通过人与语言模型交互对负面思维进行认知重构

AshishSharma[♠] **KevinRushton**[◇] **Inna Wanyin Lin**[♠] **David Wadden**[♣]**Khendra G. Lucas**[◇] **Adam S.Miner**[♥♡] **TheresaNguyen**[◇] **Tim Althoff**[♠♣]华盛顿大学保罗·G·艾伦计算机科学与工程学院[◇]美国心理健康协会 [♣]艾伦人工智能研究所[♥]斯坦福大学精神病学与行为科学系[♡]斯坦福大学生物医学信息学研究中心
{ashshar,althoff}@cs.washington.edu

arXiv:2305.02466v1 [cs.CL] 4 May 2023

摘要

一种经证实可帮助克服消极想法的治疗方法，是用更有希望的“重构后的想法”取而代之。尽管治疗可以帮助人们练习并学习这种负面思维的认知重构，但临床医生短缺和心理健康污名往往限制了人们获得治疗的机会。本文通过一项以人为本的研究，探讨语言模型如何帮助人们重构消极想法。基于心理学文献，我们定义了一个由七种语言属性构成的框架，可用于重新构建想法。我们开发了测量这些属性的自动化指标，并通过心理健康从业者的专家评判对其进行验证。我们从从业者处收集了一个包含 600 个情境、想法和重新诠释的数据集，并用它训练了一种结合检索增强的上下文学习模型，该模型能够有效生成重构后的想法，并控制其语言属性。为探究何为“高质量”的重新诠释，我们在一家大型心理健康网站上开展了一项经机构审查委员会批准的随机现场研究，参与者超过 2,000 人。除其他发现外，我们还表明，相较于过度积极的重新诠释，人们更偏好富有同理心或具体明确的重新诠释。我们的研究结果为使用语言模型帮助人们克服消极想法提供了重要启示。

1 引言

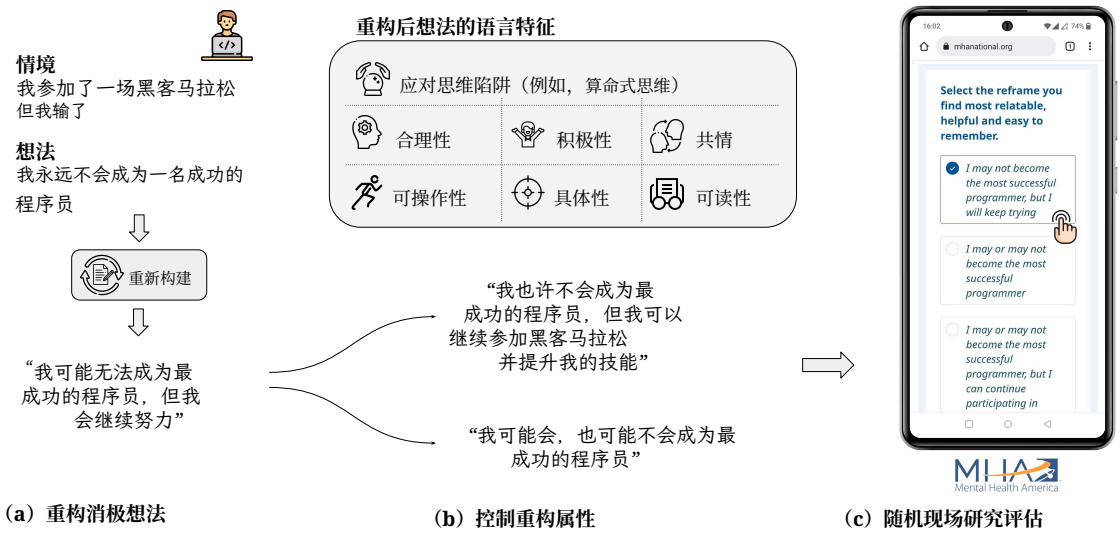
消极想法是人类认知的自然组成部分。然而，对于面临心理健康挑战的人来说，这类想法往往根深蒂固、自动出现且会引发强烈情绪，因此很难在当下克服（Beck，1976）。一种有证据支持且已确立的、用于克服消极想法的治疗干预是认知重构，即用更有希望的“重新构建后的想法”取代消极想法，从而提供一种替代对自身情境的视角（Beck，1976）。例如，设想某人的情境是“我要向 *ACL 2023* 投稿一篇研究论文”，其想法是“这篇论文肯定会被拒”。重新构建这一想法的一种可能方式是说：“这篇论文凭借新颖的方法和潜在影响，有一定机会被接收。”

心理治疗研究表明，要让重构后的想法有效，必须满足以下条件：(a) 与个人产生共鸣，(b) 有助于克服消极想法，(c) 令人印象深刻，以便在类似想法再次出现时能够回想起来（Beck，1976年；Burns，1980年）。然而，究竟是什么让重新构建后的想法能够引起共鸣、带来帮助并令人印象深刻，仍是一个难以回答且尚不明了的问题。专业治疗师能够帮助人们构思出这类高度有效的重构后的想法。然而，临床医生短缺、缺乏保险保障以及污名等障碍，往往限制了人们获得治疗师帮助的机会（Olfson，2016年；Sickel 等人，2014年）。基于自然语言处理的方法可以帮助人们在消极想法出现时即时重新构建这些想法，提供更易于使用且能够广泛普及的辅助支持。

以往的自然语言处理研究已针对多种文本重构任务开发了方法，例如情感重构和共情改写（Reif 等人，2022年；Sharma 等人，2021年），近期还研究了积极重构（Ziems 等人，2022年）。然而，如何开发能够自动生成引发共鸣、带来帮助且令人印象深刻的重构后想法的认知重构方法，仍鲜为人知。

本文研究如何利用语言模型重新构建消极想法（图1）。我们探讨消极想法可以通过哪些方式重新构建、如何利用语言模型完成这种重新构建，以及经历消极想法的人更偏好哪些类型的重新诠释。

首先，与临床心理学家合作-



（a）重构消极想法　　（b）控制重构属性　　（c）随机现场研究评估

图1：（a）我们研究了使用语言模型以不同的、更具希望的想法来重构消极想法这一任务；（b）对某一情境的不同视角可能会产生不同的重新诠释。在此，我们提出了一个包含七项重构属性的框架（见灰色框）。给定一个经过重构的想法，我们控制每项属性（例如，可操作性），以生成在该属性上得分更高或更低的重新诠释（例如，可操作性更强或更弱）；（c）我们将该模型部署在美国心理健康协会这一美国大型全国性心理健康网站上（bit.ly/changing-thoughts），并开展了一项包含2,067名参与者的随机现场研究。我们向美国心理健康协会的访客推荐语言模型生成的重构内容，并评估哪些重构属性更受欢迎，以及怎样的重新诠释更具共鸣、更有帮助且令人难忘。

通过与心理健康专业人士交流，我们开发了一个新的概念框架，用于描述想法可能被重新构建的方式。我们综合了治疗中最突出的认知重构过程，并定义了重构后的想法的七种语言属性：重新构建是否针对“思维陷阱”（有缺陷或扭曲的思维模式），是否理性、积极、富有同理心、可付诸行动、具体且易读。基于先前的研究，我们开发了自动化指标来衡量这些属性，并通过将其与心理健康从业者的判断进行相关性分析，确立了构念效度。

接下来，为了开发用于认知重新构建任务的模型，我们收集并共享[1] 一个来自心理健康从业者和临床心理学研究生的数据集。该数据集包含 600 个情境和想法，以及专家建议的重新诠释方案和对所提出的重新诠释属性的标注。利用这个数据集，我们开发了一种检索增强的上下文学习方法（Brown 等人，2020年），用来生成重构后的想法，并控制其语言属性。我们表明，与其他方法相比，该方法与专家建议的重新诠释方案重合度最高；根据心理健康专家的评估，其亲和力和有用性评分也最高，优于流行的自然语言处理基线方法。

我们探究哪些重构属性值得采用，以及怎样的重新构建才会让人感同身受、有帮助且令人印象深刻。在与心理健康专家合作（并共同署名），且经过适当的伦理审查后，我们在美国心理健康协会（MHA；一个在线分享心理健康资源和工具的热门网站）开展了为期一个月的随机现场研究，共有 2,067 名参与者在知情同意后参与。我们请 MHA 访客描述他们正在经历的情境和消极想法，随后提供由语言模型生成、具有不同语言属性的重新构建后的想法。我们发现，具体性强且共情度高的重构最受青睐，而具体性强且可付诸行动的重构则被认为最有帮助、最令人印象深刻。不过，我们发现，积极程度很高的重新构建较不受青睐。这些发现为消极想法的认知重构，以及在这一过程中运用人类与语言模型交互，提供了重要启示。

2 问题定义与目标

我们研究的任务是认知重构。给定一个情境$\mathbf{S_i}$ 和一个消极想法$\mathbf{T_i}$，任务是生成一个重新构建后的想法$\mathbf{R_i}$。

心理治疗文献（Beck，1976）强调，成功重新构建想法需要达到三个理想结果：

[1]我们的代码和数据集可在以下地址获取 https://github.com/behavioral-data/Cognitive-Reframing-of-Negative-Thoughts。

（**a**）重新构建后的想法必须与个人产生共鸣，（**b**）必须帮助他们克服负面想法，并且（**c**）当类似的负面思维模式再次出现时，这个想法必须令人难忘。

