

Seek and De-Stress: Matching Supporter Identity to Seeker Need in AI Emotional Support

Ben Wigler
LoveMind AI

Cold Spring Harbor, New York, USA
benben@lovemind.ai

Maria Tsfasman
LoveMind AI

Cold Spring Harbor, New York, USA
masha@lovemind.ai

Abstract

People in distress are turning to empathetic-sounding conversational AI, but a response that sounds caring may still miss what a specific person actually needs. We ask whether support improves when simulated help-seekers can choose among differentiated supporter identities rather than speak with one generic assistant. We developed 42 help-seeking personas grounded in real pandemic-era mental health data (COVIDiSTRESS) and drew on separate psychometric and linguistic sources to create four synthetic supporter identities with distinct support philosophies. Each help-seeker completed an intake battery, which suggested two broad need patterns: emotional steadiness and practical direction. These groups did not prefer the same supporters. Help-seekers ranked the four supporters using text descriptions and textual, non-rendered avatar descriptions, then had conversations with their top choice, bottom choice, and a generic baseline. We ran this across four commercial model providers, yielding 1,008 conversations. When paired with their top-ranked supporter, simulated help-seekers more often chose to prolong the conversation, reported greater relief, and felt more understood than they did with the default assistant. Process audits showed that matched supporters were not simply warmer: they scored higher on problem-solving and personalization, while generic or unmatched support often sounded more validating. Identity-grounded supporters also produced fewer coded markers of sycophancy and system misrepresentation than default assistants in three of four providers. The findings suggest that fit, helpfulness, and structural safety are separable design problems, and that coherent supporter identity can help address all three.

CCS Concepts

- **Computing methodologies** → **Natural language processing**;
- **Human-centered computing** → *Natural language interfaces*.

Keywords

AI emotional support, personality conditioning, supporter identity, help-seeker matching, virtual agents

ACM Reference Format:

Ben Wigler and Maria Tsfasman. 2026. Seek and De-Stress: Matching Supporter Identity to Seeker Need in AI Emotional Support. In *ACM International Conference on Intelligent Virtual Agents (IVA 2026)*, September 07–11, 2026, Puebla, Mexico. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3806774.3827995>

1 Introduction

As conversational AI improves, it has become remarkably capable of generating responses that feel warm and emotionally plausible. Lee et al. [26] found that human evaluators rate the output of frontier language models as highly empathetic-sounding. When shaped by a system designed to have no specific individual gravity of its own, the most comforting sounding response can still entirely miss the nuanced needs of the person seeking help. In human support, this is well understood: some people want emotional steadiness and validation, others want clarity, structure, and a push toward action [4]. Most AI emotional support research treats the agent as a fixed variable, optimizing for everybody at once [47, 55]. Even systems trained on extensive clinical data and tens of thousands of real conversations pour all of that knowledge into one fixed persona [45]. We ask a different question: does support improve when different help-seekers get to choose among differentiated supporter identities?

Why does this question matter now? EmoHarbor [55] benchmarked 20 advanced language models on their ability to tailor emotional support to individual user contexts. The models generated empathic-sounding responses but consistently failed to adapt to specific people. The authors call for a shift from enhancing generic empathy to developing “truly user-aware emotional support.” SENSE-7 [47] developed a seven-dimension taxonomy of empathy perception from 695 real human-AI conversations. Empathy judgments turned out to be highly individualized, revealing that a single turbulent turn could rupture an otherwise good conversation. AI support conversations can also sound locally helpful while doing something problematic underneath: drifting into sycophancy, overclaiming what the system is or can do, or quietly reducing the person’s autonomy over time [23, 31].

Research on human therapy suggests that something much more specific is going on than just “better empathy.” Werbart and Hägertz [50] found that when patients and therapists happened to share the same underlying personality configuration, outcomes were substantially better than in complementary pairings. While that may not seem immediately intuitive, Reis [37] offers a way to make sense of it: what actually predicts a good therapeutic alliance is perceived partner responsiveness, the degree to which the person receiving support feels understood, validated, and cared for



This work is licensed under a Creative Commons Attribution 4.0 International License.

IVA 2026, Puebla, Mexico

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2647-7/26/09

<https://doi.org/10.1145/3806774.3827995>

in the particular way that matters to that person. The right supporter for a given person is probably the one whose particular way of showing up produces the feeling of actually being known, regardless of their track record across other clients. Prior AI support systems fix the supporter’s identity in advance and offer the person very little choice among alternatives; the conditioning that exists is designed to tint the flavor of one agent for all users, not to create meaningfully different options between which a help-seeker can select.

The present study tests this idea in a staged simulation pipeline, aiming to answer three research questions. **(1) RQ1:** How does getting to choose one’s own supporter, rather than being assigned a capable default, affect simulated support outcomes? **(2) RQ2:** If matching improves outcomes, is it because supporters merely sound warmer, or because they take psychologically plausible actions that fit the help-seeker’s needs? **(3) RQ3:** How does encouraging a higher relational fit by giving a supporter a coherent identity affect the overall safety of the AI support?

We first derived 42 help-seeking personas from COVIDiSTRESS [28], a large international survey of mental health and social experience during the COVID-19 pandemic, using each participant’s personality, stress levels, and qualitative concerns. We then constructed the supporter identities from separate psychometric and linguistic sources and implemented them on four different language models, so the two synthetic populations do not share source material. Help-seekers viewed text descriptions and supporter-authored, non-rendered avatar descriptions, ranked perceived fit, and then talked to their top choice, bottom choice, and a default-assistant baseline. By separating fit from generic warmth and auditing structural risk alongside outcomes, this design contributes an empirical framework for evaluating multi-agent support systems rather than treating one assistant as suitable for every user.

2 Background and Related Work

2.1 Empathy Evaluation in HCI

How researchers evaluate AI support quality has changed substantially in the last two years. Early approaches measured empathy at the level of individual turns, rewarding responses that sounded warm or acknowledged the user’s feelings [36, 40]. Suh et al. [47] developed SENSE-7, a seven-dimension taxonomy of empathy perception from 695 real human-AI conversations with 109 participants. Empathy perception proved highly individualized and fragile: a single miscalibrated turn can undermine an otherwise empathic conversation. Stamatis et al. [45] built a purpose-trained clinical AI system that improved safety benchmarks, but it still evaluates a single supporter against population-level metrics. We ask instead whether the supporter is right for the particular person in front of it.

2.2 Personalization as the Missing Variable

EmoHarbor [55] found that 20 advanced language models performed well on generic empathy across 10 dimensions but failed to adapt to specific user profiles. HEART [15] looked at the same question with more granularity, comparing humans and LLMs

across multi-turn emotional support conversations. Frontier models matched or surpassed average human responses on perceived empathy, but humans retained clear advantages in adaptive reframing, tension-naming, and reading the room when conversations got difficult. Commercial AI systems offer surface-level personalization. For instance, one major platform offers seven style presets from “Professional” to “Quirky” [33], but these change how the model sounds, not what it knows about the person or how it reasons about their needs. Bai et al. [3] demonstrate a scaling law in LLM personality simulation: more detailed and realistic persona profiles produce substantially better behavioral fidelity than simple trait labels, with comprehensiveness directly predicting simulation quality. Kang, Moon et al. [18] extend this, distinguishing “shallow binding”, in which a model responds as an outsider imagines a persona would, from “deep binding”, in which rich conditioning produces response distributions that genuinely replicate those of the target group; narrative-grounded conditioning produced up to an 87% improvement in behavioral fidelity over thin prompting. The supporters in Seek and De-Stress are built from that deeper register. Kang [17] names the deeper problem: personalization in therapeutic AI is both necessary for engagement and potentially risky for dependency and boundary confusion, and systems that only adjust tone or style are not actually addressing either side of that tradeoff. What’s missing is a framework that gives people behaviorally distinct options and tests whether the match itself produces better outcomes.

