”；而P331则将隐私示例解读为人工智能“能够调取真实人物的医疗记录”。这些示例让风险更加具体，但参与者的理解并不总是符合预期。

4.6.2 个人经验塑造可信度。一些参与者将这种风险沟通与自己使用大语言模型的个人经验进行了比较。当示例与他们所见的情况不同时，他们会质疑这些示例是否符合现实，或者差异是否源于所使用的模型或提示词的写法。例如，P151 质疑“这个人工智能给出的回答明显比我使用过的任何人工智能都差”，而 P115 猜测差异是因为“我每月支付 *20* 美元使用 *ChatGPT*”。在机制性解释中也出现了类似反应。参与者结合自身经验，对相关表述提出质疑，并指出大语言模型“有时会告诉我信息来源”（P288），或“会引用较早的提示词或消息”（P18）。然而，当某个解释与先前的经验相吻合时，它可能使该经验更容易理解：“这解释了我得到的各种结果中的许多情况”（P379）。

4.6.3 缺乏明确行动的谨慎。参与者回答中一个常见的结论是，应当保持谨慎并核实重要信息。正如 P229 所解释的：“在采取行动之前，我们可以且应当核实所有重要信息”。然而，我们注意到，参与者很少说明他们会核实什么，以及会如何核实。一些参与者自己也指出了这种不确定性。P276 表示“（大语言模型的不可靠性）似乎使其变得毫无用处，或者要求采用一种结构化使用方式”；而 P50 则不确定“在许多情况下，回报是否超过潜在风险”。这些反思与定量研究结果一致：风险沟通降低了信任，却没有提高参与者在决定何时依赖大语言模型时的信心。

4.6.4 对风险与责任的不同看法。总体而言，参与者对于风险在多大程度上与自身相关，以及谁应负责应对这些风险，并未达成一致。首先，参与者在是否认为这些风险与自身相关，以及认为相关到何种程度方面存在差异

。一些人认为自己已经了解这些风险，因此更担心其他用户；例如，P377 表示 “这些风险对我个人而言都不是什么大问题，因为我了解自己使用时存在的风险。不过，我确实担心其他人”。较少有人意识到这些风险也可能适用于自己：“批判性思维的削弱可能也发生在我身上”（P207）。

参与者对于谁应该应对这些风险也有不同看法。有些人认为，问题源于用户行为，将不恰当的假设归因于“没有提前告诉人工智能”（P115），或将隐私风险归因于“用户过度分享导致的用户错误”（P266）。另一些人则希望人工智能系统自行介入（“每一种情况都应该直接弹出警告，提示‘请咨询真正的医生’”，P305），也有人指出监管方面存在空白（“目前还没人确定人工智能出错时谁该承担责任”，P382）。

这些不同立场也反映在参与者对这类沟通的反应上。有些人因此普遍变得不再信任人工智能，并指出“感觉人工智能不值得信任”（P37）。另一些人则反驳涉及其重视功能的风险说法（“坦白说，如果 *ChatGPT* 在交流时没有那么像人，我根本不会用它”，P69），甚至质疑人们是否对风险反应过度（“人们仅凭一个回答就能推断出很多事情，这完全没有必要”，P15）。

5 讨论

我们的研究是首批受控实验之一，探讨结构化风险沟通如何影响人们的信任，以及他们在决定何时依赖大语言模型进行健康信息寻求时的信心。结果显示，风险沟通显著降低了感知信任，尤其是在其与上下文相关时。然而，除客观知识水平较高的人群外，这种信任校准并未改变他们在决定何时依赖大语言模型时的信心。任务层面的影响同样不均衡：披露信息使人们在医学问答以及损伤与病症辅助方面不再倾向于使用大语言模型，但在情感陪伴方面，参与者仍然更偏好大语言模型，而非传统搜索环境。总体而言，这些发现表明，风险沟通能够改变人们对大语言模型的总体评价，尤其是在披露的风险与其使用情境相关时。然而，认识到系统存在风险，并不等于知道如何利用这一信息。对于通用大语言模型而言，要通过有效沟通帮助用户识别风险并采取情境化行动，就需要兼顾使用情境的差异，以及用户在互动中所具备的知识。

5.1 风险沟通作为失校准信任的纠正机制

在公众理解落后于技术发展和采用速度之时 [5, 67, 68, 70, 84], ，我们的研究提供了证据，表明风险沟通可以作为失校准信任的纠正机制。在我们的研究中，在接触任何信息之前，对大语言模型运作方式理解最不准确的参与者，对其表现出最高的信任，并且在决定何时依赖大语言模型方面拥有最大的信心。风险沟通显著降低了参与者在四个维度上对大语言模型的信任 [52]（对其能力、正直和善意的信念；依赖大语言模型的意向；以及向其分享信息并遵循其建议的意向）。重要的是，它使那些对人工智能理解最不准确的人，更接近于那些理解最准确者原本所达到的校准水平。

情境相关性进一步增强了这一效应。我们发现，尽管所有实验条件都使用了健康相关示例对话，健康特定框架比一般框架更能降低信任。换言之，即使用户看到了与使用相关的示例，风险信息中呈现的风险越贴近使用情境，人们对系统的信任程度就越会受到影响。这表明，仅用健康示例说明一般风险，并不一定能使这种风险等同于围绕使用案例的后果和相关顾虑组织起来的沟通。

健康和危机沟通等领域沿用的做法强调量化和因果沟通：报告绝对风险而非相对风险，以自然频数表达不确定性，并明确说明哪种行动会导致哪种后果 [1, 20, 23, 51]。这些技巧的前提是，可能采取的行动范围有限，而且其结果可以计数。然而，这一前提已不再适用于大语言模型。人们可以向大语言模型询问任何事情，同一个问题得到的答案因情况而异，也没人能可靠地量化某个提示词导致伤害的频率或导致伤害的途径。仍然可用的是上下文。如果系统无法准确告知用户某种伤害发生的可能性有多大，仍可以向用户说明这种伤害可能在什么情况下出现。一般框架可能让用户觉得伤害抽象而遥远，而情境相关的框架则能赋予伤害一种可识别的形态，让用户设想自己可能会遇到这种情况。