在这里，我们旨在了解成功的重新诠释由什么构成，以及语言模型如何在这一过程中帮助人们。为此，我们刻画重新诠释后想法的语言属性（§3），收集一个包含情境、想法和重新诠释的数据集（§4），开发生成重新诠释内容并测量和控制其属性的方法（§5；§6），并探究哪些语言属性与亲和力、有用性和记忆性这些重新诠释结果相关（§7）。

3 重构后想法的语言属性框架

我们借鉴临床治疗实践，并与心理健康专家（其中一些也是共同作者）合作，开发重构后的想法的语言属性框架。以下面的例子说明这些属性，适用于下文中的所有重新诠释——情境："我参加了黑客马拉松，但输了"；想法："我永远不会成为一名成功的程序员"。§5.1 将介绍用于衡量这些属性的自动化指标。

应对思维陷阱。 消极思维常常会陷入一些常见模式，这些模式被称为"思维陷阱"。它们也被称为认知扭曲，包括夸大且有偏差的思维模式，会导致个体不准确地感知现实（Beck，1976；Ding 等，2022年）。常见的思维陷阱包括：臆测他人的想法（"读心术"）、极端化思考（"非黑即白思维"）、专注于最坏的情形（"灾难化思维"）、试图预测未来（"算命式思维"）等。完整列表请参见附录D 。

重新诠释可能会直接应对想法中的一个或多个思维陷阱，也可能不会。例如，"我不知道未来会带来什么"这一重新诠释直接应对了"算命式思维"这一思维陷阱，而"我一定会成为一名成功的程序员"这样的重新诠释并没有应对这一思维陷阱，反而继续表达了这种思维陷阱。

合理性。 重新构建想法的另一种策略是反思支持和反对该想法的证据，并推理这些证据意味着什么（Beck，1976）。例如，未能赢得黑客马拉松是支持"我永远无法成为一名成功的程序员"这一想法的一项证据。然而，反对这一想法的证据这个想法可以是：赢得或输掉一场黑客马拉松，并不会让一个人变成失败者，这或许能促成重新构建："只输掉一场黑客马拉松并不能定义我的失败。"理性的重新构建以有力的证据为依据，而非理性的重新构建则基于不切实际的假设。

积极性。 对消极想法进行重新诠释，旨在强调情境中的积极视角，但不同的重新诠释可能具有不同程度的积极性。像"从今以后我会赢下每一场黑客马拉松"这样过度积极的重新诠释，夸大了积极视角，可能会让当事人失望，而不是帮助其取得成功（Dember 和 Penwell，1980）。另一方面，像"我可能成功，也可能失败，但我会继续尝试"这样的平衡回应，同时考虑了情境中的积极和消极视角。

共情。 承认消极想法所引发的感受可能会有所帮助（Allen andLeary，2010年；Elliott 等人，2011年）。重新构建可以通过确认一个人的感受来表达共情或自我慈悲。例如："没有赢得黑客马拉松而感到失望是可以的。"

可操作性。 为了鼓励愉悦情绪，一种常用的治疗方法是行为激活（Dimidjian 等人，2011年；Burkhardt 等人，2021年）。这包括采取可能有助于克服消极想法的行为或行动。重新构建可以建议具体行动（例如："我可以继续练习并参加黑客马拉松"），也可以不建议具体行动但具有可操作性（例如："我可能不会很成功，但我可以继续尝试"），或者完全不具有可操作性（例如："我可能会成为一名成功的程序员，也可能不会"。

具体性。 重新构建可以具体针对情境和想法（例如："一次黑客马拉松并不能定义我作为程序员的失败"），也可以足够笼统，适用于广泛的消极情境和想法（例如："我会成功"）。具体的重新构建可能在当下更有帮助，而笼统的重新构建则可能对反复出现的想法有效；这些想法通常源于一个人持有的"核心"信念（Beck，2005年；David 等人，2009年）。

可读性。 个体的语言推理能力可能存在差异（例如，不同年龄组或教育水平之间）（Kaplan 等人，1995年）。因此，重新构建的内容可能读起来简单或复杂（例如"我将来会做得很好"与）。

“我坚定地追求成功”).

4 数据收集

为促进认知重新构建的计算方法，我们收集了一个重构后的想法数据集，并为其标注了语言属性。

4.1 精选情境与负面想法

我们首先整理情境和消极想法的数据来源。

想法记录数据集（Burger 等人，2021年）。该数据集包含由 Amazon Mechanical Turk 上的众包工作者报告的假设情境和现实情境、想法及情绪过程。我们从该数据集中人工筛选出 180 对具有多样性的情境及其消极想法。

美国心理健康协会（MHA）。众包工作者提供的情境和想法可能无法反映人们在现实生活中所面临的广泛心理健康挑战。为了纳入更多贴近现实世界的情境和想法，我们在 MHA 网站（screening.mhanational.org）上开展了一项调查。我们邀请 MHA 网站的访问者（他们通常使用该网站进行精神疾病筛查）描述自己正在经历的消极想法及其相关情境。我们通过人工筛选和整理，选取了 120 对自我报告的情境与想法，以确保涵盖广泛的相关主题，并保证内容的多样性。

4.2 标注任务与流程

重构消极想法是一个在认知上颇具难度的过程，需要练习和训练，因此很难通过众包进行数据收集。为确保重新诠释和标注的高质量，我们招募了 15 名现任心理健康从业者和临床心理学研究生，他们都拥有丰富的实践经验，擅长认知重构。[2] 对于数据来源中的每一组（情境、想法）Pair（§4.1），我们请他们：(1) 写出两个不同的重构后的想法；(2) 标注每个重构后的想法所针对的思维陷阱；(3) 比较这两种重新诠释，并选出更合理、更积极、更具可操作性、更有同理心、更具体且更易读的一种。我们总共收集了 600 条重构后的想法，并对其语言属性进行了标注。我们分享这些公开数据集于 https://github.com/behavioral-data/Cognitive-Reframing-of-Negative-Thoughts。

4.3 伦理与安全性

我们的数据收集和随机现场研究（§7）是在与心理健康专家协商并合作，对参与者可能面临的潜在收益和风险进行评估后设计和开展的。两项研究均获得了华盛顿大学机构审查委员会的批准，并向参与者说明了研究目的、风险和数据收集情况。所有参与者均年满18岁，自愿提供知情同意，并获准使用危机热线。我们不评估任何临床结果。有关伦理和安全考量的扩展讨论，请参见§10 。

5 方法

我们设计用于衡量语言属性的自动化指标（§5.1），开发生成重新构建的想法的方法（§5.2），并控制这些想法属性的方法（§5.3）。

5.1 重构属性的测量

应对思维陷阱。给定一个情境 $\mathbf{S_i}$、一个消极想法 $\mathbf{T_i}$ 和一个重新构建后的想法 $\mathbf{R_i}$，我们的目标是识别重新构建后的想法所应对的思维陷阱集合。我们将此视为一项多标签分类任务，并在第[3] 4.2节收集的专家标注思维陷阱标签上微调 GPT-3 模型。

合理性。合理性是指受理由指引的特质（Damielson et al., 2004）。在此，我们将重新构建后的想法 $\mathbf{R_i}$ 的合理性操作化为其推理强度，并提出以下两个问题：(1) $\mathbf{R_i}$ 背后的推理可能是什么？；(2) 这些理由是否充分？为了理解 $\mathbf{R_i}$ 背后的推理，我们开发了一种基于溯因解释的方法（Peirce, 1974; Bhagavatula et al., 2020; Jung et al., 2022）。对于给定的 $(\mathbf{S_i}, \mathbf{T_i})$，我们使用语言模型生成 (a) 支持 $\mathbf{R_i}$ 的最可信解释，以及 (b) 反驳它的最可信解释。此外，为了检验这些解释是否充分，我们递归生成解释背后的解释，以测试其推理强度（附录 E）。令 $sup(\cdot)$ 为生成支持重新构建的解释的生成函数，并令

[2] 招募参与者时，我们通过大学邮件列表和一家心理健康组织的通讯宣传了这项研究。受邀专家的报酬为每小时 37.5 美元。

[3] 我们在本文的所有实验中使用 text-davinci-003 作为我们的 GPT-3 模型。

$ref(\cdot)$ 为生成反驳重新构建的解释的生成函数。接着，我们递归地将推理强度 $RS(\mathbf{S_i}, \mathbf{T_i}, \mathbf{R_i})$ 定义为

$$(\mathrm{P}(\mathbf{R_i} = \text{sound} | \mathbf{S_i}, \mathbf{T_i}) \times \mathbb{E}_{r \sim sup(\cdot)}[RS(\mathbf{S_i}, \mathbf{T_i}, r)]) - (\mathrm{P}(\mathbf{R_i} = \text{有缺陷} | \mathbf{S_i}, \mathbf{T_i}) \times \mathbb{E}_{r \sim ref(\cdot)}[RS(\mathbf{S_i}, \mathbf{T_i}, r)])$$