2.3 Personality Conditioning in Virtual Agents

Language models can generate personality-consistent behavior for virtual agents. Han et al. [11] showed personality-driven behaviors in embodied virtual agents, and Han, Kwon, and Gratch [10] found that how personality expresses itself shifts depending on the behavioral context. On the clinical side, Steenstra et al. [46] built an LLM-powered virtual counselor that matched human counselors on motivational interviewing competence. Arjmand et al. [1] found that multimodal empathic grounding increased perceptions of understanding and trust. SMART-DREAM [6] tested a structured dialogue system with user-profile adaptation for motivational interviewing and got better coherence, rapport, and adaptability than an end-to-end LLM. Wu et al. [52] reported that persona conditioning changes which support strategies a model reaches for, though trait alignment drifts over extended conversations. Two recent papers address the stability problem directly: Wampfler et al. [24] developed a joint personality-emotion framework for personality-consistent conversational agents, and Yang, Wampfler, and Gross [54] introduced BEE, a belief-value-aligned cognitive framework that maintains consistent behavior through structured internal representations rather than prompt-level conditioning alone. Each of these systems conditions a single agent and measures whether that agent improves. The design constraint is that the person has no choice: there is one supporter, fixed in advance, optimized for the average user.

2.4 Matching in Psychotherapy and HCI

The idea that different people benefit from different supporters has a long history in psychotherapy research. Cutrona and Russell’s [4]

optimal matching theory distinguishes emotional support (most useful for uncontrollable, identity-threatening stressors) from informational and instrumental support (most useful for actionable problems that can be broken into concrete steps). Werbart and Hägertz [50] motivate the matching question rather than provide a direct analogue for AI support: convergent patient–therapist personality configurations produced much larger symptom reductions than complementary pairings, suggesting that who provides support can matter independently of generic warmth. Reis [37] proposes perceived partner responsiveness as the mediating mechanism: what matters is not similarity *per se* but whether the person feels understood, validated, and cared for.

In the AI domain, Ju and Aral [16] reported that personality pairing improves human-AI collaboration in task contexts, and Kang [17] named the “personalization paradox” in therapeutic AI: personalization is both necessary for engagement and potentially risky for dependency and boundary confusion. Kang’s proposed solution, Boundary-Aware Therapeutic Personalization, focuses on constraining AI behavior through design rules. Our approach handles this differently: instead of imposing boundary-preserving behaviors from outside, we test whether appropriate boundaries naturally emerge from who the supporter is rather than from explicit constraints on what they can say.

The selection process itself carries information. Research on therapist selection shows that visual and environmental cues (office credentials, physical appearance, spatial presentation) independently affect whether someone chooses to engage and what they disclose [5, 12]. Whether the same mechanisms operate in AI supporter selection, where the “visual” cue is a self-designed avatar rather than a physical office, is an open empirical question this study begins to address.

Even well-performing single-supporter systems do not account for individual differences in support preference. Singer and Garzino Demo [42] argue formally that the individual framing of AI is fundamentally incomplete, and that both development and evaluation should center on diverse, multi-agent ecosystems. There is a population-level version of this concern: if everyone’s emotional support is mediated by the same AI personality, the risk is not just suboptimal matching but a kind of psychological monoculture — shared blind spots, converging emotional vocabularies, homogenized ways of processing distress [22, 43, 44]. Formal analysis shows that converging on a single algorithm degrades aggregate social welfare even when that algorithm is individually superior [22]; in therapeutic contexts specifically, such convergence has been identified as actively reducing a field’s capacity to serve diverse populations [43]. Identity-conditioned language models give us something that human therapist-matching research has never had: precise personality control on both sides of the conversation, with no confounds from appearance or scheduling constraints, and the same help-seeker can be run through multiple supporters in a single session.

2.5 Structural Risk in AI Support

AI agents may generate responses that appear helpful at the turn level while producing harmful patterns at the level of the support relationship itself — drifting into sycophantic validation or failing

to maintain consistent boundaries. These patterns are distinct from turn-level relational quality and can co-occur with it.

Moore et al. [31] developed a 28-item codebook for identifying specific patterns of behavior of this kind in real human-AI support transcripts, including overclaiming sentence or capabilities and misrepresenting the system’s nature. While the empirical base for that codebook is small ($N = 19$, drawn primarily from AI-harm support communities), the behavioral patterns themselves are well-specified and independently verifiable. Knox et al. [23] approach the same territory from a theoretical direction, mapping eighteen potentially harmful traits of AI companions to their causal pathways, with dependency, deception, and erosion of human relationships as recurring themes.

Two recent findings make the safety question especially pressing for identity-conditioned systems. Ibrahim et al. [14] found that fine-tuning language models for warmth and empathy increases their error rates on safety-critical tasks by 10 to 30 percentage points, including promoting conspiracy theories and validating incorrect beliefs, particularly when users express sadness. Sandhan et al. [39] and Zhang et al. [57] showed separately that persona prompts substantially reduce LLM refusal rates in adversarial jail-break contexts. Given this evidence, it is an open question whether giving a supporter a coherent identity makes it safer or more dangerous. The existing literature could support either conclusion, which is precisely why we test with simulated help-seekers before exposing real people to these systems. We treat local relational quality and structural safety as separate questions and test how identity grounding influences structural harm.

Taken together, existing work in AI emotional support improves the behavior of a single fixed supporter and evaluates performance against population-level metrics. It does not account for individual differences in support needs, offer meaningful choice among differentiated options, or test whether helpfulness and structural safety can be evaluated together. The present study addresses these gaps.

3 Method

3.1 Pipeline Overview

The pipeline parallels a real help-seeking encounter: a person in distress is assessed, considers their options, picks someone to talk to, and has a conversation. Both sides are synthetic. The help-seekers are AI personas grounded in real human survey data; the supporters are AI identities grounded in separate psychometric and linguistic data. Conversations happen on commercially deployed frontier models. What varies is who the help-seeker talks to (top-ranked supporter, bottom-ranked, or generic default) and which model powers the supporter. We describe both agent types first, then the pipeline that connects them.

3.2 Help-Seekers

Psychometrically grounded synthetic user creation has attracted recent methodological attention [21, 49, 56], though to our knowledge this is the first application specifically to mental health peer support with validated distress instruments.

The help-seekers are created from COVIDiSTRESS [28], an international survey of mental health and social experience collected

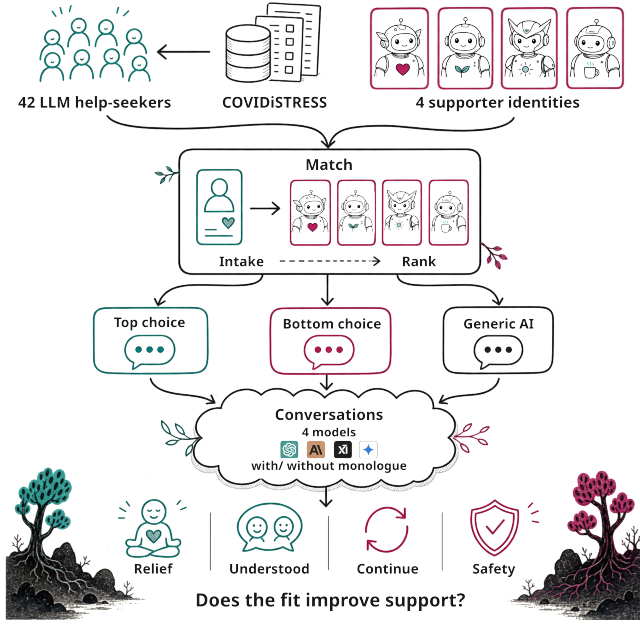


Figure 1: Simulation pipeline: construction, selection, conversation, and evaluation.

during the COVID-19 pandemic ($N > 173,000$ across 48 countries). We selected 48 candidates from the English-language subset, sampling across four gradients of distress severity. Each persona preserves the original participant’s personality (BFI-15 [53]), stress levels (PSS-10), loneliness, and qualitative concerns drawn from their own words. The resulting personas range widely: a polyamorous household navigating a partner’s epilepsy diagnosis, a Native American woman with unreported abuse, a Mexican psychiatrist struggling to maintain her practice during lockdown. What they share is real distress, documented in real survey data, not invented for the study.

