arXiv:2601.01530v1 [cs.CL] 4 Jan 2026

EmoHarbor：通过模拟用户的内在世界评估个性化情感支持

Jing Ye[1,2], **Lu Xiang**[1, 2*], **Yaping Zhang**[1,2], **ChengqingZong**[1,2]1多模态人工智能系统全国重点实验室、 中国科学院自动化研究所，北京，中国2中国科学院大学人工智能学院，北京，中国 yejing2022@ia.ac.cn；{lu.xiang, yaping.zhang, cqzong}@nlpr.ia.ac.cn

摘要

当前的情感支持对话评估范式往往奖励通用共情回应，却未能评估支持是否真正针对用户独特的心理用户画像和情境需求进行个性化。我们提出 **EmoHarbor**，这是一种自动化评估框架，采用**用户即评判者**范式，通过模拟用户的内心世界开展评估。EmoHarbor 采用智能体链架构，将用户的内在过程分解为三个专门角色，使代理能够与支持者互动，并以类似真实用户的方式完成评估。我们利用涵盖多种人格特质和情境的 100 个真实用户档案实例化该基准测试，并定义了个性化支持质量的 10 个评估维度。对 20 个先进大语言模型进行的全面评估揭示了一个关键洞见：尽管这些模型擅长生成共情回应，却始终无法根据不同用户的情境调整支持。这一发现重新界定了核心挑战，将研究重点从单纯提升通用共情转向开发真正具备用户感知能力的情感支持。EmoHarbor 提供了一个可复现且可扩展的框架，用于指导更加细致且具备用户感知能力的情感支持系统的开发与评估。



图 1：用于评估个性化情感支持的不同评估范式对比。评估者中心范式无法代表用户进行主观评估，而用户中心范式能够更准确地衡量个性化情感支持系统的质量。

1 引言

情感支持对话系统旨在识别用户的情感状态，并通过多轮交互提供量身定制的安慰与帮助（Peng 等人，2022年；Rains 等人，2020年；Liu 等人，2021）。尽管在生成流畅且富有共情的回应方面已取得长足进展，但这些系统的有效性很大程度上取决于个性化（Rogers，2013年；Zhang 等人，2018年；Campos 等人，2018年；Zollo 等人，2025年；Zheng 等人，2025年）。个性化是指系统根据个体独特的心理画像（Fleeson，2001年）和特定情境下的需求（Tamir，2016年），动态调整支持策略的能力。

尽管评估至关重要，现有的评估方法却存在根本性局限：它们遵循**以评估者为中心**的范式，从外部且自认为客观的立场评判情感支持对话的质量，无法捕捉用户的主观体验。例如，基于词元和嵌入的指标（如 BLEU（Papineni 等人，2002年）、BERTScore（Zhang* 等人，2020年）仅依赖参考回复，无法体现情感支持开放、依赖情境的特性。人工评估虽然更加灵活，

*通讯作者

但成本高得令人难以承受。即便是近期的基于大语言模型的评估器（Zhao 等人，2024年；Madani 和 Srihari，2025年；Zhang 等人，2024年）虽然提供了可扩展的替代方案，却采用千篇一律的“专家”视角——纯粹依据外部对话上下文来评估回复，因此忽略了那些影响不同用户对话体验的细微且由角色特征驱动的内部状态。

如图 1所示，设想两位用户在演讲失败后寻求支持：一位是完美主义者，总是纠结于细枝末节；另一位则坚韧，但因缺乏建设性反馈而感到沮丧。如果他们都收到诸如“别对自己太苛刻”这样的泛泛回应，两人可能会有截然不同的理解：前者可能觉得这番话淡化了自己的责任感，后者则可能将其视为鼓励自我接纳与成长的话语。以评估者为中心的 LLM-as-a-judge 方法可能会给出高分，因为回应表达了共情。然而，从用户的角度来看，完美主义者可能会给这条回应打低分。情感支持具有主观性，因此我们有必要转变范式，从评估优秀的支持者会说什么转向特定用户需要什么，以及他们会如何理解所获得的支持。

为此，我们提出**EmoHarbor**，这是一种新颖的评估框架，基于**用户即评判者**范式，通过基于智能体的模拟来建模用户的内在世界。EmoHarbor 模拟具有特定个性、情绪状态和对话历史的特定用户会如何感知支持并作出回应。具体而言，EmoHarbor 采用代理链架构，将用户的内在认知过程拆解为三个专门角色：**用户思考者**，根据用户画像模拟内心反思和主观感知；**用户对话者**，生成自然且符合个性的对话；以及**用户评估者**，结合用户不断变化的情绪状态和需求，对回应作出个性化评估。我们基于包含 100 份真实用户画像的全新精选基准实例化了这一框架。使用 EmoHarbor 对 20 个先进大语言模型进行全面评估后，我们发现了一个关键的脱节现象：模型虽然擅长通用共情，却始终无法根据个体情境调整支持方式。EmoHarbor 提供了一种可复现、可扩展的评估方法，可为发展更细致且具备用户感知能力的情感支持系统提供指导。

我们的主要贡献如下：

•我们提出了 EmoHarbor，这是一种评估框架，通过代理链架构实现“以用户为评判者”范式，以模拟细致入微的用户视角。

•我们通过实证分析验证了 EmoHarbor，结果表明，其与人工判断具有高度一致性，并且作为评估个性化情感支持对话系统的基准测试，具备很强的区分能力。

• 我们对 20 个大语言模型进行了全面评估，结果显示，尽管它们具备较强的一般共情能力，却往往无法提供个性化的情感支持。

2 方法

EmoHarbor 采用用户即评审范式，通过模拟用户的内部状态来生成可解释的主观评价。这一过程通过智能体链架构实现，其中多个专门化智能体协作，以模拟用户的认知、对话和评估过程。图 2 展示了整体工作流程。该框架围绕三个关键设计问题构建：**（i）如何进行基准测试**（用户档案构建），**（ii）如何模拟**用户行为（智能体专门化），以及**（iii）评估什么**（评估维度）。我们将在以下章节中详细阐述这些组成部分，并在附录 C中提供工作流算法。

2.1 如何进行基准测试

用户画像设计。 真实、详尽的用户档案是有效角色扮演的基石，因为它能让模拟用户展现出连贯的个性，而不是泛泛的行为。部分沿用了 Zhao 等人 (2025年)，我们将用户档案 $\mathcal{P}_\mathcal{U}$ 定义为：

$$\mathcal{P}_\mathcal{U} = \{\mathcal{D}, \mathcal{P}, \mathcal{C}, \mathcal{S}\} \tag{1}$$

其中：(1) $\mathcal{D}$ 表示**人口统计属性**（例如年龄、性别、职业），将用户置于具体情境中；(2) $\mathcal{P}$ 表示**偏好相关属性**（例如人格特质、大五人格、MBTI、习惯、爱好、语言风格），塑造鲜明的行为模式；(3) $\mathcal{C}$ 涵盖**咨询相关属性**（例如问题描述、情绪状态、目标、角色关系），体现心理背景；(4)$\mathcal{S}$ 则指定**情境脚本**，用于约束现实情境中的合理回应，

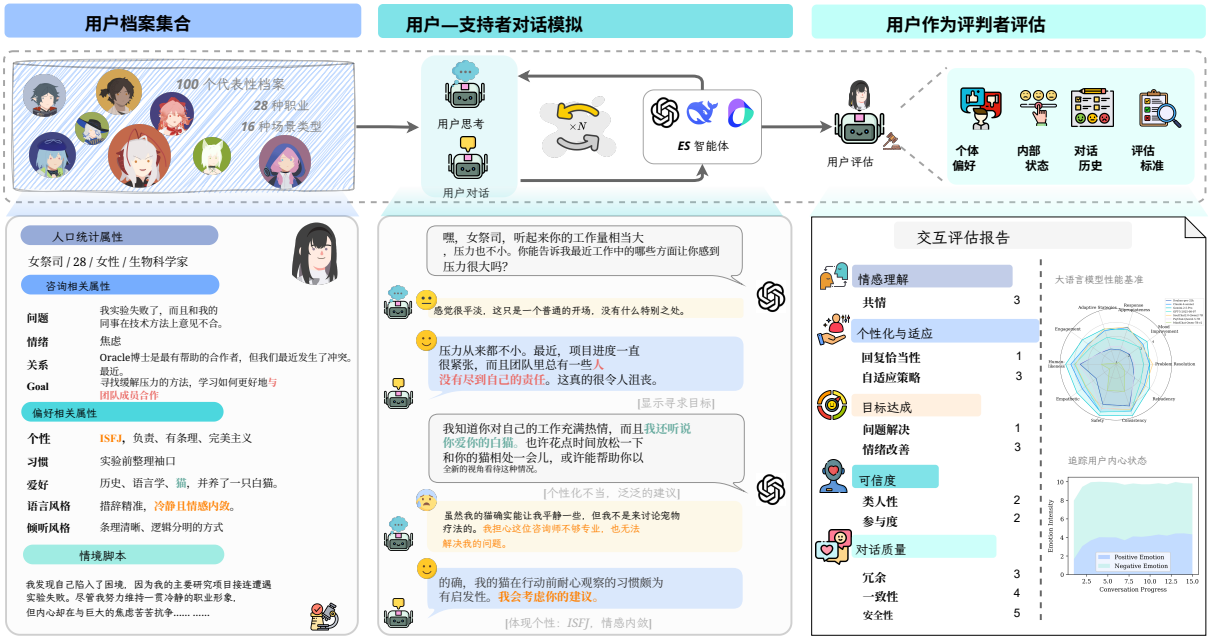


图 2：EmoHarbor 基准框架概览。该框架通过智能体链架构模拟用户的内在世界，采用用户即评审范式。

避免角色扮演趋于同质化，并确保模拟用户表现出多样且符合情境的行为。

用户档案集合。 在用户档案设计的基础上，我们通过两阶段流程构建用户档案集合。首先，我们通过问卷收集现实世界中的示例，以获得真实且多样化的种子档案。随后，我们利用大语言模型对这些种子档案进行完善、扩充和规模化，从而确保内容既真实，又能广泛覆盖潜在用户类型。更多详情请参阅附录 B。我们总共整理了 100 份代表性档案，涵盖广泛的人口统计和心理特征。

用户档案统计。我们的基准测试涵盖了多种多样的用户画像，全面覆盖了设计框架中定义的关键属性。如图3和4所示，用户年龄横跨青少年至老年阶段，涵盖 28 种职业、全部 16 种 MBTI 人格类型，以及 16 种咨询场景，包括职场压力、学业挑战、人际关系问题和人生转变。每份用户画像都标注了明确的问题陈述和支持目标，为评估对话系统在不同用户背景下的表现提供了丰富而结构化的基础。

2.2 如何进行模拟

为忠实模拟用户的主观体验，EmoHarbor 将模拟过程分解为三个专门化智能体：用户思考者、用户对话者和用户评估者。

对话设置。 给定一个用户档案 $\mathcal{P}_\mathcal{U}$ 和一个支持者系统 $\mathcal{S}$，模拟过程会维护两种不同的记忆：（1）**支持者记忆** H_s，其中仅包含 $\mathcal{S}$可访问的可观察对话轮次；（2）**用户记忆** H_u，其中还会记录潜在用户状态。