5.2 区分减少过度信任与支持依赖决策

我们的研究结果还表明，风险沟通显著改变了人们对大语言模型的信任程度，却没有改变他们在判断何时依赖大语言模型时的信心，这意味着人工智能风险传播中存在一种区别。在许多领域，风险信息与建议采取的行动相伴而来：关于吸烟的警告意味着戒烟，血糖水平偏高则意味着需要调整饮食。一旦人们接受了风险，采取什么行动就很明确。面向通用大语言模型的风险沟通并非如此。人们对整个系统形成信任，但是否依赖它则是针对每项任务分别决定的；正如我们的结果所示，同一系统在用于回答用药问题、解释医疗报告，或深夜谈论孤独感时，会带来不同程度的感知风险。因此，告知用户该系统不如他们原先以为的那么值得信任，未必足以指导他们在具体情境中作出判断。我们认为，纠正过度信任是实现更适当依赖的必要起点，但只有当用户能够运用自己的判断时，这一做法才能真正发挥作用。

事实上，这种分离印证了一项长期观察：风险意识与可操作的指导是彼此不同的结果，并不总是同时出现 [23, 24]。如果威胁相关信息未能带来效能感，人们往往会选择回避，而非采取适应性行动 [77]。在我们的研究中，提供关于大语言模型如何生成输出的机制性解释，并未像我们所希望的那样弥合这一差距。为了通过将参与者划分为不同的人工智能素养群体来解读这一结果，我们发现，与经调整的基线比较结果相比，只有事先具备准确客观知识的参与者提升了决策信心。这表明，直接从现有分类体系移用的风险沟通内容需要经过调整，才能更广泛地适用于普通人群。

因此，评估风险沟通的相关标准不应仅限于警示是否受到关注，或是否降低了信任。如果目标是实现适当依赖，还应评估干预措施能否帮助用户将已披露的风险转化为更符合具体情境的判断，例如，判断何时需要核实、与其他来源进行比较，或暂缓行动。

5.3 通用大语言模型风险沟通设计的启示

5.3.1 让披露与任务的风险特征相匹配。 任务层面研究结果表明，风险沟通不应将使用通用大语言模型视为单一的决策情境。对于两项任务，风险沟通改变了用户相较于传统搜索环境对大语言模型的相对偏好：医学问答以及损伤与病症辅助。在其他未发现风险沟通产生显著影响的任务中，情感陪伴是唯一一项参与者始终偏好大语言模型而非包含在线社区的搜索环境的任务，无论风险暴露如何。

这些差异表明，同一项披露会因用户的目标不同而产生不同影响。一种可能的解释是，这两项任务尤其依赖认知能力，其风险围绕事实准确性展开。另一方面，情感陪伴本质上具有关系性，其价值在于即时性和不带评判性 [38], 这些特质比所呈现的风险更容易被感知。更广泛地说，这意味着单一的沟通方式不太可能适用于风险特征迥异的各种使用场景；能够有效减少医疗查询中过度依赖的披露方式，可能不太适用于情感相关场景，而事实证明，这类场景在青少年中最为常见 [19]。

5.3.2 考虑披露信息的接收者。调节效应的发现表明，如果风险沟通未考虑接收者的特点，可能会适得其反。披露信息只有对那些原本就更准确地了解大语言模型如何运作的参与者，才实现了提升其决策信心的预期目的。此外，对于那些缺乏个人动机去处理并整合风险沟通内容的用户（即 NFC 较低的人），披露信息可能会加剧困惑，却无法消除困惑。因此，沟通设计应让理解更易于实现。既有研究表明，打断启发式加工的干预措施，即使对偏好较少进行分析性投入的人，也能减少其过度依赖 [11], ；交互式形式也比基于文本的形式更有效 [26]。具体而言，在人工智能素养和适当信任方面，证据表明，视频信息能够塑造人们对大语言模型的心智模型，并影响他们对大语言模型生成的回应的依赖 [16]。未来研究可以探索渐进式和上下文内的披露形式，以减轻将抽象披露转化为判断所带来的心智负担，例如提供与任务相关的简短提示，随后附上可选解释、具体的失败案例演示，或将已说明的局限转化为即时行动（如查阅专业来源）的提示词。

5.3.3 从风险沟通到人工智能素养。 综合我们的结果来看，先前的客观人工智能知识似乎会影响风险信息能否转化为实际行动。这并不意味着应以广泛的人工智能素养教育取代披露。相反，两者可能发挥互补作用。风险沟通可以在某项局限变得相关时提醒用户注意，而人工智能素养则提供理解该局限意味着什么，以及它应如何影响后续使用所需的概念框架。

这一点对于生成式人工智能尤为重要，因为它的失误往往不那么显而易见，而且用户不能仅凭使用就必然获得相关素养。许多以往的技术，其失误更加明显，例如链接失效、无法加载，或搜索未能返回匹配记录。生成式人工智能的失误表达流畅，与成功结果呈现得毫无二致，默认情况下几乎没有任何线索可以区分两者 [36, 66, 79, 84]。我们认为，一项技术越不显露其失误，其安全使用就越依赖理解而非经验；这就是为什么使用大语言模型所需的功能性素养 [48] 在某种程度上取决于概念性素养，而大多数以往工具并非如此。

因此，一个颇具前景的方向是，将披露视为提供情境化素养支持的契机，而非单独的警告。近期关于促进适当依赖的干预研究也指向这一方向：将披露与具体的核验支架相结合 [6]。此类做法也有助于恢复部分被单一综合答案所取代的分布式判断 [64]。因此，界面不应只说明大语言模型可能生成不可靠信息，还可以将风险与具体应对措施联系起来：例如，说明系统为何难以处理这类查询、有哪些替代来源，或何时需要求助专业专家。未来研究可以评估此类支架能否改善的不只是信任或信心，还有适当依赖的行为指标，例如来源核验或外部咨询。