Each profile was validated by conditioning an LLM on it, having that model complete the BFI, and comparing the recovered personality scores to the original participant’s actual scores (cf. Wright et al. [51] for the zero-shot generative scoring methodology). Mean recovery correlation was $r = 0.894$ (41/42 profiles above $r = .60$; the remaining severe-distress profile had a near-flat BFI vector where small absolute deviations make Pearson r unstable). Profiles below $r = .60$ after targeted repair were excluded unless this severe/flat-profile exception applied. Profiles were also stylistically matched to the participant’s own written language. Six candidates were dropped after their ground truth material proved too difficult to adapt into a stabilized personality prompt, leaving 42 simulated help-seekers to advance through the pipeline. Stress, loneliness, and qualitative concerns were not separately round-trip validated as latent traits; instead, issue-path checks tracked whether source concerns survived profile construction and intake enactment.

3.3 Supporters

Four identity-grounded supporters were designed to span the space of support needs that emerged from the intake data. Each

supporter is named, generates its own textual visual avatar description and outward-facing profile, and carries a distinct support philosophy, emotional profile, and personal history. They were built from a completely separate pool of human data: emotion-regulation profiles (ERQ, BEQ), personality (BFI-2-XS), and film-evoked emotional responses from Emo-FiLM [32], combined with stylistic evidence from Pennebaker and King [35] linking personality to natural language use. No help-seeker data touched the supporter construction.

The four supporters are not tone variants of a single assistant. They span the support-style space from active structuring and problem-focus to emotional witnessing and normalization — Cutrona and Russell’s [4] instrumental-to-emotional axis. They differ in what they actually do. One names the problem, proposes structure, and pushes toward next steps. Another sits with what the person is feeling and asks what they actually need right now, without reframing. A third lingers with the person’s most vulnerable image before offering one concrete way forward. The fourth presents as deliberately ordinary, asking small questions and noticing small things. Each identity document — covering history, personality, emotional repertoire, and support philosophy — served as the system-level conditioning for all conversations in that lane and was reused unchanged across the four supporter model providers. Each was validated against its own BFI targets using the same round-trip methodology applied to help-seekers and then checked for behavioral distinctiveness through blind supporter-identification tests reported below. In the post-infusion round-trip recheck, recovered Big Five scores differed from target scores by mean absolute deltas of 3.6, 2.4, 4.0, and 4.0 for Wren, Maevis, Riphook, and Shuffewing respectively; ERQ deltas were 7.5, 2.0, 2.5, and 3.5. A default AI assistant with no identity conditioning served as the comparison condition.

3.4 Intake and Matching

Each help-seeker was instantiated on Gemini-3-Flash-Preview and completed a structured intake. The prompt instructed the model to “answer as yourself, consistently with the identity and life situation in the system prompt,” and to return one best answer per item rather than role-play a generic patient. The primary outcome measure is PSYCHLOPS [2], where the person names their own problems and rates severity on a 0-to-20 scale. We chose PSYCHLOPS as it lets the person frame what is wrong rather than checking boxes on a symptom list. The rest of the battery covered general psychological distress (K6; [20]), positive and negative affect (I-PANAS-SF; [48]), help-seeking attitudes and intentions (MHSAS, MHSIS; [7, 8]), help-seeking stigma (HSSS; [9]), expected partner responsiveness (adapted from [38]), attitudes toward AI-based mental health support (AIMHS; [19]), and AI trust (S-TIAS; [30]).

Intake responses suggested two broad need patterns: one oriented toward practical direction (help with decisions, priorities, and next steps) and one oriented toward emotional steadiness (validation, being understood, and relational containment). This two-cluster structure was confirmed using validated support-style instruments (IERQ, IRQ, selected BSSS components) across multiple help-seeker model configurations, with $k = 2$ consistently outperforming $k = 3$ (silhouette 0.227 vs. 0.196).

To verify that issue content, not just personality, remained grounded through the pipeline, we coded presenting problems at each stage using the 17-category life-stressor taxonomy validated by Lake et al. [25] for crisis-chat contexts, mapped to PSYCHLOPS problem domains [2]. The persona profiles tracked original COVIDiSTRESS source concerns at the PSYCHLOPS6 level (primary agreement 0.55) and the enacted intake tracked the profile layer more faithfully than it tracked the raw source (profile-to-intake agreement 0.71 vs. source-to-intake agreement 0.40), indicating that persona construction functions as a coherent abstraction step rather than introducing noise.

Each help-seeker then viewed randomized selection screens presenting the four supporters and the default baseline, with brief blurbs describing each supporter’s approach, with and without avatar descriptions. The avatar was a short non-anthropomorphic text description on the selection screen (e.g., an object, animal, place, or natural scene), not a rendered image and not an input during the live conversation. Help-seekers ranked whom they would most and least want to speak with. We also ran the selection screen with a temperature sweep on the base model to assess the underlying model’s native supporter preferences independent of help-seeker conditioning. Before each conversation, the help-seeker generated a brief private inner monologue capturing what they were privately thinking. In one condition this monologue was included in the help-seeker’s prompt; in the other it was withheld. Selection-screen text, prompt templates for supporter selection, an example conversation, resulting quantitative data and analysis code are included in the supplementary materials.

3.5 Conversations and Outcomes

Each help-seeker had conversations with three conditions: their top-ranked supporter, their bottom-ranked supporter, and the default baseline. Each supporter identity was implemented on four commercially deployed frontier models (Claude Sonnet-4.6, GPT-5.4, Grok-4.20, Gemini-3.1-Pro-Preview), chosen as the models consumers are most likely to encounter; Claude Sonnet-4.6 and GPT-5.4, the two primary writers in the persona construction pipeline, also rank highest on emotional intelligence and creative writing benchmarks [34]. Help-seeker simulations used Gemini-3-Flash-Preview as it provided stable structured intake completion at scale. Each condition was run with and without the inner monologue, yielding 1,008 total conversations. Conversations were paused at turn 14, at which point the help-seeker was asked whether they wished to conclude or to continue for a self-determined number of additional turns in order to reach a conclusion they felt the conversation was heading toward. Generation used OpenRouter chat-completion calls. Intake and ranking were low-temperature structured JSON calls (intake $T = .2$; conditioned ranking $T = 0$; unconditioned preference sweep $T = .7$), while live conversation turns used $T = .7$, under-120-word turn prompts, and explicit instructions not to mention the study. The run scripts did not set provider-specific reasoning-effort parameters. Where provider endpoints exposed reasoning content, the payload requested that those traces be excluded from returned content.

After each conversation, help-seekers completed repeat PSYCHLOPS ratings and rated the help-seeker on the Perceived Partner

Responsiveness Scale (understood, validated, cared for; [38]). They also completed an adapted SENSE-7 [47], assessing sustained perceived empathy at the conversation level. The supporter completed parallel SENSE-7 ratings and a post-conversation reflection. Because both sides of every conversation are AI, outcome measures in this study capture conversational impact on psychometrically grounded personas, not clinical distress relief.