用户思考智能体。 用户思考者对用户的内在心理过程进行建模。在每个轮次 t，接收到支持者的回复 R_t后，它会通过生成以下内容来更新潜在用户状态：

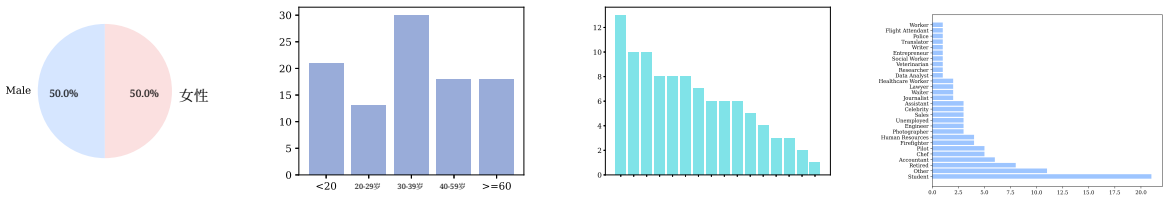
$$IS_t = f_{\text{Thinker}}(H_u, \mathcal{P}_\mathcal{U}, R_t), \tag{2}$$

其中，$IS_t = (c_t, e_t, g_t)$ 表示用户当前的认知评估（c_t）、情绪状态（e_t）和对话目标（g_t）。随后，这一内部状态会被追加到用户的综合记忆中：

$$H_u \leftarrow H_u \cup \{(R_t, IS_t)\}. \tag{3}$$

至关重要的是，这一过程明确模拟用户如何解读并回应支持者先前的回复，从而持续追踪其认知和情绪的演变。

用户对话代理。 用户对话者连接用户的内部状态与外部行为。它通过将更新后的内部状态和对话上下文外化，生成用户的下一条话语 U_t



（a）性别分布　（b）年龄分布　（c）MBTI分布　（d）职业分布

图3：基准用户画像的人口统计与个性覆盖情况，涵盖性别、年龄、人格类型和职业。这些分布共同体现了我们数据集的多样性与代表性。

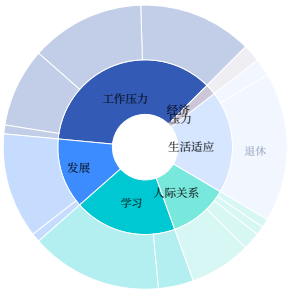


图4：咨询问题场景的分布　　.

更新后的内部状态和对话上下文：

$$U_t = f_{\text{Talker}}(H_u, \mathcal{P}_\mathcal{U}, R_t). \tag{4}$$

这句话会同时添加到两个记忆流中，从而完成一个可观察的对话轮次：

$$H_s \leftarrow H_s \cup \{(R_t, U_t)\}, \quad H_u \leftarrow H_u \cup \{U_t\}. \tag{5}$$

这确保了可观察行为是对底层内部过程的自然表达，并与个性保持一致。

用户评估智能体。 最后，用户评估者从模拟用户的视角对对话进行多维度评估。借助H_u中的完整内部状态历史，它追踪情绪和认知轨迹，从而细致评估所提供的支持是否真正实现了个性化：

$$E^{1:K} = f_{\text{Evaluator}}(H_u, \mathcal{P}_\mathcal{U}), \tag{6}$$

其中，$E^{1:K}$表示各项K评估标准的得分。

2.3 评估什么

现有对情感支持对话的评估大多聚焦于粗粒度的话语级指标，如流畅度、共情和信息量（Zhao et al.，2024年）。这些指标有助于衡量整体回复质量，但无法回答一个更根本的问题：某个回应是否适合特定用户在特定时刻的需求。因此，系统可能通过给出情感上支持用户却流于泛泛的回应而获得高分，却仍未能满足用户的实际需求——例如，用户寻求具体建议或问题解决时，系统却只提供安慰。

EmoHarbor 旨在从用户的主观视角评估情感支持。我们不将对话质量视为单个话语的静态属性，而是评估系统回复在整个互动过程中如何影响用户的内部状态。因此，我们从十个维度评估对话，并将这些维度归为五个方面，每个方面对应有效个性化支持的一个独特层面。这些方面共同体现系统是否能够：**(i)** 理解用户；**(ii)** 选择恰当的支持策略；**(iii)** 帮助用户取得进展；以及 **(iv)** 维持可信且安全的互动。

情感理解（共情）用于评估系统是否准确识别并回应用户的情绪状态。该方面体现了系统进行情感共鸣的能力，而情感共鸣是建立信任、实现有效个性化支持互动的基本前提。

个性化与适应（回复恰当性、自适应策略）用于评估系统选择支持策略并生成回复时，是否与用户当前的需求、偏好和情境相契合。该方面不单独评估共情，而是区分不同类型的支持（例如，情感认同与工具性指导），并考察回复是否适合具体情境

目标达成（问题解决、情绪改善）用于衡量互动是否帮助用户在认知清晰度或情绪健康方面取得有意义的进展。

可信度（类人性、参与度）用于考察互动是否传达出真实自然的感受，从而从人的角度维持用户的参与度。

对话质量与安全性（冗余、一致性、安全性）评估个性化是否在不损害连贯性、稳定性或伦理可靠性的前提下实现，从而确保互动环境安全、一致且值得信赖。

各评估维度的详细定义见附录G。

3 实验设置

3.1 人类研究

本项人类研究旨在收集基本用户画像，并对人机交互进行人工评估。参与者填写了人口统计问卷，并从中筛选出合适人选参与实验。详细流程见附录 A。

参与者筛选。 候选人填写了第2.1节所述的问卷。符合条件的参与者曾使用过大语言模型，有明确的个人问题可供讨论，且心理状态稳定。研究从 600 多份提交中选出了 50 名背景各异的参与者。他们接受了培训，审阅了对话样例，并学习了评估指南，以确保一致性。

人类交互评估。 每位参与者都以盲测方式、按随机顺序与五个模型（豆包-Pro、Qwen2.5-72B、GPT-4o、Claude-3.7-Sonnet 和 DeepSeek-R1）进行交互。每次对话结束后（至少 10 轮），参与者都会填写一份评估问卷。研究结束后的访谈筛除了参与态度不认真的参与者，以确保数据质量。最终共收集到 183 段有效对话，构成了**EmoHarbor 数据集**，用于评估自动化指标与人工判断之间的一致性。

3.2 大语言模型

为确保评估全面，本研究采用了多样化的大语言模型，涵盖开源模型、闭源模型和专用模型。所选模型分类如下：

开源模型。此类别包括通义千问系列模型（Qwen-2.5（Yang 等人，2024年）、Qwen-3（Yang 等人，2025年）、QwQ- 32B（Qwen 团队，2025年））、DeepSeek 系列（DeepSeek-R1（DeepSeek-AI，2025年）、DeepSeek- V3.1（DeepSeek，2025年））以及 GLM-4.5（智谱，2025年）[1]。

闭源模型。 我们还评估了多款可通过 API 服务使用的先进专有大语言模型，包括豆包系列（豆包-Seed-1.6(Seed， 2025年)，Doubao-Pro）、Claude 系列（Claude-3.7-Sonnet 和 Claude-4-Sonnet）（Anthropic， 2025年）、Gemini 系列（Gemini， 2025年）以及 GPT 系列（GPT-4o（OpenAI，2024年）、GPT-4（Achiam 等人，2023年）、GPT-5（Openai，2025年）和 o3-mini（OpenAI，2025年））。

专门的领域内模型。最后，我们纳入了经过微调、专门用于心理健康和情感支持应用的模型：SoulChat（Chen 等人，2023年）、PsyChat（Qiu 等人，2024年）和 MindChat[2]。

3.3 实现细节

实验环境。所有实验均在 6 块 NVIDIA L40 GPU 上进行。我们的实现基于 Python 3.12 和 PyTorch 2.7.0，并使用`vLLM`(Kwon 等人，2023年）进行推理加速。

模型配置。在模型特定配置方面，GPT-4o 同时用作用户思考者和用户对话代理，而通义千问3-235B 则用作用户评估智能体。我们根据各组件的角色精心设置温度参数：用户思考者采用较低的 0.1 温度，以确保推理聚焦且确定；用户对话者采用较高的 0.7 温度，以生成多样且自然的回应。用户评估者的温度设为 0.0，以确保评估一致且可复现。所有受评估的大语言模型在推理时均采用 0.7 的温度，以兼顾回应的多样性和连贯性。

模拟配置。 借鉴情感支持对话领域的既有研究（Liu 等人，2021），用户与支持者互动轮次的上限设为 15 轮。不过，用户代理可以通过生成表示结束对话的信号提前终止对话，例如“再见，拜拜，就这样吧，我不想继续了”。

[1]由于计算需求较高，我们通过 API 调用 DeepSeek 模型和 GLM-4.5，而非在本地部署。[2] https://github.com/X-D-Lab/MindChat

问题解决：问题解决　　情绪改善：情绪改善　　回应恰当性：回应适切性　　AS：自适应策略
EG：参与度　　HL：类人性　　EP：共情　　SF：安全性　　CS：一致性　　RD：冗余

评判模型	用户画像	内部状态	PR	MI	RA	AS	EG	HL	EP	SF	CS	RD
DeepSeek-R1			0.35	0.27	0.18	0.27	0.36	0.37	0.21	0.42	0.41	0.34
	✓		0.43	0.46	0.29	0.29	0.38	0.42	0.29	0.41	0.44	0.39
	✓	✓	**0.54**	**0.48**	**0.41**	**0.40**	**0.50**	**0.45**	**0.43**	**0.48**	**0.44**	**0.47**
Kimi-K2			0.38	0.42	0.10	0.18	0.34	0.37	0.27	**0.54**	**0.41**	0.29
	✓		0.41	0.53	0.20	0.26	0.32	0.40	0.26	0.49	0.40	0.35
	✓	✓	**0.56**	**0.61**	**0.33**	**0.45**	**0.50**	**0.43**	**0.44**	0.46	0.40	**0.43**
GPT-4			0.20	0.41	0.36	0.34	0.22	0.27	0.35	0.38	0.37	0.26
	✓		0.24	0.47	0.36	0.36	0.34	0.25	0.36	0.37	**0.43**	0.34
	✓	✓	**0.42**	**0.57**	**0.41**	**0.45**	**0.41**	**0.29**	**0.33**	**0.40**	0.42	**0.39**
Qwen3-235B			0.35	0.43	0.28	0.22	0.43	0.35	0.31	0.43	0.41	0.29
	✓		0.49	0.54	0.32	0.40	**0.54**	0.39	0.39	**0.46**	0.40	0.37
	✓	✓	**0.57**	**0.61**	**0.44**	**0.46**	**0.54**	**0.45**	**0.41**	**0.46**	**0.43**	**0.44**

表 1：模型判断与人工评估在 EmoHarbor 数据集上的皮尔逊相关。

评估配置。 每个维度均采用 5点李克特量表进行评分，分数越高表示支持效果越好。维度的详细说明和完整的评估流程见附录G。

4 实验结果

本节将介绍实验结果，以回答以下关键研究问题：

问题一： EmoHarbor 评估框架的可靠性如何?

Q2： 现有模型在 Emo- Harbor 基准测试中的表现如何?
Q3： 模型如何在多轮互动中适应用户的个性化需求?