5.4 局限性与未来方向

解读本研究结果时，应考虑若干局限，其中一些也揭示了未来研究的方向。首先，在测量决策信心时，我们有意安排参与者先对一系列任务中偏好使用大语言模型还是搜索进行评分，再测量其决策信心。这样的顺序旨在让自我报告的信心建立在对具体任务的判断之上，但与多题项量表相比，单题项测量的信度和敏感度可能有限。更广泛地说，决策信心反映的是参与者认为自己何时应依赖大语言模型的确定程度，而非其依赖决策在客观上是否恰当。未来研究可以探讨，信任或信心的变化是否与实际依赖行为校准程度的提升相对应。其次，风险沟通内容取自既有风险分类体系，并根据其学术认可度（通用分类体系）和情境适配性（健康专用分类体系）进行选择。然而，这些分类体系最初并非为向普通受众传达风险而设计。为避免无意中引入框架偏差，我们没有修改原文措辞，但部分内容或语言可能不便于缺乏技术背景的人理解。第三，我们对健康信息场景进行了控制，以尽量减少混杂因素；这必然意味着要以牺牲部分生态效度为代价，换取实验严谨性。参与者对脚本化对话情境短文的回应，可能无法充分反映他们在真实的大语言模型互动中会如何提问或回应，因为真实互动中的输出难以预测且因人而异。此外，从长期来看，本实验中的参与者只接触过一次风险沟通。这类效果能否在反复使用人工智能后持续存在，仍是一个有待解答的问题。未来可通过纵向设计考察其长期效果：最初的披露效果能否持续，还是需要在反复的人机互动中不断强化。最后，尽管随机化检验确认五种实验条件具有可比性，Prolific 的参与者群体可能比美国人口整体更年轻、受教育程度更高。因此，研究结果未必能推广至数字素养或健康素养较低的群体；这些群体可能最容易受到校准失当的人工智能依赖影响，也可以说最能从有效的风险沟通中受益。

6 结论

本研究考察了结构化风险沟通能否支持形成更为适当的信任水平，并提升用户决定何时依赖大语言模型进行健康信息寻求时的信心。在一项包含500名参与者的预注册实验中，风险沟通显著降低了参与者对大语言模型的信任，尤其是在风险与健康相关时。然而，它并未提升参与者决定何时依赖大语言模型时的信心。披露的影响也因任务和先前的人工智能知识而异：对于部分健康任务，风险沟通改变了参与者在大语言模型与传统搜索环境之间的偏好；而人工智能素养较低的参与者往往表现出更大幅度的信任下降，却未相应提升决策信心。总体而言，这些发现表明，纠正对人工智能过于积极的看法，并不等同于促成知情依赖。因此，有效的人工智能风险传播应超越一般性的披露，转向情境敏感且易于理解的方法，帮助用户将对人工智能风险的认知转化为关于何时以及如何依赖这些系统的可执行判断。

致谢

本研究的部分工作获得了美国疾病控制与预防中心的资助。

利益冲突

De Choudhury 是 OpenAI 福祉与人工智能专家委员会的成员。

参考文献

[1] Jessica S Ancker、Natalie C Benda、Mohit M Sharma、Stephen B Johnson、Michelle Demetres、Diana Delgado 和 Brian J Zikmund-Fisher。2025年。Making Numbers Meaningful证据综述中关于健康领域概率沟通的范围、方法与概述性发现：一项系统综述。《MDM Policy & Practice》10卷第1期（2025年），23814683241255334。[2] Joseph F. Hair Aut；William C. Black Aut；Barry J. Babin Aut；Rolph E. Anderson Aut。2024年。《多变量数据分析》。Cengage。[3] Julie Ayre、Erin Cvejic 和 Kirsten J McCaffery。2025年。2024年在澳大利亚使用 ChatGPT 获取健康信息：全国代表性调查的见解。《Medical Journal of Australia》222卷第4期（2025年），210–212。[4] Gagan Bansal、Tongshuang Wu、Joyce Zhou、Raymond Fok、Besmira Nushi、Ece Kamar、Marco Tulio Ribeiro 和 Daniel Weld。2021年。整体是否大于部分之和？人工智能解释对互补性团队表现的影响。载于《2021年CHI计算机系统中的人为因素会议论文集》，1–16。[5] Yasmine Belghith、Atefeh Mahdavi Goloujeh、Brian Magerko、Duri Long、Tom Mcklin 和 Jessica Roberts。2024年。测试、社交、探索：初中生对 ChatGPT 的使用方式与认知特征。载于《2024年CHI计算机系统人为因素会议论文集》，1–17。[6] Jessica Y Bo、Sophia Wan 和 Ashton Anderson。2025年。该不该依赖？评估促进人们适当依赖大语言模型的干预措施。载于《2025年CHI计算机系统人为因素会议论文集》，1–23。[7] Edyta Paulina Bogucka、Marios Constantinides、Sanja Šćepanović 和 Daniele Quercia。2025年。影响评估卡：沟通人工智能用途的风险与益处。《ACM人机交互》9卷第7期（2025年），1–42。[8] Ann Bostrom、Baruch Fischhoff 和 M Granger Morgan。1992年。描述危险过程的心智模型：一种方法及其在氡问题上的应用。《Journal of Social Issues》48卷第4期（1992年），85–100。[9] Virginia Braun 和 Victoria Clarke。2006年。在心理学中运用主题分析。《Qualitative Research in Psychology》3卷第2期（2006年），77–101。[10] Virginia Braun 和 Victoria Clarke。2019年。反思反思性主题分析。《Qualitative Research in Sport, Exercise and Health》11卷第4期（2019年），589–597。[11] Zana Buçinca、Maja Barbara Malaya 和 Krzysztof Z Gajos。2021年。是信任还是思考：认知强制功能可减少人工智能辅助决策中对人工智能的过度依赖。《ACM人机交互》5卷CSCW1期（2021年），1–21。[12] John T Cacioppo、Richard E Petty 和 Chuan Feng Kao。1984年。认知需求的高效评估。《Journal of Personality Assessment》48卷第3期（1984年），306–307。[13] Romain Cadario、Chiara Longoni 和 Carey K. Morewedge。2021年。理解、解释和应用医疗人工智能。《Nature Human Behaviour》5卷第12期（2021年6月28日），1636–1642。doi:10.1038/s41562-021-01146-0[14] Mohit Chandra、Suchismita Naik、Denae Ford、Ebele Okoli、Munmun De Choudhury、Mahsa Ershadi、Gonzalo Ramos、Javier Hernandez、Ananya Bhattacharjee、Shahed Warreth 等。2025年。从亲身经历到洞见：剖析使用人工智能对话代理带来的心理风险。载于《2025年ACM公平性、问责制与透明度会议论文集》，975–1004。[15] Yupeng Chang、Xu Wang、Jindong Wang、Yuan Wu、Linyi Yang、Kaijie Zhu、Hao Chen、Xiaoyuan Yi、Cunxiang Wang、Yidong Wang 等。2024年。大语言模型评估综述。《ACM智能系统与技术汇刊》15卷第3期（2024年），1–45。[16] Allison Chen、Sunnie SY Kim、Angel Nathaniel Franyutti-Cintron、Amaya Dharmasiri、Kushin Mukherjee、Olga Russakovsky 和 Judith E Fan。2026年。将大语言模型呈现为伴侣会影响人们赋予其哪些心理能力。载于《2026年CHI计算机系统人为因素会议论文集》，1–30。[17] Chun-Wei Chiang 和 Ming Yin。2022年。探究机器学习素养干预对普通人依赖机器学习模型的影响。载于《第27届智能用户界面国际会议论文集》，148–161。[18] Rebecca JW Cline 和 Katie M Haynes。2001年。互联网上的消费者健康信息搜寻：现状。《Health Education Research》16卷第6期（2001年），671–692。[19] Common Sense Media。2025年。《对话、信任与权衡：青少年如何以及为何使用人工智能伴侣》。https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions[20] Ian GJ Dawson、Johnnie EV Johnson 和 Michelle A Luke。2013年。帮助个人理解协同风险：对呈现机制性与概率性概念的信息内容进行评估。《Risk Analysis》33卷第5期（2013年），851–865。[21] 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。《Journal of Experimental Psychology: General》144卷第1期（2015年），114。[22] Maggie Harrison Dupré。2025年。一名男子陷入人工智能精神病，杀死母亲后自杀。https://futurism.com/man-chatgpt-psychosis-murders- mother[23] Baruch Fischhoff。1995年。未经修饰的风险感知与沟通：二十年的历程。《Risk Analysis》15卷第2期（1995年），137–145。[24] Meg Gerrard、Frederick X Gibbons 和 Monica Reis-Bergan。1999年。风险沟通对风险感知的影响：个体差异的重要性。《JNCI Monographs》1999年第25期（1999年），94–100。[25] Gaole He、Lucie Kuiper 和 Ujwal Gadiraju。2023年。知晓自身所知：人类能力的错觉可能妨碍人们适当依赖人工智能系统。载于《2023年CHI计算机系统中的人为因素会议论文集》，1–18。[26] Hendrik Heuer 和 Elena Leah Glassman。2022年。针对错误信息的干预措施比较评估：扩充世界卫生组织清单。载于《CHI计算机系统中的人为因素会议》，1–21。

