

大语言模型干预网络欺凌受害者心理韧性： 以自尊自悯为靶点*

岸本鹏子^{†1,2} 郝熙鸣^{†3} 艾尼卡尔·艾斯卡尔³
夏雨飞⁴ 白麒钰⁴

(¹北京师范大学心理学部; ²应用实验心理北京市重点实验室, 北京 100875)
(³南开大学社会学院社会心理学系, 天津 300350) (⁴北京大学新媒体研究院, 北京 100871)

摘要 本研究以自尊和自悯为靶点, 探索大语言模型(Large Language Models, LLM)在干预网络欺凌受害者心理韧性中的效果与作用路径。研究 1 (干预效果检验)采用随机对照设计, 比较自尊-自悯对话、心理教育对话和心理教育阅读三组的干预效果。结果显示两组对话式 LLM 均显著提升了心理韧性。研究 2 (作用路径分析)进一步采用网络干预分析揭示干预作用路径, 发现自尊-自悯对话干预除了可以直接提高心理韧性外, 还可以靶向孤独感、正念和社交维度自尊, 通过自尊自悯多维度的相互作用, 提升干预效果。本研究首次聚焦自尊与自悯路径, 基于 LLM 构建具备共情交互能力的对话干预平台, 为网络欺凌受害者提供个性化心理支持, 并为智能化心理干预工具的开发提供理论依据与实践路径。

关键词 大语言模型, 网络欺凌, 心理韧性, 自尊, 自悯

分类号 R395

1 引言

网络欺凌作为一种日益普遍的网络暴力形式, 正在对青少年和年轻人群体的心理健康造成广泛而深远的影响。社交媒体的匿名性、即时性与高度可接近性显著提升了网络欺凌的发生率, 受害者群体也在持续扩大(Kowalski et al., 2023)。公安部数据显示, 仅 2024 年我国就查处网络欺凌案件 8600 余起(何春中, 2025), 中国青年报社调查亦指出, 65.3%的受访者曾亲历或目睹网络欺凌(吴欣宇等, 2023), 凸显其在中国社交网络环境中的广泛性与紧迫性。网络欺凌会导致受害者出现抑郁、焦虑、社会回避、述情障碍、幸福感下降、睡眠障碍、物质滥用乃至慢性疼痛等多方面生理与心理损伤(Kowalski et al., 2023; Rémond et al., 2014;

Sourander et al., 2010; Ybarra, 2004)。相较于传统校园欺凌, 网络欺凌因其隐蔽性与持久性, 更难被察觉与干预, 对个体的心理冲击也更为复杂(Patchin & Hinduja, 2015)。尽管已有研究从政策、教育、平台治理等层面尝试构建多维防治体系, 但在帮助受害者恢复心理健康、提升心理复原力方面仍存在较大不足。

心理韧性(Psychological Resilience)作为个体应对逆境的关键心理资源, 被认为是缓冲网络欺凌负面影响的重要保护因素。根据 Kumpfer (1999)提出的心理韧性框架理论(Resilience Framework), 心理韧性体现了个体在认知、情绪与行为层面对压力源的适应能力。研究发现, 心理韧性水平较高的个体在面对网络欺凌时更能保持情绪稳定、采取积极应对策略, 并减少因欺凌而导致的心理困扰和后续健

收稿日期: 2025-06-11

* 国家自然科学基金(72304018), 青年人才托举工程(2023QNR001), 南开大学亚洲研究中心项目(AS2209)资助。

† 岸本鹏子和郝熙鸣为共同第一作者。

通信作者: 白麒钰, E-mail: baiqiyu_pku@163.com

康问题(Batmaz et al., 2023; Hinduja & Patchin, 2017; Lee et al., 2024)。此外,个体的心理韧性水平会影响其自我效能感、情绪调节能力和目标感(Santos et al., 2021),有助于其构建稳定的家庭支持和同伴关系(Hinduja & Patchin, 2017),这些资源为人们面对网络欺凌提供了情感支持与应对策略。因此,提升心理韧性已成为网络欺凌干预的重要方向之一。部分实践项目已尝试将心理韧性纳入干预核心,在降低网络欺凌发生率、提高情绪调节能力和促进心理恢复方面产生了积极作用(Cross et al., 2016; Navarro et al., 2015; Sontag-Padilla et al., 2018)。但相关研究仍处于发展初期,缺乏明确的机制性探索。

在提升心理韧性的心理机制中,自尊(Self-esteem)与自悯(Self-compassion)被视为两个关键且可干预的内在资源。根据 Kumpfer (1999)的心理韧性框架理论,自尊和自悯均为影响心理韧性的重要个人因素(Person Factors),人们面对压力时的自尊和自悯水平与环境中的风险和保护因素共同作用,影响人们的心理健康和幸福感。从自我结构的视角来看,自尊和自悯是构成自我认识的两个重要维度(Neff, 2011)。自尊体现个体对自身价值的认知评价,而自悯反映个体在遭遇失败或痛苦时对自己的情感态度和调节能力(Neff, 2011)。这两种与自我相关的心理资源在心理适应过程中可以相互支持,例如自悯有助于减轻低自尊对心理健康的不利影响(Marshall et al., 2015)。与此同时,自尊与自悯在增进心理健康和促进应对压力方面具有相对独立的功能。研究指出,高自尊个体在遭遇网络欺凌时更少表现出无助与情绪困扰,且更能采取有效应对策略(Kaplan, 1980; Veselska et al., 2009);而高自悯水平则有助于减少自责与情绪沉溺,增强积极情绪体验与心理恢复力(Bluth et al., 2018; Chu et al., 2018)。然而,现有网络欺凌干预多聚焦于认知教育或行为训练,较少有研究系统聚焦这两个核心心理过程,靶向提升心理韧性。因此,将自尊和自悯纳入同一干预框架中,不仅可以通过自尊途径增强个体的自我价值感和应对信心,还可以通过自悯途径培养个体的自我关怀和情绪调节能力,两种路径相互补充,形成更为全面和稳定的心理韧性提升机制。这种整合性干预设计有望产生协同增效作用,为网络欺凌受害者提供更为系统和深层的心理支持。

当前网络欺凌干预仍以线下心理教育或团体课程为主,受限于时间、人力与场景成本,其覆盖

范围和持续效果较为有限(Foody et al., 2015; Kazdin, 2019)。部分线上干预项目如“停止网络欺凌(Online Pestkoppenstoppen)”虽尝试引入社交技能训练与模块化支持,但多为通识性设计,难以应对不同受害者的个体差异和深层心理需求(Jacobs et al., 2014)。近年来,人工智能(Artificial Intelligence, AI)技术开始被探索性地应用于网络欺凌干预,但多停留在内容推送、知识问答等浅层支持,缺乏情感交互与机制导向的个性化干预设计(Chaudhary et al., 2024; Milosevic et al., 2023)。在此背景下,发展兼具靶向性与交互性的干预模式成为提升干预效能的关键突破口。本研究拟从以下方面对当前干预模式进行改进:技术层面上,研究充分利用 AI 的智能化和个性化优势,从静态的内容推送转向基于用户具体经历的动态对话生成;干预层面上,研究利用个性化 AI 工具,实现从通识性心理教育转向自尊自悯机制的精准靶向;交互层面上,通过参与者与 AI 的多轮对话,实现从单向信息传递转向建立虚拟治疗关系的双向情感交流。

大语言模型(Large Language Models, LLM)作为自然语言处理领域的重要突破,逐渐展现出在心理健康支持中的应用潜力。LLM 是基于海量文本数据训练、包含数千亿或更多参数、基于 transformer 架构的语言模型(林原 等, 2025)。LLM 能够自然地理解用户输入、生成连贯且语义丰富的回应,具备较高的语言灵活性与上下文适应能力(Milosevic et al., 2023)。作为心理干预的实施载体,LLM 高度自动化的特性使其能够实现持续服务,突破传统心理咨询在时间与人力资源上的限制(Kazdin, 2019);其次,其语言生成能力支持更具个性化和情境适应性的对话内容,有助于提升用户的交互体验与干预接受度(Chaudhary et al., 2024)。当前,尽管 LLM 在心理干预领域的应用仍处于探索阶段,但已有初步实践尝试将其用于情绪支持、压力管理、认知重构等场景,显示出其在提供心理教育、引导情绪表达、增强自我觉察等方面的潜在价值(Carlbring et al., 2023; Chiu et al., 2024; Kishimoto et al., 2025; Raile, 2024)。因此,基于 LLM 构建对话式心理干预系统,不仅具有现实可行性,也可以进一步推动心理健康服务的智能化、普及化与精准化发展。

综上所述,本研究旨在回应现有网络欺凌干预在个性化、机制导向与情感交互方面的不足,探索以 LLM 为载体、以自尊与自悯为干预靶点的对话

式干预对提升网络欺凌受害者心理韧性的有效性与作用路径。研究分为两个独立实验: 研究1(干预效果检验)采用三组平行随机对照试验设计, 比较自尊-自悯对话 LLM (靶向干预)、心理教育对话 LLM (常规干预)和心理教育阅读(对照干预)三种干预方式在提升心理韧性与改善心理健康方面的差异; 研究2(作用路径分析)则进一步考察两种 LLM 对话干预的作用路径, 整合自尊、自悯等关键变量, 采用网络干预分析方法识别不同干预方式下的变量联结结构。通过双实验设计, 本研究不仅验证了以心理机制为导向的干预策略在网络欺凌场景中的适用性, 也为人工智能驱动的个性化心理援助提供了理论依据与实践启示。