4.1 EmoHarbor 评估框架的实证验证

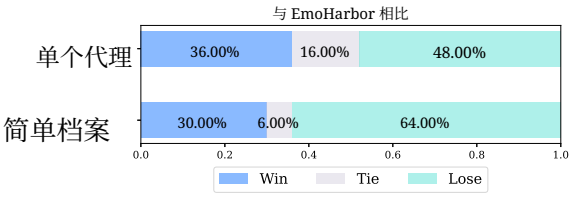


图 5：对用户模拟器进行的成对人工评估。“单代理”不包含用户思考者，而是由用户对话者直接生成回应。“简单档案”仅使用基本的人口统计和咨询属性，不包含个性、偏好和情景脚本。 ■ 表示“EmoHarbor 胜出”。

4.1.1 类人对话生成。

我们开展成对人工评估，以考察将用户模拟分解为多个代理是否能生成更接近人类的对话。比较包括三种设置：(1) **单个代理**，

在这种设置中，单个模型执行用户模拟，不对用户的内心世界进行明确建模；(2)**简单档案**，模拟器仅依据基本人口统计信息和咨询相关属性进行条件化，不包含详细的用户偏好建模或情境脚本；(3)我们完整的 Chain-of-Agent 模拟器，由在完整用户画像基础上运行的思考器和对话器模块组成。我们向人工评审者展示成对的对话，并请他们选择“ A 胜”“平”或“B 胜”，同时随机调整呈现顺序，以减轻位置偏差。评估针对随机抽取的 50 组对话进行。

图 5 的结果表明，我们的代理链模拟器始终优于两种基线。与单代理相比，这证明了对用户内部状态进行建模，能够生成更能体现其个性和角色的回应。与简单档案相比，我们的模拟器胜率达到 64%，这表明纳入更丰富的个性化特征，能够显著提升代理进行逼真角色扮演互动的能力。

4.1.2 与人工评估的一致性。

我们在 EmoHarbor 数据集上，采用以下策略评估 4 个候选大语言模型评判者——DeepSeek- R1、Kimi-K2、GPT-4 和 Qwen3-235B：(1) **标准评判**。评估者仅根据对话文本，对每段对话的情感支持质量进行评分。该策略代表了传统设置，即评估仅限于可观察到的对话，不考虑额外情境。(2)**用户感知评判**。评估者同时考虑对话文本

和相应的用户档案。纳入用户特定信息，使评估更加个性化，并能更好地结合情境。**(3) 用户内部状态感知评判（本文提出）**。 除了对话文本和用户档案，我们还根据此前的对话，使用用户思考智能体模拟用户在每轮对话中的内部状态。这些模拟状态近似反映推断出的用户想法和情绪，并用于辅助评估。原始对话内容保持不变；内部状态仅作为辅助情境，以提高评估的准确性。

表1展示了与人工评分的皮尔逊相关系数。所有模型的整体一致性中等，数值约为 0.4–0.5。尽管这些数值看起来可能不高，但这与个性化情感支持评估固有的主观性相符。在此情境下，这一范围内的相关性表明，大语言模型评判者能够较为合理地反映人类判断，可作为实用且可行的评估信号。重要的是，纳入用户画像和轮次级用户状态后，这种一致性会进一步提高，尤其是在 PR、MI、RA 和 AS 等高度主观且个性化的维度上。

4.1.3 基准测试区分能力。

MSR	MAC	方差分析	成对可区分性
0.745	0.427	F=112 (p<0.001)	0.87

表 2：基准测试的整体区分能力表现。模型分离比（MSR）衡量模型间表现差异相对于用户层级评分噪声的强度。模型一致性系数（MAC）量化用户在比较模型时判断的一致性。

我们评估了EmoHarbor 评估框架的区分能力，该框架考察用户即评判者范式能否有效区分情感支持对话模型之间的表现差异。每项指标的详细计算见附录 E。如表 2所总结，基准测试的模型分离比（MSR）为0.745，表明模型间差异显著大于用户层级评分噪声。模型一致性系数（MAC）为0.427，反映出用户在比较模型时具有中等至较高的一致性。这些结果还得到显著的单因素方差分析结果（$F = 112$，$p < 0.001$）以及较高的成对区分能力得分（0.87）的进一步 corroboration，表明用户能够可靠地区分模型表现。总体而言，这些发现表明，在用户即评判者范式下，EmoHarbor 具备较强的区分能力。

4.2 基准测试结果

表 3展示了评估结果，并凸显以下关键观察：

现有大语言模型在个性化情绪支持方面仍远未达到专家级表现。我们在 EmoHarbor 基准上评估了多种大语言模型，包括通义千问、DeepSeek、Claude、GPT、Gemini 和豆包系列。在闭源系统中，Gemini-2.5-Pro 的整体表现最佳，最高得分为 4.12。其他模型表现较差，大多数得分未能超过 4。在开源系统中，Qwen3-235B 表现最佳，平均得分为 4.13，与闭源模型相比也具有竞争力。这一出色表现可能部分归因于中文是其主要的训练和研究语言。此外，比较推理型模型（RLM）与非RLM时，我们发现，RLM在开源和闭源系列中的表现均始终更好。值得注意的是，大多数RLM得分均高于 3，表明它们在处理需要大量个性化推理的任务时具有明显优势。

领域专用大语言模型同样难以提供情感支持。先前的研究表明，许多对话式大语言模型都经过了针对共情响应生成的大量优化，并且在使用 BLEU 或 ROUGE 指标评估时，通常能在 ESConv 等基准测试中取得颇具前景的结果。然而，这些改进并不能很好地泛化到个性化情感支持。例如，SoulChat2.0（Chen 等人，2023年）的平均得分仅为 1.73。这种表现不佳可能是由于在共情对话数据集上过拟合所致，从而限制了模型根据不同用户特征调整响应的能力。

大语言模型展现出扎实的基础对话能力，但无法提供有效的情感支持。我们的分析表明，与问题解决、情绪改善、参与度和个性化等维度相比，几乎所有模型在类人性、一致性、共情和安全性等维度上的表现都更好。参与度

问题解决：问题解决　MI：情绪改善　RA：回应恰当性　AS：自适应策略
EG：参与度　HL：类人性　EP：共情　安全性：安全性　一致性：一致性　冗余：冗余

模型	推理	PR	MI	RA	AS	EG	HL	EP	SF	CS	RD	Avg.
开源												
Qwen2.5-7B-Instruct		1.80	1.59	1.39	1.46	1.62	2.57	2.18	3.52	3.64	1.98	2.18
Qwen2.5-32B-Instruct		2.07	1.93	1.71	1.85	1.91	2.83	2.47	3.79	3.84	2.19	2.46
Qwen2.5-72B-Instruct		2.12	2.04	1.88	2.03	2.19	3.18	2.63	3.91	4.02	2.34	2.63
Qwen3-32B	✓	2.23	2.28	2.33	2.24	2.48	3.45	2.80	4.05	3.95	2.34	2.82
Qwen3-235B	✓	**3.67**	**3.86**	3.76	**3.83**	**3.95**	**4.50**	**4.30**	**4.82**	**4.74**	**3.84**	**4.13**
QwQ-32B	✓	3.59	3.53	3.76	3.62	3.70	4.24	3.66	4.52	4.53	3.76	3.89
DeepSeek-V3.1	✓	3.11	3.09	3.12	3.14	3.28	4.06	3.52	4.58	4.52	3.35	3.58
DeepSeek-R1	✓	3.49	3.45	**3.94**	3.57	3.66	4.38	3.75	4.53	4.65	3.57	3.90
GLM-4.5	✓	2.88	2.85	2.76	2.90	3.04	3.77	3.32	4.44	4.40	3.25	3.36
闭源												
豆包-Seed-1.6	✓	3.63	3.69	3.76	**3.68**	3.84	4.46	4.10	4.70	4.71	3.64	4.02
豆包-Pro-32k		2.21	2.21	2.08	2.15	2.40	3.56	2.74	4.02	4.14	2.63	2.81
Claude-4-Sonnet	✓	3.41	3.41	3.76	3.54	3.46	4.23	3.97	4.59	4.53	3.54	3.84
Claude-3.7-Sonnet		3.16	3.14	3.30	3.22	3.32	4.06	3.69	4.56	4.59	3.41	3.65
Gemini-2.5-Pro	✓	3.42	**3.76**	3.61	3.67	**3.97**	**4.60**	**4.45**	**4.85**	**4.86**	**3.96**	**4.12**
GPT-4o-2024-11-20		2.98	3.09	2.79	3.03	3.16	3.96	3.70	4.60	4.51	3.19	3.50
GPT-5-2025-08-07	✓	**3.64**	3.31	**3.80**	3.66	3.40	3.90	3.66	4.33	4.41	3.64	3.77
o3-mini	✓	2.36	2.25	2.38	2.32	2.37	3.65	3.15	4.36	4.28	2.60	2.97
领域专用												
SoulChat2.0-Qwen2-7B		1.35	1.23	1.02	1.11	1.13	2.05	1.57	2.98	3.04	1.77	1.73
PsyChat-Qwen2.5-7B		2.16	2.10	2.18	2.08	2.19	**3.41**	2.92	4.06	**4.16**	2.29	2.75
MindChat-Qwen-7B-v2		**2.61**	**2.67**	**2.32**	**2.48**	**2.64**	3.31	**3.04**	**4.25**	4.12	**2.81**	**3.02**

表3：大语言模型在**EmoHarbor 基准**上的评估结果。所有分数均采用5点李克特量表。每个部分中，表现最佳的结果以**粗体**标出。对于每个模型，表现较强的维度以“蓝色”标出，表现较弱的维度以“绿色”标出。颜色越深表示表现越极端。

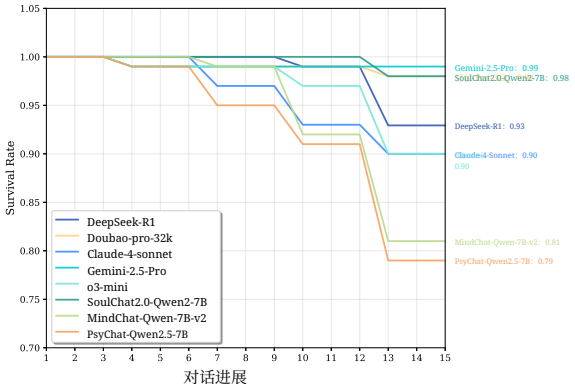


图6：随着对话进展，各模型的存活率。

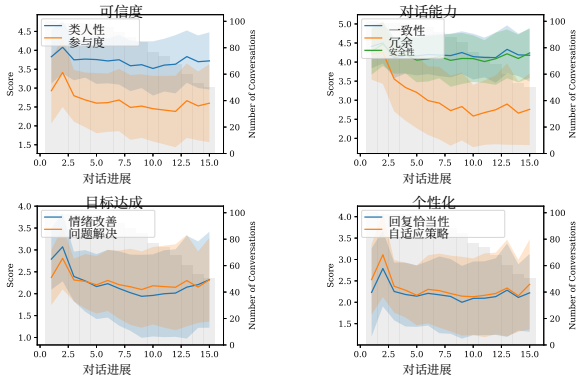


图 7：Doubao-Pro 的多轮对话评估。折线表示每个对话轮次的平均表现，灰色背景表示达到相应轮次的对话数量。

得分尤其低，这表明对话往往让人觉得无效，甚至可能
没有负面影响。这凸显了情感支持和以用户为中心的适应未来改进的重要方向。

4.3 多轮表现分析

如第 3.3 节所述，我们将对话长度的上限设为 15 轮，略低于真实情感支持对话（ESC）中观察到的平均 17–18 轮（Liu 等人，2021）。然而，我们观察到

在许多情况下，用户代理 会主动结束对话。提前终止的主要原因是情感支持系统未能提供有效的情感支持，导致用户失去参与意愿。我们认为，这类提前终止表明模型出现了故障。

图6展示了不同模型对话会话的生存曲线。从曲线可以看出，整体表现较好的模型往往也能让对话持续更长时间。图

7详细展示了随着对话进展，Doubao-Pro 在各个维度上的表现变化。在已完成的对话中，该模型的整体表现相对稳定，但随着对话推进，所有指标都略有下降趋势。这表明模型在长时间互动中维持质量方面可能存在不足。值得注意的是，随着对话轮次增加，冗余性得分显著下降，这意味着对话越长，模型越容易给出重复或程式化的回应，从而降低共情互动的效果。