为了设计解释生成函数，$sup(\cdot)$ 和 $ref(\cdot)$，我们采用上下文学习（Brown 等，2020）。我们与心理健康专家合作，设计了 10 个示范样例，涵盖情境、想法和重构后的想法，并附有支持（“这个重构后的想法是合理的，因为……”）或反驳（“这个重构后的想法有缺陷，因为……”）某个特定重新构建的解释。我们使用这些示例来提示 GPT-3。此外，为了估算概率 P($\mathbf{R_i}$ = sound) 和 P($\mathbf{R_i}$ = flawed)，我们使用在给定 $\mathbf{S_i}$、$\mathbf{T_i}$、$\mathbf{R_i}$ 以及文本“这个重构后的想法是”作为 GPT-3 输入时，分别生成“sound”和“flawed”的词元概率。[4]

积极性。为了衡量生成的重构后想法的积极性，我们采用了一个基于 RoBERTa 的情感分类器，该分类器在 TweetEval 基准上进行了微调（Barbieri 等人，2020年）。

共情。为了衡量共情，我们基于 Sharma 等人（2020b）提出的共情分类模型进行扩展。该模型基于 RoBERTa，采用一个有理论依据的框架来描述共情，该框架包含三种共情沟通机制（情绪反应、解读和探索），并以 0 到 6 的量表预测心理健康对话中的共情程度。在此基础上，我们使用一个包含 300 条重构后想法及其共情标签的人工标注数据集（由一位具备心理健康领域共情专长的作者标注），进一步在重构后想法领域对该模型进行微调。

可操作性。为了衡量可操作性，我们假设，可执行的重新构建要么（1）建议采取具体行动，要么（2）不建议采取具体行动，但容易付诸行动。

我们将行动具体性建模为二元分类任务：给定重构$\mathbf{R_i}$，预测$contains_action(\mathbf{R_i}) \in \{0,1\}$。我们通过向 GPT-3 提供 10 个重构后的想法及其可执行性评分的配对示例，进行少样本预测（详情见附录4.2A.1）。

为了确定$\mathbf{R_i}$ 的可操作性，我们考察下一组行动由$\mathbf{R_i}$蕴含的下一组行动。我们的假设是，多样化的下一组行动可能表明存在歧义，从而降低可操作性；而连贯的下一组行动可能表明思路清晰，从而提高可操作性。在此，我们指示 GPT-3 根据重新构建后的想法生成 $k = 5$ 下一步行动候选项（指令提示；零样本）。我们使用 RoBERTa（Liu 等人，2019年）对每个 k 行动候选项进行嵌入，并计算两两之间余弦相似度的平均值，从而计算下一步行动连贯性——记为 $next_action_coherence(\mathbf{R_i})$。相似度越高，表示可能的下一步行动之间越连贯。我们的总体可操作性测量定义为 $contains_action(\mathbf{R_i}) + next_action_coherence(\mathbf{R_i})$。

具体性。遵循先前研究（Xu 等人，2018；Sharma 等人，2021年），我们通过计算重构后的想法 $\mathbf{R_i}$ 与情境 $\mathbf{S_i}$ 和想法$\mathbf{T_i}$ 拼接后文本的句子嵌入相似度来衡量具体性（使用 RoBERTa 嵌入（Liu 等人，2019年）。

可读性。我们采用常用的 Coleman-Liau 指数（CLI）指标（Coleman 和 Liau，1975），该指标根据句子中的字符和词语结构评估可读性。Coleman-Liau 指数的计算公式为 $0.0588L - 0.296S - 15.8$，其中 L：每 100 个单词中的平均字母数；S 是每 100 个单词中的平均句子数。

5.2 重新框定生成

上下文学习方法可以仅凭少量示例（少样本学习）或手写指令（指令提示）来学习泛化自然语言处理任务（Brown 等人，2020年）。然而，通过对 100 个手工编写的情境和想法进行定性分析，我们发现，采用固定示例集的简单上下文学习方法，往往无法恰当地重新构建那些没有相关上下文示例的情境和想法（例如，焦虑症患者出现“思绪飞驰”）。

为了恰当地重新构建与各种情境和想法相关的想法，我们开发了一种基于检索的上下文学习方法（Liu 等人，2022b）。对于每种情境$\mathbf{S_i}$ 和消极想法$\mathbf{T_i}$，我们从数据集中检索 k个相似示例（§4）。首先，我们使用 RoBERTa 嵌入对情境和想法进行编码。然后，我们检索 k 个示例，$\{(s_1, t_1) \ldots, (s_k, t_k)\}$，

[4]我们尝试了“合理”和“有缺陷”的不同替代说法，发现结果相似。

属性	皮尔逊相关系数
应对思维陷阱	0.680***
合理性	0.448**
积极性	0.550***
共情	0.575***
可操作性	0.647***
具体性	0.427**
可读性	0.331*

表1：我们提出的属性测量与心理健康专家人工判断之间的相关性。*：$p < 0.05$；**：$p < 0.001$；***：$p < 10^{-5}$。

来自我们的数据集，基于前k 个最高$cosine_sim(s, \mathbf{S_i}) * cosine_sim(t, \mathbf{T_i})$。我们选择 $k = 5$ (附录 A.3)。

5.3 控制生成的重构思维的语言属性

虽然我们提出的方法可以生成单个重新诠释，但除了模仿检索到的示例（§3）之外，它无法直接控制重新诠释的语言属性。在此，我们旨在改变重新诠释的语言属性。

重新构建后的想法可能会，也可能不会针对原始想法 $\mathbf{T_i}$中的思维陷阱。在此，我们生成两个重新诠释$\mathbf{R_i}^{(tt,\mathrm{Y})}$ 和 $\mathbf{R_i}^{(tt,\mathrm{N})}$，其中一个针对 $\mathbf{T_i}$ 中的思维陷阱，另一个则不针对。[5] 我们从数据集中提取两组独立的上下文示例：一组至少针对一个思维陷阱，另一组则不针对（如 §4 所述）。我们使用这些示例提示 GPT-3 生成 $\mathbf{R_i}^{(tt,\mathrm{Y})}$ 和 $\mathbf{R_i}^{(tt,\mathrm{N})}$。

此外，重新构建后的想法在合理性、积极性、共情、可操作性、具体性和可读性这六项属性上的得分可能较高，也可能较低。对于这六项属性，给定一个重新构建 $\mathrm{R_i}$ 和一项语言属性 a，我们会额外生成两个重新构建 $\mathrm{R_i}^{(a,\mathrm{H})}$ 和 $\mathrm{R_i}^{(a,\mathrm{L})}$：一个在属性 a 上的得分较高，另一个得分较低（例如，可操作性较高或较低）。为此，请回想 §4.2 中的每个（情境、想法）Pair都标注了两个重新构建，并且会根据各项语言属性对这些重新构建进行比较。对于一个人工标注实例 j，设 $\mathrm{R_j}^{*(a,\mathrm{H})}$ 和 $\mathrm{R_j}^{*(a,\mathrm{L})}$ 分别为在属性 a上被判定为高和低的重新构建。为了从 $\mathrm{R_i}$生成 $\mathrm{R_i}^{(a,\mathrm{H})}$，我们使用上下文示例 $\{\mathrm{R_j}^{*(a,\mathrm{L})} \rightarrow \mathrm{R_j}^{*(a,\mathrm{H})}\}_{j=1}^{k}$提示 GPT-3，使用 $k = 5$。

[5]如果一个想法存在多个思维陷阱，我们会检查重新诠释是否至少针对其中一个。

模型	自动 BLEU	R-1	R-L	BScore	人工 Rel.	有用性
仅检索	21.6	18.8	14.2	86.7	2.58	3.14
积极重构	24.4	23.6	17.6	87.6	2.67	2.40
DialoGPT	22.5	17.4	13.5	86.3	2.49	3.21
T5	24.9	23.4	17.8	87.2	2.51	3.30
仅GPT-3	25.0	23.9	18.0	88.3	2.97	3.98
我们的模型	**27.8**	**26.0**	**19.9**	**88.6**	**3.10**	**4.11**

表 2：自动评估与人工评估结果。R- 1：ROUGE-1；R-L：ROUGE-L；BScore：BERTScore；Rel.：亲和力；Help.：有用性。

同样，为了根据$\mathbf{R_i}$生成$\mathbf{R_i}^{(a,\mathrm{L})}$，我们使用示例 $\{\mathbf{R_j}^{*(a,\mathrm{H})} \rightarrow \mathbf{R_j}^{*(a,\mathrm{L})}\}_{j=1}^{k}$提示 GPT-3。

6 实验与结果

我们评估所提出语言属性的构念效度（§6.1），并评估重构生成模型的性能（§6.2）。

6.1 语言属性的构念效度

我们通过将所提出的语言属性测量指标与心理健康专家的人工判断进行相关性分析，验证这些指标（见§4.2）。结果显示，应对思维陷阱（0.680***）和可操作性（0.647***）的皮尔逊相关系数较强；合理性（0.448**）、积极性（0.550***）、共情（0.575***）和具体性（0.427**）的相关性中等；可读性（0.331*）的相关性较弱（表1）。[6]

6.2 重新框定生成性能

我们通过自动评估和人工评估，来评估我们在 §5.2. 中开发的所提出的重新框定生成模型的性能。

实验设置。我们采用 top-p 采样，并使用$p = 0.6$ 进行文本生成（Holtzman 等人，2020年）。我们将 600 个专家标注样例（§4）按 70:30 的比例划分为训练集和测试集。