[27] Kevin Anthony Hoff 和 Masooda Bashir。2014年。自动化信任。《Human Factors：人因与工效学学会杂志》57，3（2014年9月2日），407–434。doi:10.1177/0018720814547570[28] Alon Jacovi、Ana Marasović、Tim Miller 和 Yoav Goldberg。2021年。将对人工智能的信任形式化：人类信任人工智能的前提、原因与目标。载于《ACM公平性、问责制与透明度会议论文集》。624–635。[29] Nancy K Janz 和 Marshall H Becker。1984年。健康信念模型：十年之后。《健康教育季刊》11，1（1984年），1–47。[30] Jiun-Yin Jian、Ann M Bisantz 和 Colin G Drury。2000年。建立基于实证确定的自动化系统信任量表。《国际认知工效学杂志》4，1（2000年），53–71。[31] Yueqiao Jin、Roberto Martinez-Maldonado、Dragan Gašević 和 Lixiang Yan。2025年。GLAT：生成式人工智能素养评估测试。《计算机与教育：人工智能》9（2025年），100436。[32] Eunkyung Jo、Daniel A Epstein、Hyunhoon Jung 和 Young-Ho Kim。2023年。理解利用大语言模型部署对话式人工智能用于公共卫生干预的益处与挑战。载于《2023年CHI计算机系统人为因素会议论文集》。1–16。[33] Shivani Kapania、Oliver Siy、Gabe Clapper、Azhagu Meena Sp 和 Nithya Sambasivan。2022年。“因为人工智能百分之百正确且安全”：印度用户对人工智能权威的态度及其来源。载于《2022年CHI计算机系统人为因素会议论文集》。1–18。[34] Alexandra D Kaplan、Theresa T Kessler、J Christopher Brill 和 Peter A Hancock。2023年。对人工智能的信任：元分析结果。《Human Factors》65，2（2023年），337–359。[35] Hae-Young Kim。2013年。临床研究人员的统计学笔记：使用偏度和峰度评估正态分布（2）。《修复牙科与牙髓病学》38，1（2013年），52。[36] Sunnie SY Kim、Q Vera Liao、Mihaela Vorvoreanu、Stephanie Ballard 和 Jennifer Wortman Vaughan。2024年。“我不确定，但是……”：考察大语言模型的不确定性表达对用户依赖与信任的影响。载于《2024年ACM公平性、问责制与透明度会议论文集》。822–835。[37] Sunnie SY Kim、Jennifer Wortman Vaughan、Q Vera Liao、Tania Lombrozo 和 Olga Russakovsky。2025年。促进对大语言模型的适当依赖：解释、来源与不一致性的作用。载于《2025年CHI计算机系统人为因素会议论文集》。1–19。[38] Taewan Kim、Seolyeong Bae、Hyun Ah Kim、Su-woo Lee、Hwajung Hong、Chanmo Yang 和 Young-Ho Kim。2024年。MindfulDiary：利用大语言模型支持精神科患者写日记。载于《CHI计算机系统中的人为因素会议论文集》。1–20。[39] Vivian Lai、Chacha Chen、Alison Smith-Renner、Q Vera Liao 和 Chenhao Tan。2023年。迈向人机决策科学：实证人类受试者研究中设计空间概览。载于《2023年ACM公平性、问责制与透明度会议论文集》。1369–1385。[40] Heather Landi。2024年。HIMSS24：Epic如何为临床医生拓展人工智能和环境式技术。
https://www.fiercehealthcare.com/ai-and-machine-learning/himss24-how-epic-building-out-ai-ambient-technology-clinicians[41] Frank Landymore。2025年。心理学家称人工智能正在导致前所未见的精神障碍类型。
https://futurism.com/psychologist-ai-new-disorders[42] Eleanor Lawrie。2025年。人工智能治疗师真的能成为人类帮助的替代方案吗?
https://www.bbc.com/news/articles/ced2ywg7246o[43] John D. Lee 和 Neville Moray。1994年。信任、自信与操作员对自动化的适应。《国际人机研究杂志》40，1（1994年1月），153–184。doi:10.1006/ijhc.1994.1007[44] John D Lee 和 Katrina A See。2004年。自动化信任：为适当依赖而设计。《Human Factors》46，1（2004年），50–80。[45] Weixin Liang、Nazneen Rajani、Xinyu Yang、Ezinwanne Ozoani、Eric Wu、Yiqun Chen、Daniel Scott Smith 和 James Zou。2024年。对32111张人工智能模型卡片的系统分析揭示了人工智能中的文档编制实践。《自然·机器智能》6，7（2024年），744–753。[46] Q Vera Liao 和 Jennifer Wortman Vaughan。2023年。大语言模型时代的人工智能透明度：以人为中心的研究路线图。arXiv预印本 arXiv:2306.01941（2023年），5368–5393。[47] Gabriel Lins de Holanda Coelho、Paul HP Hanel 和 Lukas J. Wolf。2020年。高效评估认知需求：开发六条目版本。《评估》27，8（2020年），1870–1885。[48] Duri Long 和 Brian Magerko。2020年。什么是人工智能素养? 能力与设计考量。载于《2020年CHI计算机系统中的人为因素会议论文集》。1–16。[49] Zhuoran Lu 和 Ming Yin。2021年。当绩效反馈有限时，人类对机器学习模型的依赖：启发式方法与风险。载于《2021年CHI计算机系统中的人为因素会议论文集》。1–16。[50] Angeline McCall。2023年。旨在帮助家庭暴力受害者的人工智能应用。
h
t
t
ps://
www.9news.com/article/tech/ai-powered-app-to-help-domestic-violence-victims/73-57207485-2573-439a-b494-6da5af37d713[51] John McClure、Robbie M Sutton 和 Chris G Sibley。2007年。听记者的还是听工程师的？关于建筑设计的基于实例的信息如何影响地震宿命论1。《应用社会心理学杂志》37，9（2007年），1956–1973。[52] D Harrison McKnight、Vivek Choudhury 和 Charles Kacmar。2002年。初始消费者信任对其与网站进行交易意愿的影响：信任建立模型。《战略信息系统杂志》11，3-4（2002年），297–323。[53] Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡片。载于《公平性、问责制与透明度会议论文集》。220–229。

