

The AI Therapist: ChatGPT Therapy Users at the Boundary of Human Therapy, Disillusionment, and AI Companionship on Reddit

Nari Yoo

School of Social Work
University of Michigan
nariyoo@umich.edu

Pengyu Chen

School of Information
University of Michigan
pyc@umich.edu

Abstract

People in distress increasingly talk to a consumer chatbot the way they would talk to a therapist: they vent to the model, work through feelings with it, and sometimes call it their therapist. We ask where that behavior is located in relation to the existing system of human care: which communities the people doing it belong to, what discourse crosses the boundary between those communities and human therapy, and in what order people arrive on the two sides. Across ten Reddit communities covering general chatbots, companion chatbots, an AI self-help community, mainstream human therapy, and therapy disillusionment, we combine the backbone of a user-by-community bipartite projection under the fixed degree sequence model, a directed reply network, a structural topic model, and an analysis of which side users reach first. Of the ten, the AI self-help community is the only one connecting the human-therapy and chatbot worlds, and its strongest backbone edge is to the therapy-disillusioned, at nine times the null expectation. Companion chatbots are as relational as the self-help community and their overlap with therapy grievance is still at chance, so relational chatbot use is not by itself adjacent to it; the adjacency is specific to purpose-built self-help. Visibility in the cross-community discourse is concentrated in one commercial account, and human-therapy participation predominantly precedes therapeutic chatbot use, though the data cannot establish that disillusionment drives the move. We interpret these patterns through the Network Episode Model.

1 Introduction

People in distress increasingly talk to a chatbot the way they would talk to a therapist (Luo et al., 2025; Rousmaniere et al., 2026). They open ChatGPT to vent, ask it to help them work through a breakup, describe it as the only thing that listens, and, in some cases, say it has replaced a human

therapist. This relational, conversational use of a general-purpose model as a substitute or complement for talk therapy is distinct from instrumental uses such as looking up a diagnosis or drafting a form (Jung et al., 2025). We refer to it throughout as therapeutic chatbot use.

How does this behavior relate to online mental health communities? Prior computational work has examined chatbot therapy inside a single corpus (Collins et al., 2025; Jung et al., 2025) and has separately characterized therapy dissatisfaction at scale (Steinbrenner et al., 2025), but no study positions therapeutic chatbot use relative to the human-therapy communities it borders, traces the discourse that crosses that border, or asks in what order people arrive. The answer matters for two debates: a growing science of AI mental health tools that has largely not characterized who its users are in the care system (Stade et al., 2024; Heinz et al., 2025), and a documented tendency for users to position the chatbot as a substitute for a human therapist (Luo et al., 2025; Siddals et al., 2024) whose driving population has not been identified. The answer also has consequences beyond research: for clinicians, whether therapeutic chatbot use signals disengagement from care; for platform designers, the commercial capture we document in spaces that present as peer support; for builders of AI mental health tools, a real-world population that differs systematically from trial samples. Methodologically, the fixed degree sequence model measures the social footprint of consumer AI at population scale without a user-level label.

RQ1 (structure): what is the position of the AI self-help community (r /therapyGPT) in the cross-community user-overlap structure, relative to mainstream therapy and to therapy disillusionment? RQ2 (discourse): in the therapeutic chatbot discourse, who occupies central and bridging positions, and what frames organize the talk? RQ3 (trajectory): among accounts active in both worlds,

does writing about therapeutic chatbot use precede or follow participation in a human-therapy community?

First, we show with the fixed degree sequence model that the AI self-help community is the only one of our ten communities that bridges the human-therapy and chatbot worlds, with over-connection strongest toward therapy disillusionment, at nine times the null expectation, against roughly twice for mainstream therapy and 1.2 times for the general-chatbot communities it grows out of. This ordering survives controls for user activity and community size, removal of commercially flagged accounts, and stricter membership definitions. Second, we characterize the discourse network and its frames, finding that visibility in the cross-community discourse is concentrated in a single commercial account and that the four community groups frame therapeutic chatbot use in different terms. Third, we examine the trajectory and find that human-therapy participation predominantly comes first, and we show why the data cannot support the stronger causal claim that disillusionment drives the move.

2 Background and Related Work

Help-seeking as a career. The Network Episode Model recast help-seeking as a socially organized career in which people cycle through illness episodes and consult a shifting set of advisors drawn from their personal networks, so that the operative question is which ties a person activates at a given moment and with what orientation toward formal care (Pescosolido, 1992; Perry and Pescosolido, 2015). A conversational chatbot fits this account as a newly available node: an advisor that can be activated at any hour, has no social cost, and can enter the career at the points where existing ties have failed. We observe public communities, not personal networks or tie activation; these are closer to lay referral structures (Freidson, 1960). The model licenses two observable implications: the order in which the chatbot is inserted into the help-seeking sequence (RQ3), and whether the venue organizing therapeutic chatbot use recruits in a patterned way from particular locations in the care system (RQ1).

Relational harm and epistemic injustice. A substantial literature establishes that psychological treatment carries genuine iatrogenic risk and that negative effects are common rather than ex-

ceptional (Lilienfeld, 2007; Barlow, 2010; Strauss et al., 2021). When clients narrate that harm, they describe relationship failures more than technical ones: feeling misunderstood, disempowered, judged, or unsafe with the therapist (Curran et al., 2019; Vybíral et al., 2024). The epistemic injustice literature identifies the same dynamic from the patient’s side, in which people in psychiatric care have their testimony discounted and lack shared concepts to make their experience legible (Crichton et al., 2017). These are the relational goods, being heard, being believed, and being met without judgment, that an always-available and non-evaluative model is positioned to supply, and compensatory-use theory supplies the motivational link from an unmet relational need to a substitute medium (Kardefelt-Winther, 2014). Survey evidence external to Reddit keeps the resulting prediction non-trivial. Among adolescents and younger adults, willingness to use nontraditional services such as apps and online support is high, but willingness is highest for traditional services (Cohen et al., 2021), so a concentration of therapeutic chatbot use next to therapy grievance would not follow from a general preference for nontraditional tools. This grounds a hypothesis that, relative to community size and member activity, the venue organizing therapeutic chatbot use may share more members with therapy grievance than chance would give it. People still in care may add the chatbot as an adjunct between sessions, and ordinary disengagement is common enough (premature termination runs near twenty percent in meta-analytic estimates; Swift and Greenberg, 2012) that arriving at a substitute after human therapy is close to the modal after-care pattern; our analysis of arrival reports only the order.

Computational studies of chatbot therapy. A first wave of empirical work, mostly on Reddit and in interviews, has shown that people use general-purpose models as a form of talk therapy and not only as an information tool. Users deliberately shape ChatGPT into a digital therapist, coach it, externalize distress, and prize it as a non-judgmental, always-available space they sometimes prefer to human clinicians (Luo et al., 2025; Jung et al., 2025; Siddals et al., 2024). Collins et al. (2025) model the co-occurrence of coded themes in ChatGPT therapy posts as a network and find that the belief that ChatGPT works as a therapist is structurally central; Lee et al. (2025) show that after ChatGPT’s release the

framing of AI mental health support diverged between AI-focused and mental health communities; [Aghakhani and Rezapour \(2026\)](#) report that explicit comparisons with human therapy are relatively rare in such narratives, and that users rarely frame AI as a replacement for professional therapy, describing it instead as complementary support when human care is unavailable; survey work confirms that this pattern extends beyond Reddit: among U.S. respondents with an ongoing mental health condition who use general-purpose LLMs for psychological support, 87.2 percent had also received psychotherapy from a human provider, and comparisons between the two were generally neutral to favorable rather than indicative of wholesale substitution ([Rousmaniere et al., 2026](#)); and [Steinbrenner et al. \(2025\)](#) give a fine-grained NLP taxonomy of psychotherapy dissatisfaction. Controlled evidence has matured in parallel, from an AI-in-the-loop tool that raised empathy inside human peer support ([Sharma et al., 2023](#)) to the first randomized trial of a purpose-built generative therapy chatbot reporting clinically meaningful symptom reductions ([Heinz et al., 2025](#)). This maturing evidence base, however, concerns purpose-built or clinician-supervised tools; general-purpose consumer assistants have a longer and still unresolved history of inadequate crisis response, first documented when consumer smartphone agents failed to reliably refer users disclosing suicidal ideation ([Miner et al., 2016](#)) and recently reconfirmed in a therapist-evaluated comparison finding that general-purpose assistants and companion chatbots alike over-rely on directive suggestion and under-perform on risk assessment relative to licensed clinicians ([Scholich et al., 2025](#)). This work establishes the phenomenon but treats each community in isolation, leaving open how the use of AI as therapy relates to the existing system of human care.

AI companionship as a distinct phenomenon.