基线。（**1**）仅检索 – 对于测试输入，我们根据 RoBERTa 嵌入的余弦相似度，检索训练集中相似度最高的示例；（**2**）积极重构 – 我们沿用 Ziems 等人（2022年）提出的基于 BART 的积极重构模型；（**3**）*DialoGPT* – 针对对话调整的 GPT-2（Zhang 等人，2020年）；（**4**）*T5* – 文本到文本转换语言模型（Raf- fel 等人，2020年）；[7]（**5**）仅*GPT-3* – 我们随机

[6]*:$p < 0.05$;**:$p < 0.001$;***:$p < 10^{-57}$仅使用 600 个样本训练 DialoGPT 和 T5 可能

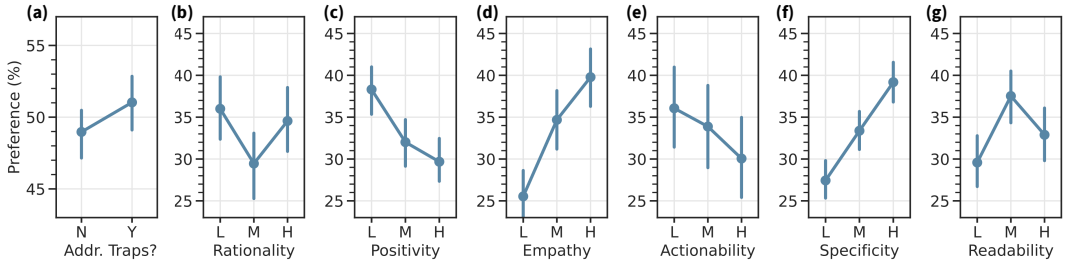


图 2：**人们更喜欢重构后的想法具有哪些语言属性？**对于 MHA 访客提供的某个情境和想法，我们向他们展示多个语言模型生成的重新诠释，这些内容在随机选取的某项属性上存在差异（例如，低、中、高可操作性）。我们发现，人们更喜欢共情度高、具体性强的重新构建。另一方面，人们较不喜欢积极性高的重新诠释。N：否；Y：是；L：低；M：中；H：高。所有图中的误差线均表示 95% 自助法置信区间。

从我们的训练集中检索 5 个示例，并用它们提示 GPT-3。

自动评估。 我们考察模型输出与上述创建的测试集划分中的真实重构表述之间的语义相似度。我们采用 BLEU（Papineni 等人，2002年）、ROUGE-1、ROUGE-L（Lin，2004年）以及 BERTScore（Zhang 等人，2019b）指标。我们发现，与表现次优的基线模型——仅GPT-3和积极重构——相比，我们提出的模型的 BLEU 得分高出 11.2%，ROUGE 得分高出 9.7%（表2）。

人工评估。我们评估了两项关键的重新构建结果指标：亲和力（重新构建后的想法有多容易引起共鸣）和有用性（重新构建后的想法在克服消极想法方面有多大帮助）。我们招募了三位心理健康从业者，请他们根据模型输出的可靠性和有用性，以 1 到 5 分的量表对测试集示例进行评分。结果发现，我们提出的模型获得了最高的亲和力和有用性评分（表2）。令人意外的是，积极重构方法的有用性评分最低，亲和力也较低，这表明，仅凭积极性来重构消极想法，可能难以引起共鸣，也不够有用。

7 在大型心理健康平台上开展的随机现场研究

接下来，我们在一个大型心理健康平台（§7.1）上部署我们的模型，并研究人们偏好的重新诠释类型（§7.2），以及亲和、有用且令人难忘的重新诠释具有哪些特征（§7.3）。

颇具挑战性。在此，我们采用过度生成策略——从收集到的数据样本出发，利用 GPT-3 的模式复现能力，再生成 10,000 个相似示例，做法与Liu 等人（2022a）类似。

7.1 模型部署

我们试图了解，我们提出的认知重构模型如何帮助经历消极想法的人。在仔细评估伦理与安全因素、与心理健康专家和临床心理学家开展积极合作（其中一些人为共同作者）并获得机构审查委员会批准后，我们将模型部署在美国心理健康协会（MHA）的网站上。这是一个大型心理健康网站，为数百万用户提供心理健康资源和工具（bit.ly/changing-thoughts）。我们开展了一项为期一个月的随机现场研究，共有 2,067 名 MHA 访客作为参与者。MHA 访客选择使用我们的模型并同意参与后，描述了自己的情境以及正在努力应对的想法。接下来，他们会看到多个按随机顺序呈现的模型生成的重构后的想法，并被要求选出自己认为最具亲和力、最有用且最令人难忘的重构后的想法，最后根据亲和力、有用性和记忆性对所选的重构后的想法进行评价（见附录 F）。

7.2 人们更喜欢哪些类型的重构后的想法？

为了了解人们偏好哪些重构属性，我们建议针对给定情境和想法，提供多个在属性取值上有所不同的语言模型生成的重构内容。首先，我们使用模型生成一个重新构建的想法。接着，随机选择一个属性（例如可操作性），并据此调整第一个重构内容（例如，生成两个可操作性更高或更低的额外重构内容），采用我们提出的可控文本生成方法（§5.3）。图2揭示了人们选择并偏好的重新诠释在语言属性上的关键差异：

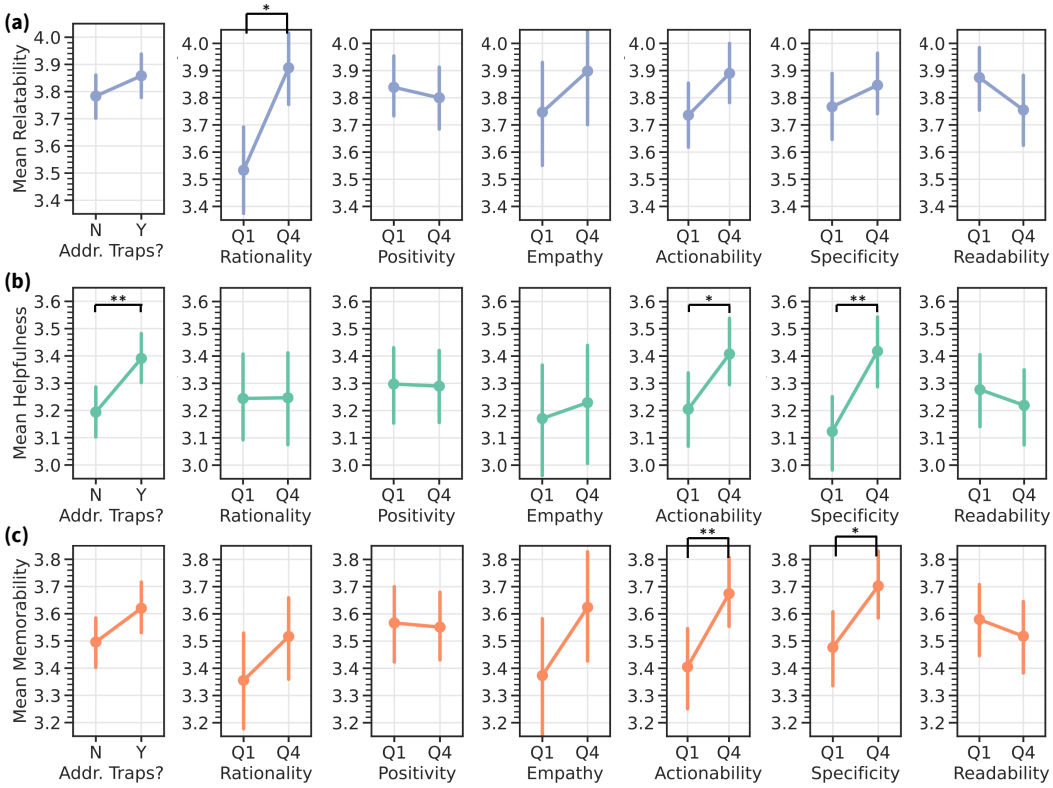


图3：**哪些语言属性与期望的认知重构结果相关？**针对给定的情境和想法，我们向 MHA 参与者展示一个由语言模型生成的重新构建，并请他们按照 1 到 5 分的量表，对其相关性、有用性和记忆性进行评分。对于每种语言属性，我们比较第一四分位数（Q1）和第四四分位数（Q4）。我们发现，**(a)** 合理性较高的重新构建更容易引起共鸣；**(b)** 针对思维陷阱、可操作性较高或具体性较高的重新构建更有用；**(c)** 可操作性较高或具体性较高的重新构建更令人印象深刻。*:$p < 0.05$;**:$p < 0.01$.