2 研究1: 干预效果检验

研究1为随机对照试验, 旨在探究自尊-自悯对话 LLM 干预对网络欺凌受害者心理韧性的提升效果。参与者被随机分配至三组干预条件: 自尊-自悯组(实验组1)、心理教育组(实验组2)和阅读组(对照组), 对比干预前后在主要与次要结果变量上的变化。根据 Chmitorz 等人(2018)对心理韧性研究测量工具的建议, 在测量心理韧性时最好选择状态性的量表, 并同时测量与心理韧性有关的变量(例如焦虑、抑郁、压力和幸福感等), 以作为次要结果支撑研究结论。因此, 本研究的主要结果包括状态和特质心理韧性, 次要结果包括焦虑、抑郁、压力和幸福感。此外, 由于需要控制参与者网络欺凌受害程度的组间差异, 因此在基线时补充测量了网络欺凌受害程度。研究假设如下:

H_1 : 自尊-自悯对话 LLM 能显著提高网络欺凌受害者的心理韧性, 即自尊-自悯组的参与者相较阅读组, 干预后心理韧性和幸福感显著提高, 焦虑、抑郁和压力显著降低。

H_2 : 心理教育对话 LLM 也有一定的干预效果, 即心理教育组的参与者相较阅读组, 干预后心理韧性和幸福感显著提高, 焦虑、抑郁和压力显著降低, 但效应量低于自尊-自悯组。

研究已通过北京师范大学心理学部伦理审查(BNU202504140106), 并在预见心理学预注册平台完成注册(202505.00007V1)。

2.1 方法

2.1.1 参与者

由于研究目标是基于自然情境中的网络欺凌受害者开展预防式微干预, 而非对具有明确临床诊

断的个体实施治疗性干预, 因此研究的人组逻辑遵循“高生态效度优先”的原则, 并未进行严格的临床筛选。研究通过抖音和小红书平台招募参与者, 入组标准为: (1)主观报告其曾经历网络欺凌; (2)年满18周岁; (3)愿意完成基线测量与干预任务。现实中的网络欺凌受害者往往伴随多重压力源(如线下欺凌、情绪困扰), 因此在招募时并未排除曾经历线下/同伴欺凌的个体, 以及自我报告有抑郁、焦虑等心理困扰的个体。

使用 G*Power 3.1 估算样本量, 参考研究预实验的效应量(见附录), 设定参数为: 重复测量方差分析, 效应量 $\eta_p^2 = 0.1$ ($f = 0.33$), $1 - \beta = 0.80$, $\alpha = 0.05$, 得到最小样本量为 27 人。为提高研究的统计稳健性并应对可能的参与者流失, 本研究在满足样本量估算要求的基础上扩大了招募规模, 最终共收到 63 份问卷。排除未通过注意力检查者 4 人, 最终共纳入参与者 59 人(男性 17 人, 占 28.8%; 女性 42 人, 占 71.2%), 平均年龄 23.05 ± 3.19 岁。参与者被随机分至三组: 自尊-自悯组(20 人)、心理教育组(19 人)和阅读组(20 人)。所有参与者均在知情同意书中被告知研究性质、可能的情绪激活风险以及可随时选择退出的权利。参与者均自愿参加实验并知情同意, 完成实验后将获得报酬 15 元。CONSORT 流程图见图 1。

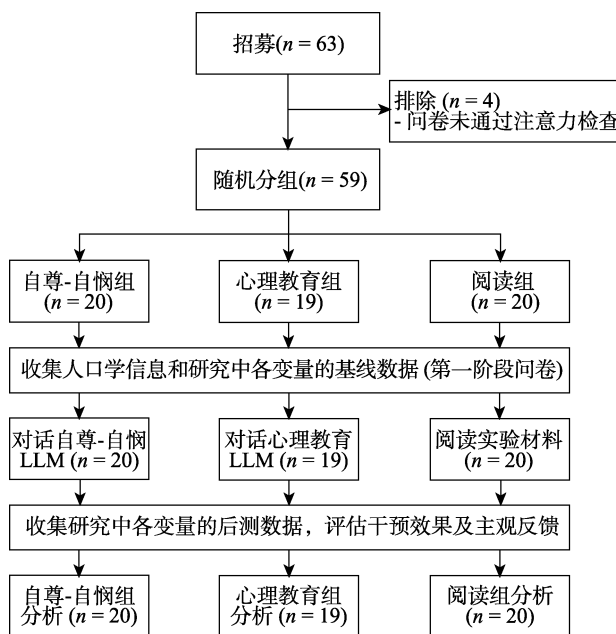


图1 研究1 CONSORT 流程图

2.1.2 测量工具

状态和特质韧性使用 Lock 等人(2020)编制的

状态-特质韧性量表,共 21 条目,1~4 级评分,分为状态韧性和特质韧性两维度,分数越高表明心理韧性越高。中文版量表由研究者遵循“翻译-回译-专家讨论和修改”流程自译,信度良好(前测 Cronbach's $\alpha = 0.82$; 后测 Cronbach's $\alpha = 0.76$)。

焦虑、抑郁和压力使用 Henry 和 Crawford (2005) 编制的 DASS-21 量表,共 21 条目,0~3 级评分,分为抑郁、焦虑和压力三维度,分数越高表明抑郁、焦虑和压力越严重。本研究采用龚栩等人(2010)修订的中文版量表,信度良好(前测 Cronbach's $\alpha = 0.96$; 后测 Cronbach's $\alpha = 0.95$)。

幸福感使用 Campbell (1976)编制的幸福感指数量表,共 9 条目,1~7 级评分,分数越高表明幸福感越高。本研究采用李艳等人(2014)修订的中文版量表,信度良好(前测 Cronbach's $\alpha = 0.88$; 后测 Cronbach's $\alpha = 0.71$)。

网络欺凌受害程度使用 Shapka 和 Maghsoudi (2017)编制的简式网络受害量表,共 12 条目,1~5 级评分,分数越高表明网络欺凌受害程度越高。本研究采用解晓龙等人(2022)修订的中文版量表,信度良好(Cronbach's $\alpha = 0.75$)。

2.1.3 实验材料

本研究实验组需要与 LLM 进行交互,交互平台为研究者自己研发的微信小程序。相较其他 LLM,有更多研究支持 ChatGPT 系列 LLM 应用于心理治疗的有效性(Carlbring et al., 2023; Raile, 2024),因此本研究使用 ChatGPT 4o 进行 LLM 干预。小程序内嵌入 ChatGPT 4o 的 API 接口,实现参与者和 LLM 的实时对话。

自尊-自悯组的 LLM 为研究者自行设计,结合当下针对网络欺凌受害者心理韧性的常规干预手段与研究人员自身临床经验。该 LLM 根据认知行为疗法原理设计,以心理治疗师的角色开展干预,将心理韧性设定为最终干预目标,同时将自尊自悯作为直接干预靶点。LLM 嵌入了针对自尊的价值澄清和成就回顾技术,以及针对自悯的自我关怀和共同人性技术,同时设置了靶点检测指令,要求 LLM 在对话中识别和回应用户的自尊自悯相关表达。心理教育组的 LLM 则以客观、中立的口吻,系统性地呈现网络欺凌的机制及应对策略,不提供情感性支持。两组 LLM 均采用半结构式对话流程,包含“创设情境与建立安全感”“回顾创伤事件”“核心干预对话”“巩固与强化”四个阶段。提示词内容涵盖干预目标、LLM 角色设定、干预技术、风险应

对策略、情感风格 and 用户体验约束等。为控制风险,“回顾创伤事件”阶段限定在三轮以内,避免情绪过度激活或二次创伤;整体对话轮数不超过十轮,以保证干预的精简性和聚焦性,减少情绪疲劳。系统亦设有自杀风险识别机制,若识别到高风险内容,将立即终止对话并提示线下求助。LLM 训练提示词和干预对话示例见附录。

阅读组阅读一段由 LLM 生成的网络欺凌科普材料(见附录),内容涵盖网络欺凌的定义、影响机制与应对建议,由本文的第一作者和通讯作者联合审核修订。

2.1.4 实验流程

实验流程见图 1。首先,参与者扫描由研究人员通过微信发送的二维码,填写第一阶段问卷,完成人口学信息与基线变量的测量。随后,将参与者随机分配至三个实验条件:自尊-自悯组与心理教育组通过微信小程序分别与不同设定的 LLM 进行文本对话;阅读组阅读与干预主题相关的文字材料。对话或阅读时间为 10~15 分钟;对话或阅读结束后,所有参与者再次扫码填写第二阶段问卷,用于评估干预效果及主观反馈。实验总时长约为 20~30 分钟。为保障伦理原则,实验过程中参与者如出现不适,可随时中止实验,研究团队将提供必要的心理疏导支持。实验结束后,研究人员会向参与者说明研究目的并致谢,同时披露研究所采用的大语言模型或文本材料来源。

2.2 结果

2.2.1 基线及描述性统计

三组参与者的性别、年龄、学历以及在基线时所测量的所有变量的描述性统计见表 1。对性别和学历进行 χ^2 检验,对基线的其他变量进行单因素方差分析,结果表明除状态韧性外其余变量组间均无显著差异。

2.2.2 操纵检验

使用 ChatGPT 4o 对参与者和 LLM 对话文本中参与者发送文本的自尊和自悯水平进行评分(提示词见附录),并进行独立样本 t 检验。由于部分参与者在输入文本时使用了表情符号(Emoji),导致最终文本不可存储,在分析时未纳入其数据(自尊-自悯组共 2 名,心理教育组共 4 名)。结果表明,自尊-自悯组文本($M = 68.50$, $SD = 9.85$)的自尊水平显著高于心理教育组文本($M = 55.33$, $SD = 18.86$), $t(20.22) = 2.44$, $p = 0.024$, $d = 0.90$;自尊-自悯组文本($M = 79.28$, $SD = 9.27$)的自悯水平显著高于心理