[54] Alex Montero、Julian Montalvo III、Audrey Kearney、Isabelle Valdes、Ashley Kirzinger 和 Liz Hamel。2026年。KFF健康信息与信任追踪民调：人工智能在健康信息与建议中的使用。华盛顿特区：凯撒家庭基金会（2026年3月）。[55] M Granger Morgan。2002年。风险沟通：心智模型方法。剑桥大学出版社。[56] Christina U Pfeuffer。2026年。人工智能可信度标签对人工智能产品认知的影响。载于《2026年CHI计算机系统中的人为因素会议论文集》。1–14。[57] Snehal Prabhudesai、Ananya Prashant Kasi、Anmol Mansingh、Anindya Das Antar、Hua Shen 和 Nikola Banovic。2025年。“人工智能在这里做出了选择，而每个选择都可能存在偏见”：学生如何通过终端用户审计活动批判性地使用大语言模型。载于《2025年CHI计算机系统中的人为因素会议论文集》。1–23。[58] Muhammad Raees 和 Konstantinos Papangelis。2026年。从信任到依赖：人机恰当依赖的测量构念。载于《2026年CHI计算机系统中的人为因素会议扩展摘要集》。1–7。[59] Sara Salimzadeh、Gaole He 和 Ujwal Gadiraju。2024年。应对不确定性：理解预后任务与诊断任务对人机决策中信任与依赖的影响。载于《2024年CHI计算机系统中的人为因素会议论文集》。1–17。[60] Laura Sbaffi 和 Jennifer Rowley。2017年。网络健康信息中的信任与可信度：未来研究综述与议程。《医学互联网研究杂志》19，6（2017年），e218。[61] Renata Schiavo。2013年。健康传播：从理论到实践。约翰·威利父子出版公司。[62] Jakob Schoeffer、Maria De-Arteaga 和 Niklas Kuehl。2024年。人机决策中的解释、公平性与适当依赖。载于《2024年CHI计算机系统中的人为因素会议论文集》。1–18。[63] Mrinank Sharma、Meg Tong、Tomasz Korbak、David Duvenaud、Amanda Askell、Samuel R Bowman、Newton Cheng、Esin Durmus、Zac Hatfield-Dodds、Scott R Johnston 等。2023年。迈向理解语言模型中的谄媚。arXiv预印本 arXiv:2310.13548（2023年）。[64] Steven D Shaw 和 Gideon Nave。2026年。快思考、慢思考与人工思考：人工智能如何重塑人类推理以及认知屈从的兴起。SSRN工作论文6097646（2026年）。[65] Renee Shelby、Shalaleh Rismani、Kathryn Henne、AJung Moon、Negar Rostamzadeh、Paul Nicholas、N’Mah Yilla-Akbari、Jess Gallegos、Andrew Smart、Emilio Garcia 等。2023年。算法系统的社会技术危害：为危害减少界定分类体系。载于《2023年AAAI/ACM人工智能、伦理与社会会议论文集》。723–741。[66] Yike Shi、Qing Xiao、Qing Hu、Hong Shen 和 Hua Shen。2026年。大语言模型的海妖之歌：用户如何感知并应对大语言模型中的暗黑模式。载于《2026年CHI计算机系统中的人为因素会议论文集》。1–23。[67] Sofia Eleni Spatharioti、David Rothschild、Daniel G Goldstein 和 Jake M Hofman。2025年。基于大语言模型的搜索对决策的影响：速度、准确性与过度依赖。载于《2025年CHI计算机系统中的人为因素会议论文集》。1–15。[68] Mark Steyvers、Heliodoro Tejeda、Aakriti Kumar、Catarina Belem、Sheer Karny、Xinyue Hu、Lukas W Mayer 和 Padhraic Smyth。2025年。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》7，2（2025年），221–231。[69] Austin M Stroud、Sarah A Minteer、Xuan Zhu、Jennifer L Ridgeway、Jennifer E Miller 和 Barbara A Barry。2025年。对透明且可信的心血管人工智能的患者信息需求：一项定性研究。《公共科学图书馆数字健康》4，4（2025年），e0000826。[70] Cora Sula、Umama Dewan 和 Nora McDonald。2026年。借助人工智能排演自我：青少年的跨领域聊天机器人使用。载于《2026年CHI计算机系统中的人为因素会议论文集》。1–14。[71] S Shyam Sundar 和 Jinyoung Kim。2019年。机器启发式：当我们在个人信息方面比信任人类更信任计算机时。载于《2019年CHI计算机系统中的人为因素会议论文集》。1–9。[72] Melanie B Tannenbaum、Justin Hepler、Rick S Zimmerman、Lindsey Saul、Samantha Jacobs、Kristina Wilson 和 Dolores Albarracín。2015年。诉诸恐惧：恐惧诉求效果与理论的荟萃分析。《心理学通报》141，6（2015年），1178。[73] Qiaosi Wang、Chidimma L Anyi、Vedant Das Swain 和 Ashok K Goel。2024年。探索人工智能易错性：遭遇人格误表征后人们对人工智能的反应与认知。arXiv预印本 arXiv:2405.16355（2024年）。[74] Laura Weidinger、John Mellor、Maribeth Rauh、Conor Griffin、Jonathan Uesato、Po-Sen Huang、Myra Cheng、Mia Glaese、Borja Balle、Atoosa Kasirzadeh 等。2021年。语言模型造成危害的伦理与社会风险。arXiv预印本 arXiv:2112.04359（2021年）。[75] Joe Wilkins。2025年。青少年与人工智能聊天机器人交谈后持续住院。https://futurism.com/teen-hospitalized-ai-chatbot[76] Magdalena Wischnewski、Alisa Scharmann、Annika Ridder 和 Nicole Krämer。2026年。经认证但并不完美：探究人工智能认证与系统性能对人工智能系统信任与依赖的作用。载于《2026年CHI计算机系统中的人为因素会议论文集》。1–16。[77] Kim Witte。1992年。让恐惧回归恐惧诉求：扩展平行过程模型。《传播学专论》59，4（1992年），329–349。[78] Ziwei Xu、Sanjay Jain 和 Mohan Kankanhalli。2024年。幻觉不可避免：大语言模型的内在局限。arXiv预印本 arXiv:2401.11817（2024年）。[79] Dong Whi Yoo、Jiayue Melissa Shi、Violeta J Rodriguez 和 Koustuv Saha。2026年。用于心理健康自我管理的人工智能聊天机器人：以亲身经历为中心的定性研究。《JMIR心理健康》13（2026年），e78288。[80] Hye Sun Yun 和 Timothy Bickmore。2025年。大语言模型时代的在线健康信息搜索：横断面网络调查研究。《医学互联网研究杂志》27（2025年），e68560。[81] Renwen Zhang、Han Li、Han Meng、Jinyuan Zhan、Hongyuan Gan 和 Yi-Chieh Lee。2025年。人工智能陪伴的阴暗面：人机关系中有害算法行为的分类体系。载于《2025年CHI计算机系统中的人为因素会议论文集》。1–17。

