Taking a turn for the better: Conversation redirection throughout the course of mental-health therapy

Vivian Nguyen[•] **Sang Min Jung**[•] **Lillian Lee**[•]
Thomas D. Hull[TS] **Cristian Danescu-Niculescu-Mizil**[•*]
vn72@cornell.edu sj597@cornell.edu llee@cs.cornell.edu
derrick@talkspace.com cristian@cs.cornell.edu
[•] Cornell University [TS]Talkspace

Abstract

Mental-health therapy involves a complex conversation flow in which patients and therapists continuously negotiate what should be talked about next. For example, therapists might try to shift the conversation's direction to keep the therapeutic process on track and avoid stagnation, or patients might push the discussion towards issues they want to focus on.

How do such patient and therapist redirections relate to the development and quality of their relationship? To answer this question, we introduce a probabilistic measure of the extent to which a certain utterance immediately *redirects* the flow of the conversation, accounting for both the intention and the actual realization of such a change. We apply this new measure to characterize the development of patient-therapist relationships over multiple sessions in a very large, widely-used online therapy platform. Our analysis reveals that (1) patient control of the conversation's direction generally increases relative to that of the therapist as their relationship progresses; and (2) patients who have less control in the first few sessions are significantly more likely to eventually express dissatisfaction with their therapist and terminate the relationship.

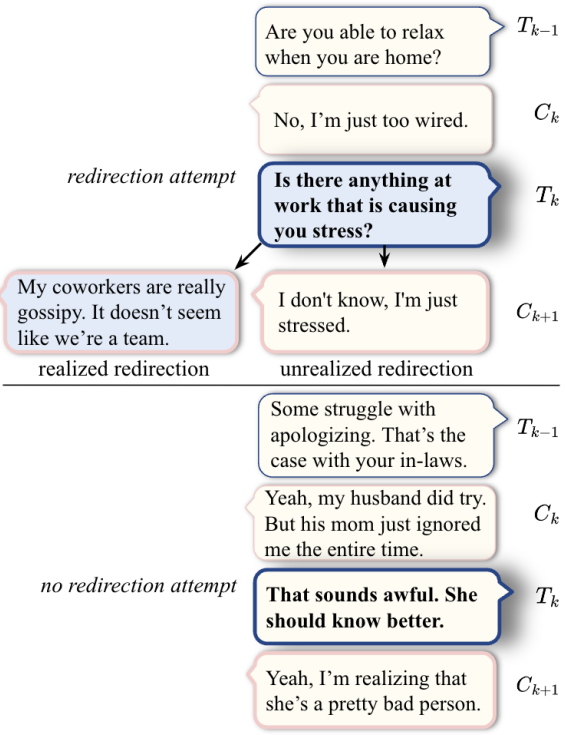

Figure 1: **Top:** Examples of attempted redirection, both realized (left) and unrealized (right). **Bottom:** Example where redirection is not attempted. T_k and C_k refer to the therapist's and patient's k^{th} utterance, respectively. Note that in general, both parties can redirect.

1 Introduction

Mental-health therapy conversations are remarkably and consequentially complex. They involve an ongoing negotiation between a patient and a therapist regarding what should be talked about, how it should be talked about, and by whom. Conversational bids to shift the direction of the discussion (henceforth *redirections*) are a common and essential aspect of any therapeutic relationship. However, we lack an understanding of how these dynamics play out throughout the course of the therapeutic relationship and how they might relate to its quality.

One key aspect of redirections that may explain why they have yet to be studied in a systematic and rigorous way is their joint, rather single-utterance, manifestation. In contrast to other commonly studied strategies and conversational acts (Malgaroli et al., 2023a) that tend to be executed by either the therapist (e.g., empathy (Sharma et al., 2020)) or the patient (e.g., change talk (Park et al., 2019)), redirection is a truly shared patient-therapist act: to be realized, a bid to redirect the conversation must be accepted by the other participant.