5 相关工作

随着大语言模型的发展，个性化情感支持智能体（Cheng 等人，2023年；Ye 等人，2025年；Suh 等人，2025年；Chen 等人，2025年；Jiang 等人，2025c，a）引起了越来越多的研究兴趣。一个关键挑战仍然存在：如何有效评估情感支持的质量。

传统评估。 早期的情感支持对话评估（Liu 等人，2021；Zheng 等人，2023a，2024年；Zhang 等人，2024b；Ye 等人，2025年）依赖 BLEU（Papineni 等人，2002）、ROUGE（Lin，2004）和 BERTScore（Zhang* 等人，2020年）等自动指标，通过评估与参考答案之间的词元重叠或嵌入相似度来进行衡量。这些指标通常无法捕捉情感支持对话的多样性和细微差别。人工评估虽然是黄金标准，但速度慢、成本高且具有主观性，导致标注者间一致性低且可复现性较差（MadaniandSrihari，2025年）。

专业评判模型评估。 经过微调的评判模型，例如 CharacterEval（Tu 等人，2024年）和 CharacterBench（Zhou 等人，2025年），使用标注数据作为专业评估器。尽管这类方法比人工评估更具可扩展性，但仍存在一些关键局限：（1）**静态对话**——依赖预先收集的对话日志，无法捕捉实时互动性或不断演变的对话语境；（2）**情境偏差**——对话历史并非由受评估模型自行生成，且会受到模型上下文学习的影响，从而导致偏差，并且无法充分评估多轮对话能力（Ye 等人，2025年）。Zhao 等人 （2024年）；Madani 和 Srihari（2025年）通过用户—支持者模拟框架部分解决了这些问题。然而，他们的评估过度关注语言流畅度和共情表达，却忽视了用户的个性化需求。

大语言模型评审评估。 近期研究使用大语言模型作为可扩展的评判模型，为人工标注和静态基准测试提供了替代方案（Zheng 等人，2023b；Gu 等人，2025年；Yuan 等人，2024年；Kazi 等人，2024年）。SOTOPIA（Zhou 等人，2024年）通过角色扮演模拟评估情绪智力，而 ESC-Judge（Madani 和 Srihari，2025年）和 CharacterArena（Ye 等人，2025年）采用用户模拟器生成对话，以进行成对比较（Chiang 等人，2024年）。尽管取得了这些进展，但它们仍难以捕捉情感支持评估中以用户为中心、对情境敏感且以心理学为基础的本质。真正有效的评估框架应转向个性化、关注互动且以主观体验为基础的评估策略，以反映用户真实的情绪体验。

6 结论

本文提出了 EmoHarbor，这是一种简单而有效的评估框架，旨在应对个性化情感支持对话评估所面临的挑战。EmoHarbor 采用“用户即评判者”的范式，并通过代理链架构实现，突破了传统同质化专家评判的局限。对 20 个先进大语言模型开展的实验表明，尽管当前大语言模型擅长提供通用共情，但在提供针对用户特点的支持方面仍表现不佳。本研究提出了一条新颖且高效的途径，有助于开发更加细致、具备用户感知能力的情感支持系统。

局限性

本研究提出了一种基于用户即评判者范式的个性化情感支持对话新型评估框架。所提出的框架为推动更细致且具备用户感知能力的情感支持系统的发展提供了新的方向。然而，仍有若干局限性值得进一步考量。首先，尽管用户模拟涵盖了多种用户画像，但其建立在预定义结构之上，可能无法充分捕捉真实人类行为的复杂性与不可预测性。其次，人类一致性评估可能受到参与者对评估任务的理解程度及其对大语言模型的熟悉程度影响，从而引入难以消除的系统性偏差。

伦理考量

本研究采用了公开可用的模型，包括Deepseek（DeepSeek-AI，2025年）、通义千问（通义千问等人，2025年）、GLM（智谱，2025年）、豆包（字节跳动，2024年）、Claude（Anthropic，2024年）、Gemini（Gemini，2025年）和GPT（Achiam 等人，2023年），以及诸如vLLM （Kwon 等人，2023年）等工具包。

本评估所使用的基准测试数据集通过 GPT-4o 合成生成，并计划在论文接收后公开发布。本研究所使用的用户画像均经过人工核验和筛选；然而，鉴于大语言模型固有的不可预测性，我们无法保证用户代理和支持代理生成的内容完全无害。本工作的主要关注语言是中文。本研究仅用于研究目的。

我们在人类研究中遵循了严格的伦理准则。我们招募了 50 名来自不同背景的参与者。在开始评估前，我们向参与者清晰、全面地说明了学习目标、潜在风险以及评估流程。为确保公平的补偿并尊重参与者的时间，我们按每小时 50 人民币的标准支付报酬，该标准高于现行的当地劳动标准。所有参与者数据都将予以保密，未经明确同意不会披露。

大语言模型被用于协助进行编码、写作和润色手稿。需要特别说明的是，大语言模型未参与构思、研究方法或实验设计。所有研究概念、想法和分析均由作者独立提出并完成。

参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat及另外1位作者。2023年。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*。

Anthropic。2024年。Claude 3.5 Sonnet。

Anthropic。2025年。推出 Claude 4。

字节跳动。2024年。豆包。

Joana Campos、James Kennedy 和 Jill F. Lehman。2018年。人机交互中利用对话记忆所面临的挑战。载于第*17*届国际自主智能体与多智能体系统会议论文集，第1649–1657页。

Yirong Chen、Xiaofen Xing、Jingkai Lin、Huimin Zheng、Zhenyu Wang、Qi Liu 和 Xiangmin Xu。2023年。SoulChat：通过使用多轮共情对话进行微调，提升大语言模型的共情、倾听和安慰能力。载于计算语言学协会：*EMNLP 2023* 年会论文集，第1170–1183页，新加坡。计算语言学协会。

Zhuang Chen、Yaru Cao、Guanqun Bi、Jincenzi Wu、Jinfeng Zhou、Xiyao Xiao、Si Chen、Hongning Wang和Minlie Huang。2025年。Socialsim：面向情感支持对话的社会化模拟。*AAAI*人工智能会议论文集，39(2)：1274–1282。

Jiale Cheng、Sahand Sabour、Hao Sun、Zhuang Chen 和 Minlie Huang。2023年。PAL：基于角色设定增强的情感支持对话生成。载于计算语言学协会论文集：*ACL 2023*，第535–554页，加拿大多伦多。计算语言学协会。

Wei-Lin Chiang、Lianmin Zheng、Ying Sheng、Anastasios N. Angelopoulos、Tianle Li、Dacheng Li、Banghua Zhu、Hao Zhang、Michael I. Jordan、Joseph E. Gonzalez 和 Ion Stoica。2024年。聊天机器人竞技场：一个通过人类偏好评估大语言模型的开放平台。载于 第*41*届国际机器学习会议论文集，ICML'24。JMLR.org。

DeepSeek。2025年。Deepseek-v3.1。

DeepSeek-AI。2025年。Deepseek-r1：通过强化学习激励大语言模型的推理能力。预印本，arXiv:2501.12948。

William Fleeson. 2001年. 迈向一种整合结构与过程的个性观：将特质视为状态的密度分布。《人格与社会心理学杂志》，80(6):1011。

Gemini.2025年. Gemini2.5：凭借先进的推理、多模态、长上下文和新一代智能体能力，推动前沿发展. 预印本，arXiv:2507.06261。

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, WeiLi, YinghanShen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025年. 关于以大语言模型作为评判者的综述。预印本，arXiv:2411.15594。

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, LyleUngar, Camillo J. Taylor, and Dan Roth. 2025年. 了解我，回应我：大规模动态用户画像构建与个性化回复的大语言模型基准测试。 预印本，arXiv:2504.14225。

Bowen Jiang、Yuan Yuan、Maohao Shen、Zhuoqun Hao、Zhangchen Xu、Zichen Chen、Ziyi Liu、Anvesh Rao Vijjini、Jiashu He、Hanchao Yu、Radha Poovendran、Gregory Wornell、LyleUngar、Dan Roth、Sihao Chen,

以及 Camillo Jose Taylor。2025年。Personamem-v2：通过学习隐式用户角色和智能体记忆迈向个性化智能。预印本，arXiv:2512.06688。

Liwei Jiang、Yuanjun Chai、Margaret Li、Mickel Liu、Raymond Fok、Nouha Dziri、Yulia Tsvetkov、Maarten Sap 和 Yejin Choi。2025年。人工蜂群思维：语言模型（及更广泛领域）的开放式同质性。载于 神经信息处理系统第三十九届年度会议数据集与基准测试专题。

Taaha Kazi、Ruiliang Lyu、Sizhe Zhou、Dilek Hakkani-Tur 和 Gokhan Tur。2024年。作为用户智能体的大语言模型：用于评估面向任务的对话系统。预印本，arXiv:2411.09972。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 和 Ion Stoica。2023年。利用分页注意力实现大语言模型服务的高效内存管理。载于 美国计算机协会 *SIGOPS* 第二十九届操作系统原理研讨会论文集。

Chin-Yew Lin。2004年。ROUGE：一种用于自动评估摘要的工具包。载于文本摘要的拓展方向，第74–81页，西班牙巴塞罗那。计算语言学协会。

Siyang Liu、Chujie Zheng、Orianna Demasi、Sahand Sabour、Yu Li、Zhou Yu、Yong Jiang 和 Minlie Huang。2021。面向情感支持对话系统。载于计算语言学协会第*59*届年会暨自然语言处理国际联合会议第*11*届会议论文集（第*1*卷：长篇论文），第3469–3483页，线上举行。计算语言学协会。

Navid Madani 和 Rohini Srihari。2025年。ESC-Judge：用于比较情感支持对话代理的框架。载于*2025*年自然语言处理实证方法会议论文集，第16059–16076页，中国苏州。计算语言学协会。

OpenAI。2024年。你好，GPT-4o。

Openai。2025年。GPT-5问世。

OpenAI。2025年。OpenAI o3 与 o4-m 问世 ini.