(1) 人们更偏好共情度高且具体的重构表达。我们发现，人们偏好共情度高的重新诠释的频率比偏好共情度较低的高出55.7%（39.7% 对 25.5%；$p < 10^{-5}$）；偏好具体性高的重新诠释的频率比偏好具体性较低的高出43.1%（39.2% 对 27.4%；$p < 10^{-5}$）。以往研究表明，在心理健康支持对话中，共情以及较少“模板化”的回应十分重要（Sharma 等人，2020年b；Althoff 等人，2016年）。在此，我们表明，语言模型生成的重构内容所具备的共情性和具体性，或许能帮助人们重构消极想法。

(2) 过于积极的重新诠释不太受青睐。另一方面，积极性较高的重新诠释比积极性较低的重新诠释受到的偏好低 22.7%（29.6% vs. 38.3%；$p < 10^{-5}$）。这可能是因为，对于那些已经正在经历引发情绪的消极想法的个体而言，采纳一种过于积极的重新构建后的想法可能具有挑战性（Dember 和 Penwell，1980）。

参与者也更倾向于选择中等可读性的回复-框架过于简单或过于复杂的重新诠释，可能是因为它们的语言适用于更广泛的受众。

7.3 语言属性重构后的想法与认知重新构建的期望结果有何关联？

我们评估什么样的重新构建具有共鸣、有帮助且令人难忘。我们向参与者展示一个模型生成的重构，并请他们根据三项结果指标在五点李克特量表上对其进行评分（1：非常不同意；5：非常同意）。为了避免选择效应（§7.2），我们没有在这项实验中让参与者从多个重新构建中进行选择。图3揭示了重构后的想法的哪些属性与不同的预期结果相关：

（1）合理性更高的重新诠释更容易引起共鸣。我们发现，合理性更高的重新诠释比合理性较低的重新诠释更容易引起共鸣，幅度高出 10.8%（3.91 对 3.53；$p < 0.05$）。这可能是因为合理性更高的重新诠释

顾名思义，更有可能基于理由，也不太可能做出不切实际的假设，因此更容易让人产生共鸣。

(2) 能够应对思维陷阱、且可操作性和具体性更强的重新诠释更有帮助。 能够应对思维陷阱的重新诠释，比未能应对思维陷阱的重新诠释更有帮助，提升了 6.3%（3.39 对 3.19； $p < 0.01$）。这类重新诠释会针对思维模式中的认知偏差（例如“算命式思维”；附录 D）提出质疑。心理治疗研究表明，这种方法在应对消极想法方面更有效（Beck，1976；Burns，1980）。此外，我们发现，可操作性较高的重新诠释比可操作性较低的重新诠释更有帮助，提升了 6.6%（3.41 对 3.20； $p < 0.05$）；具体性较高的重新诠释比具体性较低的重新诠释更有帮助，提升了 9.6%（3.42 对 3.12； $p < 0.01$）。

(3) 可操作性和具体性更强的重新诠释更令人难忘。 我们发现，可操作性较高的重新诠释比可操作性较低的重新诠释更令人难忘，提升了 7.9 %（3.67 对 3.40； $p < 0.01$）；具体性较高的重新诠释比具体性较低的重新诠释更令人难忘，提升了 6.3%（3.70 对 3.48； $p < 0.05$）。

8 相关工作

若干用于心理健康领域的人类与语言模型交互工具旨在协助支持服务提供者，例如临床医生（Tanana 等人，2019年； Shen 等人，2020年）或同伴（Sharma 等人，2023）。我们的研究揭示了人类与语言模型交互如何通过认知重构，直接为面临心理健康挑战的人们提供支持。关于认知重构的计算研究一直依赖小规模众包研究（Smith 等人，2021年；Morris 等人，2015年）。我们的研究开发了可扩展的认知重构方法，并在一个大型心理健康平台上开展了随机现场研究。以往的文本重构研究已针对风格、情感、礼貌表达和共情迁移等相关任务开发方法（Reif 等人，2022年；Madaan 等人，2020年；Sharma 等人，2021年），也研究了积极重构（Ziems 等人，2022年）。我们的研究开发了用于认知重构的文本重构方法，并表明，应对思维陷阱、合理性、可操作性、具体性和可读性等语言属性对于实现高质量的重新诠释至关重要。从更广泛的角度看，我们的研究也与自然语言处理在心理健康和心理健康福祉领域不断发展的研究相关（Althoff 等人，2016年；Sharma 和 De Choudhury，2018；Gaur 等人，2019年；Lee 等人，2019年；Miner 等人，2019年；Pendse 等人，2019年；Pérez-Rosas 等人，2019年；Pruksachatkun 等人，2019年；Yang 等人，2019年；Zhang 等人，2019a；Jaidka 等人，2020年；Saha 和 Sharma，2020年；Sharma 等人，2020a、b；Wadden 等人，2021年；Welch 等人，2020年；Zhang 和 Danescu-Niculescu-Mizil，2020年；Lahnala 等人，2021年；Lin 等人，2022年；Naseem 等人，2022年；Pérez-Rosas 等人，2022年；Shah 等人，2022年；Shen 等人，2022年；Stewart 等人，2023年）。

9 结论

在本文中，我们开展了一项研究，探讨人类-语言模型交互如何支持人们对消极想法进行认知重构。我们定义了一个包含七种认知重构语言属性的框架，开发了用于测量这些属性的自动化指标，并邀请心理健康专家验证其测量结果。我们从心理健康专家处收集并分享了一个包含600个情境、想法和重新诠释的数据集，并利用该数据集训练了一个基于GPT-3的检索增强型上下文学习模型。我们将该模型部署在美国心理健康协会网站上，并对2,067名参与者开展了随机现场研究。我们发现，正在与消极想法作斗争的人更偏好高度共情或具体的重新诠释，但并不偏好高度积极的重新诠释。

10 道德声明

在心理健康等高风险场景中进行干预，必须考虑与安全、隐私和偏见相关的伦理问题。AI 在尝试提供帮助时，可能会对正在应对心理健康挑战的人产生适得其反的影响。为此，我们与心理健康专业人士和临床心理学家积极合作并进行咨询，采取了多项措施来尽量降低这些风险。

机构审查委员会批准。 我们获得了华盛顿大学机构审查委员会对数据收集（机构审查委员会编号 STUDY00015882）以及随机实地研究（机构审查委员会编号 STUDY00016783）的批准。我们所在机构要求所有开展涉及人类受试者研究的研究人员通过在线 CITI 课程完成受试者保护培训。开展这些研究的研究生均已获得机构审查委员会认证。

参与者知情同意。我们获

得了随机现场研究中所有参与者的知情同意（附录 H）。所有参与者的年龄均为18岁及以上。我们告知参与者，他们将与一个基于人工智能的模型互动，该模型会自动生成重新构建的想法，且没有人工监控。此外，我们还告知他们，生成的部分内容可能会令人不安或困扰。

危机资源。 我们明确说明，模型不应用作“求助”渠道，也不应用于存在自杀意念或自我伤害的情况。此外，在研究开始时，我们向参与者提供了两项危机援助资源——危机短信热线（crisistextline.org）和 988 自杀与危机生命热线（988lifeline.org）。

安全措施。 为尽量减少语言模型生成的有害认知重构，我们过滤掉了任何包含自杀意念或与自我伤害相关词语或短语的回应。为此，我们使用自杀风险评估词汇表（如 有自杀倾向的感觉”“希望死去”“伤害自己”）创建了包含 50 个正则表达式的列表。所用词汇表包括 Gaur 等人 (2019年）。凡是匹配任一正则表达式的语言模型生成回应，都会被过滤掉，不会展示给参与者。此外，参与者还可以通过“举报不当内容”按钮标记不恰当的认知重构建议（附录 C）。

隐私。 在我们的随机现场研究中，我们没有收集任何可识别个人身份的信息，并在开展数据分析前删除了所有用户标识符。所有研究数据均存储在独立的安全计算环境中，只有受过培训的研究人员才能访问这些数据。§4.1中收集的情境和想法经过了匿名化处理：我们手动删除了所有用户标识符，并将地点、姓名等具体可识别信息替换为更通用的表述，遵循 Matthews 等人 (2017年)。

11 局限性

我们在单一平台（美国心理健康协会）和单一语言（英语）中开展了随机现场研究。然而，MHA 是心理健康资源中尤其受欢迎的来源，每年访问者超过五百万。

此外，我们注意到，一系列社会文化因素可能会影响消极想法应如何重新构建，以及用于协助这一过程的语言模型应如何开发。针对特定社群开展研究，包括代表性不足的社群和少数群体，超出了本研究的范围。确保这些工具能够公平地被获取，并使其适应各种社会文化背景，还需要进一步研究。

并非所有认知重构的实现方式都会引出情境，但我们认为，这对于让重新构建与个人产生关联至关重要。未来，当个人使用该系统处理多个情境和想法时，研究如何随着时间推移更有效地学习其背景信息将会很有趣。出于隐私方面的考虑，我们目前不会收集可用于关联多个会话的信息。不过，在充分考虑伦理因素并获得用户同意的前提下，这种方法可能会有所裨益。

本文主要聚焦于创建一种能够即时奏效的干预措施。这一研究动机源于近期临床心理学研究，该研究指出，此类单次、即时干预能够带来显著且积极的长期心理健康结果（Schlei-der 等人，2022年）。为纳入对较长期效果的部分考量，我们评估了重新构建的记忆性，这可能是其未来实用性的关键因素。尽管如此，评估长期结果至关重要，也是未来研究的重要方向。最后，我们强调，本研究并未考察短期或长期临床结果。

致谢

我们感谢参与数据标注的心理健康从业者和临床心理学研究生，以及参与我们实地研究的 MHA 访客。我们感谢 UW行为数据科学小组成员在本项目开展过程中提出的建议和反馈。我们还感谢 Justin Evans 在模型部署方面提供的协助，感谢 Xiang Lorraine Li 就数据收集提出的意见，以及感谢 Sebastin Santy 就工具界面提出的意见。T.A.、A.S. 和 I.W.L. 部分获得了 NSF 资助项目 IIS-1901386、NSF CAREER IIS-2142794、NSF 资助项目 CNS-2025022、美国国立卫生研究院资助项目 R01MH125179、比尔及梅琳达·盖茨基金会（INV-004841）、美国海军研究办公室（#N00014-21-1-2154）、微软人工智能无障碍项目资助和 Garvey研究所创新资助的支持。