表 1 研究 1 基线及描述性统计

变量	自尊-自悯组	心理教育组	阅读组	统计量	<i>p</i>
性别(<i>n</i> , %)				χ^2	
男	7 (35)	5 (26.32)	5 (25)	0.573	0.751
女	13 (65)	14 (73.68)	15 (75)		
学历 (<i>n</i> , %)				χ^2	
专科及以下	2 (10)	1 (5.26)	1 (5)	2.564	0.633
本科	18 (90)	16 (84.21)	17 (85)		
硕士	0 (0)	2 (10.53)	2 (10)		
基线 (<i>M</i> , <i>SD</i>)				<i>F</i>	
年龄	22.85 (2.62)	23.42 (3.63)	22.90 (3.39)	0.185	0.832
网络欺凌受害	30.50 (10.71)	29.32 (9.07)	32.15 (8.84)	0.431	0.652
状态韧性	3.26 (0.28)	3.01 (0.46)	3.38 (0.38)	4.723	0.013
特质韧性	3.09 (0.44)	3.04 (0.53)	3.35 (0.56)	2.110	0.131
焦虑	5.90 (5.56)	5.58 (5.32)	4.90 (5.05)	0.184	0.832
抑郁	5.35 (5.67)	5.21 (4.96)	3.50 (5.06)	0.766	0.469
压力	7.80 (5.36)	6.58 (5.19)	5.75 (5.92)	0.702	0.500
幸福感	40.52 (9.01)	35.64 (12.52)	42.97 (11.73)	2.112	0.131

教育组文本($M = 68.40$, $SD = 11.99$), $t(31) = 2.94$, $p = 0.006$, $d = 1.03$ 。以上表明, 自尊-自悯对话干预相较于心理教育对话干预, 显著提高了对话的自尊和自悯水平, 达到了干预目的。

2.2.3 主要结果

以组别(自尊-自悯组, 心理教育组, 阅读组)为被试间变量, 时间(干预前、后)为被试内变量, 对参与者的状态韧性和特质韧性进行重复测量方差分析, 在分析状态韧性时将基线的状态韧性数据作为协变量控制。结果表明, 状态韧性的时间主效应显著, $F(1, 55) = 6.41$, $p = 0.014$, $\eta_p^2 = 0.104$, $f = 0.34$, 组别主效应显著, $F(2, 55) = 3.23$, $p = 0.047$, $\eta_p^2 = 0.105$, $f = 0.34$, 时间与组别的交互作用显著(见图 2a), $F(2, 55) = 3.23$, $p = 0.047$, $\eta_p^2 = 0.105$, $f = 0.34$; 特质韧性的时间主效应显著, $F(1, 56) = 16.54$, $p < 0.001$, $\eta_p^2 = 0.228$, $f = 0.54$, 组别主效应和时间与组别的交互作用均不显著。

多重比较结果表明, 相较于干预前, 自尊-自悯组($p < 0.001$, $d = 1.32$)和心理教育组($p = 0.001$, $d = 1.06$)在干预后状态韧性得分显著提升, 阅读组无显著变化。对状态韧性交互作用的进一步简单效应分析发现, 在控制基线后, 干预后自尊-自悯组的状态韧性显著高于阅读组($p = 0.019$, $d = 0.76$), 表明自尊-自悯对话干预效果显著优于无交互的文本阅读干预。心理教育组与其他组之间差异不显著, 说明心理教育对话干预对状态韧性的提升效应相

对有限。此外, 虽然自尊-自悯组与心理教育组在干预后状态韧性上的差异未达统计显著水平, 但自尊-自悯组的提升幅度和效应量更高($d = 1.32$ vs. $d = 1.06$), 提示其干预效力可能更强。

2.2.4 次要结果

以组别(自尊-自悯组, 心理教育组, 阅读组)为被试间变量, 时间(干预前、后)为被试内变量, 对参与者的焦虑、抑郁、压力和幸福感进行重复测量方差分析。结果表明, 焦虑的时间主效应显著, $F(1, 56) = 17.33$, $p < 0.001$, $\eta_p^2 = 0.236$, $f = 0.56$, 组别主效应和时间与组别的交互作用均不显著; 抑郁的时间主效应显著, $F(1, 56) = 14.75$, $p < 0.001$, $\eta_p^2 = 0.208$, $f = 0.51$, 组别主效应和时间与组别的交互作用均不显著; 压力的时间主效应显著, $F(1, 56) = 23.01$, $p < 0.001$, $\eta_p^2 = 0.291$, $f = 0.64$, 组别主效应不显著, 时间与组别的交互作用显著(见图 2f), $F(2, 56) = 5.74$, $p = 0.005$, $\eta_p^2 = 0.170$, $f = 0.45$; 幸福感的时间主效应显著, $F(1, 56) = 31.50$, $p < 0.001$, $\eta_p^2 = 0.360$, $f = 0.75$, 组别主效应不显著, 时间与组别的交互作用显著(见图 2c), $F(2, 56) = 4.60$, $p = 0.014$, $\eta_p^2 = 0.141$, $f = 0.41$ 。

多重比较结果表明, 相较于干预前, 自尊-自悯组($p < 0.001$, $d = 1.76$)在干预后压力的得分显著降低, 而心理教育组和阅读组变化不显著; 自尊-自悯组($p < 0.001$, $d = 1.07$)和心理教育组($p < 0.001$, $d = 0.94$)在干预后幸福感的得分显著提高, 而阅读

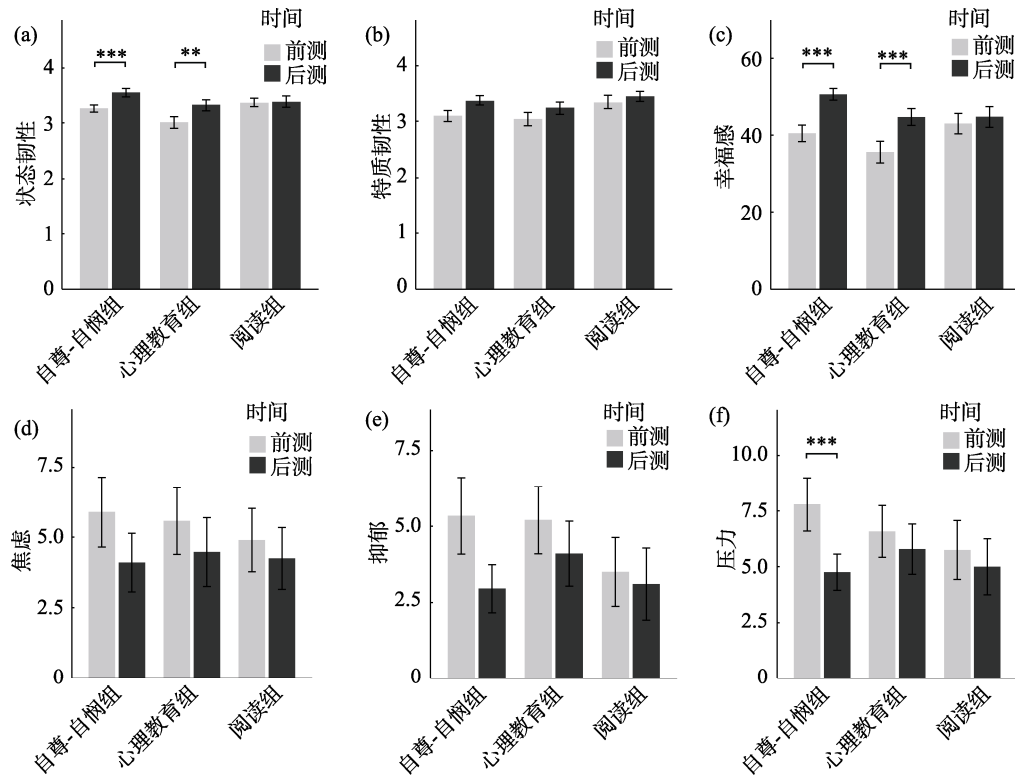


图2 干预前后状态韧性、特质韧性、焦虑、抑郁、压力和幸福感的指标变化

注: “***” $p < 0.001$, “**” $p < 0.01$, “*” $p < 0.05$ 。误差棒为标准误。

组变化不显著。对显著的交互作用进行简单效应分析结果表明,干预后三组之间的压力水平和幸福感均无显著的组间差异。此外,虽然自尊-自悯组与心理教育组在干预后幸福感上的差异未达统计显著,但自尊-自悯组的提升幅度和效应量更高($d = 1.07$ vs. $d = 0.94$),提示其干预效力可能更强。

2.3 讨论

研究1结果表明,两种基于LLM的对话式干预均在一定程度上提升了网络欺凌受害者的心理韧性和幸福感,减轻了压力水平。其中,自尊-自悯对话LLM组在多个指标的效应量上表现出更明显的变化趋势,但与心理教育对话组之间的差异未达统计显著,需谨慎解释其干预优势。上述发现初步支持了以自尊与自悯为干预靶点的可行性,也提示具备情感交互能力的AI干预工具在心理韧性提升中的潜力。然而,现有结果尚难揭示两类干预在具体作用路径上的差异。因此,研究2将进一步采用网络分析方法,从自尊、自悯等核心变量入手,比较不同LLM干预路径的结构特征。

3 研究2: 作用路径分析

尽管研究1发现两种LLM干预在提升心理韧性方面均有效,但其作用机制仍不明确,尤其是自

尊-自悯对话干预与心理教育对话干预在心理过程上的结构差异尚待厘清。研究2旨在聚焦自尊和自悯,进一步探讨这两种干预方式在核心心理机制上的不同路径。

由于本研究的作用路径涉及的变量及其维度较多,且研究定位于探索心理干预的起效路径,使用传统的中介或调节方法存在逻辑推理上的局限性(温忠麟等, 2024; Fishbein et al., 2023)。网络分析(Network Analysis)将心理和行为描述为相互作用的节点之间的因果系统(Borsboom & Cramer, 2013; Epskamp et al., 2018),并通过可视化节点之间的关系来帮助识别核心节点并解释它们的相互作用(Wang et al., 2022)。网络干预分析(Network Intervention Analysis, NIA)建立在网络分析的基础上,将干预类型作为节点纳入网络,通过构建混合图形模型(Mixed Graphical Model, MGM)估算网络,能够反映干预措施对特定变量的唯一直接效应,同时考虑到干预措施对网络的全局影响以及其他测量变量之间的关系(Haslbeck et al., 2021)。研究使用网络干预分析,旨在识别自尊-自悯对话LLM与心理教育对话LLM两组干预在心理机制网络中的差异特征,以揭示干预背后的潜在作用路径。