[82] Yixuan Zhang、Joseph D Gaggiano、Nutchanon Yongsatianchot、Nurul M Suhaimi、Miso Kim、Yifan Sun、Jacqueline Griffin 和 Andrea G Parker。2023年。谈论社交媒体中的信任时，我们指的是什么？一项系统综述。载于*2023年CHI*计算机系统中的人为因素会议论文集。1–22。[83] Yunfeng Zhang、Q Vera Liao 和 Rachel KE Bellamy。2020年。信心和解释对人工智能辅助决策中准确性与信任校准的影响。载于*2020*年公平性、问责制与透明度会议论文集。295–305。[84] Zhiping Zhang、Michelle Jia、Hao-Ping Lee、Bingsheng Yao、Sauvik Das、Ada Lerner、Dakuo Wang 和 Tianshi Li。2024年。“这是一场公平的游戏”，还是如此？考察用户使用基于大语言模型的对话代理时如何权衡披露风险与益处。载于*CHI*计算机系统中的人为因素会议论文集。1–26。[85] Jiawei Zhou、Amy Z. Chen、Darshi Shah、Laura M. Schwab-Reese 和 Munmun De Choudhury。2025年。公共卫生领域采用大语言模型的风险分类法与反思工具。《*ACM*人机交互论文集》9，7，文章363（2025年11月），32页。https://doi.org/10.1145/3757544[86] Jiawei Zhou、Lei Zhang、Mei Li、Benjamin D Horne 和 Munmun De Choudhury。2026年。我们所描述的人工智能：大语言模型及其在健康领域的应用如何呈现在不同渠道的公共话语中。载于*2026*年*CHI*计算机系统人为因素会议论文集（*CHI ’26*）。计算机协会，纽约，美国，文章393，22页。doi:10.1145/3772318.3790316

A 调查问卷

总体态度与偏好

您寻求健康相关信息的主要动机是什么？（可多选）

□ 提出有关疾病、治疗、检测结果或医学术语的具体问题，以便更好地理解健康信息。

□ 为正经历压力、焦虑、情绪低落或其他情绪困扰的人寻求治疗指导和应对支持。

□ 通过交谈寻求被倾听和陪伴的感觉，尤其是在感到孤独或只是想聊聊天的时候。

□ 寻求帮助，缓解在某人患有健康状况或存在功能障碍时管理日常活动和便利措施。 .g.