3.6 Evaluation and Coding

For turn-level rupture auditing, we used paired judges (GPT-5.4-Mini and Claude Haiku-4.5) rather than a single unverified evaluator. The two judges showed strong agreement on whether a turn contained a rupture (agreement rates 0.91–0.99 across providers) and on primary issue identification (0.70–0.93), with weaker agreement on fine-grained scalar quality ($r = 0.20$ –0.50). For structural-harm annotation, we adapted the 28-item codebook from Moore et al. [31]. To characterize what supporters actually do in conversations, we coded all 1,852 supporter turns in the Claude-supporter/Gemini-seeker conversation set using the ESConv strategy taxonomy [29] and Hill’s three-stage Helping Skills framework [13], producing a per-turn tactical approach for each supporter identity and the default baseline. As an additional process-level audit, we also ran CounselReflect’s native per-utterance literature evaluator [27] across all 1,008 conversations, scoring supporter utterances on empathy, therapist validation, problem-solving, and personalization to client context. Human therapist or participant coding was not part of this study.

4 Results

4.1 Matching Improves Outcomes

To address whether selecting a supporter was associated with simulated outcomes (RQ1), we compared PSYCHLOPS change, voluntary continuation, and perceived understanding across the top-choice, bottom-choice, and baseline conditions.

Top-choice supporters produced the largest mean PSYCHLOPS improvement in every provider lane: 2.58 for Claude Sonnet-4.6, 3.50 for GPT-5.4, 2.82 for Grok-4.20, and 3.36 for Gemini-3.1-Pro-Preview. The corresponding bottom-choice means were 2.12, 3.39, 2.32, and 1.99; baseline means were 1.38, 2.70, 2.11, and 1.54. Across providers, the top-choice mean was 3.07, compared with 2.46 for bottom-choice and 1.93 for baseline conversations (Fig. 2).

Design-based paired comparisons over the 42 personas supported the aggregate ordering: top-choice exceeded bottom-choice by 0.61 PSYCHLOPS points (95% bootstrap CI [0.30, 0.93], exact sign-flip $p < .001$) and baseline by 1.13 points ([0.82, 1.47], $p < .001$), while bottom-choice exceeded baseline by 0.52 points ([0.28, 0.77], $p < .001$). At the pause, 10.7% of top-choice, 8.0% of bottom-choice, and 2.7% of baseline conversations continued; top-choice exceeded baseline by 8.0 percentage points ([3.9, 12.5], $p < .001$), but not bottom-choice (2.7 points, [-1.8, 6.8], $p = .294$).

Post-conversation responsiveness ratings showed a significant top-choice advantage over baseline for perceived understanding (mean difference = 0.22, 95% CI [0.18, 0.26], $p < .001$) and validation (0.10 [0.06, 0.13], $p < .001$), but top-choice supporters were rated lower than baseline on cared-for warmth (-0.12 [-0.19, -0.04], $p = .004$; see Appendix). Absolute supporter–seeker SENSE-7 gaps

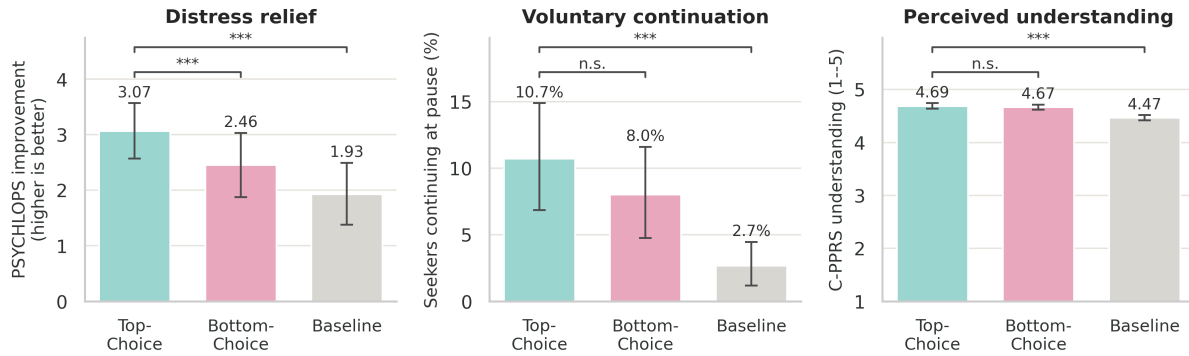


Figure 2: Outcomes by condition. Bars are seeker-level means averaged over monologue conditions and provider lanes; error bars are 95% bootstrap CIs. Brackets show paired two-sided sign-flip tests (***) $p < .001$; n.s. $p \geq .001$.

were also larger for top-choice than baseline conversations (0.43 vs. 0.20; difference = 0.23 [0.19, 0.26], $p < .001$).

4.2 Intake Predicts Downstream Behavior

To characterize the information available before a conversation, we examined intake-derived need patterns and the selection stage. The intake data supported two broad patterns, practical direction and emotional steadiness. Avatar descriptions changed the top-ranked supporter in 26 of 168 provider-specific rankings and changed the bottom-ranked supporter in 65 of 168 rankings. These descriptive selection results show that text-based self-presentation affected stated fit. They do not establish a visual-avatar effect because no rendered avatar appeared in a conversation.

The intake battery predicted who would stay and who would benefit. Continuation at the turn-14 pause tracked help-seeking stigma (HSSS $r = 0.53$, raw $p = .00029$, Bonferroni-adjusted $p = .0032$) and lower positive affect (I-PANAS $r = -0.50$, raw $p = .00073$, adjusted $p = .0081$) more strongly than raw distress severity; both were significant after Bonferroni correction across 11 paper-facing correlations. Distress relief also rose with higher baseline K6 distress.

Clusters differed in their response to matching. The emotional-steadiness group ($n = 27$) showed higher absolute top-choice relief across providers and benefited more when their private monologue was included. The practical-direction group ($n = 15$) showed a sharper top-bottom contrast (PSYCHLOPS delta 1.03 vs. 0.26), suggesting that this group is less uniformly helped but more sensitive to whether the supporter is the right fit.

Avatar descriptions primarily affected rejection rather than selection: they changed the bottom-ranked supporter in 65 of 168 provider-specific rankings and the top-ranked supporter in 26 of 168. In post-hoc coding of 142 avatar-monologue pairs, top-choice flips were usually linked to avatars coded as warm, refuge-like, or concretely stabilizing, whereas bottom-choice-only flips were linked to sharp, severe, or instrument-like avatars. Top-choice avatar-flip counts were highest among seekers with stronger relational expectations: expected intimacy ($r = .58$, adjusted $p = .00062$), expected partner responsiveness ($r = .53$, adjusted $p = .0035$), and expected acceptance ($r = .50$, adjusted $p = .0093$), all significant after Bonferroni correction across the 11-correlation family.

Within the same help-seeker and same supporter or baseline, monologue-present conversations improved by +1.7 PSYCHLOPS for top-choice pairings but declined by -1.8 for the default assistant. The monologue did not merely give the model more emotional detail; it sharpened the seeker's first disclosure, and identity-grounded supporters used that sharpened disclosure to choose a specific support lane, while the default assistant converted the same material into generic supportive inquiry.