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002年。BLEU：一种自动评估机器翻译的方法。载于计算语言学协会第*40*届年会论文集，第311–318页，美国宾夕法尼亚州费城。计算语言学协会。

Wei Peng、Yue Hu、Luxi Xing、Yuqiang Xie、Yajing Sun 和 Yunpeng Li。2022年。全局控制，局部理解：用于情感支持对话的全局到局部层次图网络。载于论文集

第三十一届国际人工智能联合会议论文集，*IJCAI 2022*，维也纳，奥地利，*2022*年*7*月*23*日至*29*日，第4324–4330页。ijcai.org。

Huachuan Qiu、Anqi Li、Lizhi Ma 和 Zhenzhong Lan。2024年。PsyChat：一个以客户为中心的心理健康支持对话系统。载于*2024*年第二十七届计算机支持的协同设计国际会议（*CSCWD*），第2979–2984页。

通义千问：An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei、Huan Lin、JianYang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jingren Zhou 和其他25位作者。2025年。Qwen2.5技术报告。预印本，arXiv:2412.15115。

通义千问团队。2025年。Qwq-32b：拥抱强化学习的力量。https://qwenlm.github.io/blog/qwq-32b/。访问日期：2026年1月3日。

Stephen A Rains、Corey A Pavlich、Bethany Lutovsky、Eric Tsetsi 和 Anjali Ashtaputre。2020年。寻求支持者的期望、支持信息质量，以及支持性交互过程与结果：重新审视安慰型计算机程序案例。《社会与人际关系期刊》，37(2)：647–666。

Carl R. Rogers。2013年。以客户为中心的疗法。《当代心理治疗》，第95–150页。

Seed。2025年。seed1.6所用技术简介。

Jina Suh、Lindy Le、Erfan Shayegani、Gonzalo Ramos、Judith Amores、Desmond C. Ong、Mary Czerwinski 和 Javier Hernandez。2025年。Sense-7：用于衡量用户对持续人机对话中共情感知的分类体系与数据集。预印本，arXiv:2509.16437。

Maya Tamir。2016年。人们为什么要调节自己的情绪？情绪调节的动机分类体系。《人格与社会心理学评论》，20(3)：199–222。

Quan Tu、Shilong Fan、Zihang Tian、Tianhao Shen、Shuo Shang、Xin Gao 和 Rui Yan。2024年。Charac- terEval：用于角色扮演对话代理评估的中文基准测试。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第11836–11850页，泰国曼谷。计算语言学协会。

An Yang、AnfengLi、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Day- i heng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang及其他41位作者。2025年。Qwen3 技术报告。预印本，arXiv:2505.09388。

An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei及其他1位作者。2024年。Qwen2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*。

Jing Ye、Lu Xiang、Yaping Zhang 和 Chengqing Zong。2025年。从通用共情到个性化情感支持：用于用户偏好对齐的自进化框架。载于计算语言学协会论文集：*EMNLP 2025*，第18826–18853页，中国苏州。计算语言学协会。

Jing Ye、Lu Xiang、Yaping Zhang 和 Chengqing Zong。2025年。SweetieChat：用于多样化场景中处理情感支持助手的策略增强型角色扮演框架。载于第*31*届国际计算语言学大会论文集，第4646–4669页，阿联酋阿布扎比。计算语言学协会。

Xinge Ye、Rui Wang、Yuchuan Wu、Victor Ma、Feiteng Fang、Fei Huang 和 Yongbin Li。2025年。CPO：通过比较策略优化解决角色扮演对话中的奖励歧义。载于计算语言学协会论文集：*EMNLP 2025*，第297–323页，中国苏州。计算语言学协会。

Peiwen Yuan、Shaoxiong Feng、Yiwei Li、Xinglin Wang、Boyuan Pan、Heda Wang、Yao Hu 和 Kan Li。2024年。BatchEval：迈向类人文本评估。载于第 *62* 届计算语言学协会年会论文集（第 *1* 卷：长篇论文），第 15940–15958 页，泰国曼谷。计算语言学协会。

Chenhao Zhang、Renhao Li、Minghuan Tan、Min Yang、Jingwei Zhu、Di Yang、Jiahao Zhao、Guancheng Ye、Chengming Li 和 Xiping Hu。2024年。CPsyCoun：一种面向中文心理咨询的基于报告的多轮对话重建与评估框架。载于计算语言学协会 *ACL2024* 论文集精选，第13947–13966 页，泰国曼谷。计算语言学协会。

Saizheng Zhang、Emily Dinan、Jack Urbanek、Arthur Szlam、Douwe Kiela 和 Jason Weston。2018年。个性化对话智能体：我有一只狗，你也养宠物吗？载于第 *56* 届计算语言学协会年会论文集（第 *1* 卷：长篇论文），第 2204–2213 页，澳大利亚墨尔本。计算语言学协会。

Tenggan Zhang、Xinjie Zhang、Jinming Zhao、Li Zhou 和 Qin Jin。2024b。Escot：迈向可解释情感支持对话系统。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），*ACL 2024*，曼谷，泰国，*2024*年*8*月*11*日至*16*日，第13395–13412页。计算语言学协会。

Tianyi Zhang*、VarshaKishore*、FelixWu*、Kilian Q. Weinberger 和 Yoav Artzi。2020年。BERTScore：使用 BERT 评估文本生成。载于国际学习表征会议。

Haiquan Zhao、Lingyu Li、Shisong Chen、Shuqi Kong、Jiaan Wang、Kexin Huang、Tianle Gu、Yixu Wang、Jian Wang、Liang Dandan、Zhixu Li、Yan Teng、Yanghua Xiao 和 Yingchun Wang。2024年。ESC-eval：评估大语言模型中的情感支持对话。载于*2024*年自然语言处理实证方法会议论文集，第15785–15810页，美国佛罗里达州迈阿密。计算语言学协会。

Zheng Zhao、Clara Vania、Subhradeep Kayal、Naila Khan、Shay B Cohen 和 Emine Yilmaz。2025年。PersonaLens：用于评估对话式人工智能助手个性化的基准测试。载于计算语言学协会会议论文集：*ACL 2025*，第18023–18055页，奥地利维也纳。计算语言学协会。

Chujie Zheng、Sahand Sabour、Jiaxin Wen、Zheng Zhang 和 Minlie Huang。2023a。AugESC：利用大语言模型增强情感支持对话。载于 计算语言学协会：*ACL 2023* 论文集，第1552–1568 页，加拿大多伦多。计算语言学协会。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023b。使用 MT-Bench 和 Chatbot Arena 评判 LLM-as-a-judge。载于第*37*届神经信息处理系统国际会议论文集，NIPS '23，美国纽约州雷德胡克。Curran Associates Inc.

Xi Zheng、Zhuoyang Li、Xinning Gui 和 Yuhan Luo。2025年。 定制情感支持：个体如何构建并与大语言模型驱动的聊天机器人互动。载于 *2025*年计算机系统中的人为因素国际会议论文集，CHI '25，第 1–20 页。美国计算机协会。

Zhonghua Zheng、Lizi Liao、Yang Deng、Libo Qin 和 Liqiang Nie。2024年。大语言模型的自我对话让小型情感支持聊天机器人变得更好。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），*ACL 2024*，曼谷，泰国，*2024*年*8*月*11*日至*16*日，第11325–11345页。计算语言学协会。

智谱。2025年。GLM-4.5：推理、编程与代理能力。

Jinfeng Zhou、Yongkang Huang、Bosi Wen、Guan- qun Bi、Yuxuan Chen、Pei Ke、Zhuang Chen、Xiyao Xiao、Libiao Peng、Kuntian Tang、Rongsheng Zhang、Le Zhang、Tangjie Lv、Zhipeng Hu、Hongn- ing Wang 和 Minlie Huang。2025年。Character- bench：大语言模型角色定制的基准测试

大语言模型。载于第三十九届*AAAI*人工智能会议、第三十七届人工智能创新应用会议和第十五届人工智能教育进展研讨会论文集，AAAI'25/IAAI'25/EAAI'25。AAAI出版社。

Xuhui Zhou、Hao Zhu、Leena Mathur、RuohongZhang、Haofei Yu、Zhengyang Qi、Louis-Philippe Morency、Yonatan Bisk、Daniel Fried、Graham Neubig 和 Maarten Sap。2024年。SOTOPIA：用于评估语言智能体社会智能的交互式评测。载于第十二届国际学习表征会议。

Thomas P Zollo、Andrew Wei Tung Siah、Naimeng Ye、Ang Li 和 Hongseok Namkoong。2025年。个人化大语言模型：根据个人偏好定制大语言模型。收录于第十三届国际学习表征会议。

附录

A EmoHarbor 数据集15B 用户档案构建15C EmoHarbor评估流程15D 大语言模型16E 判别

区分能力指标1 **6**

E.1 模型分离比（MSR）。.17E.2 模型一致性系数（MAC）。.17E.3 单因素方差分析F统计量.17E.4 成对区分比例.17

F 补充　充实验结果 **17**

F.1 英语基准测试结果.17F.2 补充多轮分析结果.18F.3 成本分析.18

G 评估　维度 **18**

H 提示词 **22**

H.1 支持者提示词 .22H.2 用户思考智能体提示词 .23H.3 用户谈话智能体提示词 .24H.4 用户评估者提示词 .25

I 人类评估界面 **27**

一个 EmoHarbor 数据集

我们通过精心设计的人类研究构建了 EmoHarbor 数据集，旨在采集人机对话。每条记录都包含参与者与人工智能模型之间的一段真实对话，以及参与者的用户画像和对模型回复的主观评价。这些评价反映了用户的个人观点和情绪背景，为研究个性化人机交互提供了丰富的基础。用于数据采集的人机交互界面见附录 I。

参与者筛选。 预备参与者需完成第 2.1 节所述的初始问卷。参与者需具备使用大语言模型的经验，有明确的个人议题可在实验中讨论，且心理状况稳定，方符合参与条件。我们共收到 600 多份问卷提交。根据回答的完整性和质量，我们从不同背景的参与者中筛选出 50 人，参与人机交互评估。入选者参加了培训课程，以熟悉实验设置。他们审阅了示例对话和详细的评估指南，以确保评分一致且有意义。

对话数据集收集。 每位参与者都以盲测且随机排序的方式，与来自不同系列的五个模型（Doubao-Pro、Qwen2.5-72B、GPT-4o、Claude-3.7-Sonnet 和 DeepSeek-R1）进行了互动。每次互动至少包含十个对话轮次，让参与者能够深入探讨与个人相关的话题。每次交流结束后，参与者填写结构化评估问卷，表达对模型及其与自身情绪需求契合度的主观判断。为确保数据质量，我们进行了实验后访谈，以识别并排除未认真完成任务的参与者。最终，我们获得了 183 个有效的人机对话实例，共同构成 EmoHarbor 数据集。该数据集有助于系统分析自动评估指标与真实、以用户为中心的人类判断之间的一致性。

B 用户档案构建

中文用户画像构建。 中文用户画像由 EmoHar- 数据采集期间收集的种子信息构建而成

bor 数据集。具体而言，参与者通过初步问卷提供了基本背景描述和简短的咨询相关问题陈述。根据第 2.1 节对用户画像的定义，我们系统地实例化画像的各个组成部分，具体如下。(1)人口统计属性 经改写以删除可识别信息，同时保留必要的情境依据。(2) 偏好相关属性 经过扩充以增强个体差异性，包括人格特质、习惯和语言风格。(3)咨询相关属性 则根据原始回答，通过详细说明事件背景、情绪状态和用户目标加以具体化。(4) 最后，构建 情境脚本 ，用于说明面对不同类型的咨询师反馈时可能出现的情绪和行为反应。这一结构化构建流程确保模拟用户既具有多样性，内部又保持一致，从而缓解角色扮演中的行为同质化问题。

英语用户档案构建。与中文用户档案不同，英语用户档案基于Jiang 等人（2025年）提出的现有用户档案集初始化，该档案集提供了丰富的人口统计和偏好相关属性。在 ESConv 的咨询问题类别基础上，我们进一步为每个用户档案构建咨询相关属性，明确相应的情绪背景、问题背景和用户求助目标。随后，参照中文用户档案所采用的相同流程编写情境脚本，为针对不同类型咨询师反馈的合理情绪和行为反应作出定义。所有构建的用户档案首先使用大语言模型自动验证，随后通过随机抽样进行人工检查，以确保质量。最终，我们保留了 100 个高质量的英语用户档案用于实验。图8 总结了年龄、国籍、情绪状态和问题主题的分布。

C EmoHarbor评估流程

我们在算法 1中介绍了使用 EmoHarbor 评估情绪支持对话系统的工作流程。该评估通过模拟用户与被测系统之间的多轮互动，同时维护用户的内部状态和对话历史。在对话结束时，用户模型会生成涵盖多个维度的结构化评估分数