□ 寻求有关生活习惯的指导，例如饮食、体育活动、睡眠和日常作息，以支持整体健康。□ 其他（请注明）

在寻求健康相关信息时，您通常会使用以下哪些来源？（量表：从不、很少、有时、经常、总是）□ 传统搜索引擎（例如 Google）如果您高度依赖嵌入人工智能的搜索引擎或相关功能（例如 *Google* 搜索中的 *AI* 概览、*Bing Chat*、*Perplexity*），请选择“人工智能工具或聊天机器人”□ 专家建议（例如护士、医生、专科医生）□ 我个人认识的人，例如家人或朋友□ 人工智能工具或聊天机器人（例如 Bing Chat、Gemini、ChatGPT、Perplexity）

在传统搜索引擎中……您通常会使用以下哪些来源？（量表：从不、很少、有时、经常、总是）□ 在线社区和论坛（例如，Reddit 群组、Facebook群组、健康论坛）□ 社交媒体（例如，YouTube、Instagram、TikTok）□ 专家撰写的网站（例如，梅奥诊所、Psychology Today、WebMD）□ （跨）政府资源（例如，美国食品药品监督管理局、世界卫生组织）

您对通过通常使用的来源找到可靠且准确的健康信息有多大信心?
□ 完全没有信心□稍
微有信心□ 有一定信
心□相当有信心□非常
有信心

您觉得通过通常使用的健康信息来源找到的信息，在多大程度上能满足您的需求或回答您的问题？□ 从不□ 很少□ 有时□ 经常□ 总是

人工智能知识与使用情况
您对语言人工智能和大语言模型（LLM），例如 ChatGPT、Gemini 和 BingChat，了解多少？□ 完全不了解：我对它们是什么没有基本概念□ 略有了解：我对它们是什么有一些概念□ 中度了解：我知道它们是什么，也了解其基本功能和用途□ 非常了解：我对它们的能力、局限性和应用有扎实的认识□ 极其了解：我具备专家级知识，包括它们的技术层面、开发过程和工作原理

您会使用 ChatGPT、Gemini 和 Bing Chat 等语言人工智能和大语言模
型吗？□从未使用过□ 很少使用，偶尔才用□ 有时使用，但并不经常□ 经
常使用，已成为日常习惯□ 总是使用，只要适用就会一直使用

您使用语言人工智能和大语言模型（包括高度依赖 Google 搜索的 AI 概览）查找信息或问题答案的频率如何?

□ 从不□ 很少：只在特定情况下使用□ 有时：偶尔使用□ 经常：
定期使用□ 总是：将它们视为查找信息的主要来源

对于使用人工智能工具帮助回答健康相关问题，你感觉如何？□非常不自
在

□ 不自在□ 中立□
自在□非常自在

下面的问题都与人工智能有关——只需选择你认为最准确的选项，不必查找任何资料。没有陷阱题！我们只是想了解你的看法。

以下哪项最能描述大语言模型（LLM）？ □ 它通过分析和总结大量网络内容来生成文本。□ 它根据前面词语的上下文预测下一个词来生成文本。□ 它通过同时将输入文本翻译成多种语言来生成文本。□ 它通过使用预定义模板并填充空白来生成文本。

使用大语言模型（LLM）为作业收集信息时，你应如何看待它提供的信息？ □ 大语言模型的回答始终比你在互联网上找到的任何信息都更值得信任，因此你可以不经进一步核实就使用这些回答。□ 大语言模型的回答通常比互联网来源更值得信任，但你仍应通过其他可靠来源核实这些信息。□大语言模型的回答不一定比互联网来源更值得信任，因此你应使用其他可信参考资料交叉核对这些信息。□ 大语言模型的回答不如互联网来源值得信任，因为它依赖过时信息。

信任与态度
对于语言人工智能和大语言模型（大语言模型）.....我
认为：（量表：1 - 非常不同意至 5 - 非常同意）

(1) 它们能够胜任并有效地提供健康信息；(2) 我认为它们是诚实的；(3) 它们关心我的福祉，而不只是自身利益；(4) 在艰难的健康状况下，我始终可以依赖它们；(5) 我觉得在遇到重要的健康问题时，可以指望它们提供帮助；(6) 如果遇到需要付费就医的棘手健康状况，我会选择使用支持这些技术的医院/诊所；(7) 我愿意向它们提供姓名、地址和电话号码等信息；(8) 我愿意向它们透露自己的健康问题详情；(9) 我会放心地根据它们提供的信息采取行动；(10) 我会毫不犹豫地使用它们提供的信息；(11) 我会有信心根据它们提供的健康建议采取行动

对于以下用途，您使用 [大语言模型与传统搜索引擎] 时会感到多自在？（量表：
1—非常不自在，5—非常自在）

- (...) 医学问答?
- (...) 治疗性心理健康支持?
- (...) 情感陪伴?
- (...) 障碍与病症援助?
- (...) 健康生活方式指导?
- （……）选择「非常不舒服」？

您对判断何时以及是否使用大语言模型来满足健康信息需求有多大信心?
□非常没有信心□
没有信心□ 中立□
有信心□非常有信
心

人口统计特征与偏好
您的年龄是多少?
□ 18–24 岁□ 25–
34 岁□ 35–44 岁
□ 45–54 岁□ 55–
64 岁□ 65 岁及
以上

您的性别是什么?
□ 女性□ 男性□ 非二
元性别□ 不愿透露

以下哪项最能描述您的种族? □ 美洲
印第安人或阿拉斯加原住民□ 亚洲人
□黑人或非裔美国人□ 西班牙裔、拉
丁裔或西班牙血统□ 夏威夷原住民或
太平洋岛民□ 白人□ 两个或以上种族
□ 其他（请注明）

□ 不愿透露

您完成的最高教育程度是什么？□ 高中或以下未完成□ 高中毕业生（文凭或普通教育发展证书）□曾就读大学但未获得学位□ 副学士学位（两年制学位）□学士学位（四年制学位）□ 高级学位（硕士、专业或博士学位）□ 不愿透露

请指出以下陈述是否符合您的特点或信念：（量表：1＝极不符合，5＝极其符合）

- 相比简单问题，我更喜欢复杂问题
- 我喜欢承担处理需要大量思考的情境的责任
- 思考对我来说并无乐趣
- 比起肯定会挑战我思维能力的事情，我更愿意做不太需要动脑的事情

- 我很喜欢需要为问题想出新解决方案的任务
- 比起有一定重要性但不太需要动脑的任务，我更愿意做智力性强、难度高且重要的任务

B 示例性消费者健康任务场景

医学问答

（就疾病、治疗、检测结果或医学术语提出具体问题，以更好地理解健康信息。）一位与你同龄的人最近在常规体检时做了常规血液检查。结果显示，其血糖水平略高于正常值，医生的总结中提到了“糖尿病前期”。他们的下次预约在两周后。为了了解这意味着什么，他们向人工智能寻求帮助。