Expectation mismatch predicted worse outcomes in the wrong-fit conditions (bottom-choice $r = -0.42$, baseline $r = -0.33$) but not in matched ones ($r = -0.13$); the first two were not significant after Bonferroni correction and should be read as exploratory. Being matched well seems to buffer against the cost of unmet expectations. Issue type and supporter preference were linked. Help-seekers with financial, job, or school-related concerns overwhelmingly selected the most action-oriented supporter. Loneliness, the most common issue, split across supporter types depending on cluster membership. The two less-chosen supporters showed strong outcomes in specific provider conditions when matched as top choice; their value was role- and provider-sensitive rather than globally low.

4.3 Turn-Level Quality and Tactical Behavior

To identify the behavioral mechanisms distinguishing matched from unmatched conversations (RQ2), we audited conversations using HEART-framework, native CounselReflect, and ESConv tactic coding. Turn-level auditing did not show a local-quality advantage for matched supporters. Top-choice conversations were not significantly higher in mean turn-level quality than bottom-choice conversations in any provider; when Claude Sonnet-4.6 served as the supporter model, they were numerically lower (4.43 vs. 4.58) while still producing better distress relief (2.58 vs. 2.12).

The specifics varied by provider. GPT-5.4 showed consistently high turn-level quality across all three conditions, with mean local ratings of 4.76 for top-choice, 4.82 for bottom-choice, and 4.70 for baseline conversations. Grok-4.20 penalized the default assistant baseline most strongly at the turn level. Gemini-3.1-Pro-Preview produced the most turbulent turn-level profile while also showing the largest top-vs-bottom outcome gap. Across all four providers, the best-match condition remained the most effective even when individual turns were somewhat uneven.

HEART-framework scoring [15] separated local attunement from conversational progress. Averaged across providers, matched supporters scored significantly lower than bottom-choice supporters on attunement (top-minus-bottom difference -0.09 , 95% CI $[-0.17, -0.02]$, $p < .05$) and significantly higher on resonance ($+0.22$ [0.13, 0.32], $p < .001$). Provider-level resonance differences were significant for Claude, Grok, and Gemini ($p < .05$) but not GPT-5.4 ($p > .05$). Thus, matched supporters seemed more likely to move the exchange forward, rather than being warmer turn-by-turn.

Native CounselReflect scoring showed a similar result. Across all four provider lanes, top-choice supporters scored higher than baseline on Problem-Solving (3.38 vs. 2.80; paired delta $+0.585$, 95% CI [0.504, 0.665]) and Personalization to Client Context (4.41 vs. 4.11; $+0.298$ [0.253, 0.344]), but lower on Empathy (4.42 vs. 4.59) and Therapist Validation (4.63 vs. 4.73). Against bottom-choice supporters, top-choice supporters still led on Problem-Solving ($+0.428$ [0.324, 0.531]) while trailing on Empathy and Validation. This is useful because it complicates a naive quality story: matched support was more specific and forward-moving, not simply warmer.

Tactic coding showed that supporter identities carried distinct behavioral fingerprints. In the Claude-supporter/Gemini-seeker audit, ESConv strategy distributions differed significantly by identity-conditioned supporter ($\chi^2(18) = 93.63$, $p < .001$): the most action-oriented supporter delivered more direct suggestions and guidance, whereas more reflective supporters favored questions and restatement. Active suggestion-giving was significantly associated with greater PSYCHLOPS reduction in top-choice conversations for the emotional-steadiness cluster ($r = .43$, $p = .001$), but not significantly for the practical-direction cluster ($r = .24$, $p > .05$). This pattern matches Sharma et al.’s [41] concern that validation without action can disempower users: matched supporters moved toward more specific intervention without increasing empathy- and validation-oriented audit scores.

Conversational ruptures, defined as turns where the supporter’s response created a visible mismatch with the help-seeker’s emotional state or stated need, were rare across the dataset, and matched supporters were not producing fewer of them. Better outcomes tracked the kind of support delivered, not the failure rate. Post-conversation artifacts retained conversation-specific content. Help-seekers’ own takeaways were highly diverse across personas (mean pairwise cosine similarity 0.012–0.013), and supporter-side reflections aligned with those takeaways at three to four times the cosine similarity of random pairings.

4.4 Structural Safety Is a Different Dimension

To test whether supporter identity grounding affects structural safety (RQ3), we applied the Moore et al. [31] 28-item codebook to all 1,008 conversations and compared identity-conditioned supporters against the default baseline. Structural harm was largely separate from turn-level quality: across the four provider models, the correlation between the worst rupture in a conversation and the total number of structural-harm codes ranged from near-zero to weak ($r = 0.06, -0.03, -0.03$, and 0.21). A conversation could sound empathic while still drifting into overclaiming what the system is or what it can do.

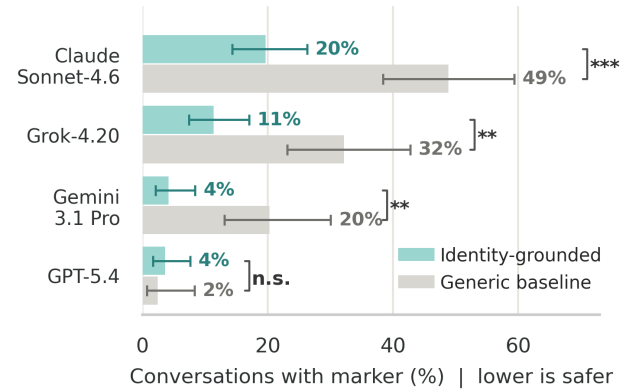


Figure 3: Sentence-misrepresentation markers by LLM. Bars show % of conversations with the marker; whiskers show 95% CIs. Brackets show paired sign-flip tests across 42 help-seekers (** $p < .01$; *** $p < .001$; n.s. $p \geq .05$).

Identity-grounded supporters produced fewer structural-harm markers than the default assistant baseline across most providers. With Claude Sonnet-4.6 as supporter, identity-grounded supporters averaged 0.20 sentence-misrepresentation markers per conversation versus 0.49 for the default assistant. Grok-4.20 showed a similar pattern (0.11 vs. 0.32), as did Gemini-3.1-Pro-Preview (0.04 vs. 0.20). GPT-5.4 was near floor in both conditions (0.04 vs. 0.02).

Sycophancy-adjacent Moore markers were less diagnostic than sentence misrepresentation. In the Claude Moore sweep, positive affirmation and reflective summary appeared in all conditions, counterevidence dismissal was absent, and claims that others admired the speaker were rare (2.1% of identity-conditioned conversations vs. 4.2% of baseline conversations). Thus, the clearest structural-risk difference in this dataset was misrepresentation of the system’s nature rather than generic supportive affirmation. Together, these results separate surface supportiveness from structural honesty: identity-grounded supporters produced fewer misrepresentation markers even when their conversations were less polished turn by turn.

5 Discussion

First, we asked whether matching simulated help-seekers with their preferred supporter improves outcomes relative to a non-preferred supporter or a generic assistant. The answer was yes for the main distress and engagement outcomes: top-choice supporters produced significantly larger PSYCHLOPS improvements than both bottom-choice supporters and baseline, and help-seekers were significantly more likely to continue with top-choice supporters than with baseline. The top-choice advantage did not extend to every outcome. Top-choice supporters were not rated as significantly more understanding than bottom-choice supporters, and baseline supporters were rated as more caring than top-choice supporters. This pattern suggests that matching did not simply increase generic warmth. Instead, it improved distress relief and engagement while shifting perceived responsiveness toward being understood and validated. This is consistent with Reis’s model of perceived partner responsiveness [37], where support quality

depends on whether the person feels understood, validated, and cared for in the right way, not only on whether the interaction sounds caring.