A separate literature documents AI companionship as relationship development for its own sake. Bonds on platforms such as Replika deepen through reciprocal self-disclosure and accrue affective value independent of any clinical complaint ([Skjuve et al., 2021](#)), with attachment driven by perceived anthropomorphism and authenticity and emotional dependence as a recognized downside ([Pentina et al., 2023](#)). Companion use can reduce loneliness through feeling heard ([De Freitas et al., 2026](#)), yet companionship-oriented use is associ-

ated with lower well-being for users with thin offline networks ([Zhang et al., 2025](#)), and a computational study of Reddit’s companionship community shows intimate human-AI bonds forming unintentionally through routine use ([Pataranutaporn et al., 2025](#)). Users themselves sometimes collapse the roles, treating one app as friend, therapist, and mirror at once ([Maples et al., 2024](#)). Consistent with this, discourse-level analysis finds that companionship-oriented and reassurance-seeking uses of AI more often involve misaligned therapeutic alliance and reported risks such as dependence and symptom escalation, in contrast to task- and goal-aligned uses ([Aghakhani and Rezapour, 2026](#)). Companionship and therapeutic use can therefore look alike because both are relational and emotional. We show that at the level of community structure the boundary nonetheless holds, which is a distinction the prior literature has not drawn structurally.

Cross-community structure, brokerage, and trajectories.

A shared-user projection of subreddits cannot be interpreted directly, because large or highly active communities share many users for reasons unrelated to topical affinity, so structural claims require a null that holds nodal degree fixed ([Maslov and Sneppen, 2002](#); [Zweig and Kaufmann, 2011](#); [Neal, 2014](#)). We use the curveball algorithm ([Strona et al., 2014](#)), whose trade chain provably samples binary matrices with fixed margins uniformly ([Carstens, 2015](#)) and whose fast variants make degree-preserving backbone extraction tractable at scale ([Godard and Neal, 2022](#)); because backbone results shift with the null and the threshold ([Cimini et al., 2022](#)), we report signed z -scores under a Benjamini-Hochberg decision rule. The closest prior work is [Evkoski et al. \(2026\)](#), who build a disorder-association network from Reddit coposting under a bipartite configuration-model null; we share their platform and null family but examine the human-AI therapy boundary instead of ICD categories and add a directed reply network that their design and the broader Reddit mental health literature do not address ([De Choudhury and De, 2014](#); [Waller and Anderson, 2021](#)). For that reply network we engage the online-support brokerage literature, which operationalizes betweenness as bridging social capital and ties brokerage to support provision ([Pan et al., 2020](#); [Lee and Peng, 2025](#)), while following the warning that betweenness is the centrality most prone to false positives

and must be assessed against a degree-preserving null (Farine and Carter, 2022). Community detection uses the Louvain method (Blondel et al., 2008) and discourse frames use the structural topic model (Roberts et al., 2014). Finally, multi-community wandering is the empirical norm on Reddit (Tan and Lee, 2015), which both supplies the mechanism behind our projection edges and licenses treating movement between therapy disillusionment and AI self-help as a trajectory.

3 Data

We collected the full available post and comment history of ten subreddits from publicly available Reddit archives (Pushshift and Arctic Shift dumps), sorted into five groups: general chatbots (r/ChatGPT, r/OpenAI, r/ClaudeAI), companion chatbots (r/Replika, r/CharacterAI), AI self-help (r/therapyGPT), mainstream human therapy (r/therapy, r/talktherapy, r/asktherapist), and disillusionment (r/therapycritical). Table 1 reports the size and coverage of each. We call r/therapy, r/talktherapy and r/asktherapist the human-therapy communities, and mainstream-therapy communities where the contrast with r/therapycritical is at issue, since that community is also about human therapy; the two names denote the same three subreddits. Coverage extends through mid-2026 for all ten communities. After removing deleted and removed authors and the moderation bots, the corpus contains roughly 2.2 million unique accounts. One community’s provenance matters for interpretation. r/therapyGPT was created in spring 2025, its founding posts frame it around disappointment with human therapy, and developers of therapy chatbots, including the commercial account analyzed in Section 5.2, were active from its first weeks. We therefore report the structural results with and without these accounts.

Focal construct. We flag therapeutic chatbot use by keyword match over post and comment text. A narrower keyword list of relational and substitution phrases (“ai therapist,” “as my therapist,” “instead of therapy,” “process my feelings,” “vent to it,” “emotional support”) matches 34,979 general-chatbot users; a broader list that also accepts any emotional or mental health word matches 104,124. We use the narrower list; both lists are in Appendix D. Because the flag is a lexicon, we characterize what it captures by attributing every flagged document to its trigger families (Appendix D). The

Community	Group	Authors	Docs
r/therapy	human	129,136	4,812
r/talktherapy	human	80,825	3,783
r/asktherapist	human	65,281	2,093
r/therapycritical	disillusion	1,245	255
r/therapyGPT	self-help	9,083	4,583
r/OpenAI	general	349,568	4,517
r/ChatGPT	general	1,241,147	45,687
r/ClaudeAI	general	222,277	3,921
r/Replika	companion	62,623	7,268
r/CharacterAI	companion	344,230	15,601

Table 1: Corpus. “Docs” counts documents flagged as therapeutic chatbot discourse (AI mentions in human-therapy communities; therapeutic-use language in chatbot communities). All communities cover posts and comments through mid-2026.

mix differs by community. Substitution and own-therapy phrases dominate in r/therapyGPT (70 percent of flagged documents) while half of the companion communities’ flagged documents fire only on venting or feeling-heard phrases. We also validate the flag against coded documents, model-coded at scale and hand-coded by two authors on a 300-document subsample, with the two coders agreeing at $\kappa = .80$ across the six categories and .98 on the therapeutic-use contrast (Appendix E). It is a good detector of therapeutic chatbot discourse and a weak detector of the writer’s own use. Of flagged documents, 92 percent in r/therapyGPT and 56 percent in the general-chatbot communities are about therapeutic chatbot use by someone, while 44 and 31 percent are the writer’s own use, the remainder being mostly third-party discussion of the same subject. We therefore treat a flagged document as evidence of therapeutic chatbot discourse and say so wherever the distinction matters.

4 Methods

Bipartite projection and backbone (RQ1). We form the user-by-community incidence matrix and extract the backbone of its one-mode projection (Neal, 2014) under the fixed degree sequence model (FDSM), in which a tie is judged against the distribution of projections whose bipartite degree sequences match the observed one. Restricting the matrix to users who belong to two or more communities, we sample that distribution with the curveball algorithm, which holds each user’s number of communities and each community’s number of multi-community members fixed. From 1,000 samples, thinned at six trades per user between draws, we compute, for each of the 45 commu-

nity pairs, the observed co-membership, the null mean and standard deviation, a z score, a ratio of observed to null, and a two-sided empirical p value corrected across pairs with Benjamini and Hochberg at $p < .05$. Conditioning on the multi-membership margins isolates whether two communities share more members than their sizes and their members' breadth would predict. Because z scores are not comparable across pairs of very different size, we take tie strength from the ratio of observed to null and significance from the corrected p . The backbone is recomputed with commercially flagged accounts removed and with membership redefined as at least two contributions (Appendix A).

Reply network (RQ2). We build directed reply edges from a comment's author to the author of its parent, restricted to threads that contain therapeutic chatbot content, and we pool the human-therapy communities with r/therapyGPT, whose comment archives are complete. On the giant component we compute weighted in-degree as a measure of audience, betweenness as a measure of bridging (estimated with 3,000 source pivots under an identical estimator for observed and null values; Appendix B), and Louvain communities, and we profile the highest-centrality accounts. We analyze the general-chatbot reply network separately, because it is a near-disconnected population and because in a community of that size a thread flagged by a single therapeutic comment tends to return the original poster of a popular unrelated thread rather than a therapeutic interlocutor; there we identify central accounts by therapeutic-content volume instead of by reply centrality.

Discourse frames (RQ2). We fit a structural topic model with $K=15$ on 64,658 flagged documents (corpus construction and preprocessing in Appendix C), with topic prevalence modeled as a function of community group and a smooth function of time. Documents enter by different rules on the two sides of the boundary, by construction. Chatbot-community documents are flagged by therapeutic-use language, human-therapy documents by any AI mention, so the human column describes how AI is discussed inside therapy communities, not therapeutic use specifically. We report topic content as the highest-probability and FREX terms and report mean prevalence by group, with all pairwise group contrasts estimated from the fitted model and corrected for multiplicity.

Trajectory (RQ3). For users active in both a human-therapy community and the general-chatbot communities, we compare the timestamp of first human-therapy activity with the timestamp of the first post in which the user writes about chatbot therapy, restricting to the window in which both could occur, that is after the public release of ChatGPT in late 2022. We report the share who arrive at the chatbot first across window anchors and minimum-activity thresholds. We also fit a Poisson rate model for crossing into therapeutic chatbot use, with disillusionment-community membership, account age, and total activity as covariates and the log of person-time at risk as an offset, and we state its identification limits.