3.1 方法

3.1.1 参与者

参考前人 NIA 研究样本量(Blanken et al., 2019; Fishbein et al., 2023), 计划招募参与者 100 人。研究通过抖音和小红书平台另分两轮次招募主观报告其曾经历网络欺凌的参与者, 第一轮收到 73 份问卷, 排除未通过注意力检查者 4 人, 纳入参与者 69 人; 第二轮收到 38 份问卷, 排除未通过注意力检查者 2 人, 纳入参与者 36 人。研究最终纳入参与者 105 人(男性 52 人, 占 49.52%; 女性 53 人, 占 50.48%), 平均年龄 22.12 ± 2.77 岁。参与者被随机分至两组: 自尊-自悯组(54 人)和心理教育组(51 人)。所有参与者均自愿参加实验并知情同意, 完成实验后参与者将获得报酬 15 元。CONSORT 流程图见图 3。

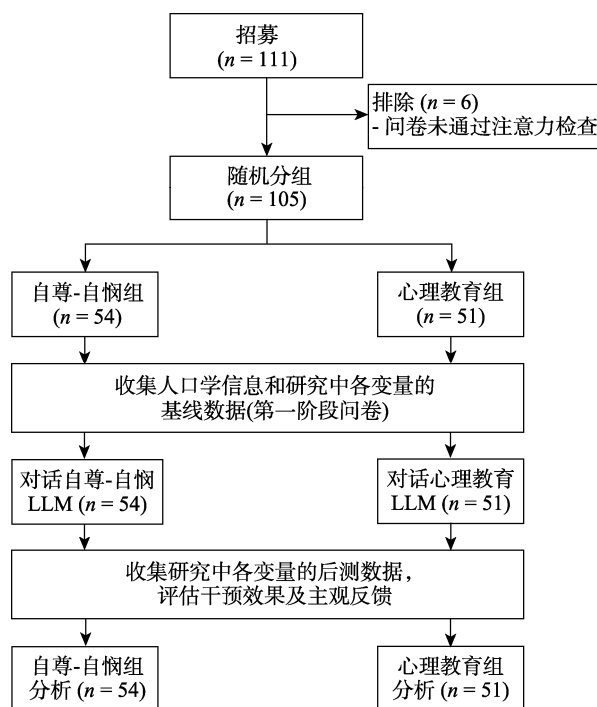


图3 研究2 CONSORT 流程图

3.1.2 测量工具

自尊使用 Heatherton 和 Polivy (1991)编制的状态自尊量表, 共 20 条目, 1~7 级评分, 分为表现、社交和外表三维度, 分数越高表明自尊水平越高。本研究采用钱旭宏(2024)修订的中文版量表, 信度良好(前测 Cronbach's $\alpha = 0.93$; 后测 Cronbach's $\alpha = 0.93$)。

自悯使用 Neff (2003)编制的自我怜悯量表, 共 26 条目, 1~5 级评分, 分为自我友善、自我批判、普遍人性感、孤立、正念及过度认同六维度, 分数

越高表明自悯水平越高。本研究采用陈健等人(2011)修订的中文版量表, 信度良好(前测 Cronbach's $\alpha = 0.82$; 后测 Cronbach's $\alpha = 0.84$)。

其余测量工具同研究 1。

3.1.3 实验流程

除测量工具不同外, 其余同研究 1。

3.1.4 网络干预分析

使用 R 4.4.2 进行相关分析、网络干预分析和网络可视化。在网络干预分析中, 为了反映干预前后各变量的变化趋势, 所有数据均使用后测与前测的差值。

首先使用 R 4.4.2 计算各变量与其维度之间的 Spearman 相关矩阵和权重相关矩阵。然后在 R 4.4.2 中选择 MGM 构建网络(Haslbeck et al., 2021), 将干预类型作为节点(自尊-自悯组 = 1, 心理教育组 = 0)纳入网络。为精简网络结构、突出较强效应并方便解读, 在绘制网络结构图时, 使用 LASSO 去除权重矩阵中相关系数小于 0.08 的边。评估网络中节点重要性的典型方法是计算网络结构的中心性指数(Costantini et al., 2015; Newman, 2010; Opsahl et al., 2010), 常用的中心性指数包括节点强度(Node Strength)、中介度(Betweenness)、接近度(Closeness)和预期影响(Expected Influence) (Epskamp et al., 2018), 本研究使用相关稳定性(Correlation Stability, CS)系数较好的中心性指数作为主要参照指标。最后, 使用 bootnet 软件包对网络结构的准确性和稳定性进行了分析。采用非参数 bootstrap 法(bootstrap = 1000)计算边权值的 95%置信区间, 检验边权值的准确性。采用 CS 系数评价中心性指数的稳定性。CS 系数最低不小于 0.25, 一般在 0.50 以上更好(Epskamp et al., 2018)。同时, 通过基于数据子集的网络模型估计, 研究了中心性指数的稳定性。

3.2 结果

3.2.1 描述性统计

两组参与者的性别、年龄、学历以及在基线时所测量的所有变量的描述性统计见表 2。对性别和学历进行 χ^2 检验, 对基线的其他变量进行 t 检验, 结果表明所有变量组间均无显著差异。各变量间的 Spearman 相关矩阵和权重相关矩阵热力图见附录。

3.2.2 干预效果

研究 2 使用问卷测量的自尊和自悯水平作为操纵检验指标, 结果同研究 1。在状态韧性、特质韧性、焦虑、抑郁和压力上的干预效果同研究 1, 幸福感前后变化在研究 2 中不再显著。总体效应量与

表 2 研究 2 基线及描述性统计

变量	自尊-自悯组	心理教育组	统计量	<i>p</i>
性别 (<i>n</i> , %)			χ^2	
男	24 (44.44)	28 (54.90)	1.147	0.284
女	30 (55.56)	23 (45.10)		
学历 (<i>n</i> , %)			χ^2	
专科及以下	9 (16.67)	2 (3.92)	4.550	0.103
本科	43 (79.63)	47 (92.16)		
硕士	2 (3.70)	2 (3.92)		
基线 (<i>M</i> , <i>SD</i>)			<i>t</i>	
年龄	21.76 (1.97)	22.51 (3.40)	-1.394	0.166
网络欺凌受害	33.83 (10.13)	31.20 (8.58)	1.436	0.154
状态韧性	2.71 (0.91)	2.89 (0.74)	-1.129	0.262
特质韧性	2.76 (0.92)	2.81 (0.75)	-0.274	0.785
焦虑	7.00 (5.62)	5.84 (4.90)	1.121	0.265
抑郁	7.13 (5.99)	6.00 (5.48)	1.007	0.316
压力	8.96 (6.65)	7.82 (5.92)	0.925	0.357
幸福感	35.15 (2.82)	36.24 (4.95)	-1.375	0.173
状态自尊	4.08 (1.53)	4.62 (1.31)	-1.925	0.057
自悯	3.30 (0.59)	3.36 (0.42)	-0.540	0.590

研究 1 相似, 提示干预效果具有一定稳定性。干预效果的详细分析见附录。

3.2.3 网络干预分析结果

变量各维度(节点)之间估计的网络结构见图 4。

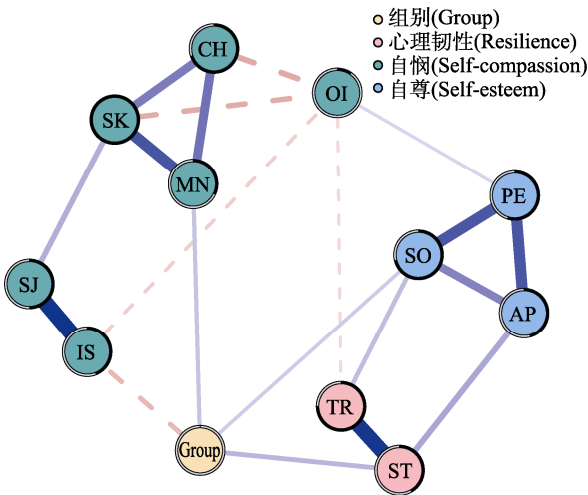


图 4 变量各维度(节点)之间估计的网络结构

注: 所有变量的计算均为后测与前测值之差, 干预类型编码为自尊-自悯组 = 1, 心理教育组 = 0。组别(Group), 表现(Performance, PE), 社交(Social, SO), 外表(Appearance, AP), 自我友善(Self-Kindness, SK), 自我评判(Self-Judgment, SJ), 普遍人性感(Common Humanity, CH), 孤立感(Isolation, IS), 正念(Mindfulness, MN), 过度沉迷(Over-Identified, OI), 状态(State, ST), 特质(Trait, TR)。在网络中, 蓝色边表示正相关, 红色边表示负相关, 去除权重矩阵中相关系数小于 0.08 的边。边的厚度代表相关的强度(边越厚, 相关值越大)。每个节点周围的可预测性值用一个环表示(环黑色占比越高, 变量对所有连接的节点的可预测性越高)。彩图见电子版。

网络中各节点的预期影响见图 5。

基于网络的中心性指数稳定性分析结果显示, 该网络预期影响的 CS 系数为 0.68, 高于其他中心性指数, 且其值大于 0.50, 表明该中心性指数稳定, 因此本研究主要使用预期影响来评估节点中心性。