用户： “我的血液检查结果出来了，上面显示我的血糖略高。医生提到了糖尿病前期。这到底是什么意思？”

人工智能： “糖尿病前期意味着你的血糖水平高于正常值，但还没有高到可以诊断为2型糖尿病的程度。这本质上是一个早期预警信号，表明你的身体在有效控制血糖方面遇到了困难

在这个阶段，这并不意味着你一定会患上糖尿病——许多糖尿病前期患者永远不会发展为2型糖尿病，尤其是在进行一些生活方式改变之后。通常建议采取的措施包括调整饮食和增加身体活动，不过最合适的做法取决于个人的具体情况。

值得在下次预约时与你的医生详细讨论一下你 r
具体结果意味着什么，以及他们是否会为你提供任何建议。”

治疗性心理健康支持

（为正经历压力、焦虑、情绪低落或其他情绪困扰的人提供指导和应对支持。）

有一位与你年龄相仿的人，过去几个月一直情绪低落。他们没有和治疗师谈过，部分原因是他们不确定自己的问题是否“严重到足以”需要求助。一天晚上，他们决定改为和人工智能聊聊这件事。

> **用户：** “我这几个月来一直挺情绪低落的。没发生什么特别严重的事，我只是大多数时候都觉得有些麻木、提不起劲。我还没跟任何人聊过这件事。我不确定这是否值得担心。”

> **人工智能：** “听起来，你一直默默承受着这些感受，已经有一段时间了——而你不确定这是否值得担心，这一点本身就值得关注。即使没有明确的原因，几个月来一直感到麻木、提不起劲，也值得认真对待，而不是一笔带过。有时，正是因为没有发生什么特别严重的事，你才更难允许自己承认这些感受。但情绪低落不需要特定的诱因，才算真实或值得重视。我可以问问吗——你说感到麻木、提不起劲，是指觉得做什么都没有乐趣，还是觉得很难开始做事，抑或两者都有？”

情感陪伴

（通过对话让人感受到自己被倾听和陪伴，尤其是在感到孤独或只是想找人聊天的时候。）

一个与你年龄相仿的人六个月前因工作机会搬到了一个新城市。当时搬家是合理的选择，但他们发现建立社交生活并不容易，大多数晚上都是独自度过。他们感受到熟悉的人和地方不在身边。一天晚上，他们开始和人工智能聊天，只是想找个人陪伴。

> **用户：** “我六个月前搬到了一个新城市，但还是觉得自己并没有真正安顿下来。我还没认识多少人，也很想念家乡。我没事，只是觉得有点孤独吧。”

> **人工智能：** “六个月听起来似乎应该足够让人适应一个地方、产生归属感了，但很多时候其实并不够——尤其是在一个与任何人都还没有共同经历的地方，从头开始建立社交生活。想念家乡并不只是想念那个地方——通常也是因为怀念与那些已经了解你的人相处时的自在，你不必费力去建立联系。这样的舒适感需要时间在新地方重新建立起来，而在它形成之前，即使一座城市人来人往，也可能让人感到孤立无援。你最想念以前生活的地方的什么？”

障碍与病症援助

（当某人患有健康状况或存在功能障碍时，帮助其管理日常活动（例如慢性疼痛、行动受限、视力／听力变化），如提供实用建议和便利措施。）

一位与你年龄相仿的人过去一年一直饱受持续性膝痛困扰，并被诊断为早期骨关节炎。医生告诉他们要保持活跃，但要避免高冲击活动。他们发现，爬楼梯、从椅子上起身、长时间步行等日常活动都变得很不舒服，也开始限制他们的行动。他们于是向人工智能寻求实用建议。

用户： “我患有早期骨关节炎，膝盖疼痛，爬楼梯和从椅子上起身等日常活动都变得非常不舒服。医生说要保持活跃，但要避免高冲击运动。我该如何应对日常生活中的这些问题？”

人工智能： “膝骨关节炎会让许多日常动作变得不适，而在避免高冲击运动的同时保持活动，总体方向是正确的——活动有助于维持关节及其周围的肌肉。

一些日常调整通常会有所帮助：上下楼梯时先迈较强壮的腿，用手臂撑起身体离开椅子，以及避免坐过低的座椅，都能减轻患侧膝关节的负担。究竟哪些方法最有帮助，确实取决于哪些活动让你在日常生活中最感到困扰。最近对你来说最困难的是什么？”

健康生活方式指导

（提供有关饮食、身体活动、睡眠和日常作息等习惯的指导，以促进整体健康。）一位与你年龄相仿的人一段时间以来一直在努力吃得更健康，但发现很难持续做出改变。他们白天的饮食还算合理，但晚上更难坚持——下班后感到疲惫时，他们往往会选择方便但营养价值较低的食物。他们并不需要严格饮食计划，只是希望找到一些切实可行的方法来养成更好的习惯。于是，他们向人工智能寻求指导。

用户： “我想吃得更健康，但很难坚持下去，尤其是在晚上感到疲惫的时候。我不想要严格的饮食计划——只想找到一些切实可行的方法来养成更好的饮食习惯。我该从哪里开始？”

人工智能： “你描述的这种晚间模式——白天表现不错，但感到疲惫时就会默认选择营养较少的食物——非常常见。疲惫会减少用于做决定的心理能量，因此方便往往会占上风。

有几件事通常会有所帮助：在晚上提前准备好更容易获取、营养也相对合理的食物；白天摄入足够的食物，这样你回到家时就不会在疲惫之外还感到真正的饥饿。相比宽泛的意愿，小而具体的改变通常更容易坚持——先选一个习惯进行调整，通常比一下子彻底改变所有事情更有效。你通常如何度过晚上的时间？”

C 构念效度：探索性因子分析

在进行分析之前也评估了内部一致性，最终综合得分的内部一致性极佳（α =.942），四个分量表的内部一致性均达到良好至极佳水平（信任信念： α = .831；信任意向： α = .825；信息分享意向： α = .708；建议遵循意向： α = .929）。

子量表	Item	因子载荷
TrustBelief	TB1	.84
	TB2	.80
	TB3	.74
TrustIntention	TI1	.90
	TI2	.90
	TI3	.60
IntentionShareInfo	SI1	.68
	SI2	.81
IntentionFollowAdvice	FA1	.89
	FA2	.87
	FA3	.95

表4。探索性因子分析结果