Integration of the four analyses. The four analyses answer one question at different levels of the same data. The backbone establishes, at the level of populations, which communities share members beyond what their sizes and their members' breadth predict, and therefore where the boundary between human therapy and consumer chatbots is open at all. The reply network and the topic model then ask what passes through the one opening it identifies: who is heard there, who connects otherwise separate communities, and in whose terms the chatbot is described. The third level adds the order in which the same accounts appear on the two sides. The levels constrain each other's interpretation. An over-connected pair with no shared discourse would indicate co-presence without contact, and a shared frame between communities that share no members would indicate parallel talk with no traffic. We find neither. At the one boundary the null identifies there are shared members, a shared reply network and a consistent order of arrival, while the frames diverge across the four groups, so the traffic is real and the vocabularies on either side of it are not the same.

Unit of observation. Throughout, the unit is participation in a public community and text posted about a chatbot. We never observe a chatbot conversation, a clinical encounter, or a treatment episode. Community membership is evidence about who is present in a discourse; we do not treat it as evidence that a person used a chatbot therapeutically, entered or left treatment, or was dissatisfied with care. Where we report a sequence, it is the order in which accounts became active in communities.

5 Results

5.1 Structural position (RQ1)

The backbone (Figure 1; full 45-pair table in Appendix A) yields two results. First, the boundary between the human-therapy and chatbot worlds is closed everywhere except at one community. Every direct general-chatbot and companion-chatbot pairing with a human-therapy community is significantly under-connected, with z as low as -193 for the largest pair, and the disillusionment community is itself under-connected to the general chatbots. The AI self-help community is the only community with significant positive ties into both worlds, over-connecting to all three mainstream-therapy communities, to the disillusionment community, and to r/ChatGPT and r/ClaudeAI.

Second, that bridge is steeply ordered. The disillusionment tie is by far the strongest, at 77 shared members against a null mean of 8.2, nine times the null expectation ($z=+26.5$, corrected $p<.01$). The tie rests on 77 accounts, so the ratio itself is uncertain. An exact Poisson interval on the observed count places it between 7.4 and 11.7 times the null, which leaves it well clear of the mainstream-therapy ties either way. The mainstream-therapy ties run about twice their null expectation, and the general-chatbot ties, though significant for r/ChatGPT and r/ClaudeAI, are only 1.2 times chance; the overlap with r/OpenAI, whose members discuss the same underlying models, is not distinguishable from chance at all ($z=-1.0$). Per capita, the contrast is the same, with 6.2 percent of the disillusionment community's members appearing in the AI self-help community against 0.5 percent in r/therapy. This ordering is robust. Removing all 133 commercially flagged accounts and bots (four are among the 77 shared members) leaves the disillusionment tie at 73 against a null of 8.0 ($z=+23.9$), and requiring at least two contributions for membership sharpens it to 14 times the null while the mainstream ratios remain near two (Appendix A).

The companion-chatbot communities behave differently from the AI self-help community. r/Replika and r/CharacterAI form their own tight cluster ($z=+104$) and tie to the general chatbots (CharacterAI-ChatGPT $z=+52$), but their co-membership with the disillusionment community is at chance ($z=-1.9$ and -0.9 , not significant) and with mainstream therapy is below chance. So relational, emotional chatbot use is not by itself ad-

jacent to therapy grievance; this ordering is specific to the purpose-built therapy self-help community. The co-membership network in Figure 3 (appendix) shows the same structure spatially. r/therapyGPT lies between the human-therapy and chatbot clusters as the only bridge, while companion chatbots attach to the general-chatbot cluster. Because the null holds user activity and community size fixed, none of these positions can be explained away as an effect of either.

This pattern appears in the simple crossover rates (Table 3). The share of a community's users who also use a general chatbot for therapeutic purposes is near 0.6 percent in mainstream-therapy communities, 3.45 percent in the disillusionment community (43 of 1,245 members; 95 percent CI 2.57 to 4.62), and 12.51 percent in the AI self-help community. Against the pooled mainstream-therapy communities the disillusionment rate ratio is 5.7 (CI 4.2 to 7.7). Two base rates keep these contrasts in proportion: in absolute terms the mainstream communities dominate (r/therapy alone contributes 632 crossers against 43 from the small disillusionment community, 96.6 percent of whose members show no therapeutic chatbot use), and 93.5 percent of the 34,979 flagged general-chatbot users appear in no therapy-world community at all, so the crossover analysis characterizes the observable overlap, not the typical flagged user.

5.2 Discourse network (RQ2)

The pooled reply network over therapeutic discourse in the human-therapy communities and r/therapyGPT has 27,712 nodes and 79,182 edges, a giant component covering 98 percent of nodes, and strong community structure, with a Louvain modularity of 0.639 across 55 communities. Two results follow. First, the single account with the highest audience and highest betweenness (B1 in Table 2) is the commercial founder of a custom therapy chatbot, whose activity spans the AI self-help and mainstream-therapy communities. The cross-community discourse is thus partly a matter of commercial capture, so we report network statistics with and without the commercially flagged accounts. Second, the accounts that bridge into the disillusionment community are a small set of organic multi-community users (B2, B4, B5), each active in three or four communities including r/therapycritical, which is the individual-level counterpart of the structural adjacency in RQ1. The detected communities correspond to the groups.

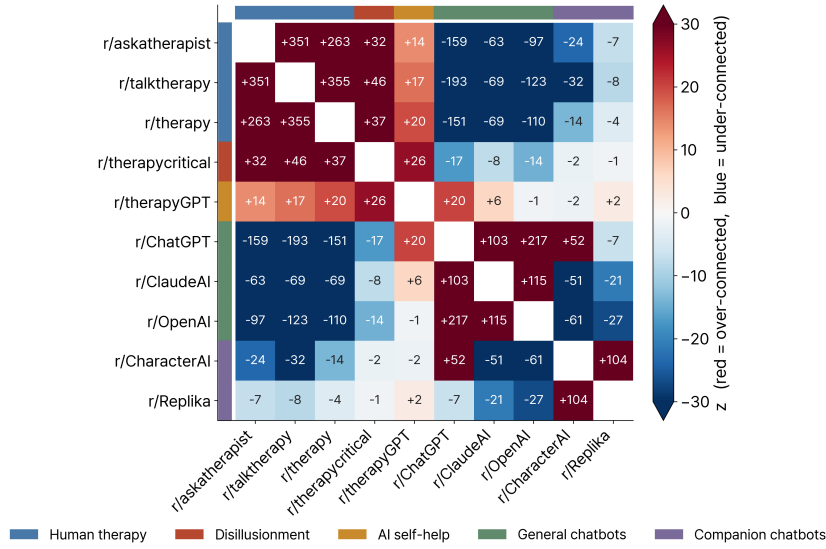


Figure 1: Cross-community co-membership under the fixed degree sequence model (z). Red marks pairs that share more members than would be expected by chance; blue marks pairs that share fewer. The human-therapy block, the general-chatbot block, and the companion block are each internally over-connected; r/therapyGPT is the one community that over-connects into both worlds, most strongly to r/therapycritical, while every direct human-therapy $\times$ chatbot pair is under-connected.

r/therapyGPT forms its own dense blocks, separate from the mainstream-therapy thread clusters, so therapeutic chatbot discourse is structurally distinct from the discussion of AI that occurs inside mainstream therapy communities. The general-chatbot reply network, analyzed separately, has no comparable central actor. On r/OpenAI (the general-chatbot community with complete comments) reply audience concentrates on the original posters of large non-therapeutic threads (the top account draws 41,515 replies yet writes no flagged document), so we rank by therapeutic-content volume, which returns dispersed organic individuals (the most prolific writes 187 posts) and no dominant or commercial node. Commercial capture is thus specific to the self-help and human-therapy discourse.

Betweenness is the centrality most prone to false positives, so we test the brokerage claims against the degree-preserving null such a measure requires (Farine and Carter, 2022), using one hundred degree-preserving rewirings of the directed graph, with the identical betweenness estimator for observed and null values. The commercial account’s betweenness falls significantly *below* its degree-preserving expectation ($z = -8.1$, empirical $p = .01$), so its visibility is a volume effect and not a brokerage position. The organic multi-community accounts, by contrast, hold betweenness above the

Acct	In-deg	Betw	#Com	Role
B1	1419	0.087	3	commercial founder
B2	431	0.050	4	bridge to disillusion
B3	324	0.029	3	mainstream peer
B4	126	0.018	4	bridge to disillusion
B5	116	0.018	3	bridge to disillusion

Table 2: Top accounts in the reply network by audience and bridging (anonymized). “#Com” is the number of communities the account spans; B2, B4, B5 are also active in r/therapycritical.

null (z from +4.0 to +10.2) and span genuine structural holes (Burt constraint below .01, effective size ranging from the hundreds to just over one thousand), so the real cross-community brokerage belongs to them. The community partition is stable across random seeds (mean pairwise adjusted Rand index 0.65; modularity 0.64), and removing the commercially flagged and bot accounts leaves modularity essentially unchanged (0.640 to 0.642) while an organic bridge becomes the top broker, so the group structure does not depend on the commercial actor.