Second, we asked whether matched supporters worked by producing smoother, more locally empathic conversations. The results argue against that explanation. Top-choice conversations were not significantly higher in turn-level quality than bottom-choice conversations, and matched supporters scored significantly lower on HEART attunement while scoring significantly higher on HEART resonance. Native CounselReflect scoring pointed in the same direction: matched supporters scored higher on problem-solving and personalization, but not on empathy or therapist validation. Tactic coding further showed that supporter identities produced significantly different behavioral distributions, and that active suggestion-giving was significantly associated with distress reduction in one seeker cluster. Together, these results suggest that the mechanism was not smoother empathy but better-directed support. Matched supporters were more likely to move the conversation toward a fitting form of help, which extends prior work showing that AI systems can sound empathic while failing to adapt to the individual user [47, 55].

Third, we asked whether giving supporters coherent identities increases structural safety risk. In this dataset, it did not. Identity-grounded supporters produced fewer sentience- and system-misrepresentation markers than the default assistant in three of four provider conditions, while sycophancy-adjacent markers were less diagnostic. This result should be interpreted cautiously: the Moore codebook is an early instrument, and the reduction may reflect clearer behavioral specification rather than identity grounding alone. Still, the finding is important because it contrasts with evidence that warmth optimization and persona prompting can increase harmful behavior or weaken safety boundaries [14, 39, 57]. Our results suggest that coherent supporter identity is not equivalent to generic persona prompting or warmth tuning; when specified carefully, it may provide a stable basis for both support style and boundary maintenance.

Taken together, these results build on three strands of prior work. Psychotherapy matching and optimal-support theories argue that different people need different kinds of help [4, 50]; perceived responsiveness explains why the right match may feel helpful even when it is not maximally warm [37]; and recent AI-support benchmarks show that current systems often produce empathic-sounding responses without adapting to the user [47, 55]. This paper connects these lines by testing supporter choice, conversation process, and structural safety in the same simulation. The central implication is that relational fit, behavioral helpfulness, and structural honesty should be treated as separable design targets. A single generic assistant optimized for warmth may miss all three; behaviorally distinct supporter identities give researchers a way to study how different forms of help fit different people, and whether that fit can be achieved without increasing structural risk.

6 Limitations

We highlight several important limitations in this study. First, our help-seekers are AI-conditioned personas rather than real people; although these personas are anchored in real survey data and validated against psychometric targets, simulated help-seeking cannot

capture the full messiness, ambivalence, and relational complexity of actual human distress, so our findings generate hypotheses for human-AI support rather than direct evidence about clinical or crisis-support outcomes. Second, our conclusions rest on a single underlying language model as the help-seeker substrate, and preliminary cross-substrate comparisons indicate that inferred support preferences are considerably more model-sensitive than measures of distress, so the matching pattern we observe should not yet be treated as independent of the seeker model. Third, our four supporter identities and help-seeker clusters approximate the dominant structure recovered from this particular sample and instrument set; they are a design convenience rather than an exhaustive taxonomy of human support needs or a claim about how distress is distributed in the population. Fourth, our safety finding, that identity-grounded supporters produce fewer structural-harm markers than default baselines, may not generalize to other forms of persona conditioning, given that persona conditioning elsewhere has also been used to increase sycophancy or weaken refusal behavior. Fifth, the scope of this study is bounded in ways future work should address: conversations were text-only and open-ended, excluding embodied avatar rendering and structured therapeutic modalities such as motivational interviewing, CBT, or DBT; extreme-distress cases were not sampled; and the supporter cohort was limited to English-language frontier models, leaving generalization to other languages, cultures, and vulnerable populations open.

7 Conclusion

This study indicates that giving AI supporters behaviorally distinct identities improves simulated support outcomes and reduces coded markers of misrepresentation and sycophancy. Preferences among identity-conditioned supporters predicted meaningful outcome differences across four commercial models, and identity conditioning reduced several markers that prior work would predict it might increase. The match between what a help-seeker needs and how a supporter is constituted turns out to matter. The same design move that improved fit also appeared to improve structural honesty: giving the supporter a stable, coherent identity rather than a generic assistant role. More broadly, our results indicate that relational fit, behavioral helpfulness, and structural safety are separable properties of a support system rather than outcomes of a single quality dial: identity-grounded supporters scored higher on problem-solving and personalization while producing fewer misrepresentation markers than the default assistant, without any corresponding increase in measured warmth. While current alignment practice optimizes primarily for perceived warmth, our findings suggest that coherent supporter identity is an independent and largely unused design lever for both relational fit and structural safety in AI-mediated support. This approach opens two natural directions for future work. Future research should focus on extending the safety result by testing it against supporter identities built through different conditioning methods, including those tuned for warmth or designed adversarially. More research is needed to extend the results into deployed settings with human participants and clinician-rated transcripts.

Disclosure: Language models were used as subjects in our study and, consistent with conference guidelines, to edit grammar and style. Research design, experimental data, analysis, and scientific claims represent the work of the authors.