Our **first contribution** is to introduce a measure to quantify an utterance's redirection effect in

*Senior corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 9507–9521

a way that explicitly takes into account this two-part nature: it jointly captures both the *intention* to change the course of the conversation and the actual *realization* of this change through a reply that complies with this intention. Figure 1 (top, left branch) illustrates an example of an utterance with high redirection effect: the therapist intends to switch the focus towards specific sources of stress, and the patient conforms by mentioning frustration at work. In contrast, if the patient resists the switch (Figure 1, top, right branch), the redirection is not realized. There is a third alternative: Figure 1 (bottom) shows an utterance with low redirection effect due to the lack of intention to redirect.

Key to our work is capturing this type of interplay between two participants at a specific utterance juncture, rather than a discourse-level notion of shift, where a single person can change the focus even on their own (e.g., topic control (Nguyen et al., 2014, inter alia)). Our redirection measure's inherent joint nature is particularly well-suited for examining the therapeutic relationship, and thus adds to the toolkit of computational methods available for studying therapeutic practice (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020; Imel et al., 2024; Yang and Jurgens, 2024). We illustrate this in collaboration with Talkspace—a large online text-based therapy platform—by applying it to characterize the development and perceived quality of the patient-therapist relationship.

Our analysis reveals that as their relationship develops, both patients and therapists redirect to a lesser extent, suggesting an overall smoother and more focused conversation flow. However, in relative terms, patients gain increasing control of the direction of the conversation in comparison to the therapists; **this discovery is our second contribution**. Furthermore, as **our third contribution**, we find that patients are more likely to eventually express dissatisfaction with the therapeutic relationship and terminate it when they are not able to redirect the conversation in the first few sessions. Cumulatively, these results underline the importance of considering patient agency in psychotherapy (Carey, 2010; Huber et al., 2021), or, concomitantly, therapist willingness to follow the patient's lead.

While here we focus on the mental-health domain, **one final contribution** is to highlight the possibilities in examining any sort of longer-term relationships developed through multiple conversations over time. Future work could apply our measure and analysis framework to study the conversational process in other conversation-rich settings where complex long-term relationships are developed, such as education (e.g., advisor-advisee relationships). To encourage such work, we make our code publicly available as part of ConvoKit (Chang et al., 2020), together with a demo on a publicly-available dataset in a different domain, US Supreme Court oral arguments (Appendix C).[1]

2 Therapy Setting

Long-term text-based therapy. We develop our methodology in the context of Talkspace, a telehealth therapy platform. Patients can choose between different plans, which may include video therapy alongside messaging therapy. After a consultation in which they answer a few questions about their symptoms and describe their preferences, patients are matched by the platform to a suitable therapist who is licensed in their state.

In this work, we use the (English-language) text-based conversations (after redaction of personally identifiable information by Talkspace) from a five-year span.[2] During this time, Talkspace hosted over 18,000 licensed therapists providing services to over 300,000 patients. In total, over 65 million messages were exchanged.

Therapies can be sustained over long periods of time, with more than 26,000 therapies lasting over a year and 17,000 therapies comprising over 500 messages. This setting thus provides an opportunity to study the long-term conversational dynamics as the therapeutic relationship develops.

At any point, patients can either cancel their therapy or switch to a different therapist. When doing so, they are asked to provide a reason, either by selecting from a drop-down menu (e.g., "The treatment provided by my therapist was not helpful," "I met my goal / I feel better") or by entering free text. We have some assurance that these reasons are authentic because patients are informed that therapists cannot access them. These reasons serve as imperfect indicators of the perceived quality of the therapeutic relation (for a broader discussion of difficulties in operationalizing the quality of the therapeutic relationship, see Section 8).

Session identification. One challenge in studying the development of therapeutic relationships in this

[1] https://convokit.cornell.edu/

[2] The use of the therapist and patient data is done with their consent, and this research was approved by the Cornell IRB.

text-based setting is accounting for the wide variety of durations, tempos, and volubility exhibited in different therapies. A longitudinal analysis requires a methodology that captures the progression of the therapeutic process while allowing for meaningful aggregation across therapies with these vastly different interaction patterns.