5.3 Discourse frames (RQ2)

The groups frame therapeutic chatbot use through different topics (Figure 2; exact prevalences in Table 5, appendix). Of the 66 pairwise group contrasts across the eleven reported frames, estimated

from the fitted model and Benjamini-Hochberg corrected, 58 are significant at $p < .001$; the exceptions are the self-help and human-therapy groups, indistinguishable on the feeling-heard and family-distress frames, the two groups our structural results place closest together. The therapy topic itself, clinicians, sessions, CBT, insurance, and cost, is the single largest frame in both r/therapyGPT (.200) and the human-therapy communities (.192), the two groups closest to formal therapy. They add different things. r/therapyGPT pairs the therapy frame with emotional support and reflection (.106) and with talk of harm, danger, and AI psychosis (.102), the discourse closest to using the model as talk therapy, while the human-therapy communities add privacy, clinical notes, and consent (.126). The general-chatbot communities frame the chatbot instrumentally, and the companion communities through casual venting (.187), feeling heard (.105), and roleplay (.097), with the therapy frame nearly absent (.027). Each group renders the chatbot in its own terms. Companion use, though relational and emotional, is not framed as a substitute for therapy the way the self-help community frames it.

5.4 Trajectory (RQ3)

Among users active in both worlds, human-therapy participation predominantly comes first (Table 9). With the content-anchored sequence restricted to the post-release window, 34 percent arrive at the chatbot first; this is stable across window anchors and a minimum-activity threshold (29 to 35 percent, Appendix F), and among users whose chatbot activity is in r/OpenAI, the general-chatbot community with continuously complete comment coverage, the chatbot-first share is lowest, at 24 percent. Human first is thus the robust direction, between 65 and 76 percent across specifications. The same pattern holds for the AI self-help community specifically (Figure 4, appendix), where 1,112 users are active in both r/therapyGPT and a human-therapy community, and among the 924 of them whose first human-therapy activity falls in the ChatGPT era, 77 percent (708) were active in the human community first, against 23 percent (216) active in r/therapyGPT first.

We caution against the stronger causal interpretation that disillusionment drives the move. A Poisson rate model with corrected exposure gives a raised but modest association between disillusionment-community membership and later therapeutic chatbot use (incidence rate ratio 3.4,

Appendix G), but when membership is restricted to precede the window of risk the estimate is not identifiable, because only 31 disillusionment-community members ever cross, and about a fifth joined the disillusionment community after their first therapeutic chatbot post. The same order ambiguity holds inside the AI self-help community, where 60 percent of the 77 shared users were active in the disillusionment community first, a majority but not the near-unanimity a simple story of grievance and exit would predict. The disillusionment community is small and largely a creation of the consumer-AI era, so we report the cross-sectional pattern as the finding and decline the temporal claim; recovering it would require within-user evidence of rising dissatisfaction before the first chatbot-therapy post, which we leave as future work.

6 Discussion

Within the Network Episode Model, these community-level patterns place the venues that organize therapeutic chatbot talk inside ordinary help-seeking. The self-help community is the one crossing point we observe between mainstream therapy and the chatbot world, and it recruits most intensively, at roughly nine times chance, from the disillusionment community; the order of arrival in these communities is predominantly human therapy first, though the data cannot settle whether disillusionment drives the move, and community membership is not itself evidence of treatment or of leaving it. The discourse analysis adds two cautions for treating these communities as organic peer support: the cross-community conversation is concentrated in a commercial account, so apparent grassroots diffusion is partly marketing, and the four groups talk past one another, maintaining the boundary in their vocabularies as much as in their interactions.

These results address two debates. For the growing science of AI mental health tools, which has concentrated on efficacy and safety in trial and survey samples (Heinz et al., 2025; Sharma et al., 2023; Stade et al., 2024), our account describes who is observably present in these discourses: most people our construct flags appear in no therapy community, mainstream-therapy communities dominate the observable minority in absolute numbers, and the community organizing therapeutic chatbot use recruits far out of proportion from the therapy-disillusioned. A tool can be clinically ef-

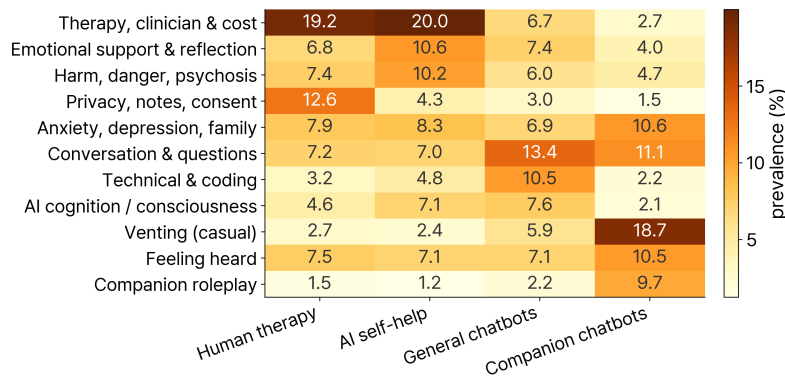


Figure 2: Structural topic model frame prevalence by community group. The therapy frame (clinician, session, cost) dominates both the AI self-help and human-therapy groups; the self-help group adds emotional support, reflection, and talk of harm and AI psychosis, the human-therapy group adds privacy and clinical notes, the general-chatbot group is instrumental and conversational, and the companion group is venting, feeling heard, and roleplay.

fective and still be discussed most intensively in a community that draws disproportionately on people who write about disappointment with human treatment. For the parallel debate over whether AI will replace human therapists, our findings make the replacement claim precise. Casting the chatbot as a digital therapist is well-documented (Luo et al., 2025; Siddals et al., 2024; Collins et al., 2025); we add that this framing is concentrated in the self-help and human-therapy groups and is nearly absent from the companion group (Figure 2). Yet that absence is not an absence of therapeutic function. The companion communities peak on the feeling-heard and family-distress frames, the felt benefits reported in the companionship literature (Siddals et al., 2024; De Freitas et al., 2026), so what separates them is framing, not emotional substance. This is consistent with discourse-level evidence that companionship-oriented use is where therapeutic alliance is weakest and where dependence and symptom escalation are most often reported (Aghakhani and Rezapour, 2026). Two implications follow: the relational failures the disillusionment community records, feeling unheard, judged, or dismissed, are repairable features of human care; and the therapy frame that dominates the self-help and human-therapy groups is saturated with cost and insurance language, so access, not only grievance, shapes who arrives at the substitute.

7 Limitations

Identity and measurement. An account is a proxy for a person and a post is a proxy for use; the same person may hold several accounts and a single account may be shared, so every co-membership

and crossover count is an estimate, not an exact head count (De Choudhury and De, 2014). Inferring mental health-relevant states from social-media proxies has well-documented validity limits (Chancellor and De Choudhury, 2020). We never observe the chatbot conversations themselves, only what people write about them on Reddit, and we do not quote users (see Ethics). Deleted and removed content is non-random, and we drop it, which can move the prevalence in either direction.

Construct measurement. “Using AI as talk therapy” is identified by keyword match (Appendix D), and the validation in Appendix E sets what that costs. Precision for the writer’s own therapeutic use is .43 in r/therapyGPT and .23 to .34 elsewhere, so the flagged-document counts in Table 1 are counts of discourse and overstate first-person use by a factor of two to four. Recall against the broader lexicon is about .39, so the flag also misses documents that do describe therapeutic chatbot use. The co-membership analysis behind RQ1 uses community membership and never the flag, so it is unaffected. The crossover rates are a ratio of two quantities measured with the same flag, and precision does not differ detectably by the crossing user’s origin community (Appendix E), so the contrast survives even though the levels are inflated. The topic model describes flagged discourse.

History of r/therapyGPT. r/therapyGPT was created in spring 2025 and its founding posts frame the community around disappointment with human therapy, so the disillusionment tie may partly follow from how the community advertised itself instead of from an unprompted convergence. Three

checks bound the concern without removing it: the ordering holds after removing all 133 commercially flagged accounts including the founder, it strengthens under a stricter membership definition, and the same adjacency appears in the reply network among organic multi-community accounts. The design cannot separate a founding frame that attracts disillusioned users from a disillusioned population that would have converged on any community of this kind.

Inference. The disillusionment community is small and largely post-dates the consumer-AI era, so the rate model cannot identify whether grievance temporally precedes adoption; we report the cross-sectional adjacency and decline the causal claim. Small counts also set the precision of the headline ratios, which rest on 77 shared members and 43 crossers.

Generalizability. Reddit communities are self-selected and are not representative of people in therapy or of chatbot users at large; the corpus is English-language; and we report no clinical outcomes, so nothing here establishes whether therapeutic chatbot use helps or harms its users. The structural topic model is interpretive and includes off-topic register and technical topics, which we omit from the reported frame table (Appendix C).