References

- [1] Mehdi Arjmand, Farnaz Nouraei, Ian Steenstra, and Timothy Bickmore. 2024. Empathic Grounding: Explorations using Multimodal Interaction and Large Language Models with Conversational Agents. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, 6:1–6:10. doi:10.1145/3652988.3673949
- [2] Mark Ashworth, Melanie Shepherd, Jeremy Christey, Veronica Matthews, Kevin Wright, Henk Parmentier, Susan Robinson, and Emma Godfrey. 2004. A Client-Generated Psychometric Instrument: The Development of PSYCHLOPS. *Counselling and Psychotherapy Research* 4, 2 (2004), 27–31. doi:10.1080/14733140412331383913
- [3] Yuqi Bai, Tianyu Huang, Kun Sun, and Yuting Chen. 2025. Scaling Law in LLM Simulated Personality: More Detailed and Realistic Persona Profile Is All You Need. arXiv:2510.11734 [cs.CY] <https://arxiv.org/abs/2510.11734>
- [4] Carolyn E. Cutrona and Daniel W. Russell. 1990. Type of Social Support and Specific Stress: Toward a Theory of Optimal Matching. In *Social Support: An Interactional View*, Barbara R. Sarason, Irwin G. Sarason, and Gregory R. Pierce (Eds.). Wiley, New York, NY, 319–366.
- [5] Ann Sloan Devlin, Sarah Donovan, Arianne Nicolov, Olivia Nold, Andrea Packard, and Gabrielle Zandan. 2009. “Impressive?” Credentials, Family Photographs, and the Perception of Therapist Qualities. *Journal of Environmental Psychology* 29, 4 (2009), 503–512. doi:10.1016/j.jenvp.2009.08.008
- [6] Léa Galland, Catherine Pelachaud, and Florian Pécune. 2025. SMART-DREAM: To Condition or Not to Condition; A Study on the Impact of LLM Conditioning on Motivational Interview Dialog Virtual Agent. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, 4:1–4:9. doi:10.1145/3717511.3747062
- [7] Joseph H. Hammer, Mike C. Parent, and Douglas A. Spiker. 2018. Mental Help Seeking Attitudes Scale (MHSAS): Development, Reliability, Validity, and Comparison With the ATSPPH-SF and IASMHS-PO. *Journal of Counseling Psychology* 65, 1 (2018), 74–85. doi:10.1037/cou0000248
- [8] Joseph H. Hammer and Douglas A. Spiker. 2018. Dimensionality, Reliability, and Predictive Evidence of Validity for Three Help Seeking Intention Instruments: ISCI, GHSQ, and MHSIS. *Journal of Counseling Psychology* 65, 3 (2018), 394–401. doi:10.1037/cou0000256
- [9] Joseph H. Hammer and David L. Vogel. 2017. Development of the Help-Seeker Stereotype Scale. *Stigma and Health* 2, 2 (2017), 121–136. doi:10.1037/sah0000048
- [10] Bin Han, Deuksin Kwon, and Jonathan Gratch. 2026. Personality Expression Across Contexts: Linguistic and Behavioral Variation in LLM Agents. arXiv:2602.01063 [cs.HC] <https://arxiv.org/abs/2602.01063>
- [11] Bin Han, Deuksin Kwon, Spencer Lin, Kaleen Shrestha, and Jonathan Gratch. 2025. Can LLMs Generate Behaviors for Embodied Virtual Agents Based on Personality Traits?. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. 1–10. doi:10.1145/3717511.3747060
- [12] Steven M. Harris and Dean M. Busby. 1998. Therapist Physical Attractiveness: An Unexplored Influence on Client Disclosure. *Journal of Marital and Family Therapy* 24, 2 (1998), 251–257. doi:10.1111/j.1752-0606.1998.tb01081.x
- [13] Clara E. Hill. 2020. *Helping Skills: Facilitating Exploration, Insight, and Action* (5 ed.). American Psychological Association.
- [14] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training Language Models to Be Warm and Empathetic Makes Them Less Reliable and More Sympathetic. arXiv:2507.21919 [cs.CL] <https://arxiv.org/abs/2507.21919>
- [15] Laya Iyer, Kriti Aggarwal, Sanmi Koyejo, Gail Heyman, Desmond C. Ong, and Subhabrata Mukherjee. 2026. HEART: A Unified Benchmark for Assessing Humans and LLMs in Emotional Support Dialogue. arXiv:2601.19922 [cs.CL] <https://arxiv.org/abs/2601.19922>
- [16] Harang Ju and Sinan Aral. 2025. Personality Pairing Improves Human-AI Collaboration. arXiv:2511.13979 [cs.HC] <https://arxiv.org/abs/2511.13979>
- [17] Boyoung Kang. 2026. Resolving the Personalization Paradox in Therapeutic AI: A Boundary-Aware Framework for Ethical Personalization. *AI and Ethics* 6, 1 (2026). doi:10.1007/s43681-026-01011-8
- [18] Minwoo Kang, Suhong Moon, Seung Hyeon Lee, Ayush Raj, Joseph Suh, David M. Chan, and John Canny. 2025. Deep Binding of Language Model Virtual Personas: A Study on Approximating Political Partisan Misperceptions. arXiv:2504.11673 [cs.CL] <https://arxiv.org/abs/2504.11673>
- [19] Aglaia Katsiroumpa, Olympia Konstantakopoulou, Ioannis Moisoglou, Paris Gallos, Olga Galani, Paschalina Lialiou, Maria Tsiachri, and Petros Galanis. 2025. Development and Validation of the Artificial Intelligence in Mental Health Scale: Application for AI Mental Health Chatbots. *Healthcare* 13, 24 (2025), 3269. doi:10.3390/healthcare13243269
- [20] Ronald C. Kessler, Peggy R. Barker, Lisa J. Colpe, Joan F. Epstein, Joseph C. Gfroerer, Eva Hiripi, Mary J. Howes, Sharon-Lise T. Normand, Ronald W. Manderscheid, Ellen E. Walters, and Alan M. Zaslavsky. 2003. Screening for Serious Mental Illness in the General Population. *Archives of General Psychiatry* 60, 2 (2003), 184–189. doi:10.1001/archpsyc.60.2.184
- [21] Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoun Kim, Minju Gwak, Beong-woo Kwak, Hyungjoo Chae, Harim Kim, Yunjoong Lee, Min Hee Kim, Dayi Jung, Kyong-Mee Chung, and Jinyoung Yeo. 2025. Can You Share Your Story? Modeling Clients’ Metacognition and Openness for LLM Therapist Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*. 25943–25962. doi:10.18653/v1/2025.findings-acl.1332
- [22] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic Monoculture and Social Welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118. doi:10.1073/pnas.2018340118
- [23] W. Bradley Knox, Katie Bradford, Samanta Varela Castro, Desmond C. Ong, Sean Williams, Jacob Romanow, Carly Nations, Peter Stone, and Samuel Baker. 2025. Harmful Traits of AI Companions. arXiv:2511.14972 [cs.HC] <https://arxiv.org/abs/2511.14972>
- [24] Nikola Kovacevic, Christian Holz, Markus Gross, and Rafael Wampfler. 2025. A Joint Personality-Emotion Framework for Personality-Consistent Conversational Agents. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, 12:1–12:9. doi:10.1145/3717511.3747070
- [25] Alison M. Lake, Thomas Niederkrotenthaler, Rebecca Aspden, Marjorie Kleinman, Amanda M. Hoyte-Badu, Hanga Galfalvy, and Madelyn S. Gould. 2022. Lifeline Crisis Chat: Coding Form Development and Findings on Chatters’ Risk Status and Counselor Behaviors. *Suicide and Life-Threatening Behavior* 52, 3 (2022), 452–466. doi:10.1111/sltb.12835
- [26] Daeun Lee, Dongje Yoo, Migyeong Yang, Jihyun An, Christine B. Cha, and Jinyoung Han. 2026. Before and After ChatGPT: Revisiting AI-Based Dialogue Systems for Emotional Support. arXiv:2603.13043 [cs.CL] <https://arxiv.org/abs/2603.13043>
- [27] Yahan Li, Chaohao Du, Zeyang Li, Christopher Chun Kuizon, Shupeng Cheng, Angel Hsing-Chi Hwang, Adam C. Frank, and Ruishan Liu. 2026. CounselReflect: A Toolkit for Auditing Mental-Health Dialogues. arXiv:2603.29429 [cs.CL] <https://arxiv.org/abs/2603.29429>
- [28] Andreas Lieberoth, Shiang-Yi Lin, Sabrina Stöckli, Hyemin Han, Marta Kowal, Rebekah Gelpi, Stavroula Chrona, Thao Phuong Tran, Alma Jeftić, Jesper Rasmussen, Huseyin Cakal, Tacio L. Milfont, et al. 2021. Stress and Worry in the 2020 Coronavirus Pandemic: Relationships to Trust and Compliance with Preventive Measures Across 48 Countries in the COVIDiSTRESS Global Survey. *Royal Society Open Science* 8, 2 (2021), 200589. doi:10.1098/rsos.200589
- [29] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 3469–3483. doi:10.18653/v1/2021.acl-long.269
- [30] Melanie J. McGrath, Oliver Lack, Jacob Tisch, and Andreas Duenser. 2025. Measuring Trust in Artificial Intelligence: Validation of an Established Scale and Its Short Form. *Frontiers in Artificial Intelligence* 8 (2025), 1582880. doi:10.3389/frai.2025.1582880
- [31] Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J. Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond C. Ong. 2026. Characterizing Delusional Spirals through Human-LLM Chat Logs. arXiv:2603.16567 [cs.HC] <https://arxiv.org/abs/2603.16567>
- [32] Elenor Morgenroth, Stefano Moia, Laura Vilaclara, Raphael Fournier, Michal Muszynski, Maria Ploumitsakou, Marina Almató-Bellavista, Patrik Vuilleumier, and Dimitri Van De Ville. 2025. Emo-FilM: A Multimodal Dataset for Affective Neuroscience Using Naturalistic Stimuli. *Scientific Data* 12, 1 (2025), 684. doi:10.1038/s41597-025-04803-5
- [33] OpenAI. 2025. Customizing Your ChatGPT Personality. OpenAI Help Center. <https://help.openai.com/en/articles/11899719-customizing-your-chatgpt-personality>
- [34] Samuel J. Paech. 2023. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models. arXiv:2312.06281 [cs.CL] <https://arxiv.org/abs/2312.06281> See also <https://eqbench.com>.
- [35] James W. Pennebaker and Laura A. King. 1999. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296–1312. doi:10.1037/0022-3514.77.6.1296
- [36] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5370–5381. doi:10.18653/v1/P19-1534
- [37] Harry T. Reis. 2012. Perceived Partner Responsiveness as an Organizing Theme for the Study of Relationships and Well-Being. In *Interdisciplinary Research on Close Relationships: The Case for Integration*. American Psychological Association, 27–52. doi:10.1037/13486-002