参考文献

Ashley Batts Allen 和 Mark R Leary。2010 年。《自我关怀、压力与应对》。社会与人格心理学指南。

Tim Althoff、Kevin Clark 和 Jure Leskovec。2016年。咨询对话的大规模分析：自然语言处理在心理健康领域的应用。计算语言学协会会刊。

Francesco Barbieri、Jose Camacho-Collados、Luis Espinosa Anke 和 Leonardo Neves。2020年。TweetE- val：推文分类的统一基准与比较评估。载于*EMNLP Findings*。

Aaron T Beck。1976年。认知疗法与情绪障碍。InternationalUniversitiesPress。

Judith S Beck。2005年。针对棘手问题的认知疗法：当基础方法不起作用时该怎么办。Guilford Press。

Chandra Bhagavatula、Ronan Le Bras、Chaitanya Malaviya、Keisuke Sakaguchi、Ari Holtzman、Hannah Rashkin、Doug Downey、Wen-tau Yih 和 Yejin Choi。2020年。溯因常识推理。载于*ICLR*。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、JaredD Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。*NeurIPS*。

Franziska Burger、Mark A Neerincx 和 Willem-Paul Brinkman。2021年。用于认知疗法的自然语言处理：从想法记录中提取图式。*PloSone*。

Hannah A Burkhardt、George S Alexopoulos、Michael D Pullmann、Thomas D Hull、Patricia A Areán 和 Trevor Cohen。2021年。行为激活与抑郁症状：基于文本的治疗会话中语言指标的纵向评估。*JMIR*。

David D Burns。1980年。感觉良好：新情绪疗法。*NewYork*。

Meri Coleman 和 Ta Lin Liau。1975年。一种用于机器评分的计算机可读性公式。*Journal ofApplied Psychology*。

Peter Damielson、Robert Audi、Cristina Bicchieri 等。2004年。《牛津理性手册》。美国牛津大学出版社。

Daniel David、Steven Jay Lynn 和 Albert Ellis。2009年。《理性与非理性信念：研究、理论与临床实践》。牛津大学出版社。

William N. Dember 和 Larry Penwell。1980 年。幸福感、抑郁症与波莉安娜原则。心理学学会公报。

Sona Dimidjian, Manuel B阿雷拉 Jr, Christopher Martell、Ricardo F Muñoz 和 Peter M Lewinsohn。2011年。抑郁症行为激活治疗的起源与现状。《临床心理学年度评论》。

Xiruo Ding、Kevin Lybarger、Justin Tauscher 和 Trevor Cohen。2022年。改进低频认知扭曲的分类：领域专用模型与数据增强。在《*2022*年北美计算语言学协会年会论文集：人类语言技术：学生研究研讨会》。

Robert Elliott、Arthur C Bohart、Jeanne CWatson 和 Leslie S Greenberg。2011年。共情。《心理治疗》。

Manas Gaur、Amanuel Alambo、Joy Prakash Sain、Ugur Kursuncu、Krishnaprasad Thirunarayan、Ramakant Kavuluru、Amit Sheth、Randy Welton 和 Jyotishman Pathak。2019年。面向早期干预的知识感知型自杀风险严重程度评估。发表于 *WWW*。

Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。2020年。神经文本退化的奇特案例。发表于 *ICLR*。

Kokil Jaidka、Niyati Chhaya、Saran Mumick、Matthew Killingsworth、Alon Halevy 和 Lyle Ungar。2020年。超越积极情绪：基于写作提示解构快乐时刻。发表于 *ICWSM*。

Jaehun Jung、Lianhui Qin、Sean Welleck、Faeze Brahman、Chandra Bhagavatula、Ronan Le Bras 和 Yejin Choi。2022年。助产式提示：通过递归解释进行逻辑一致的推理。发表于*EMNLP*。

Carole A Kaplan、Anne E Thompson 和 Sheila M Searson。1995年。儿童和青少年的认知行为疗法。《儿童疾病档案》。

Allison Lahnala、Yuntian Zhao、Charles Welch、Jonathan K Kummerfeld、Lawrence C An、Kenneth Resnicow、Rada Mihalcea 和 Verónica Pérez- Rosas。2021年。探索在线支持论坛中自我认定的咨询专长。载于*ACL-IJCNLPFindings*。

Fei-Tzin Lee、Derrick Hull、Jacob Levine、Bonnie Ray 和 Kathleen McKeown。2019年。识别不同心理治疗方法中的治疗师的对话行为。载于第六届计算语言学与临床心理学研讨会论文集。

Rensis Likert。1932年。《态度测量》技术。《心理学档案》。

Chin-Yew Lin。2004年。ROUGE：用于摘要自动评估的软件包。载于文本摘要拓展。

Inna Lin、Lucille Njoo、Anjalie Field、Ashish Sharma、Katharina Reinecke、Tim Althoff 和 Yu- lia Tsvetkov。2022年。掩码语言模型中的性别化心理健康污名。载于*EMNLP*。

Alisa Liu、Swabha Swayamdipta、Noah A Smith 和 Yejin Choi。2022a。Wanli：面向自然语言推理数据集创建的人工与 AI 协作。*arXiv* 预印本 *arXiv:2201.05955*。

Jiachang Liu、Dinghan Shen、Yizhe Zhang、William B Dolan、Lawrence Carin 和 Weizhu Chen。2022b。什么样的上下文示例适合 GPT-3？载于*Deep Learning Inside Out*（*DeeLIO 2022*）论文集：第三届深度学习架构的知识提取与集成研讨会。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Aman Madaan、Amrith Setlur、Tanmay Parekh、Barnabás Poczós、Graham Neubig、Yiming Yang、Ruslan Salakhutdinov、Alan W Black 和 Shrimai Prabhumoye。2020年。礼貌风格迁移：一种标记并生成方法。载于*ACL*。

Tara Matthews、Kathleen O'Leary、Anna Turner、Manya Sleeper、Jill Palzkill Woelfer、Martin Shelton、Cori Manthorne、Elizabeth F Churchill 和 Sunny Consolvo。2022年。幸存者的故事：应对亲密伴侣虐待时的隐私与安全实践。载于*CHI*。Adam S Miner、Nigam Shah、Kim D Bullock、Bruce A Arnow、Jeremy Bailenson 和 Jeff Hancock。2022年。将对话式人工智能纳入心理治疗时的关键考量。*Frontiers in Psychiatry*。Robert R Morris、Stephen M Schueller 和 Rosalind W Picard。2022年。基于网络、众包的点对点认知重评平台治疗抑郁症的疗效：随机对照试验。*JMIR*。Usman Naseem、Adam G Dunn、Ji nman Kim 和 Matloob Khushi。2022年。使用序数分类在 Reddit 上早期识别抑郁症严重程度。载于*WWW*。Mark Olfson。2022年。建设治疗患有严重精神疾病成人所需的心理健康 workforce 能力。*Health Affairs*。Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。2022年。BLEU：一种自动评估机器翻译的方法。载于*ACL*。Charles Sanders Peirce。2022年。《查尔斯·桑德斯·皮尔士文集》。哈佛大学出版社。Sachin R Pendse、Kate Niederhoffer 和 Amit Sharma。2022年。在线心理健康支持论坛使用方面的跨文化差异。*CSCW*。

Verónica Pérez-Rosas、Kenneth Resnicow、Rada Mihalcea 等。2022年。Pair：用于动机式访谈中咨询师反思评分的提示感知边际排序。载于*EMNLP*。

Verónica Pérez-Rosas、Xinyi Wu、Kenneth Resnicow 和 Rada Mihalcea。2019年。什么样的咨询师才算优秀？学习区分高质量和低质量的咨询对话。载于*ACL*。

Yada Pruksachatkun、Sachin R Pendse 和 Amit Sharma。2019年。改变的时刻：分析在线心理健康论坛中基于同伴的认知支持。载于*CHI*。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020年。探索统一文本到文本转换器的迁移学习极限。*JMLR*。

Emily Reif、Daphne Ippolito、Ann Yuan、Andy Co-enen、Chris Callison-Burch 和 Jason Wei。2022年。利用大型语言模型进行任意文本风格迁移的方法。在 *ACL*。