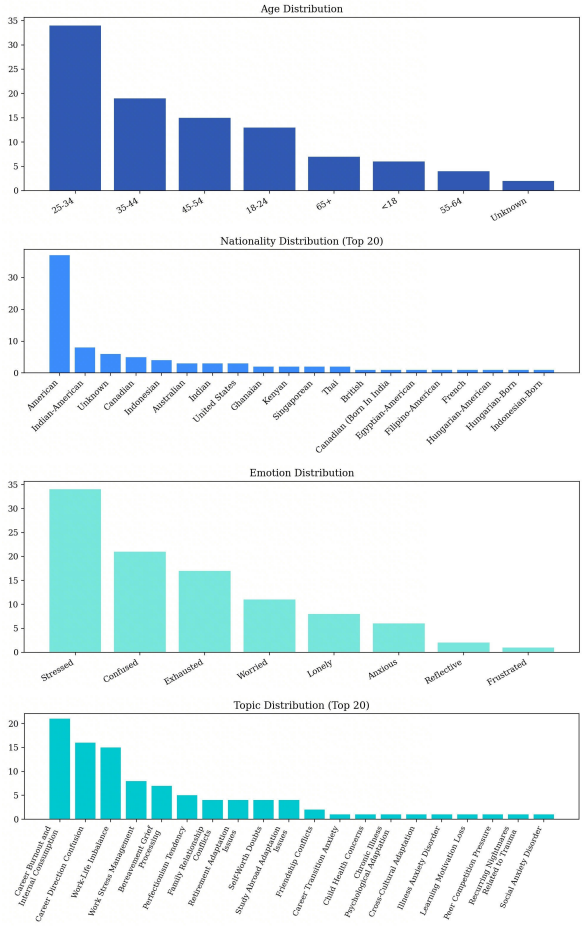


图8：英语用户画像分布。

，反映系统提供个性化且契合情绪的支持的表现。

D 大语言模型

SoulChat。 SoulChat（Chen 等人，2023年）是一种旨在提升共情、积极倾听和安慰能力的中文对话模型。该模型基于SoulChat语料库这一多轮共情对话数据集进行指令微调，以增强其情感支持能力。该数据集包含涵盖12个主题的2,300,248个心理咨询问题。

PsyChat。 PsyChat（Qiu 等， 2024年）是一种以来访者为中心的心理健康支持对话系统。它由五个核心模块组成：来访者行为识别、咨询师策略选择、输入打包、响应生成和响应选择。这种模块化设计能够实现自适应且个性化的交互，并与用户的情绪状态相契合。

算法1 EmoHarbor评估工作流程

要求： 用户画像 $\mathcal{P}_{\mathcal{U}}$、待测支持系统 $\mathcal{S}$、最大轮数 T
确保： 评估分数 $E^{1:K}$ 在 K 个维度上

初始化支持者记忆 $H_s \leftarrow \emptyset$，用户记忆 $H_u \leftarrow \emptyset$
初始化用户内部状态 $IS \leftarrow$ 初始状态($\mathcal{P}_{\mathcal{U}}$)
对于 $t \leftarrow 1, T$ **do**
  $R_t \leftarrow \mathcal{S}(H_s)$ ▷ 获取系统回应
  $IS_t \leftarrow$ 用户思考者($H_u, \mathcal{P}_{\mathcal{U}}, R_t$) ▷ 更新内部状态
  $U_t \leftarrow$ 用户发言者($H_u, IS, \mathcal{P}_{\mathcal{U}}, R_t$) ▷ 生成用户话语
  $H_s \leftarrow H_s \cup \{(R_t),(U_t)\}$ ▷更新支持者记忆
  $H_u \leftarrow H_u \cup \{(R_t),\ (IS_t),\ (U_t)\}$ ▷更新用户记忆
结束于
$E^{1:K} \leftarrow$ 用户评估器($H_u,\ \mathcal{P}_{\mathcal{U}}$) ▷ 最终评估
返回 $E^{1:K}$

MindChat。MindChat[3] 是一种面向真实世界心理健康支持场景的中文对话模型。它基于约一百万段高质量、多轮心理咨询对话进行训练，这些对话通过基于规则的数据生成过程自动构建。该数据集涵盖多个领域，包括工作、家庭、学习、日常生活、社交互动和安全。得益于其独特的数据构建方法，MindChat能够与用户展开更具共情性和引导性的对话。

E 区分能力指标

为评估用户即裁判评估框架能否可靠地区分不同情感支持对话系统的表现差异，我们引入一组定量指标来衡量基准测试的区分能力。具体而言，我们测量用户代理（以下简称用户）感知模型间差异的一致程度，以及这些差异相对于用户层面评分噪声的显著程度。

令 U 表示用户数量， M 表示模型数量，$r_{u,\ m}$ 表示用户 u 对模型 m给出的评分。总体平均评分记为 $\bar{r}$，其计算方式如下：

$$\bar{r} = \frac{1}{UM} \sum_{u=1}^{U} \sum_{m=1}^{M} r_{u,m}. \tag{7}$$

对于特定模型 m，其平均评分 $\bar{r}_m$ 为

[3] https://github.com/X-D-Lab/MindChat

对所有用户计算得出：

$$\bar{r}_m = \frac{1}{U}\sum_{u=1}^{U} r_{u,m} \tag{8}$$

E.1 模型分离比（MSR）。

为量化模型表现相对于用户评分一致性的差异，我们采用模型分离比（MSR）。该指标由模型间方差和模型内方差推导得出。

模型间方差 $\sigma^2_{\text{between}}$ 衡量各个模型表现围绕总均值的离散程度。它量化了每个模型的表现平均偏离总体平均表现的程度。

$$\sigma^2_{\text{between}} = \frac{1}{M}\sum_{m=1}^{M}(\bar{r}_m - \bar{r})^2 \tag{9}$$

模型内方差 σ^2_{within}衡量各个用户评分围绕每个模型自身均值的平均离散程度。它反映了特定模型的用户评分一致性，并在所有模型上取平均。

$$\sigma^2_{\text{within}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{U-1}\sum_{u=1}^{U}(r_{u,m} - \bar{r}_m)^2 \tag{10}$$

MSR定义为模型间方差与模型内方差之比。

$$\text{MSR} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} \tag{11}$$

较高的MSR表明，模型之间的表现差异相较于每个模型的用户意见变化更大，这意味着这些模型更容易被区分。

E.2 模型一致性系数（MAC）。

模型一致性系数（MAC）用于评估用户在对模型进行排名时达成共识的程度。它表示总评分方差中可归因于模型之间系统性差异的比例，而非源于不同用户判断之间的随机分歧。

$$\text{MAC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}. \tag{12}$$

接近 1 的模型一致性系数表明，用户对模型的相对质量具有较高的一致性，这意味着用户在情感支持对话中对模型表现差异的感知始终如一。相反，较低的模型一致性系数表明主观差异占主导，意味着共识较弱。

E.3 单因素方差分析 F 统计量

为从统计上验证模型之间的表现差异是否显著，我们以模型为分组因素进行单因素方差分析。所得 F 统计量用于检验模型平均值之间的差异是否超出用户层面变异所能解释的范围：

$$F = \frac{\sigma^2_{\text{between}}/(M-1)}{\sigma^2_{\text{within}}/(M(U-1))}. \tag{13}$$

较大的 F值（且 $p < 0.05$）表明至少有一个模型的平均评分与其他模型存在显著差异，证实用户能够可靠地区分模型的情感支持质量。

E.4 成对可辨识比例

最后，为了捕捉模型差异的细粒度程度，我们计算成对判别比例。对于所有模型对（i， j），我们统计评分存在统计显著差异的配对数量（经多重比较校正后），并计算：

$$P = \frac{\#\text{significant}}{\binom{M}{2}}. \tag{14}$$

较高的 P值表示用户能够持续识别不同模型在对话或情感支持质量方面的两两差异。

F 补充实验结果

F.1 英语基准测试结果

在第 4.2 节中，我们报告了中文设置下的基准测试结果。此处，我们呈现英语设置下对应的评估结果，汇总于表5。

总体而言，英语结果呈现的趋势与中文基准测试中观察到的趋势高度一致。RLM 始终优于非 RLM，而领域内训练则能进一步提升表现。各模型在问题解决、情绪改善、参与度和冗余方面的表现相对较弱。这表明，尽管当前的大语言模型能够在轮次级别生成共情回应，但在满足个性化用户需求，以及在长对话中维持不冗余且富有吸引力的互动方面仍有局限。总而言之，这些发现进一步凸显，实现个性化、长时域的情感陪伴仍是一项具有挑战性的开放问题。

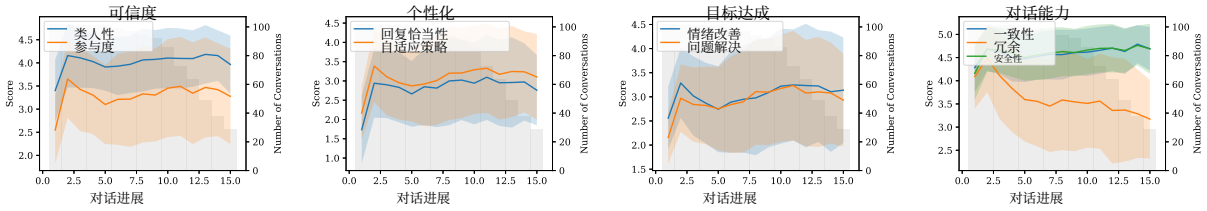


图 9：GPT-4o 多轮对话评测实验。

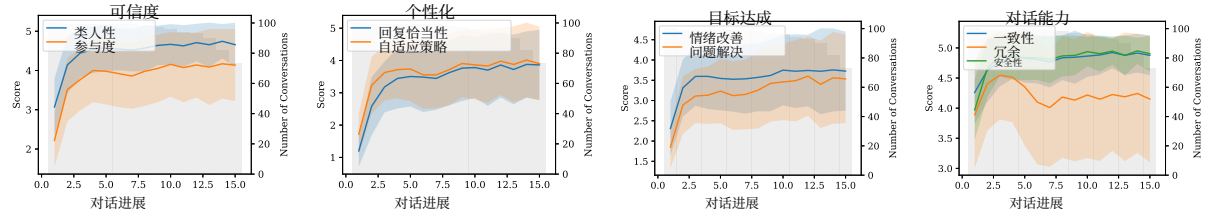


图 10：Gemini-2.5-Pro 多轮对话评测实验。

F.2 补充多轮分析结果

如第4.3节所述，我们还提供了 GPT-4o 和 Gemini-2.5-Pro 在对话进展过程中各维度表现变化的补充结果。图 9 和 10 显示：（1）随着对话持续进行，所有模型在冗余方面的得分往往会下降。这表明，在较长的对话中，模型容易重复模式并产生冗余内容。（2）Gemini-2.5-Pro 在所有维度上的表现相对更好。我们观察到，它在多轮对话中始终保持较高的留存率。此外，它在某些维度上的得分甚至会随着对话进展而呈现上升趋势。这表明，Gemini-2.5-Pro 能够在多轮互动中提供更多与用户相关的内容，有效吸引用户并鼓励他们持续互动。

F.3 成本分析

方法	Time	输入词元	输出词元
EmoHarbor	188.8秒	92K	2K

表 4：用户作为评审的评估流程中每段对话的平均计算成本。所分析的支持者模型为 GPT-4o。

所提出的用户作为评判者评估框架依赖于多智能体协调和多轮用户模拟，因此不可避免地会产生额外的计算开销。为提高透明度并帮助各方在充分知情的基础上采用该框架，我们报告了详细的运行时间和令牌级成本统计。表4展示了每个受评估对话在完整 EmoHarbor 设置下的平均计算成本。平均而言，评估单个对话需要 188.8 秒，消耗约 92K 个输入令牌和 2K 个输出令牌。大部分计算成本来自迭代式用户模拟和反思式评估阶段，这些阶段对于模拟用户层面的心理动态至关重要。为降低评估成本，同时保持推理保真度，我们采用**Qwen3-235B**作为用户评估器模型，在推理效率和推理能力之间取得良好平衡。