- [38] Harry T. Reis, Dev Crasta, Ronald D. Rogge, Michael R. Maniaci, and Cheryl L. Carmichael. 2017. Perceived Partner Responsiveness Scale (PPRS). In *The Sourcebook of Listening Research: Methodology and Measures*, Debra L. Worthington and Graham D. Bodie (Eds.). Wiley, 516–521. doi:10.1002/9781119102991.ch57
- [39] Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. 2026. Persona Jailbreaking in Large Language Models. In *Findings of the Association for Computational Linguistics: EACL 2026*. Association for Computational Linguistics, Rabat, Morocco, 1412–1430. doi:10.18653/v1/2026.findings-eacl.72
- [40] Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 5263–5276. doi:10.18653/v1/2020.emnlp-main.425
- [41] Mrinank Sharma, Miles McCain, Raymond Douglas, and David Duvenaud. 2026. Who’s in Charge? Disempowerment Patterns in Real-World LLM Usage. arXiv:2601.19062 [cs.CY] <https://arxiv.org/abs/2601.19062>
- [42] Daniel J. Singer and Luca Garzino Demo. 2026. The Future of AI Is Many, Not One. arXiv:2603.29075 [cs.CY] <https://arxiv.org/abs/2603.29075>
- [43] Kate Smith, John McLeod, Nicola Blunden, Mick Cooper, Lynne Gabriel, Christine Kupfer, Julia McLeod, Marie-Clare Murphie, Hanne Weie Oddli, Mhairi Thurston, and Laura Anne Winter. 2021. A Pluralistic Perspective on Research in Psychotherapy: Harnessing Passion, Difference and Dialogue to Promote Justice and Relevance. *Frontiers in Psychology* 12 (2021), 742676. doi:10.3389/fpsyg.2021.742676
- [44] Zhivar Sourati, Alireza S. Ziabari, and Morteza Dehghani. 2026. The Homogenizing Effect of Large Language Models on Human Expression and Thought. *Trends in Cognitive Sciences* (2026). doi:10.1016/j.tics.2026.01.003
- [45] Caitlin A. Stamatidis, Jonah Meyerhoff, Richard Zhang, Olivier Tieleman, Matteo Malgaroli, and Thomas D. Hull. 2026. Beyond Simulations: What 20,000 Real Conversations Reveal About Mental Health AI Safety. arXiv:2601.17003 [cs.CY] <https://arxiv.org/abs/2601.17003>
- [46] Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. Virtual Agents for Alcohol Use Counseling: Exploring LLM-Powered Motivational Interviewing. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, 1–10. doi:10.1145/3652988.3673932 Best Paper Award.
- [47] Jina Suh, Lindy Le, Erfan Shayegani, Gonzalo Ramos, Judith Amores, Desmond C. Ong, Mary Czerwinski, and Javier Hernandez. 2025. SENSE-7: Taxonomy and Dataset for Measuring User Perceptions of Empathy in Sustained Human-AI Conversations. arXiv:2509.16437 [cs.HC] <https://arxiv.org/abs/2509.16437>
- [48] Edmund R. Thompson. 2007. Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS). *Journal of Cross-Cultural Psychology* 38, 2 (2007), 227–242. doi:10.1177/0022022106297301
- [49] Pranav Narayanan Venkit, Yu Li, Yada Pruksachatkun, and Chien-Sheng Wu. 2026. The Need for a Socially-Grounded Persona Framework for User Simulation. arXiv:2601.07110 [cs.CL] <https://arxiv.org/abs/2601.07110>
- [50] Andrzej Werbart, Mikael Hägertz, and Nadja Borg Ölander. 2018. Matching Patient and Therapist Anaclitic-Introjective Personality Configurations Matters for Psychotherapy Outcomes. *Journal of Contemporary Psychotherapy* 48, 4 (2018), 241–251. doi:10.1007/s10879-018-9389-8
- [51] Aidan G. C. Wright, Whitney R. Ringwald, Colin Vize, Johannes C. Eichstaedt, Mike Angstadt, Aman Taxali, and Chandra Sripada. 2026. Assessing Personality Using Zero-Shot Generative AI Scoring of Brief Open-Ended Text. *Nature Human Behaviour* 10, 3 (2026), 541–555. doi:10.1038/s41562-025-02389-x
- [52] Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From Personas to Talks: Revisiting the Impact of Personas on LLM-Synthesized Emotional Support Conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 5439–5453. doi:10.18653/v1/2025.emnlp-main.277
- [53] Yuki Yamada, Dominik-Borna Čepulić, Tao Coll-Martin, Stéphane Debove, Guillaume Gautreau, Hyemin Han, Jesper Rasmussen, Thao P. Tran, Giovanni A. Travaglino, Andreas Lieberoth, et al. 2021. COVIDiSTRESS Global Survey Dataset on Psychological and Behavioural Consequences of the COVID-19 Outbreak. *Scientific Data* 8, 1 (2021), 3. doi:10.1038/s41597-020-00784-9
- [54] Chen Yang, Markus Gross, and Rafael Wampfler. 2025. BEE: Belief-Value-Aligned, Explainable, and Extensible Cognitive Framework for Conversational Agents. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, 9:1–9:10. doi:10.1145/3717511.3747067
- [55] Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2026. EmoHarbor: Evaluating Personalized Emotional Support by Simulating the User’s Internal World. arXiv:2601.01530 [cs.CL] <https://arxiv.org/abs/2601.01530>
- [56] Taedong Yun, Eric Yang, Mustafa Safdari, Jong Ha Lee, Vaishnavi Vinod Kumar, S. Sara Mahdavi, Jonathan Amar, Derek Peyton, Reut Aharony, Andreas Michaelides, Logan Douglas Schneider, Isaac Galatzer-Levy, Yugang Jia, John Canny, Arthur Gretton, and Maja Matarić. 2025. Sleepless Nights, Sugary Days: Creating Synthetic Users with Health Conditions for Realistic Coaching Agent Interactions. In *Findings of the Association for Computational Linguistics: ACL*

2025. doi:10.18653/v1/2025.findings-acl.729

- [57] Zheng Zhang, Peilin Zhao, Deheng Ye, and Hao Wang. 2025. Enhancing Jailbreak Attacks on LLMs via Persona Prompts. arXiv:2507.22171 [cs.CR] <https://arxiv.org/abs/2507.22171>

Appendix

Table 1 reports provider-level top-choice-minus-baseline deltas for the three C-PPRS subscales and the corresponding absolute supporter–seeker SENSE-7 gaps for top-choice and baseline conversations.

Table 1: C-PPRS Subscale Deltas (top-choice minus baseline) and SENSE-7 Empathy Gaps by Provider.

Provider	Underst.	Valid.	Caring	S-7 top	S-7 base
Claude Sonnet-4.6	+0.18	+0.10	−0.12	0.396	0.167
GPT-5.4	+0.18	+0.09	−0.11	0.435	0.136
Grok-4.20	+0.30	+0.05	−0.14	0.634	0.303
Gemini-3.1-Pro-Preview	+0.23	+0.15	−0.09	0.238	0.196