Koustuv Saha 和 Amit Sharma。2020年。在线心理健康社区中实现有效社会心理结果的因果因素。在*ICWSM*。

Jessica L Schleider、Michael C Mullarkey、Kathryn R Fox、Mallory L Dobias、Akash Shroff、Erica A Hart 和 Chantelle A Roulston。2022年。新冠肺炎期间针对青少年抑郁症的在线单次干预随机试验。《自然·人类行为》。Raj Sanjay Shah、Faye Holt、Shirley Anugrah Hayati、Aastha Agarwal、Yi-Chia Wang、Robert E Kraut 和 Diyi Yang。2022年。对在线点对点咨询平台中的动机式访谈策略进行建模。*CSCW*。Ashish Sharma、Monojit Choudhury、Tim Althoff 和 Amit Sharma。2020a。心理健康平台上点对点互动的参与模式。载于*ICWSM*。Ashish Sharma、Inna W Lin、Adam S Miner、David C Atkins 和 Tim Althoff。2021年。促进在线心理健康支持中的共情对话：一种强化学习方法。载于*WWW*。Ashish Sharma、Inna W. Lin、Adam S. Miner、David C. Atkins 和 Tim Althoff。2023年。人机协作能在基于文本的点对点心理健康支持中促成更多共情对话。《自然·机器智能》。Ashish Sharma、Adam S Miner、David C Atkins 和 Tim Althoff。2020b。理解基于文本的心理健康支持中所表达共情的一种计算方法。载于*EMNLP*。Eva Sharma 和 Munmun De Choudhury。2018年。心理健康支持及其与在线社区中语言适应的关系。载于*CHI*。

Siqi Shen、Verónica Pérez-Rosas、Charles Welch、Sou-janya Poria 和 Rada Mihalcea。2022年。面向咨询对话的知识增强型反思生成。载于*ACL*。

Siqi Shen、Charles Welch、Rada Mihalcea 和 Verónica Pérez-Rosas。2020年。利用增强语境的生成式预训练转换器生成咨询风格的反思。发表于*SIGDIAL*。

Amy E Sickel、Jason D Seacat 和 Nina A Nabors。2014年。心理健康污名更新：后果综述。心理健康进展。

C Estelle Smith、William Lane、Hannah Miller Hillberg、Daniel Kluver、Loren Terveen 和 Svetlana Yarosh。2021年。众包驱动的认知重评系统的有效策略：面向心理健康的 Flip* Doubt 网络应用的实地部署。*CSCW*。

Ian Stewart、Charles Welch、Lawrence An、Kenneth Resnicow、James Pennebaker 和 Rada Mihalcea。2023年。支持健康相关行为改变的表达式访谈代理：一项关于新冠肺炎行为的研究。*JMIR* 形成性研究。

Michael J Tanana、Christina S Soma、Vivek Srikumar、David C Atkins 和 Zac E Imel。2019年。Clientbot 的开发与评估：用于训练基础咨询技能的类患者对话代理。*JMIR*。

David Wadden、Tal August、Qisheng Li 和 Tim Al-thoff。2021年。审核对在线心理健康对话的影响。载于*ICWSM*。

Charles Welch、Allison Lahnala、Verónica Pérez-Rosas、Siqi Shen、Sarah Seraj、Larry An、Kenneth Resnicow、James Pennebaker 和 Rada Mihalcea。2020年。《表达性访谈：一种应对新冠肺炎的对话系统》。载于*EMNLP 2020* 上第 *1* 届新冠肺炎自然语言处理研讨会论文集（第 *2* 部分）。

Xinnuo Xu、Ondˇrej Dušek、Ioannis Konstas 和 Ver- ena Rieser。2018 年。通过建模、过滤和优化连贯性与多样性来改善对话。在 *ACL*。

Diyi Yang、Zheng Yao、Joseph Seering 和 Robert Kraut。2019年。渠道很重要：在线癌症支持小组中的自我披露、互惠与社会支持。载于*CHI*。

Justine Zhang 和 Cristian Danescu-Niculescu-Mizil。2020年。咨询对话中的目标平衡：向前推进还是回顾过去。发表于*ACL*。

Justine Zhang、Robert Filbin、Christine Morrison、Jaclyn Weiser 和 Cristian Danescu-Niculescu-Mizil。2019a。《找到自己的声音：心理健康咨询师的语言发展》。载于*ACL*。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。2019b。Bertscore：使用 BERT 评估文本生成。*arXiv* 预印本 *arXiv:1904.09675*。

Yizhe Zhang、Siqi Sun、Michel Galley、Yen-Chun Chen、Chris Brockett、Xiang Gao、Jianfeng Gao、Jingjing Liu 和 Bill Dolan。2020年。DialoGPT：用于对话回应生成的大规模生成式预训练。载于ACL，系统演示</style)。

Caleb Ziems、Minzhi Li、Anthony Zhang 和 Diyi Yang。2022年。通过文本重构引导积极视角。发表于 *ACL*。

一种方法

A.1 重构后想法的语言特征

我们对 § 中描述的方法进行了补充说明5.1。

可操作性。 如§5.1所述，我们使用以下指标衡量可操作性： $contains_action(\mathbf{R_i})$，以及 $next_action_coherence(\mathbf{R_i})$。

对于 $contains_action(\mathbf{R_i})$，我们的少样本上下文学习方法如下。我们使用收集的数据中被标注为高或低可操作性的重构后想法（§4.2），手动创建 10 个示例。如果一个重构后的想法包含某个动作，我们便要求 GPT-3 从中提取该动作；否则，则要求其生成文本"*No Action*"。附录A.2 展示了示例。随后，我们将这 10 个示例用作上下文示例，再输入待分类的重新构建 $\mathbf{R_i}$ 。如果 GPT-3 预测$\mathbf{R_i}$中存在一个动作，我们就赋予$contains_action(\mathbf{R_i}) = 1$；否则赋值为 0。

对于 $next_action_coherence(\mathbf{R_i})$，我们指示 GPT-3 根据一个重构后的想法生成 $k = 5$ 个可能的后续动作。给定 $(\mathbf{S_i}, \mathbf{T_i}, \mathbf{R_i})$，令$\mathbf{A_i} = a_{i1}$、$a_{i2}$、……、$a_{ik}$为生成的后续动作集合。令 $emb(\cdot)$ 表示 RoBERTa 嵌入。然后，我们将 $next_action_coherence(\mathbf{R_i})$定义为 $emb(a_i)$与 $emb(a_j)$在所有 a_i、$a_j \in \mathbf{A_i}$上的平均余弦相似度。

A.2 行动生成提示

我们使用以下提示模板，通过 GPT-3 提取行动：

> **陈述**："我的银行卡可能在很多不同的地方，我想先逐一查看，再得出任何结论。"**建议行动**："查看银行卡。"

> **陈述**："我取消了那次旅行，因为不得不这么做。虽然这样做让我很难受，但这是正确的决定。"**建议行动**：无

此外，我们使用以下指令提示，通过 GPT- 3 生成下一组行动："根据以下陈述，建议这个人可以采取的 *5* 项行动："

A.3 我们提出的检索增强上下文学习方法的超参数选择

对于要检索的示例数量，我们测试了 $k = 1$, 5，10 和 20，并基于对 100 条人工编写的情境和想法进行的定性评估，发现 $k = 5$能生成最有效的重构后的想法。

B 可复现性

论文中创建的代码和数据集将在 https://anonymous 下以仅需署名的学术许可协议共享。现有人工制品的使用符合其预期用途。对于基于 GPT-3 的模型，我们将使用 OpenAI 库。对于其他深度学习模型，我们将在两块 NVIDIA Titan RTX GPU 上训练这些模型。我们使用 evaluate Python 库（pypi.org/project/evaluate）计算 BLEU 和 ROUGE 分数，并使用 scipy 进行统计检验。

C 被标记的重构建议

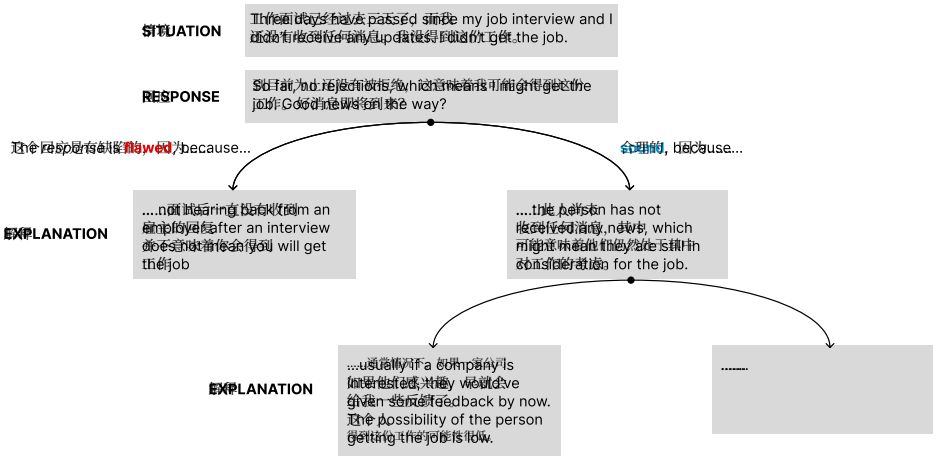
在 5,760 条重新诠释建议中，有 32 条被标记（0.56%）。其中 19 条较为笼统（59%），5 条对当事人的情境做出了错误假设（16%），另有 8 条可能无法引起当事人的共鸣（25%）。重要的是，我们没有发现任何有害或不安全的被标记重构建议；在这些情境中，这一点至关重要。未来，探索创建更个性化的重新诠释建议的方法，或许有助于避免建议过于笼统、带有假设性或不够贴合当事人。