8 Ethics

Data and consent. We use data that members posted to public subreddits, obtained from publicly available Reddit archives (Pushshift and Arctic Shift); no data were collected through the Reddit API for this study and none are redistributed. Deleted and removed content is excluded from analysis. The text is already public and we collect no new data from participants, but the people involved are discussing mental health distress, so we treat the data as sensitive throughout. Because the study is a secondary analysis of already public text, involves no interaction or intervention with individuals, and uses no private identifiable information, it does not meet the regulatory definition of human-subjects research and did not require institutional review board review.

Anonymization. We report no personally identifying information. Usernames are withheld, the accounts in Table 2 are anonymized and described only by role, and example content is paraphrased rather than quoted, because verbatim text from a

small community is quasi-identifying and can be matched to an author on a public platform. We make no attempt to deanonymize accounts, to recover real-world identities, or to link accounts across platforms. The single commercial actor we discuss is characterized by its role, not named. Role-level description of a highly visible account is only partial anonymization; we accept it for a commercial actor analyzed in its commercial capacity, and no other account is described at that granularity.

Vulnerable users and dual use. Any method that locates and profiles people who are in distress or disillusioned with care carries a dual-use risk, including targeting by bad actors or by commercial interests of exactly the kind our results document. To limit that risk, we release only aggregate network and topic outputs and the coding scheme; we do not release user-level classifications, crossover user lists, or any account labeling by group. We frame the commercial-capture and disillusionment findings as cautions for clinicians, platform designers, and researchers, not as a way to locate the population.

Availability. The full analysis code, the keyword lists, and the aggregate outputs reported here (the 45-pair null tables and robustness variants, the trigger-family attribution, and the topic prevalences and terms) are available from the authors on request; raw Reddit data are not redistributed, and the release documents where each public archive can be obtained. We withhold a public code release because the code reproduces user-level community and construct labels.

Stakes. Because a therapy chatbot can be clinically effective and still, as we show, be discussed most intensively in a community that overlaps unusually with people writing about disappointment with human care, most of whom were active in that care first, the appropriate response is to repair the relational failures they describe rather than to route distressed users toward unsupervised tools.

References

Elham Aghakhani and Rezvaneh Rezapour. 2026. [Like a therapist, but not: Reddit narratives of ai in mental health contexts](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11716–11736, San Diego, California. Association for Computational Linguistics.

- David H. Barlow. 2010. [Negative effects from psychological treatments: A perspective](#). *American Psychologist*, 65(1):13–20.
- Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. [Fast unfolding of communities in large networks](#). *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- C. J. Carstens. 2015. [Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm](#). *Physical Review E*, 91(4):042812.
- Stevie Chancellor and Munmun De Choudhury. 2020. [Methods in predictive techniques for mental health status on social media: A critical review](#). *npj Digital Medicine*, 3:43.
- Giulio Cimini, Alessandro Carra, Luca Didomenicantonio, and Andrea Zaccaria. 2022. [Meta-validation of bipartite network projections](#). *Communications Physics*, 5:76.
- Kimberly A. Cohen, Colleen Stiles-Shields, Nathaniel Winkquist, and Emily G. Lattie. 2021. [Traditional and nontraditional mental healthcare services: Usage and preferences among adolescents and younger adults](#). *Journal of Behavioral Health Services & Research*, 48(4):537–553.
- Amanda C. Collins, Damien Lekkas, Michael V. Heinz, Janelle Annor, Franklin Ruan, and Nicholas C. Jacobson. 2025. [Chatgpt as therapy: A qualitative and network-based thematic profiling of shared experiences, attitudes, and beliefs on reddit](#). *Journal of Psychiatric Research*, 191:277–284.
- Paul Crichton, Havi Carel, and Ian James Kidd. 2017. [Epistemic injustice in psychiatry](#). *BJPsych Bulletin*, 41(2):65–70.
- Joe Curran, Glenys D. Parry, Gillian E. Hardy, Jennifer Darling, Ann-Marie Mason, and Eleni Chambers. 2019. [How does therapy harm? a model of adverse process using task analysis in the meta-synthesis of service users’ experience](#). *Frontiers in Psychology*, 10:347.
- Munmun De Choudhury and Sushovan De. 2014. [Mental health discourse on reddit: Self-disclosure, social support, and anonymity](#). In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 8, pages 71–80.
- Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. 2026. [Ai companions reduce loneliness](#). *Journal of Consumer Research*, 52(6):1126–1148.
- Bojan Evkoski, Srebrenka Letina, and Petra Kralj Novak. 2026. [The structure of psychopathology on reddit: Network analysis of mental health communities in relation to the icd diagnostic system](#). *Journal of Medical Internet Research*, 28:e80958.
- Damien R. Farine and Gerald G. Carter. 2022. [Permutation tests for hypothesis testing with animal social network data: Problems and potential solutions](#). *Methods in Ecology and Evolution*, 13(1):144–156.
- Eliot Freidson. 1960. [Client control and medical practice](#). *American Journal of Sociology*, 65(4):374–382.
- Karl Godard and Zachary P. Neal. 2022. [fastball: A fast algorithm to randomly sample bipartite graphs with fixed degree sequences](#). *Journal of Complex Networks*, 10(6):cnac049.
- Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. 2025. [Randomized trial of a generative ai chatbot for mental health treatment](#). *NEJM AI*, 2(4).
- Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. 2025. [“i’ve talked to chatgpt about my issues last night”: Examining mental health conversations with large language models through reddit analysis](#). *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–25. Article CSCW356.
- Daniel Kardefelt-Winther. 2014. [A conceptual and methodological critique of internet addiction research: Towards a model of compensatory internet use](#). *Computers in Human Behavior*, 31:351–354.
- Sanguk Lee and Tai-Quan Peng. 2025. [Bridging and bonding: The roles of brokerage and closure in mobilizing support provision in online support groups](#). *PLOS ONE*, 20(6):e0325108.
- Sanguk Lee, Minjin Rheu, and Jie Zhuang. 2025. [The chatgpt effect: Investigating shifting discourse patterns, sentiment, and benefit–challenge framing in ai mental health support](#). *Behavioral Sciences*, 15(9):1172.
- Scott O. Lilienfeld. 2007. [Psychological treatments that cause harm](#). *Perspectives on Psychological Science*, 2(1):53–70.
- Xiaochen Luo, Smita Ghosh, Jacqueline L. Tilley, Patricia Besada, Jinqiu Wang, and Yangyang Xiang. 2025. [“shaping chatgpt into my digital therapist”: A thematic analysis of social media discourse on using generative artificial intelligence for mental health](#). *Digital Health*, 11:20552076251351088.
- Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. [Loneliness and suicide mitigation for students using gpt3-enabled chatbots](#). *npj Mental Health Research*, 3(1):4.
- Sergei Maslov and Kim Sneppen. 2002. [Specificity and stability in topology of protein networks](#). *Science*, 296(5569):910–913.
- Adam S. Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. [Smartphone-based conversational](#)

- agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, 176(5):619–625.
- Zachary Neal. 2014. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39:84–97.
- Wenjing Pan, Bo Feng, and Cuihua Shen. 2020. Examining social capital, social support, and language use in an online depression forum: Social network and content analysis. *Journal of Medical Internet Research*, 22(6):e17365.
- Pat Pataranutaporn, Sheer Karny, Chayapatr Archiraranguprok, Constanze Albrecht, Auren R. Liu, and Pattie Maes. 2025. “my boyfriend is ai”: A computational analysis of human-ai companionship in reddit’s ai community. *Preprint*, arXiv:2509.11391.
- Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, 140:107600.
- Brea L. Perry and Bernice A. Pescosolido. 2015. Social network activation: The role of health discussion partners in recovery from mental illness. *Social Science & Medicine*, 125:116–128.
- Bernice A. Pescosolido. 1992. Beyond rational choice: The social dynamics of how people seek help. *American Journal of Sociology*, 97(4):1096–1138.
- Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.
- Tony Rousmaniere, Yimeng Zhang, Xu Li, and Sid-dharth Shah. 2026. Large language models as mental health resources: Patterns of use in the united states. *Practice Innovations*, 11(2):139–155.
- Till Scholich, Maya Barr, Shannon Wiltsey Stirman, and Shriti Raj. 2025. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health*, 12:e69709.
- Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:46–57.
- Steven Siddals, John Torous, and Astrid Coxon. 2024. “it happened to be the perfect thing”: Experiences of generative ai chatbots for mental health. *npj Mental Health Research*, 3(1):48.
- Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzæg. 2021. My chatbot companion - a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, 149:102601.
- Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12.
- Tobias Steinbrenner, Christopher Lalk, Alin Kabjesz, Drin Ferizaj, Juan Segundo Pena Loray, Flavio Iovoli, and Julian Rubel. 2025. Exploring negative experiences in psychotherapy using an nlp approach on online forum data. *npj Mental Health Research*, 4(1):59.
- Bernhard Strauss, Romina Gawlytta, Andrea Schleu, and Dominique Frenzl. 2021. Negative effects of psychotherapy: Estimating the prevalence in a random national sample. *BJPsych Open*, 7(6):e186.
- Giovanni Strona, Domenico Nappo, Francesco Boc-cacci, Simone Fattorini, and Jesús San-Miguel-Ayanz. 2014. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5:4114.
- Joshua K. Swift and Roger P. Greenberg. 2012. Pre-mature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 80(4):547–559.
- Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 1056–1066.
- Zbyněk Vybíral, Benjamin M. Ogles, Tomáš Řiháček, Barbora Urbancová, and Veronika Gocieková. 2024. Negative experiences in psychotherapy from clients’ perspective: A qualitative meta-analysis. *Psychotherapy Research*, 34(3):279–292.
- Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600:264–268.
- Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. 2025. The rise of ai companions: Interaction with ai companions and psychological well-being. *arXiv preprint*. ArXiv:2506.12605.
- Katharina A. Zweig and Michael Kaufmann. 2011. A systematic approach to the one-mode projection of bi-partite graphs. *Social Network Analysis and Mining*, 1(3):187–218.