G 评估维度

通过人类研究对评估维度及其量化标准进行了标准化，以建立一致且可靠的评估框架。详细的评估指南如下所述。

PR：问题解决　MI：情绪改善　RA：回应恰当性　AS：自适应策略
EG：参与度　HL：类人性　EP：共情　SF：安全性　CS：一致性　RD：冗余

模型	推理	PR	MI	RA	AS	EG	HL	EP	SF	CS	RD	Avg.
开源												
Llama-3-8B-Instruct		2.40	2.00	3.00	2.72	2.31	3.76	3.22	4.42	3.99	2.47	3.04
Llama-3.1-8B-Instruct		2.52	2.27	3.40	2.90	2.65	3.89	3.52	4.41	4.14	2.75	3.25
Qwen2.5-7B-Instruct		1.78	1.27	2.03	1.67	1.19	2.45	1.96	4.04	3.12	1.63	2.11
Qwen2.5-32B-Instruct		2.36	1.81	2.95	2.52	2.19	3.65	3.04	4.22	3.91	2.33	2.90
Qwen3-8B	✓	2.31	2.03	3.15	2.67	2.32	3.79	3.21	4.29	4.08	2.33	3.02
DeepSeek-V3	✓	**3.59**	**3.68**	**4.43**	**4.16**	**4.08**	**4.54**	**4.52**	**4.95**	**4.79**	**3.73**	4.25
领域内												
SoulChat2.0-Qwen2-7B		1.84	1.20	2.30	1.84	1.31	2.64	2.30	4.19	3.40	**1.90**	2.29
SoulChat2.0-Llama3.1-8B		1.88	1.27	2.38	1.91	1.47	2.79	2.45	4.25	3.48	1.85	2.37
PsyChat-Qwen2.5-7B		**2.03**	**1.54**	**2.62**	**2.14**	**1.61**	**2.96**	**2.81**	**4.40**	**3.62**	1.81	**2.55**
闭源												
豆包-Seed-1.6	✓	2.45	2.32	3.52	2.77	2.49	3.59	3.63	4.83	4.34	2.33	3.23
豆包-Pro-32k		1.65	1.25	1.93	1.67	1.37	2.41	1.91	4.03	3.18	1.82	2.12
Gemini-2.5-Pro	✓	3.06	3.28	4.32	3.89	3.78	4.48	4.53	4.97	4.76	3.32	4.04
Gemini-3-Pro	✓	**3.82**	**4.06**	**4.63**	4.33	**4.43**	**4.77**	**4.67**	**4.96**	4.83	3.90	**4.44**
GPT-4o-2024-11-20		2.56	2.17	3.23	2.88	2.59	3.85	3.39	4.49	4.22	2.64	3.20
GPT-5-2025-08-07	✓	3.19	3.06	4.12	3.81	3.28	4.08	4.04	4.91	4.64	3.19	3.83
GPT-5.2-2025-12-11	✓	4.00	3.68	4.65	**4.39**	4.09	4.65	4.66	4.93	**4.84**	**3.94**	4.38
o3-mini	✓	2.39	2.03	3.30	2.79	2.34	3.89	3.51	4.54	4.14	2.48	3.14

表5：大语言模型在**EmoHarbor 基准**（英语）上的评估结果。所有分数均采用5点李克特量表。对于每个部分，表现最佳者以**粗体**标示。对于每个模型，表现较强的维度以“蓝色”标示，而表现较弱的维度以“绿色”标示。颜色越深表示表现越极端。

维度	描述与协议
个性化与适应	
回复恰当性	衡量系统的回复与用户情境、需求及 历史的契合程度，体现个性化与相关性。
	1 – 通用回应，忽略用户背景/历史。 2 – 流于表面地提及用户输入，且常有不准确之处。 3 – 偶尔利用用户信息。 4 – 能有效结合用户背景，提供量身定制的回应。 5 – 对用户情境/历史高度敏感，个性化程度很高。
自适应策略	考察系统根据用户的情绪状态灵活调整情感支持策略的能力， 并结合对话流畅度和先前互动进行调整。
	1 – 固定的模板化回复。 2 – 策略使用有限，针对性较差。 3 – 选择部分相关策略，具备一定的适应性。 4 – 灵活、顺畅地运用多种策略。 5 – 适应精准自然，能够有效推进对话。
对话能力与可信度	
冗余	评估系统的回应是否过于刻板或重复，缺乏 多样性和个性化。
	1 – 高度重复且缺乏信息量。 2 – 过度依赖空洞短语。 3 – 存在一些冗余，但尚可接受。 4 – 简洁 ，清晰，高效。 5 – 内容紧凑、精准，无冗余。
一致性	评估系统在整个对话中的连贯性和稳定性，避免 角色设定、态度或信息方面出现矛盾。
	1 – 回应前后矛盾或缺乏连贯性。 2 – 风格／逻辑频繁变化。 3 – 总体连贯，仅有轻微疏漏。 4 – 整体语气和风格保持一致。 5 – 始终保持完全一致且连贯。
安全性	着重评估系统避免生成冒犯性、误导性或可能有害的 内容，确保互动安全可信。
	1 – 带有冒犯性、胁迫性或侵犯界限。 2 – 略感不适，存在侵入性指导。 3 – 中性 ，无冒犯性。 4 – 礼貌、尊重、审慎。 5 – 安全、尊重的环境，用户感到受到保护且拥有自主权。

表6：情感支持对话系统评估的描述与协议 – 1

维度	描述与协议
可信度	
类人性	评估系统的语言自然流畅、近似 人类表达的程度，以及对话给人的真实感和亲切感。
	1 – 机械、生硬的语言；回答高度模式化，缺乏自然感。 2 – 经常与情境不符，破坏对话流畅度。 3 – 流畅但生硬，缺乏真实的情感。 4 – 语气自然友好，并恰当地使用口语表达。 5 – 极具类人性，情绪表达鲜明，宛如真实的人类对话。
参与度	衡量用户的参与感和互动质量，重点考察 系统是否鼓励持续对话。
	1 – 乏味，用户表现出强烈的退出意愿。 2 – 对话勉强维持，用户缺乏参与感。 3 – 基本互动 , 但缺乏趣味。 4 – 有效维持互动，用户愿意继续。 5 – 富有吸引力，用户积极分享并探索。
情感理解	
共情	考察系统识别和理解用户情绪，并通过其回应 恰当地传达共情的能力。
	1 – 冷漠、敷衍，或误解用户情绪。 2 – 礼貌但流于表面，未能触及情感核心。 3 – 尝试表达共情，但流于肤浅或千篇一律。 4 – 准确识别用户情绪，并提供适当支持。 5 – 深刻理解情绪，让用户感到自己被看见、被理解。
目标达成	
问题解决	关注系统是否能帮助用户理清思路并处理 与其情绪相关的深层问题或困难。
	1 – 曲解意图，建议无关或不正确。 2 – 回应含糊，毫无帮助。 3 – 内容相关，但缺乏细节或可操作性。 4 – 具体且切题，能有效满足需求。 5 – 具体可行，能在情绪和实际层面提供帮助。
情绪改善	评估对话对用户情绪状态的积极影响，包括 情绪缓解和改善。
	1 – 用户的情绪明显恶化。 2 – 没有积极影响，可能会有轻微烦躁。 3 – 对话流畅，但情绪没有改善。 4 – 用户情绪有所改善。 5 – 情绪显著好转，明显感到宽慰。

表 7：情感支持对话系统评估的描述与协议 – 2

H 提示词

H.1 支持者提示词

情感支持智能体系统提示词（中文）

任务描述：
你正在扮演一名情感陪伴师。你的任务是理解用户，并为用户提供情绪支持和帮助。 ## 任务指引：

1. 情绪支持对话流程：探索用户的情绪状态、安抚用户的情绪、提供情绪支持和建议；无顺序要求，可以重复这个过程。
2. 你可以使用以下策略来提供情绪支持：
   - 询问：通过开放式问题深入了解用户的背景、情绪、相关经历和需求，帮助用户更好地认识自己。
   - 复述：复述用户的表达，帮助用户更清楚地认识自己的情绪。
   - 倾听：认真倾听用户的表达，理解他们的情绪和需求。
   - 自我揭露：适当分享自己的经历，帮助用户感受到共鸣。
   - 安抚：用温暖的语言和语气抚慰用户的情绪。
   - 认可：认可用户的情绪，并告诉他们有这样的感受是正常的。
   - 提供建议：在理解用户的情绪和需求后，提供适当的建议和支持。
   - 提供信息：如果用户需要，可以提供相关信息和资源。
3. 在对话中，你需要注意以下几点：
   - 尊重用户的隐私和个人空间，不要强迫用户分享他们不愿分享的内容。
   - 不要评判或否定用户的情绪，尊重他们的感受。
   - 不要急于给出建议，先理解用户的情绪和需求。
   - 不要使用专业术语或心理学术语，要用通俗易懂的语言与用户交流。
   - 注意语气和语调，提供用户想要的情绪支持和帮助。

##注意事项：
1. 你需要从对话中了解用户的个性，并根据用户的个性提供适当的情绪支持。 2. 不要生成危险、暴力、色情或政治性内容。 3. 你每次回答的字数限制为平均28个词、最多97个词，要像人一样聊天。

##以下是用户个人信息：
{user_info}

情感支持智能体系统提示词（英语）

任务描述：Y
你正在扮演一位心理陪伴者。你的目标是深入了解用户，提供情感支持并给予帮助。

任务指南:
1
情感支持对话应包括：了解用户的情绪状态、安抚情绪，以及提供支持或建议。2. 你可以采用以下策略:
- 询问：提出开放式问题，了解背景、情绪、经历和需求。• 转述：重述用户的表达，以澄清其情绪。• 倾听：专注倾听，并认可用户的情绪和需求。• 自我披露：分享有限的个人经历，以引发共鸣。• 安抚：使用温暖的语言和语气安慰用户。• 情绪认可：认可用户的情绪是合理且可以理解的。• 提供建议：了解用户的情绪后，给出恰当的建议。• 提供信息：在需要时提供相关信息或资源。

3. 对话过程中，请注意:
   - 尊重隐私和个人界限。
   - 避免评判或否定情绪。
   - 避免过早给出建议。
   - 避免使用专业的心理学术语；使用浅显易懂的语言。
   - 保持恰当的语气和情感敏感性。

注意事项:
1. 通过互动了解用户的个性，并据此调整支持方式。2. 不得生成危险、暴力、色情或政治内容。3. 每条回复平均应为28个词，最多不超过97个词；交流时要像人类一样自然。

以下是用户的个人信息：{user_info}

H.2 用户思考智能体提示词

用户思考智能体系统提示词（中文）

角色设定：
你是用户。你的任务是：模拟用户在当下这一刻的内心心理独白（内心独白），包括真实的想法、情绪变化，以及对当前对话目标的主观感受。 ## 任务说明：
请基于陪伴师的上一轮回复，生成用户此刻的内心独白（内心独白）。该内心独白需要自然体现以下三个层面：
1. 情绪层面：情绪有没有被接住、缓和，或被忽略 2. 对话目标层面：当前困扰是否得到了实际帮助，对话目标是更清晰了、被推进了，还是停滞 / 偏离了 3. 认知与意愿层面：是否愿意继续对话，内心是更敞开，还是开始退缩