在网络结构层面, 各变量的变化呈现出清晰的关联集群与跨构念传递特征。首先, 自尊的表现、社交和外表三节点之间以较厚的蓝色边连接, 提示三者的变化呈较强正相关, 即任一自尊维度的提升, 可同向带动另外两个维度的提升。其次, 自悯的自我友善与普遍人性感、正念呈正相关, 自我评判与孤立感呈正相关, 而正念、普遍人性感与过度沉迷呈负相关, 反映出自悯内部存在“正向促进-负向抑制”的关联逻辑。同时, 状态韧性和特质韧性呈较强正相关, 反映二者同步变化的趋势。最后, 干预同时与自悯的孤立感和正念、自尊的社交维度以及状态韧性相连接, 自悯的过度沉迷维度与自尊的表现维度和特质韧性相连接, 自尊的社交维度与特质韧性相连接, 自尊的外表维度和状态韧性相连接, 说明干预对心理症状网络的影响并非局限于单一构念, 而是存在跨维度的传递性。

在干预作用路径层面, 自尊-自悯对话 LLM 相较心理教育对话 LLM, 降低了参与者的孤立感, 提高了参与者的正念和状态自尊的社交维度, 并直接提高了参与者的状态韧性。孤立感的降低和正念的提高, 通过过度沉迷这一节点影响到状态自尊的表

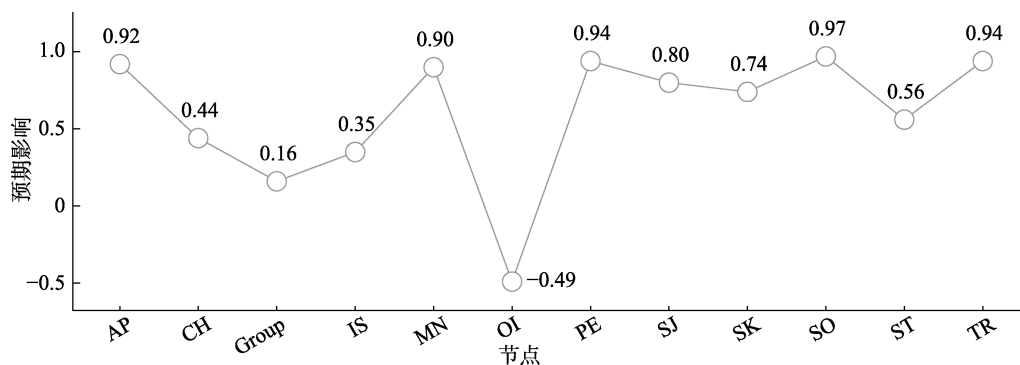


图5 网络中各节点的预期影响

现维度, 并和状态自尊的其他维度共同作用, 提高参与者的心理韧性。

网络中上述节点的预期影响绝对值较高, 表现出其是网络中关联影响的核心因素, 进一步表明干预通过上述节点间的关系影响心理韧性。同时, 组别的预期影响绝对值最低, 且其可预测性也最低, 表明其作为网络中干预驱动变量的独立性。

网络边权估计的 bootstrap 置信区间以及移除部分样本后采样得到的网络与原始样本间中心性指数的平均相关性图示见附录。bootstrap 法评估边权值的准确性结果显示, 边权值的置信区间较窄, 说明其具有足够的准确性。同时, 上文提到的重要节点连接的置信区间均不含 0, 说明节点连接的关系显著。随着样本量的减少, 中心性指数中的节点强度、中介度和接近度逐渐表现出不稳定性, 而预期影响则保持了较高的稳定性。这一趋势与 CS 系数的结果一致, 进一步验证了预期影响的稳健性。

3.3 讨论

研究 2 进一步揭示了自尊-自悯对话 LLM 相较于心理教育对话 LLM 在干预路径上的差异。结果表明, 自尊-自悯对话 LLM 相较于心理教育对话 LLM, 降低了参与者的孤独感, 提高了参与者的正念和状态自尊的社交维度, 并直接提高了参与者的状态韧性。孤独感的降低和正念的提高, 通过过度沉迷这一节点影响到状态自尊的表现维度, 并和状态自尊的其他维度共同作用, 提高参与者的心理韧性。以上表明, 自尊-自悯对话干预除了可以直接提高心理韧性外, 还可以靶向孤独感、正念和社交维度自尊, 通过自尊自悯多维度的相互作用, 促进干预效果的进一步提升。研究 2 的结果弥补了研究 1 中作用路径层面解释的不足, 进一步支持了以自尊自悯为靶点的个性化干预设计的理论价值。

4 讨论

本研究通过两项实验验证了由 LLM 驱动的对话式干预在提升网络欺凌受害者心理韧性方面的有效性及其作用路径。研究发现: (1)自尊-自悯对话 LLM 显著提升了网络欺凌受害者的心理韧性, 降低了其焦虑、抑郁和压力水平, 提升了幸福感; (2)自尊-自悯对话干预除了可以直接提高心理韧性外, 还可以靶向孤独感、正念和社交维度自尊, 通过自尊自悯多维度的相互作用, 促进干预效果的进一步提升。

4.1 对话式 LLM 干预的短期效果与作用路径

研究 1 (干预效果检验)表明, 即便是一次性的短期干预, 也能显著提升状态性的心理韧性, 尤其在自尊-自悯对话组中效果最为突出。这提示 LLM 模拟心理治疗师提供情感支持与认知引导, 有助于在面对网络欺凌时提供即时心理缓冲和认知重塑, 从而增强个体当下的适应能力。这一发现与 Fredrickson (2001) 提出的拓展-建构理论 (Broaden-and-Build Theory)相符, 强调积极情绪能够拓展个体的思维与行为反应模式, 并促进心理资源的建构。

研究 2 (作用路径分析)进一步通过网络干预分析发现, 自尊自悯在干预起效的作用路径中呈现多维度交互的特征。自尊各维度(表现、社交、外表)在干预中均与心理韧性呈现稳定的正相关, 提示其通过增强自我价值与社会认同感, 促进心理恢复。而自悯影响则更具特异性, 其通过“过度沉迷”维度与特质韧性和自尊的表现维度相连, 表明个体对负面情绪沉溺的减少可以提升其对自我表现的评价, 并促进其情绪复原与心理适应。

本研究通过一次性对话干预, 观察到心理韧性(尤其是状态韧性)干预前后的显著变化以及组间差

异,验证了状态韧性在短期数字化干预中的可塑性。然而,当前针对心理韧性的理论和测量工具仍将其作为特质变量看待(Bonanno & Diminich, 2013; Bonanno et al., 2015),缺乏适用于单次干预的、结果导向的直接测量工具(Chmitorz et al., 2018)。尽管本研究使用的测量工具包含心理韧性的状态和特质维度(Lock et al., 2020),且遵循Chmitorz 等人(2018)的建议,同时测量了与心理韧性有关的变量(焦虑、抑郁、压力和幸福感)以作为次要结果支撑结论,但仍无法保证研究结果能反映出心理韧性长期且真实的改变。此外,研究 1 和研究 2 关于心理韧性的干预效果存在一定差异,干预效果的持续性有待进一步验证。未来针对心理韧性的线上干预研究应更多关注干预效果在不同时间维度上的变化,验证状态与特质韧性的动态关系(Chmitorz et al., 2018)以及干预效果的持续性和真实性。

4.2 自尊与自悯的多维度交互

本研究以自尊自悯为核心干预靶点,揭示了其在提升网络欺凌受害者心理韧性中的多维度交互式作用路径,进一步补充了拓展-建构理论。该理论强调,积极情绪不仅是心理健康的结果,更是其成因,能够拓展个体的认知-行为反应模式,并在长期时间内建构心理与社会资源。

研究发现,自尊与自悯在提升心理韧性的作用路径层面呈现多维度交互的特征。自尊的社交和外表维度稳定地正向预测心理韧性,说明其主要通过增强个体的自我价值感与社会认同,提升其面对网络欺凌时的信心与应对能力。这一发现与“自尊保护机制”理论(Kaplan, 1980)和“自尊的心理保护作用”理论(Orth & Robins, 2013)一致,进一步验证了自尊作为心理干预靶点的理论价值。

干预除了直接作用于状态韧性外,还靶向自尊的社交维度以及自悯的孤独感和正念维度。自尊的社交维度进一步影响到特质韧性,这与上述“自尊保护机制”和“自尊的心理保护作用”理论相符(Orth & Robins, 2013)。自悯的作用路径则更为复杂和多元。干预通过 LLM 对话设计,带来真实的对话式治疗体验,让参与者认识到“自己的体验是他人共有的”,进而降低孤独感,带动自我评判的减少,进而为自我友善和普遍人性感的提升腾出空间(Neff, 2003),最终启动自悯的正向循环,并降低过度沉迷。同时,LLM 靶向自悯的设计可以让参与者减少对当下困境的评判,提高情绪调节能力,进一

步促进自我友善和普遍人性感的提升,强化自悯的正向循环,降低过度沉迷。上述路径通过过度沉迷这一节点与自尊的表现维度和特质韧性相关,在直接影响韧性的同时,通过参与自尊的作用路径进一步强化干预效果。

上述路径体现出,在干预中个体通过与 LLM 对话逐步识别并调整自身的情绪反应模式,减少对痛苦经历的情绪卷入,这表现出自尊和自悯在情绪加工与自我接纳方面的积极作用。“过度沉迷”是自悯结构中一个具有负向特征的维度,指个体在面对痛苦或失败时,过度陷入自我批评、情绪泛化和情绪卷入之中,难以从中抽离并进行理性反思(Neff, 2003)。高水平的“过度沉迷”通常与抑郁、焦虑等负性心理状态密切相关(Bluth et al., 2018),并显著削弱个体的心理适应力(Muris et al., 2016)。干预在靶向自悯时通过孤独感和正念两节点起效,最终减少个体的过度沉迷,这一变化反映出个体在面对网络欺凌情境时,逐渐学会以更理性、温和的视角看待自身遭遇,进而增强其心理复原能力。这一发现为拓展-建构理论在数字化心理干预情境下的应用提供了新的实证支持,同时也提示未来干预应关注自尊与自悯的协同激活,以实现更全面的心理恢复与成长。