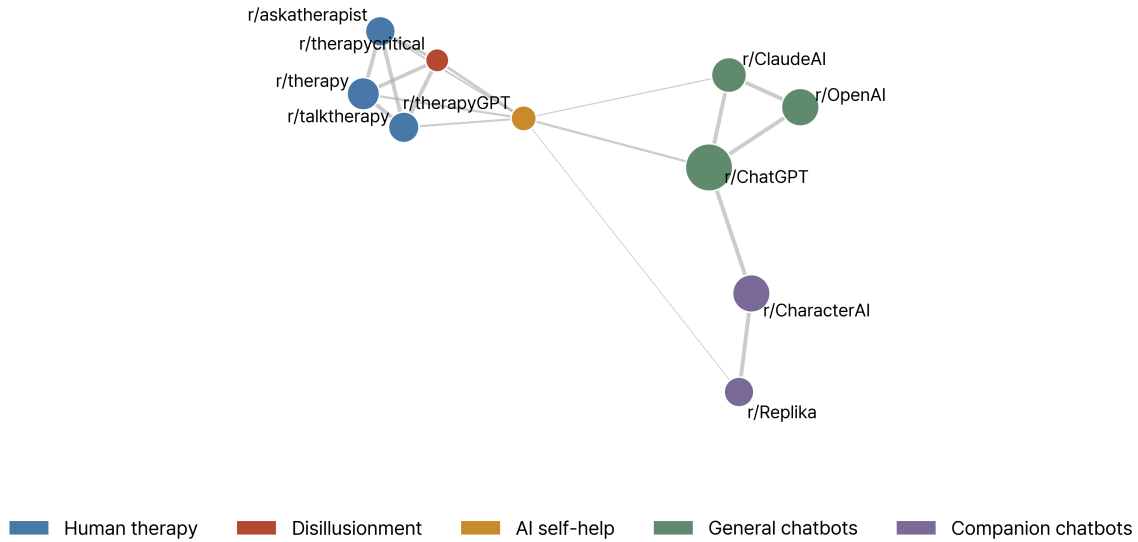


Figure 3: Co-membership network of significant positive overlap ties. Nodes are communities, sized by users and colored by group. r/therapyGPT is the only community linking the human-therapy and chatbot worlds.

Community	Gen. chatbot	Therap. use	Companion
r/therapy	6.0%	0.49%	1.43%
r/talktherapy	7.1%	0.75%	1.27%
r/askatherapist	7.1%	0.66%	1.45%
r/therapycritical	17.0%	3.45%	3.37%
r/therapyGPT	47.3%	12.51%	4.79%

Table 3: Share of each community’s users who also appear in a general-chatbot community (any use), who write about chatbot therapy there (therapeutic use), and who appear in a companion-chatbot community. The disillusionment and self-help communities cross over at the highest rates on every measure.

A Bipartite backbone, all pairs

Table 4 reports all 45 community pairs under the fixed degree sequence model (1,000 curveball samples, thinned at six trades per user between draws, mean lag-one autocorrelation 0.04; observed and null-mean co-membership; z ; Benjamini-Hochberg at $p < .05$, ns otherwise; the empirical threshold is corrected $p = .002$). Within-group pairs are strongly over-connected, as expected. r/therapyGPT over-connects to both worlds, with the r/therapycritical tie far above every other cross-cluster ratio (9.4 times the null against 1.9–2.1 for the mainstream-therapy pairs and 1.2 for ChatGPT and ClaudeAI); its ties to r/OpenAI and r/CharacterAI are at chance. Every direct general-chatbot and companion pairing with a human-therapy community is significantly under-connected.

Robustness. Two full reruns of the null probe the headline pattern. First, removing all 133 commercially flagged accounts and bots from the incidence matrix, four of which are among the 77 disillusionment–self-help shared members, leaves the tie at 73 observed against a null of 8.0 ($z = +23.9$), with every other tie’s sign and significance unchanged. Second, redefining membership as at least two contributions in a community strengthens the pattern: the disillusionment–self-help tie rises to 14 times its null expectation (43 vs. 3.1, $z = +21.7$) while the mainstream-therapy ratios remain near two, and the self-help community’s tie to r/OpenAI turns modestly negative ($z = -2.9$). This ordered pattern is therefore not attributable to promotional accounts or to drive-by single-post membership.

B Discourse-network robustness

We test the brokerage claims of Section 5.2 against a degree-preserving null and check the stability of the community partition (Farine and Carter, 2022). The brokerage and stability analyses are computed on the giant component of the focal discourse network (human-therapy and r/therapyGPT, AI-discussion threads), which holds 27,187 of its 27,712 nodes and 78,739 edges. Betweenness is estimated with 3,000 source pivots under identical settings for the observed graph and for each of 100 degree-preserving directed rewirings (ten success-

Community A	Community B	Obs	Null	z
therapy	talktherapy	15007	1605	+355
talktherapy	asktherapist	12833	1257	+351
therapy	asktherapist	9201	1230	+263
OpenAI	ChatGPT	136116	116777	+217
OpenAI	ClaudeAI	42644	30282	+115
Replika	CharacterAI	3246	704	+104
ChatGPT	ClaudeAI	67020	56664	+103
ChatGPT	CharacterAI	27255	23261	+52.4
talktherapy	therapycritical	309	38	+45.8
therapy	therapycritical	244	37	+36.6
asktherapist	therapycritical	192	29	+31.9
therapycritical	therapyGPT	77	8	+26.5
therapyGPT	ChatGPT	3947	3291	+20.2
therapy	therapyGPT	618	298	+19.5
talktherapy	therapyGPT	579	304	+16.6
asktherapist	therapyGPT	443	235	+14.2
therapyGPT	ClaudeAI	1041	877	+6.2
therapyGPT	Replika	133	110	+2.4
therapycritical	Replika	11	14	-0.9 ns
therapyGPT	OpenAI	1738	1773	-1.0 ns
therapyGPT	CharacterAI	342	376	-1.9 ns
therapycritical	CharacterAI	34	46	-1.9 ns
therapy	Replika	466	554	-3.9
ChatGPT	Replika	5920	6222	-6.7
asktherapist	Replika	284	436	-7.3
therapycritical	ClaudeAI	34	105	-7.7
talktherapy	Replika	375	565	-8.4
therapy	CharacterAI	1442	2004	-13.8
therapycritical	OpenAI	46	212	-14.3
therapycritical	ChatGPT	194	389	-17.3
ClaudeAI	Replika	881	1652	-21.1
asktherapist	CharacterAI	710	1566	-23.7
OpenAI	Replika	2127	3350	-26.9
talktherapy	CharacterAI	705	2045	-32.4
ClaudeAI	CharacterAI	2609	6085	-50.8
OpenAI	CharacterAI	7462	12458	-60.5
asktherapist	ClaudeAI	462	3715	-62.7
therapy	ClaudeAI	724	4756	-68.7
talktherapy	ClaudeAI	595	4858	-69.0
asktherapist	OpenAI	923	7564	-97.3
therapy	OpenAI	1549	9714	-110
talktherapy	OpenAI	1149	9920	-123
therapy	ChatGPT	7015	18122	-151
asktherapist	ChatGPT	4180	14098	-159
talktherapy	ChatGPT	5232	18508	-193

Table 4: All 45 community pairs under the fixed degree sequence model, sampled with the curveball algorithm, 1,000 samples.

ful swaps per edge; a replicate that cannot complete its swaps aborts rather than entering the null under-mixed). Commercial accounts are flagged by promotional share, at least two promotional posts, and at least two percent of the account’s contributions; one previously flagged high-activity account (B2, promotional share 0.2 percent) reclassifies as organic under this rule. The top accounts span genuine structural holes (Burt constraint $\leq .008$; effective size 168–1,062). The commercial account B1’s betweenness falls significantly below the null (observed .087 against $.132 \pm .006$, $z = -8.1$, empirical $p = .01$): its audience reflects volume, not bridging position. The organic accounts exceed the null (B2 $z = +4.0$, B3 $z = +8.0$, B4 $z = +10.2$, B5 $z = +8.3$; empirical $p = .01$). The Louvain partition is stable across random seeds (mean pairwise adjusted Rand index 0.65, SD 0.04; modularity 0.64), and removing the commercial

and bot accounts present in this network (129 accounts) changes modularity only from 0.640 to 0.642 and makes B2 the top broker, so the group structure is not a product of the commercial actor. This is fewer than the 133 removed from the co-membership incidence matrix in Section 5.1, because the stricter adjudication above reclassifies B2 as organic and only accounts that appear in this discourse network are dropped. Finally, excluding r/therapyGPT entirely leaves a reply network of the mainstream and disillusionment communities (19,490 nodes) whose top brokers are again organic multi-community accounts with betweenness above the same null (50 rewirings; z up to +4.1, including B3 and B5), while the commercial account no longer ranks among the top brokers at all: its centrality is bound to the community it promotes in, and the organic brokerage into the disillusionment community is not.