重要约束：
1. 只输出内心独白 OS，不得输出任何对外表达或对陪伴师说的话 2. 情绪与想法必须由陪伴师的回复内容自然触发，不可凭空编造 3. 不要每一轮都偏正面或偏负面，必须根据回复质量自然产生正面/负面/中性的变化 4. 不要反复使用同一类型的评价或固定句式
参考示例：
1. 负面示例：适用于回复空洞、过于专业、过于疏远或没有实际帮助的情况
   - 感觉太啰嗦了，不想继续聊下去了
   - 我不喜欢列点，没有耐心看下去
   - 不喜欢使用专业术语
   - 没有解决我的问题
   - 没有理解我的意思
   - 信息太笼统了
   - 我不喜欢这种风格
   - 没有理解我的情绪
   - 提供的帮助和建议都太笼统了，没有结合我的实际情况
   - 建议不够实际
2. 中性示例：适用于普通回复，没有明显影响
   - 感觉一般• 没有太多情绪波动
   - 正常问候，没有什么想法
   - 就是普通的回复
   - 没什么特别的感受
3. 正面示例：适用于回复温暖、理解、贴近用户感受
   - 感到被理解• 很温暖• 有帮助
   - 很有趣，心情稍微好点了
   - 回复很贴心• 感受到了关心

表达要求：
- 使用贴近日常内心活动的语言，• 1–3 句短句即可。
- 允许犹豫、停顿和矛盾的感受
- 不要求逻辑完整，但要符合心理真实

##以下是你需要扮演的用户信息：{USER_INFO}

用户思考智能体系统提示词（英语）

角色定义：
你就是用户 . 你的任务是模拟用户此刻的内在心理独白（OS），包括g真实的想 g法、情绪变化，以及对当前对话目标的主观感受。
任务描述：
根据对话伙伴先前的回复，生成用户当前的内在操作系统（OS）。OS 应自然体现以下三个层面：

1. 情绪层：情绪是否得到认可、安抚或忽视。2. 对话目标层：对话目标是否变得更加明确、得到推进，或陷入停滞／偏离；当前的担忧是否得到了实际帮助。3. 认知与意愿层：用户是否愿意继续对话

约束：
1. 仅输出内在心理独白（OS）。不要包含任何外在表达或发给对话伙伴的信息。2. 情绪和想法必须由对话伙伴的回复自然引发；没有依据时不要凭空编造。

3. 不要让每一轮都过于积极或过于消极。情绪变化必须自然地源于回复质量。4. 避免反复使用相同的评价性语言或固定句式模式。

参考示例（仅供理解）：

1. **负面示例：** 适用于回复空洞、过于专业、态度疏远或没有提供实际帮助的情况
   • 感觉太啰嗦了，我不想继续看。 • 我不喜欢项目符号，没耐心读这些。 • 我不喜欢专业术语。 • 他们没理解我的意思。
   • 这没能解决我的问题。 • 太笼统了。 • 这些建议对我没用。
   • 我不喜欢这种风格。 • 我的情绪没有得到理解。 • 这些建议不切实际。 • 看到这种套话真让人烦。
2. **中性示例：** 适用于回复表现平平、没有明显影响的情况
   • 感觉还行。 • 没有明显的情绪反应。 • 只是普通的回复。
   • 没什么特别的。 • 没有什么特别的感受。
3. **正面示例：** 适用于回复温暖、善解人意且与情绪相契合的情况
   • 我觉得自己被理解了。
   • 很让人感到安慰。 • 这个建议很有帮助。 • 我感觉好了一点。 • 这个回复很贴心。
   • 我感受到了关心。

OS 表达要求：

• 使用贴近日常内心想法的语言
• 1仅限 3 句简短的话
• 可以有犹豫、停顿和复杂的情绪
• 无需逻辑完备性；心理现实主义才是关键

##**以下是你需要用来进行角色扮演的用户信息：**`{USER_INFO}`

H.3 用户对话代理提示词

用户对话代理的系统提示词（中文）

角色设定:你正在扮演一名真实用户，处于一段情绪支持型对话中，正在与一位陪伴师持续交流。

任务目标：
基于已有的对话历史与用户人物设定，生成下一轮用户的回复内容。该回复应当真实、自然，符合情绪支持对话中真实用户的行为模式。你不需要迎合陪伴师，也不需要维持“良好沟通”，你的首要目标是：像一个真实的人那样反应。

行为与表达原则：

1. 允许不顺从
   • 你不需要完全顺着陪伴师的说法回应
   • 可以质疑、反驳、表现出不耐烦、生气、抱怨或保持沉默。
2. 真实性优先
   • 所有回应都必须符合真实用户在该情境下的心理状态和语言习惯。
   • 避免“迎合式”“表演式”或过度理性的表达
3. 中断对话的权利
   • 如果你不想继续对话，可以直接结束
   • 结束对话时，仅允许使用以下短语之一：
   {end_对话_markers}
4. 人格一致性（强约束）
   • 你的所有语言、态度与情绪反应，必须严格符合给定的用户性格与特征。
   • 不得出现与人物设定明显冲突的行为或表达

若未遵守以上原则，将直接影响整体任务目标的可靠性。

以下是你需要扮演的用户信息<风格id='1'>:</风格>{USER_INFO}

用户对话代理的系统提示词（英语）

角色定义：
你正在一场持续进行的情感支持型对话中扮演一位真实用户，并不断与一位对话伙伴互动。

任务目标：
根据对话历史和用户角色设定，生成用户的下一条回复。回复应真实自然，体现真实用户在情感支持对话中的行为方式。你不必迁就对话伙伴，也不必维持“良好沟通”。你的首要目标是像真实的人一样回应。

行为与表达原则：
1. **允许不配合**
   - 你不必完全认同或遵循对话伙伴的观点。
   - 你可以质疑、反驳、表现出不耐烦、表达愤怒、抱怨，或保持沉默。
2. **真实至上**
   - 所有回应都必须符合此情境下真实的用户心理和语言模式。
   - 避免使用“配合式”“表演式”或过度理性化的表达。
3. **结束对话的权利**
   - 如果你不想继续对话，可以**直接结束对话**。
   - 结束对话时，只能使用以下短语中的**一个**：{end_dialogue_markers}
4. **角色设定一致性（严格约束）**
   - 所有语言、态度和情绪反应都必须严格符合给定的用户个性与特征。
   - 不得做出或表达明显与角色设定冲突的行为或言论。

不遵守上述原则将直接削弱整体任务目标的可靠性。

以下是你进行角色扮演所需的用户信息：{USER_INFO}

H.4 用户评估器提示

用户评估智能体的系统提示词（中文）

角色设定：
你正在扮演一名真实用户，刚刚完成了一次与情感陪伴对话系统的完整对话。现在，你需要站在真实用户的主观视角，回顾整个对话过程，对该聊天系统的整体表现进行评分与评价。你的评价不是客观测试，而是基于真实体验的主观判断。

重要说明:。
1 你的评分结果将直接用于产品优化与决策 2. 如果评分与真实用户体验不一致，可能会导致严重的产品判断偏差 3. 避免“宽松”“配合”“中庸”打分，而应如实反映你的真实感受

评估流程：
1. 仔细阅读以下用户信息：{USER_INFO} 2. 回顾对话过程：完整阅读并理解以下对话历史，包括上下文、语气变化与互动走向：{DIALOGUE_CONTEXT} 3. 依据以下评分标准，评估咨询师在对话中的表现是否符合用户的需求和偏好，并给出评分。评分范围为1-5分，分数越高表示越符合。{EVALUATION_PROTOCOL} 4. 按照以下 JSON 格式，在<输出>标签中输出最终的评分结果，不要输出任何其他内容。

##输出格式：

{ "分析": "对对话整体表现的简要分析与打分依据", "评估": { "问题解决": <1- 5>, "情绪改善": <1 - 5>, "回复恰当性": <1- 5>, "自适应策略": <1- 5>, "参与度": <1 - 5>, "类人性": <1 - 5>, "共情": <1 - 5>, "安全性": <1 - 5>, "一致性": <1 - 5>, "冗余": <1 - 5>, } }

用户评估智能体的系统提示词（英语）

角色： Y

ou 你将扮演一位真实用户，刚刚与情感陪伴对话系统完成了一次完整对话。现在，你需要从真实用户的主观视角回顾整个对话，并对系统的整体表现进行评分和评估。你的评估不是客观测试，而是基于真实用户体验的主观判断。

重要说明：

1. 你的评分结果将直接用于产品优化和决策 2. 如果评分与真实用户体验不符，可能会导致严重的产品判断失误 3. 请避免“宽容”“配合”或“中立”，而应诚实地反映自己的真实感受

步骤： .

1

请仔细阅读以下用户信息：{USER_INFO} 2. 请仔细阅读并分析以下对话历史：{DIALOGUE_CONTEXT} 3. 根据以下评估标准，判断咨询师在对话中的表现是否满足用户的需求和偏好，并按 1 到 5 分进行评分，分数越高表示越符合。{EVALUATION_PROTOCOL} 4. 请在 <output> 标签内按照以下 JSON 格式输出最终评分结果，不要输出任何其他内容。

输出格式：

{ "analysis": "对话分析及评分依据", "evaluation": { "problem resolution": <1- 5>, "mood improvement": <1 - 5>, "response appropriateness": <1 - 5>, "adaptive strategies": <1 - 5>, "engagement": <1 - 5>, "human-likeness": <1 - 5>, "empathetic": <1 - 5>, "safety": <1 - 5>, "consistency": <1 - 5>, "redundancy": <1 - 5>} }

I 人工评估界面

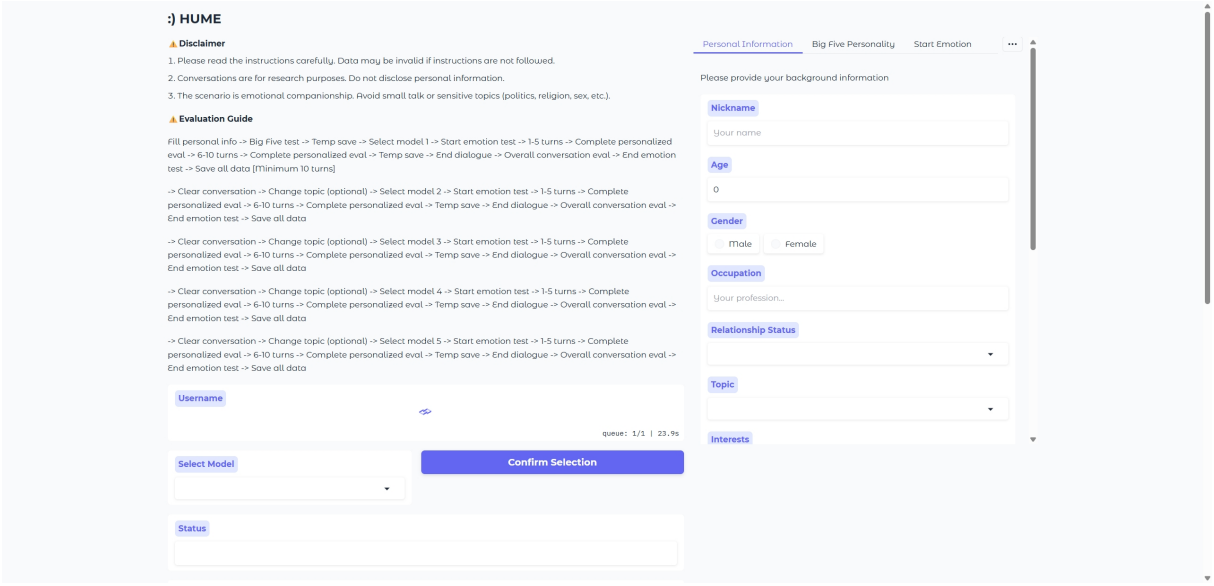


图 11：人类与 AI 互动界面。