4.3 虚拟治疗关系的相关探索性分析

在现有针对 LLM 心理干预的理论框架中,虚拟治疗关系是一个重要的影响变量。虚拟治疗关系是指用户在与 AI 互动过程中,主观感受到的被理解、被支持与被接纳的关系体验,其本质上是传统治疗关系在虚拟环境中的延伸与重构(Bickmore & Picard, 2005; Fitzpatrick et al., 2017)。已有研究表明,虚拟治疗关系的建立不仅能够增强用户对干预内容的信任与参与度,还能显著提升干预的情感支持质量与心理效果(Miner et al., 2016; Provoost et al., 2017),增加数字化心理干预程序的可接受性和有效性(Newman et al., 2011)。在心理服务可及性较低的情境中,虚拟治疗关系有望成为 AI 辅助心理干预的重要支撑。鉴于虚拟治疗关系在 LLM 心理干预中的理论价值,研究 2 (作用路径分析)中也对其进行了探索性分析。

结果表明,自尊-自悯组和心理教育组参与者感受到的虚拟治疗关系在总分和各维度均无显著差异(见附录)。因此,在本研究设计的靶向自尊自悯的 LLM 对话干预中,不同 LLM 的设置并未影响虚拟治疗关系的建立。作为心理治疗起效的共同因

素之一, (虚拟)治疗关系可以加强咨询师和来访者之间的情感联结, 明确治疗目标和节奏, 提高来访者对心理治疗的认同度和依从性(Cuijpers et al., 2019)。本研究针对 LLM 的差异化提示词设计, 即靶向自尊自悯的心理韧性干预及尽可能不带共情和治疗元素的心理教育, 并未对虚拟治疗关系造成显著影响, 这在一定程度上表明对话式心理干预本身即可通过关注、支持与陪伴, 同来访者建立治疗关系, 其存在一定的治疗价值。未来研究可在此基础上, 进一步探讨 LLM 干预中虚拟治疗关系作为单独治疗成分的干预效果及其具体起效机制, 明确作为共同因素的治疗关系在 LLM 对话式干预中的独特作用。

4.4 理论意义

本研究验证了状态韧性在短期数字化干预中的可塑性, 表明心理韧性不仅是稳定的人格特质, 也是一种可以被激活和增强的动态性心理资源。同时, 本研究通过网络干预分析揭示了自尊和自悯在干预路径中的多维度交互式作用路径: 自尊增强个体的自我价值感与社会认同, 自悯作用于自我接纳与情绪加工, 自尊-自悯对话干预除了可以直接提高心理韧性外, 还可以靶向孤立感、正念和社交维度自尊, 通过自尊自悯多维度的相互作用, 促进干预效果的进一步提升。这一路径丰富了对自尊与自悯功能机制的理解, 为理解其功能机制提供了新视角。研究也补充和扩展了拓展-建构理论, 首次在 AI 干预背景下实证检验积极情绪的触发与心理资源建构过程, 为未来干预设计提供了理论借鉴。

4.5 实践价值

随着 AI 技术的快速发展, 基于 LLM 的心理干预方式为网络欺凌受害者提供了新的支持路径。本研究验证了 LLM 在无需专业人员实时介入的前提下, 依然能够实现有效的心理支持, 展现出其在网络欺凌应对中的广泛应用前景。相较于传统心理干预依赖面对面服务、受限于时间和资源的局限性, LLM 具备高度可扩展性与低成本优势, 可嵌入社交平台、小程序或移动应用中, 提供即时响应与个性化干预(Demszky et al., 2023)。这种技术特性可以有效提升心理服务资源的可及性, 尤其适用于大规模心理健康需求尚未被满足的群体。传统的心理干预往往依赖于人际关系中的信任、共情与持续互动, 而 Bickmore 和 Picard (2005)等人关于关系型智能体(Relational Agents)的研究表明, 计算系统若具备

一定的社会情感表达能力(如使用社会性语言、非语言行为、共情反馈等), 也能够与用户建立并维持长期的社会情感关系, 从而提升用户的信任感、参与度与干预接受度。基于此, 本研究通过构建具备共情语言风格与互动策略的 LLM 对话系统, 模拟关系型智能体的行为模式, 通过实验验证了在无须专业人员实时介入的前提下, 对话 LLM 依然能够实现有效的心理支持与干预效果。

此外, 本研究构建的自尊-自悯对话干预路径, 打造了“情感交互+认知引导”的综合干预方式。通过激活个体的自我价值感与情绪调节能力, LLM 能够促进参与者在短时间内实现心理状态的积极转变。这种干预方式不仅具备技术可行性, 也为未来构建智能化、个性化的心理健康服务体系提供了实践参考。

4.6 研究局限与展望

在预实验和研究 1 (干预效果检验)中, 由于缩短研究时长的需要, 并未直接测量自尊和自悯两个干预靶点, 而是使用 LLM 对文本的自尊和自悯水平评分作为替代指标进行操纵检验。该推断较为间接, 可能受到模型对话风格等因素的影响。研究 2 (作用路径分析)中对自尊和自悯进行测量并直接分析, 操纵检验的显著结果表明靶向干预的有效性, 弥补了上述间接推断的不足。

当前 LLM 的干预内容仍然依赖研究者设定的提示词, 尚未实现真正意义上的自主识别与动态适配, 个性化水平仍有待提升(Carlbring et al., 2023)。未来研究可结合用户画像建模、伦理机制嵌入与风险识别系统, 推动 AI 心理干预走向更高效、安全与智能化的发展路径(Jain et al., 2024)。

不同温度参数的 LLM 对心理干预的效果可能产生不同影响(Lin, 2025)。由于不同温度的 LLM 对干预效果的影响并非是本研究关注的核心问题, 因此我们在干预中使用了 ChatGPT 4o 的默认温度参数, 并未检验过在不同温度设置下, 提示词或模拟对话的稳定性或一致性。未来可以进一步探讨该干预设计中不同温度对干预效果的影响。

最后, 由于有更多研究支持 ChatGPT 系列 LLM 应用于心理治疗的有效性(Carlbring et al., 2023; Raile, 2024), 因此本研究使用 ChatGPT 4o 进行 LLM 干预。然而, ChatGPT 系列 LLM 主要采用英文语料训练, 在中文语境下可能存在文化偏差(Tao et al., 2024)。未来可以进一步研究中文语境下 LLM 干预的文化偏差及其对结果造成的影响。

5 结论

本研究表明,自尊-自悯对话 LLM 和心理教育对话 LLM 均可有效提升网络欺凌受害者的心理韧性,但前者具有更高的干预效力与路径特异性。具体而言,自尊-自悯对话干预除了可以直接提高心理韧性外,还可以靶向孤立感、正念和社交维度自尊,通过自尊自悯多维度的相互作用,促进干预效果的进一步提升。这提示具备情感引导与认知干预功能的 LLM,尤其是以自尊与自悯为靶点的干预,可能在 AI 辅助心理支持中发挥更为独特和深入的作用。