C Structural topic model

We fit the structural topic model with $K = 15$ on the pooled focal-discourse documents, with topic prevalence a function of community group and a smooth spline of time (ten basis terms). The corpus derives from the 92,520 flagged documents in Table 1 as follows: documents shorter than eight words and exact duplicate texts are dropped, each community is capped at 20,000 documents by random sample (only r/ChatGPT exceeds the cap, sampled from 45,687), and vocabulary is thresholded at ten documents, leaving 64,658 documents and 8,448 stemmed terms; the model converged in 51 EM iterations. The r/therapycritical documents (245 after cleaning) are pooled into the human-therapy group for this analysis. Preprocessing lowercases, removes punctuation, numbers, and stopwords, and stems; the stopword list also removes the AI product and tool names (for example, ChatGPT, GPT, Claude, OpenAI, Gemini, Replika, Character.AI, and the bare token AI) so that the topics reflect content and not brand mentions. Four of the fifteen topics are omitted from Table 5: two contraction and register topics, a roleplay and everyday topic, and a loneliness topic that overlaps the feeling-heard frame; the table reports raw mean topic prevalences by group for the eleven construct-relevant frames without renormalizing over the remaining topics. Group contrasts in the text come from the fitted model’s covariate effects, estimated for all pairwise group comparisons and

corrected with Benjamini-Hochberg across the 66 tests on the reported frames. As a robustness check on commercial contamination, re-aggregating topic prevalence with all commercially flagged authors excluded (1,156 documents) moves no reported group mean by more than .01. Table 6 lists the top FREX terms (stemmed) for each frame; profanity and sexually explicit terms are elided from the printed lists. Topic content comes from the highest-probability and FREX terms; the labels are the authors' summaries and are interpretive.

Model selection and robustness. A held-out comparison of $K \in \{10, 15, 20, 25\}$ on a 20,000-document subsample shows held-out likelihood and residuals improving with K while semantic coherence declines; the held-out gain beyond $K=15$ is small (-6.915 at 15 against -6.894 at 20) relative to the coherence cost, and $K=15$ is the smallest K at which the construct-relevant frames separate. The reported patterns are robust to preprocessing and to K : refits with apostrophes normalized and contraction stems stopworded, which removes the two register topics, reproduce every headline peak at $K=15$ and $K=20$ (therapy and clinician frame dominant in the human and self-help groups, harm and emotional-support peaks in self-help, privacy peak in human, venting and roleplay peaks in companion), and at $K=10$ the coarser model merges frames but preserves their group direction. The sharpest check masks every trigger phrase out of the documents before fitting, so the flag lexicon cannot echo into the topics: the clinician and therapy frame still dominates the human (.20) and self-help (.15) groups, reflective emotional support (.12) and a harm, safety, and guardrail frame (.09) still peak in self-help, and romance and roleplay still peak in the companion group. One reported frame does not survive this masking as a distinct topic: casual venting dissolves into general casual register once the venting phrases themselves are removed, so that label is partly lexicon-derived.

D Focal-construct keywords

We flag therapeutic chatbot use by matching post and comment text against two keyword lists. The narrower list, which we use, matches explicit relational, venting, or substitution phrases; the broader list also matches any emotional or mental health word.

Narrower list. “ai therap”; “as” or “like” a/my therapist; “instead of,” “replace,” “better than,” or “cheaper than” therapy; “my therapist,” “in therapy,” “seeing a therapist,” “been to therapy”; “vent,” “venting”; “talking to it / ChatGPT / GPT / Claude / the AI”; “emotional support”; “process my emotions/feelings”; “opened up to”; “non-judgment”; “someone to talk to”; “feels heard,” “less lonely,” “less alone”; “listens to me,” “here for me”; “my anxiety / depression / trauma / grief”; “helped me through / cope / process”; “comforts me”; “talk through”; “judgment-free,” “never judges,” “doesn’t judge”; “poured my heart out”; “trauma dump”; “rant to it”; “understands me”; “validates me”; “safe space”; “always there for me”; “lean on it”; “emotional crutch”; substitution for a “counselor,” “psychiatrist,” “psychologist,” or “social worker”; “my counselor / social worker / psychiatrist / psychologist.”

Broader list. “therapy / therapist,” “counsel,” “psycholog,” “psychiatr,” “mental health,” “depress,” “anxiet / anxious,” “panic attack,” “trauma,” “ptsd,” “suicid,” “self-harm,” “lonely,” “grief,” “heartbreak,” “breakup,” “emotional support,” “venting,” “my feelings,” “coping,” “emotional,” “self-esteem,” “burnout,” “overwhelmed,” “psychotherapy,” “mental illness,” “social work / social worker,” “case worker,” “well-being,” “self-care.”

The narrower list matches 34,979 general-chatbot users and the broader list 104,124.

Trigger-family attribution. Table 7 attributes every narrow-flagged document to the keyword families that fire in it. The flag means different things in different communities: in r/therapyGPT, 70 percent of flagged documents contain a substitution or own-history phrase and only 14 percent rest solely on venting or feeling-heard phrases, while in the companion communities that pattern reverses (50 percent venting or feeling-heard only, about 12 percent substitution or history). The loosest family, the addressee phrases, is also the one that fires most often in the general-chatbot communities (37 percent in r/ChatGPT, 58 percent in r/ClaudeAI), which is where the flag’s precision is weakest and where the validation in Appendix E concentrates. Cross-community contrasts in Tables 1 and 3 inherit this mix, and the reversal itself is construct-level evidence for the boundary the structural analysis finds: flagged companion-community talk is companionship register without substitution framing.

Frame	Human	Self-help	Gen.	Comp.
Therapy, clinician & cost	.192	.200	.067	.027
Emotional support & reflection	.068	.106	.074	.040
Harm, danger, psychosis	.074	.102	.060	.047
Privacy, notes & consent	.126	.043	.030	.015
Anxiety, depression, family	.079	.083	.069	.106
Conversation & questions	.072	.070	.134	.111
Technical & coding	.032	.048	.105	.022
AI cognition/consciousness	.046	.071	.076	.021
Venting (casual)	.027	.024	.059	.187
Feeling heard	.075	.071	.071	.105
Companion roleplay	.015	.012	.022	.097

Table 5: Mean structural topic model prevalence by community group for the eleven construct-relevant frames (Self-help = r/therapyGPT, Gen. = general chatbots, Comp. = companion chatbots); bold marks each frame’s peak. Columns are in the order Figure 2 plots them. These are the exact values visualized in Figure 2.

Frame	Top FREX terms (stemmed)
Therapy, clinician & cost	therapi, therapist, cbt, journal, afford, therapist, psychologist, session
Emotional support & reflection	emot, reflect, empathi, support, meaning, growth, deeper, explor
Harm, danger, psychosis	harm, danger, psychosi, suicid, delus, abus, blame, mental
Privacy, notes & consent	record, privaci, note, consent, law, legal, email, polici
Anxiety, depression, family	depress, year, mom, dad, husband, mother, anxiety, sleep
Conversation & questions	talk, ask, answer, convers, rememb, question, write, respond
Technical & coding	api, code, workflow, token, opus, agent, web, project
AI cognition/consciousness	logic, languag, conscious, analysi, predict, interpret, intellig, frame
Venting (casual)	vent, gonna, bro, lmao, wanna, site (profanity elided)
Feeling heard	feel, listen, thank, anyon, alway, sometim, judg, love
Companion roleplay	rep, luka, erp, women, nsfw, filter (explicit terms elided)

Table 6: Top FREX terms (stemmed, as output by the model) for the eleven construct-relevant frames in Table 5. Product and tool names were removed before fitting.