D 思维陷阱列表

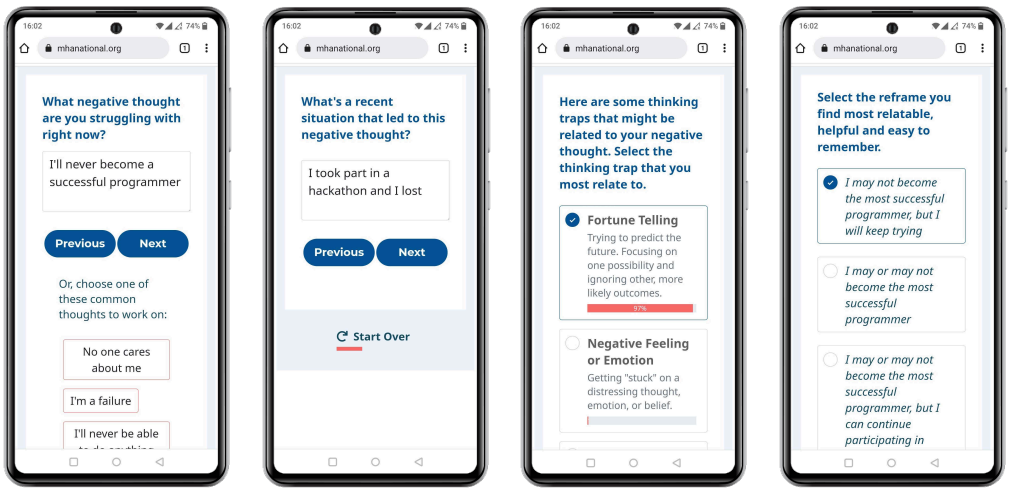
思维陷阱	描述	示例
非黑即白思维	极端思考。	如果不够完美，那我就是失败了。根本不存在所谓的“差不多就好”。
过度概括	仅凭一次经-ence.	他们没回我消息。根本没人会回我消息回到我身边。
贴标签	根据一个行为或特征来定义一个人。特征。	我说了些令人尴尬的话。我真是个失败者。
算命式思维	试图预测未来。只关注一种可能性，而忽视其他更有可能出现的结果。	我开会要迟到了。我会出洋相的。
读心术	假设自己知道别人正在想什么。	她没跟我打招呼。她一定是在生我的气。
情绪化推理	把自己的感受当成事实。	我醒来时感到焦虑。我就是知道有些—今天会有不好的事情发生。
“应该”式陈述	对自己设定不切实际的期望。	我不应该需要寻求帮助。我应该是独立的。
个人化	把事情归咎于自己，或认为事情与自己有关you.	他今天很安静。我想知道自己做错了什么。
否定积极因素	当好事发生时，你会忽视它或认为它不算数。	我只是运气好才赢的。
灾难化思维	只关注最坏的情况。	老板问我有没有几分钟时间谈谈。我要被开除了！
比较与沮丧	拿自己的最糟情况与别人的最佳状态相比。	我侄女的生日派对上，人数是.....的两倍
责怪	把自己的力量拱手交给别人。	我大吼不是我的错，是你惹我生气的！
负面感受或情绪	陷入令人苦恼的想法中，情绪-tion, 或信念。	我感到很孤独。

E 示例：说明我们的合理性测量

图4：为测量推理强度，我们为每个重新构建生成两个解释——一个说明它为何可能合理，另一个说明它为何可能有缺陷。为了检查这些解释本身是否推理充分，我们递归地为这些解释生成解释。这里，我们选择的递归树深度为 3。此外，在每一步中，我们都会生成三个支持某种重新构建的解释，以及三个反对它的解释。



F 随机实地研究界面



(a) 想法 **(b) 情境** **(c) 认知扭曲** **(d) 重构后的想法**

图 5：我们随机现场研究所用界面的示意图：**(a)** 参与者首先写下此刻困扰自己的负面想法；**(b)** 我们请参与者描述一个可能引发该想法的近期情境；**(c)**人工智能模型识别该想法中可能存在的认知扭曲。参与者选择自己最有共鸣的认知扭曲；**(d)** 人工智能模型生成并推荐三个不同的重构后的想法，帮助克服负面思维及相关的认知扭曲。参与者选择自己认为最有共鸣、最有帮助且最容易记住的重新构建。为简洁起见，此图省略了提供给参与者的部分说明，包括知情同意和评估。

G 数据收集说明

图 6：在与心理健康专家进行数据收集期间显示的
说明。续见下一页（1/3）。

认知重构

研究目标

本研究的目标是收集用于认知重构的数据集。

定义

情境	发生在个人身上或情境中的任何事情 这个人所处的（例如："我的老板从我身边走过 穿过走廊时没有向我打招呼"）
想法	这个人在情境中的想法 （例如："他们为什么生我的气？"）。
思考陷阱	~~思维陷阱，也~~称为认知扭曲，是 **夸大、偏颇且不理性的**想法， 会导致人们对现实产生不准确的认知。
思维Trap类别	思维陷阱的类别包括臆测 他人的想法（"读心术"）、非黑即白地思考（"全盘 皆是或一无所有的思维"）、只关注最糟糕的 情况（"灾难化思维"）、只关注坏的一面 （"否定积极因素"）等。

思维陷阱示例

情境： 老板从走廊经过我身边，却没有跟我打招呼

想法： 他们为什么生我的气？

思维陷阱： 读心术

图7：数据收集期间向心理健康专家展示的说明。续见下一页（2/3）。

在这里，我在揣测老板的想法，并假设他们对我不满
。除非我去问他们，否则我无法弄清楚这件事。

想法 思维回应是自我对话（与自己交谈），旨在挑战
回应 原始想法中的思维陷阱
原始想法

思维回应示例

情境： 老板从走廊经过我身边，却没有跟我打招呼。

想法： 他们为什么生我的气？

回应 1： 我无从揣测他们可能在想什么。
也许她当时有很多事情要操心

回应 2： 他们是我认识的最棒的人。他们一定没有
注意到我。

回应 3： 他们可能生我的气了，但至少他们什么也没说。

认知 认知重构是一个帮助人们**重构**觉察自己想法中的思维陷阱，
并理性地应对这些陷阱的过程。

研究步骤

在这项研究中，你将完成 20 项认知重构任务。每项任务中，你会看到一个情境和一个想法。你需要识别想法中的思维陷阱，并写下思维回应并为其评分。

注意： 不建议使用“两者相似”选项（如适用）。只有当两个回应确实完全相同，且没有任何可区分之处时，才使用此选项。

图8：与心理健康专家进行数据收集时展示的说明（3/3）。

内容警告

本研究包含的情境和想法涉及但不限于自我伤害和自杀意念，可能会令人不安。如有任何疑问或顾虑，请发送电子邮件联系我们。如果某些内容令您产生强烈的负面反应，您可以联系危机短信热线，也可以发送短信 HOME 至 741741。

如果您对作为研究参与者所享有的权利有任何疑问，或希望获取信息、提出问题，或与研究人员以外的人员讨论对本研究的任何疑虑，请联系人体受试者部，电话：xxx。

H 随机现场研究中使用的 MHA 知情同意书

图9：向 MHA 访客展示的知情同意书。续见下一页（1/2）。

使用条款

此工具使用人工智能生成重构后的想法，是一项研究的一部分。

目的：本研究旨在了解数字工具如何帮助人们识别思维陷阱，并练习重构消极想法。

流程：我们会请您描述一个想法，以及一个让您困扰的情境。然后，您将找出该想法中可能存在的“思维陷阱”（或认知扭曲），并以更积极、现实或有帮助的方式重新构建这个想法。最后，我们会请您填写一份可选的人口统计调查，您可以根据意愿跳过。预计完成此工具需要 ~5 分钟。

益处：使用此工具，您可能会了解思维陷阱，并练习识别这些陷阱以及重构消极想法和情境。不过，我们无法保证此工具一定能帮助您重新构建自己的想法。

数据收集与共享：我们不会询问您的姓名或任何可识别身份的个人信息。我们会尽可能对使用数据进行匿名化处理，并对其进行分析以改进此工具；这些数据也可能会用于未来的研究并与他人共享。

风险：谈论您正在努力应对的情境和想法可能会让您感到不安，并引发负面情绪反应。此外，该工具使用人工智能生成重构后的想法。我们已采取适当措施，以避免产生有害的重新诠释，但生成的内容仍有可能令您感到不适。此外，可选的人口统计调查会询问一些可能涉及敏感内容的信息，并可能让您感到不自在（例如：“目前导致您出现心理健康问题的主要因素是什么？”）。这

图10：向 MHA 访客展示的知情同意书（2/2）。

该工具并非由人工实时监控，不应将其用作“求助信号”渠道。如果某些内容引起你强烈的负面反应，可以发送短信至 741741 联系 MHA，或拨打或发送短信至 988。

参与： 参与本研究完全出于自愿。您不会因参与而获得任何报酬。您可以拒绝参与，也可以随时停止参与，且不会因此受到惩罚，也不会失去原本有权享有的任何福利。

联系我们： 如果您对本研究有任何疑问或顾虑，或者认为自己因参与本研究而受到伤害，请发送电子邮件至 XXX 联系我们。如果您对自己作为研究参与者的权利有任何疑问，可以致电人体受试者部，电话号码为 XXX。

☐ 勾选此框即表示您同意使用此工具，也确认您已年满18岁。请确保您关于此工具的问题已得到解答，并且您了解需要进行的操作。如果之后想到问题，可以联系我们。您可以随时停止使用此工具。如需保存本知情同意书的副本，请使用此链接。

上一步 Next