参 考 文 献

- Batmaz, H., Türk, N., Kaya, A., & Yıldırım, M. (2023). Cyberbullying and cyber victimization: Examining mediating roles of empathy and resilience. *Current Psychology*, 42(35), 30959–30969. <https://doi.org/10.1007/s12144-022-04134-3>
- Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, 12(2), 293–327. <https://doi.org/10.1145/1073341.1073346>
- Blanken, T. F., Van Der Zweerde, T., Van Straten, A., Van Someren, E. J., Borsboom, D., & Lancee, J. (2019). Introducing network intervention analysis to investigate sequential, symptom-specific treatment effects: A demonstration in co-occurring insomnia and depression. *Psychotherapy and Psychosomatics*, 88(1), 52–54. <https://doi.org/10.1159/000495045>
- Bluth, K., Mullarkey, M., & Lathren, C. (2018). Self-compassion: A potential path to adolescent resilience and positive exploration. *Journal of Child and Family Studies*, 27(9), 3037–3047. <https://doi.org/10.1007/s10826-018-1125-1>
- Bonanno, G. A., & Diminich, E. D. (2013). Annual research review: Positive adjustment to adversity—trajectories of minimal-impact resilience and emergent resilience. *Journal of Child Psychology and Psychiatry*, 54(4), 378–401. <https://doi.org/10.1111/jcpp.12021>
- Bonanno, G. A., Romero, S. A., & Klein, S. I. (2015). The temporal elements of psychological resilience: An integrative framework for the study of individuals, families, and communities. *Psychological Inquiry*, 26(2), 139–169. <https://doi.org/10.1080/1047840X.2015.992677>
- Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, 9(1), 91–121. <https://doi.org/10.1146/annurev-clinpsy-050212-185608>
- Campbell, A. (1976). Subjective measures of well-being. *American Psychologist*, 31(2), 117–124. <https://doi.org/10.1037/0003-066X.31.2.117>
- Carlbring, P., Hadjistavropoulos, H., Kleiboer, A., & Andersson, G. (2023). A new era in Internet interventions: The advent of ChatGPT and AI-assisted therapist guidance. *Internet Interventions*, 32, 100621. <https://doi.org/10.1016/j.invent.2023.100621>
- Cuijpers, P., Reijnders, M., & Huibers, M. J. (2019). The role of common factors in psychotherapy outcomes. *Annual Review of Clinical Psychology*, 15(1), 207–231. <https://doi.org/10.1146/annurev-clinpsy-050718-095424>
- Chaudhary, P. K., Yalamati, S., Palakurti, N. R., Alam, N., Kolasani, S., & Whig, P. (2024, July). Detecting and preventing child cyberbullying using generative artificial intelligence. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* (pp. 1–5). IEEE. <https://doi.org/10.1109/APCIT59408.2024.10206797>
- Chen, J., Yan, L., & Zhou, L. (2011). Reliability and validity of Chinese version of self-compassion scale. *Chinese Journal of Clinical Psychology*, 19(6), 734–736. <https://doi.org/10.16128/j.cnki.1005-3611.2011.06.006>
- [陈健, 燕良弼, 周丽华. (2011). 中文版自悯量表的信效度研究. *中国临床心理学杂志*, 19(6), 734–736.]
- Chiu, Y. Y., Sharma, A., Lin, I. W., & Althoff, T. (2024). A computational framework for behavioral assessment of LLM therapists. arXiv: 2401.00820.
- Chmitorz, A., Kunzler, A., Helmreich, I., Tüscher, O., Kalisch, R., Kubiak, T., ... Lieb, K. (2018). Intervention studies to foster resilience—A systematic review and proposal for a resilience framework in future intervention studies. *Clinical Psychology Review*, 59, 78–100. <https://doi.org/10.1016/j.cpr.2017.11.002>
- Chu, X. W., Fan, C. Y., Liu, Q. Q., & Zhou, Z. K. (2018). Cyberbullying victimization and symptoms of depression and anxiety among Chinese adolescents: Examining hopelessness as a mediator and self-compassion as a moderator. *Computers in Human Behavior*, 86, 377–386. <https://doi.org/10.1016/j.chb.2018.04.039>
- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mottus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the art personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, 54, 13–29. <https://doi.org/10.1016/j.jrp.2014.07.003>
- Cross, D., Shaw, T., Hadwen, K., Cardoso, P., Slee, P., Roberts, C., ... Barnes, A. (2016). Longitudinal impact of the Cyber Friendly Schools program on adolescents' cyberbullying behavior. *Aggressive Behavior*, 42(2), 166–180. <https://doi.org/10.1002/ab.21609>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11), 688–701. <https://doi.org/10.1038/s44159-023-00241-5>
- Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1), 195–212. <https://doi.org/10.3758/s13428-017-0862-1>
- Fishbein, J. N., Haslbeck, J., & Arch, J. J. (2023). Network intervention analysis of anxiety-related outcomes and processes of acceptance and commitment therapy (ACT) for anxious cancer survivors. *Behaviour Research and Therapy*, 162, 104266. <https://doi.org/10.1016/j.brat.2023.104266>
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>
- Foody, M., Samara, M., & Carlbring, P. (2015). A review of cyberbullying and suggestions for online psychological therapy. *Internet Interventions*, 2(3), 235–242. <https://doi.org/10.1016/j.invent.2015.05.002>
- Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56(3), 218–226. <https://doi.org/10.1037/0003-066X.56.3.218>
- Gong, X., Xie, X., Xu, R., & Luo, Y. (2010). Psychometric properties of the Chinese versions of DASS-21 in Chinese

- college students. *Chinese Journal of Clinical Psychology*, 18(4), 443–446. <https://doi.org/10.16128/j.cnki.1005-3611.2010.04.020>
- [龚栩, 谢嘉瑶, 徐蕊, 罗跃嘉. (2010). 抑郁-焦虑-压力量表简体中文版(DASS-21)在中国大学生中的测试报告. *中国临床心理学杂志*, 18(4), 443–446.]
- Haslbeck, J. M. B., Borsboom, D., & Waldorp, L. J. (2021). Moderated network models. *Multivariate Behavioral Research*, 56(2), 256–287. <https://doi.org/10.1080/00273171.2019.1677207>
- He, C. (2025, January 10). Ministry of Public Security: In 2024, public security organs handled more than 8, 600 cases of cyber violence. *China Youth Daily*. https://s.cyol.com/articles/2025-01/10/content_Q42PLECW.html
- [何春中. (2025-1-10). 公安部: 2024 年公安机关共办理网络暴力案件 8600 余起. *中国青年报*.]
- Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology*, 60(6), 895–910. <https://doi.org/10.1037/0022-3514.60.6.895>
- Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 44(2), 227–239. <https://doi.org/10.1348/014466505X29657>
- Hinduja, S., & Patchin, J. W. (2017). Cultivating youth resilience to prevent bullying and cyberbullying victimization. *Child Abuse & Neglect*, 73, 51–62. <https://doi.org/10.1016/j.chiabu.2017.09.010>
- Jacobs, N. C., Völlink, T., Dehue, F., & Lechner, L. (2014). Online Pestkoppenstoppen: Systematic and theory-based development of a web-based tailored intervention for adolescent cyberbully victims to combat and prevent cyberbullying. *BMC Public Health*, 14(1), 1–16. <https://doi.org/10.1186/1471-2458-14-396>
- Jain, G., Pareek, S., & Carlbring, P. (2024). Revealing the source: How awareness alters perceptions of AI and human-generated mental health responses. *Internet Interventions*, 36, 100745. <https://doi.org/10.1016/j.invent.2024.100745>
- Kaplan, H. B. (1980). Deviant behavior in defense of self. *American Journal of Sociology*, 87(5). <https://doi.org/10.1086/227584>
- Kazdin, A. E. (2019). Annual research review: Expanding mental health services through novel models of intervention delivery. *Journal of Child Psychology and Psychiatry*, 60(4), 455–472. <https://doi.org/10.1111/jcpp.12937>
- Kishimoto, T., Hao, X., Chang, T., & Luo, Z. (2025). Single online self-compassion writing intervention reduces anxiety: With the feedback of ChatGPT. *Internet Interventions*, 39, 100810. <https://doi.org/10.1016/j.invent.2025.100810>
- Kowalski, R. M., Giumetti, G. W., & Feinn, R. S. (2023). Is cyberbullying an extension of traditional bullying or a unique phenomenon? A longitudinal investigation among college students. *International Journal of Bullying Prevention*, 5(3), 227–244. <https://doi.org/10.1007/s42380-022-00127-1>
- Kumpfer, K. L. (1999). Factors and processes contributing to resilience: The resilience framework. In M. D. Glantz & J. L. Johnson (Eds.), *Resilience and development: Positive life adaptations* (pp. 179–224). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/0-306-47167-1_9
- Lee, H., Bozek, D., Park, J., & Lee, J. M. (2024). The role of perceived social support and resilience in college cyberbullying victimization and mental health: A serial mediation model. *Journal of College Student Mental Health*, 39(4), 719–736. <https://doi.org/10.1080/28367138.2024.2407431>
- Li, Y., He, W., Zhang, X., Guo, Q., & Ma, B. (2014). Correlation between psychological capital and well-being index of parental rearing style, coping style of college students. *Chinese Journal of School Health*, 35(1), 128–131. <https://doi.org/10.16835/j.cnki.1000-9817.2014.01.051821>
- [李艳, 何畏, 张贤, 郭巧红, 马波. (2014). 大学生心理资本与父母教养方式应对方式幸福感指数的相关性. *中国学校卫生*, 35(1), 128–131.]
- Lin, Y., Sun, R., Zhang, J., Xu, K., & Yu, S. (2025). Discussion topics, sentiments, and usage intentions: A three-dimensional perspective of public attitudes toward large language models. *Journal of Psychological Science*, 48(6), 1359–1369. <https://doi.org/10.16719/j.cnki.1671-6981.20250606>
- [林原, 孙睿桐, 张君, 许侃, 于硕. (2025). 议题、情绪和使用倾向: 大语言模型公众态度的三维视角——基于网络评论的计算文本分析. *心理科学*, 48(6), 1359–1369.]
- Lin, Z. (2025). *Large language models as psychological simulators: A methodological guide*. arXiv:2506.16702.
- Lock, S., Rees, C. S., & Heritage, B. (2020). Development and validation of a brief measure of psychological resilience: The state trait assessment of resilience scale. *Australian Psychologist*, 55(1), 10–25. <https://doi.org/10.3758/s13428-017-0862-1>
- Marshall, S. L., Parker, P. D., Ciarrochi, J., Sahdra, B., Jackson, C. J., & Heaven, P. C. L. (2015). Self-compassion protects against the negative effects of low self-esteem: A longitudinal study in a large adolescent sample. *Personality and Individual Differences*, 74, 116–121. <https://doi.org/10.1016/j.paid.2014.09.013>
- Milosevic, T., Verma, K., Carter, M., Vigil, S., Laffan, D., Davis, B., & O'Higgins Norman, J. (2023). Effectiveness of artificial intelligence-based cyberbullying interventions from youth perspective. *Social Media and Society*, 9(1), <https://doi.org/10.1177/20563051221147325>
- Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, 176(5), 619–625. <https://doi.org/10.1001/jamainternmed.2016.0400>
- Muris, P., Meesters, C., Pierik, A., & de Kock, B. (2016). Good for the self: Self-compassion and other self-related constructs in relation to symptoms of anxiety and depression in non-clinical youths. *Journal of Child and Family Studies*, 25(2), 607–617. <https://doi.org/10.1007/s10826-015-0235-2>
- Navarro, R., Ruiz-Oliva, R., Larrañaga, E., & Yubero, S. (2015). The impact of cyberbullying and social bullying on optimism, global and school-related happiness and life satisfaction among 10-12-year-old schoolchildren. *Applied Research in Quality of Life*, 10(1), 15–36. <https://doi.org/10.1007/s11482-013-9292-0>
- Neff, K. D. (2003). Self-compassion: An alternative conceptualization of a healthy attitude toward oneself. *Self and Identity*, 2(2), 85–101. <https://doi.org/10.1080/15298860309032>
- Neff, K. D. (2011). Self-compassion, self-esteem, and well-being. *Social and Personality Psychology Compass*, 5(1), 1–12. <https://doi.org/10.1111/j.1751-9004.2010.00330.x>
- Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.
- Newman, M. G., Szkodny, L. E., Llera, S. J., & Przeworski, A. (2011). A review of technology-assisted self-help and minimal contact therapies for anxiety and depression: Is human contact necessary for therapeutic efficacy? *Clinical Psychology Review*, 31(1), 89–103. <https://doi.org/10.1016/j.cpr.2010.09.008>

- Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3), 245–251. <https://doi.org/10.1016/j.socnet.2010.03.006>
- Orth, U., & Robins, R. W. (2013). Understanding the link between low self-esteem and depression. *Current Directions in Psychological Science*, 22(6), 455–460. <https://doi.org/10.1177/0963721413492763>
- Patchin, J. W., & Hinduja, S. (2015). Measuring cyberbullying: Implications for research. *Aggression and Violent Behavior*, 23, 69–74. <https://doi.org/10.1016/j.avb.2015.05.013>
- Provoost, S., Lau, H. M., Ruwaard, J., & Riper, H. (2017). Embodied conversational agents in clinical psychology: A scoping review. *Journal of Medical Internet Research*, 19(5), e151. <https://doi.org/10.2196/jmir.6553>
- Qian, X. (2024). Reliability analysis and Rash model measurement of the state self-esteem scale in a Chinese sample. *Psychological Monthly*, 19(23), 87–89. <https://doi.org/10.19738/j.cnki.psy.2024.23.023>
- [钱旭宏. (2024). 状态自尊量表在中国样本中的信效度分析与 Rash 模型测评. *心理月刊*, 19(23), 87–89.]
- Raile, P. (2024). The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11(1), 47. <https://doi.org/10.1057/s41599-023-02567-0>
- Rémond, J. J., Kern, L., & Romo, L. (2014). A cyberbullying study: Analysis of cyberbullying, comorbidities and coping mechanisms. *L'encéphale*, 41(4), 287–294. <https://doi.org/10.1016/j.encep.2014.08.003>
- Santos, D., Mateos-Pérez, E., Cantero, M., & Gámez-Guadix, M. (2021). Cyberbullying in adolescents: Resilience as a protective factor of mental health outcomes. *Cyberpsychology, Behavior, and Social Networking*, 24(6), 414–420. <https://doi.org/10.1089/cyber.2020.0337>
- Shapka, J. D., & Maghsoudi, R. (2017). Examining the validity and reliability of the cyber-aggression and cyber-victimization scale. *Computers in Human Behavior*, 69, 10–17. <https://doi.org/10.1016/j.chb.2016.12.015>
- Sontag-Padilla, L., Dunbar, M. S., Ye, F., Kase, C., Fein, R., Abelson, S., ... Stein, B. D. (2018). Strengthening college students' mental health knowledge, awareness, and helping behaviors: The impact of active minds, a peer mental health organization. *Journal of the American Academy of Child & Adolescent Psychiatry*, 57(7), 500–507. <https://doi.org/10.1016/j.jaac.2018.03.019>
- Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., ... Helenius, H. (2010). Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of General Psychiatry*, 67(7), 720–728. <https://doi.org/10.1001/archgenpsychiatry.2010.79>
- Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), 346. <https://doi.org/10.1093/pnasnexus/pgae346>
- Veselska, Z., Geckova, A. M., Orosova, O., Gajdosova, B., van Dijk, J. P., & Reijneveld, S. A. (2009). Self-esteem and resilience: The connection with risky behavior among adolescents. *Addictive Behaviors*, 34(3), 287–291. <https://doi.org/10.1016/j.addbeh.2008.11.005>
- Wang, M. C., Deng, J., Shou, Y., & Sellbom, M. (2022). Cross-cultural examination of psychopathy network in Chinese and U.S. prisoners. *Journal of Psychopathology and Behavioral Assessment*, 44(3), 620–635. <https://doi.org/10.1007/s10862-022-09960-0>
- Wen, Z., Wang, Y., Ma, P., & Meng, J. (2024). The influence relationship among variables and types of multiple influence factors working together. *Acta Psychologica Sinica*, 56(10), 1462–1471. <https://doi.org/10.3724/SP.J.1041.2024.01462>
- [温忠麟, 王一帆, 马鹏, 孟进. (2024). 变量之间的影响关系和多重影响因素的共同作用类型. *心理学报*, 56(10), 1462–1471.]
- Wu, X., Wang, P., Yang, W., Ding, X. (2023, June 20). 65.3% of the respondents said that they or those around them had experienced cyberbullying. *China Youth Daily*. https://news.cyol.com/gb/articles/2023-06/20/content_zx5pJpiYlq.html
- [吴欣宇, 王品芝, 杨万淑, 丁香雪钰. (2023-6-20). 65.3%受访青年表示自己或周围人遭遇过网络暴力. *中国青年报*.]
- Xie, X., Zheng, H., Ju, K., Xiao, B., Liu, J., & Shapka, J. D. (2022). Examining the validity and reliability of the short form of cyberbullying and cybervictimization (CAV) scale. *Chinese Journal of Clinical Psychology*, 30(5), 1155–1159. <https://doi.org/10.16128/j.cnki.1005-3611.2022.05.029>
- [解晓龙, 郑宏, 鞠康, 肖博文, 刘俊升, Jennifer D. Shapka. (2022). 简式网络欺凌与网络受害量表中文版的信效度检验. *中国临床心理学杂志*, 30(5), 1155–1159.]
- Ybarra, M. L. (2004). Linkages between depressive symptomatology and Internet harassment among young regular Internet users. *CyberPsychology & Behavior*, 7(2), 247–257. <https://doi.org/10.1089/109493104323024500>

Large language model-facilitated interventions for enhancing psychological resilience in cyberbullying victims: Targeting self-esteem and self-compassion

KISHIMOTO Tomoko^{1,2}, HAO Ximing³, ASKAR Ankar³, XIA Yufei⁴, BAI Qiyu⁴

(¹ Faculty of Psychology, Beijing Normal University, Beijing 100875, China) (² Beijing Key Laboratory of Applied Experimental Psychology, Beijing Normal University, Beijing 100875, China) (³ Department of Social Psychology, School of Sociology, Nankai University, Tianjin 300350, China) (⁴ School of New Media, Peking University, Beijing 100871, China)

Abstract

Cyberbullying is a prevalent form of online aggression that threatens mental health, especially among adolescents and young adults. Compared with traditional bullying, it is more persistent, less visible, and harder to

control, often leading to anxiety, depression, stress, social withdrawal, and reduced well-being. Psychological resilience is an important protective resource that helps victims coping with adversity. Among the psychological factors related to resilience, self-esteem and self-compassion are especially relevant because they shape how individuals evaluate themselves and respond to suffering. Existing interventions for cyberbullying victims often lack personalization, emotional interaction, and scalability. Against this background, the present research examined whether large language model (LLM)-facilitated dialogue interventions could enhance resilience in cyberbullying victims, and whether an intervention targeting self-esteem and self-compassion would be more effective than general psychoeducation.

The research consisted of two independent studies. Study 1 used a three-arm randomized controlled design with 59 participants who reported previous cyberbullying experiences. They were randomly assigned to a self-esteem/self-compassion dialogue condition, a psychoeducational dialogue condition, or a psychoeducational reading control condition. Participants completed baseline measures, received a 10- to 15-minute intervention, and then completed posttest assessments. Primary outcomes were state resilience and trait resilience, and secondary outcomes included anxiety, depression, stress, and well-being. Study 2 recruited a new sample of 105 participants and compared the two dialogue conditions only. In addition to the outcomes used in Study 1, Study 2 assessed state self-esteem and self-compassion and applied network intervention analysis to examine the mechanisms underlying the two interventions. Both studies used a WeChat mini-program connected to ChatGPT-4o, and all procedures were completed with informed consent and ethical approval.

In Study 1, repeated-measures analyses showed that both dialogue-based LLM interventions improved resilience-related outcomes more clearly than the reading control. For state resilience, the time $\times$ group interaction was significant, indicating that change over time differed across conditions. After baseline state resilience was controlled, posttest state resilience in the self-esteem/self-compassion group was significantly higher than that in the reading group. Both dialogue groups showed significant within-group increases in state resilience, but the self-esteem/self-compassion group showed the largest improvement. For trait resilience, only the time effect was significant, suggesting a general pre-post increase without a clear differential group effect. Regarding secondary outcomes, anxiety and depression decreased over time across groups. Stress and well-being showed significant time $\times$ group interactions: stress significantly decreased only in the self-esteem/self-compassion group, whereas well-being significantly increased in both dialogue groups but not in the reading group. In Study 2, the overall intervention pattern was broadly consistent with Study 1, supporting the stability of the main findings. Network intervention analysis further showed that, compared with the psychoeducational dialogue, the self-esteem/self-compassion dialogue was uniquely linked to lower isolation, higher mindfulness, higher social self-esteem, and a direct increase in state resilience. These nodes formed a pathway through which changes in self-esteem and self-compassion jointly contributed to resilience enhancement.

Taken together, the findings showed that brief LLM-facilitated dialogue interventions can improve the psychological functioning of cyberbullying victims, and that a targeted intervention focusing on self-esteem and self-compassion has greater specificity in both outcomes and mechanisms. The results suggest that resilience, particularly state resilience, can be strengthened in short-term digital intervention contexts. They also indicate that self-esteem and self-compassion contribute to resilience through interrelated processes involving social self-evaluation, mindfulness, and reduced isolation. In practical terms, LLM-based interventions may provide scalable and responsive support for individuals affected by cyberbullying.

Keywords large language models, cyberbullying, psychological resilience, self-esteem, self-compassion