E Construct validation

A stratified document sample, coded against the study codebook, measures what the keyword flag costs in precision and recall. Documents are coded by a language model (Claude Opus 5) given the codebook and the document text alone, with no indication of which keyword fired or which stratum the document came from. Two authors then independently hand-coded a 300-document subsample of the same draws, also blind to stratum, and we report both codings without forming a consensus label. Agreement between the two coders is $\kappa = .80$ on the six-category judgment and $\kappa = .98$ on the binary contrast between therapeutic use and everything else. Human-coded precision on the flagged documents is .24 (95 percent CI .18 to .31) and .23 (.17 to .30), and the model reproduces the two coders’ six-category labels on 89 and 80 percent of the subsample, agreeing with each on the therapeutic-use contrast at $\kappa = .92$. We report the human figures as the estimate and the model coding of the full sample as its extension.

Design. One exhaustive scan of the raw dumps assigns every chatbot-community document to one of three strata: NARROW, matched by the flag we use; BROAD-ONLY, matched by the broader emotional and mental health lexicon but not by the narrow one; and NEITHER. The scan reproduces every published document count in Table 1 exactly, which is our check that the validation measures the same flag the analysis used. Stratum sizes are counted in full, so recall is design-based rather than assumed: true positives are the NARROW count times its precision, and missed positives are each unflagged stratum’s size times the share of it that describes therapeutic chatbot use. We draw 200 flagged documents per community group for precision, 150 flagged users with up to three documents each for user-level precision, and about 115 documents per community in each of the two unflagged strata for recall.

Precision. Table 8 reports two definitions of the construct, because they behave differently. Almost every flagged document does refer to an AI system (.89 to .98), and the great majority in the therapy-

Community	Flagged	Subst.	Own hist.	Venting	Heard	Process.	Talk-to-AI	Vent/heard	Subst./hist.
r/ChatGPT	45,687	.18	.11	.14	.13	.15	.37	.25	.27
r/OpenAI	4,517	.15	.06	.11	.13	.14	.44	.23	.20
r/ClaudeAI	3,921	.08	.04	.12	.11	.08	.58	.23	.12
r/Replika	7,268	.08	.05	.23	.29	.24	.17	.50	.12
r/CharacterAI	15,601	.05	.07	.38	.14	.15	.25	.50	.12
r/therapyGPT	4,583	.48	.26	.09	.09	.14	.10	.14	.70

Table 7: Share of narrow-flagged documents containing each trigger family (families are not exclusive). Subst. = explicit substitution phrases; Own hist. = the writer’s own human-therapy history; Heard = feeling-heard/non-judgment phrases; Process. = emotional-processing phrases; Talk-to-AI = the addressee phrases (“talking to it / ChatGPT / the AI”), the loosest family and the largest one in the general-chatbot communities. “Vent/heard” = documents whose only triggers are venting or feeling-heard phrases; “Subst./hist.” = documents containing a substitution or own-history phrase.

adjacent communities are about therapeutic chatbot use by someone: .92 in r/therapyGPT and .75 in the human-therapy communities, against .56 in the general-chatbot communities and .41 in the companion communities. First-person use is a stricter target and the flag is correspondingly weaker on it, from .44 in r/therapyGPT down to .23 for the AI-mention flag in the human-therapy communities. The dominant non-target category is third-party discussion of the same subject rather than an unrelated document: it accounts for .46 of flagged documents in r/therapyGPT and .36 in the human-therapy communities. In the companion communities the three non-target categories are of one size, with .30 describing no relevant chatbot use at all and .27 companionship framing, which is the boundary the structural results are about, so the flag’s errors there are themselves informative. r/therapyGPT is the highest on both definitions, and the companion communities are the lowest on the discourse definition, which tracks the separation the structural results find.

User level. Table 1 counts documents, but the crossover rates and the co-membership network count users, and a user is flagged if any one document matches. User-level precision is .58 in r/therapyGPT, .46 in the companion communities and .18 in the general-chatbot communities. It does not run far above document-level precision because 76.6 percent of flagged chatbot-side users are flagged by exactly one document (mean 1.60).

Crossover contrast. A low precision does not by itself move a ratio of two quantities measured the same way. The ratio would move if precision depended on where the crossing user came from. We test this directly: every one of the 43 disillusionment-community members who

also write flagged general-chatbot content, and 90 mainstream-therapy crossers, coded blind to origin. User-level precision is .58 (95 percent CI .43 to .72) for the disillusionment crossers and .53 (.43 to .63) for the mainstream crossers, the two intervals overlap across most of their range, so we find no sign that the flag is more credulous about one group of crossing users than the other, though 43 and 90 users cannot rule out a small difference. Applying each estimate to its own crossover rate raises the contrast rather than lowering it, from 5.7 to 6.2 times. Both figures are well above the .18 precision of the general flagged-user pool, because a user who appears in a therapy community and writes flagged chatbot content is a much cleaner case than a user drawn from the flagged pool at large; the crossover analysis therefore uses a better-measured population than the corpus-wide precision suggests.

Recall. Against the broader lexicon, which is the estimable comparison, recall is about .39: the narrow flag recovers roughly two of every five documents the broad lexicon would also select, and the miss rate in the BROAD-ONLY stratum runs from .07 in r/ClaudeAI to .34 in r/therapyGPT. The NEITHER stratum is a different matter. Documents matching no mental health word at all are almost never on construct in the general-chatbot and companion communities, at 1 of 533 coded, but they are in r/therapyGPT, at 11 of 110, where the community context supplies what the words do not. That stratum holds 15.3 million documents, so no feasible sample bounds its contribution tightly, and we do not report a recall figure over the whole corpus.

Group	AI	Discourse	Own use	User
AI self-help	.93	.92	.44	.58
General chatbots	.94	.56	.31	.18
Companion chatbots	.89	.41	.26	.46
Human therapy	.98	.75	.23	—

Table 8: Precision of the flag, 200 coded documents per group. “AI” = the document refers to an AI system; “Discourse” = it is about therapeutic chatbot use by anyone; “Own use” = it is the writer’s own therapeutic use; “User” = the account has at least one own-use document among up to three coded (50 users per group; the AI-mention flag is the one applied to the human-therapy communities, for which a user-level estimate is not used anywhere in the paper). Wilson intervals are in the text.

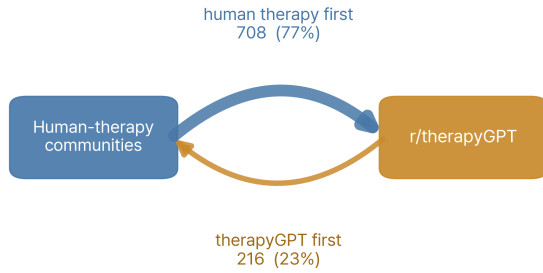


Figure 4: Users shared between human-therapy communities and r/therapyGPT, and the order in which they arrived. Among the 924 shared users whose first human-therapy activity falls in the ChatGPT era, 77 percent were active in human therapy first.

F Directionality robustness

Among users active in both a human-therapy community and the general chatbots, the chatbot-first share is stable across window anchors and a minimum-activity threshold (Table 9), and falls to its lowest in the single community with complete comment coverage, so cleaner measurement moves the estimate toward human-first. For the AI self-help community specifically, of the 924 users (among the 1,112 active in both r/therapyGPT and a human-therapy community) whose first human-therapy activity falls in the ChatGPT era, 77 percent were active in the human community first (Figure 4). The Poisson rate model for crossing into therapeutic chatbot use, which is estimated on a different and much larger set of accounts, is in Appendix G.

G Rate model

Table 10 reports the full Poisson rate model for crossing into therapeutic chatbot use (outcome:

Specification	<i>n</i>	Chatbot first
Post-release window	959	33.7%
+ activity threshold	444	29.3%
Window anchored +3 months	932	34.7%
Window anchored +6 months	891	35.2%
Complete coverage (r/OpenAI)	54	24.1%

Table 9: Share arriving at a chatbot before human therapy, across window anchors and a minimum-activity threshold. Human first is the robust direction across all specifications.

first therapeutic chatbot post; offset: log person-time at risk; $n=93,751$ accounts, 963 events, 58,880 person-years), with all covariates shown. M1 treats disillusionment-community membership as ever-observed and gives a raised but modest association ($IRR \approx 3.4$); this specification is endogenous, since 31 disillusionment-community members cross at all and 19 percent of them join that community only after their first therapeutic chatbot post. M2 restricts membership to precede the at-risk window, and the membership effect is then not identifiable (too few pre-window crossers). Activity and account age are stable across both models. We therefore report the cross-sectional pattern (Table 4) and decline the temporal claim.

Covariate	M1 IRR (95% CI)	M2 IRR (95% CI)
Disillusion. member	3.41 (2.4–4.9)	n.i.
Log activity	0.83 (0.79–0.88)	0.84 (0.80–0.89)
Account age	0.83 (0.78–0.89)	0.83 (0.78–0.88)

Table 10: Poisson rate model for crossing into therapeutic chatbot use ($n=93,751$; 963 events; log person-time offset). M1: disillusionment membership ever-observed (endogenous, since membership can postdate the outcome). M2: membership restricted to precede the at-risk window, where the membership effect is not identifiable (n.i.). Activity and account age are significant at $p < .001$ in both models.