To address this challenge, we need to account for several factors. First, each therapist-patient pair develops a unique interaction style and tempo, including frequency of interaction, response time, and how often turns are split into multiple messages. Second, exchanges between a therapist and a patient include not only therapeutic conversation, but also short check-ins that deal with brief updates, scheduling, or survey completions. Last, an important component of many therapies is the occurrence of spans of *approximately synchronous* conversations, or *sessions*, over the therapy's duration. These three factors render conventional units of progression in conversational analysis—such as time or number of utterances—unsuited for this complex interactional setting.

Explicit session boundaries are lacking in text-based conversations and so must be inferred. We start by distinguishing high-activity periods interwoven with periods of little or no activity, using the method from Kushner and Sharma (2020) to capture these bursts. For each therapy, we define session-split points as moments when an utterance's reply time exceeds $N\times$ (*median reply time of that entire therapy*).[3] We further ensure that each session represents an actual conversation between the patient and the therapist (as opposed to, for example, a short notification and acknowledgement regarding scheduling) by filtering out short exchanges—any burst of activity with fewer than four turns (so that what remains includes at least two nonconsecutive utterances from each speaker)—or includes a video session. We further remove any automated messages related to video sessions, survey completion, or scheduling.

To validate the quality of the session splits, we sampled therapies with at least 10 sessions and ran a Bayesian distinguishing-word analysis (Monroe et al., 2017) on the first and last utterance of each session. The distinguishing words are intuitive: sessions start with greetings ("hi", "morning", "hey", "how") and end with a farewell or expression of gratitude ("enjoy", "welcome", "thank", "thanks").

[3]We settle on $N = 100$; see Appendix A.1 for details.

3 Redirection Measure

3.1 What redirection is (not)

We aim to quantify the redirection effect of a given utterance in a conversation: the extent to which it alters the immediate focus of the conversation. We start from the realization that for an utterance to have a high redirection effect, it must (a) attempt to put the conversation on a different course (intention of redirection), and (b) receive a reply that is compliant with this redirection (realization of redirection). We will examine previous methods that capture these two conditions separately and explore how they guide us toward a design for a measure that jointly accounts for both of them.

For concreteness, we employ the convention (and notation) from Figure 1 of centering the discussion on a therapist's utterance T_k; but the analogous definitions apply to a patient's utterance C_k.

Orientation. Orientation (Zhang and Danescu-Niculescu-Mizil, 2020) captures the degree to which an utterance *intends* to move the conversation away from what was already discussed. A "forwards-oriented" utterance (high orientation) is one that intends to advance the conversation towards a specific target, and thus is *expected* to be followed by a reply centered on that target. A "backwards-oriented" utterance (low orientation) is one that intends to address what was previously said, and thus is *expected* to follow a specific utterance. Thus, orientation characterizes an utterance's intended objective, regardless of whether that intention is realized or not.

In Figure 1 (top), T_k is a high-orientation utterance since the therapist intends to redirect the conversation by asking the patient about a specific source of stress at work. This remains true regardless of whether the patient conforms in C_{k+1} by aligning themselves with the newly suggested direction (by introducing their frustration with gossip; left branch), or resists the redirection attempt (by maintaining the focus on their general state of stress, rather than delving into specifics; right branch).

Similarity difference. To incorporate the actual reply, one may simply compute the similarity of the reply C_{k+1} with the original utterance T_k and compare it against a reference point to account for the ongoing direction of the conversation. A potential reference point is the extreme scenario when the therapist wishes to ensure no redirection will take place by simply repeating their previous

utterance T_{k-1}. In other words, the redirection of utterance T_k could be formalized as the difference between the similarity of C_{k+1} and T_k and that of C_{k+1} and T_{k-1}.

Unlike orientation, this measure considers the actual reply; but it hinges on the assumption that semantic similarity can sufficiently capture the redirection realization. Successful redirection, however, is not synonymous with similarity between utterance and reply. Two examples in Figure 1 (top, left vs. right branch) illustrate this: while they intuitively show different levels of redirection, in both cases the similarity difference is high.

Uptake. For a more nuanced take on redirection realization that goes beyond similarity, we could aim to measure the reply's "uptake": how much it answers, acknowledges, or builds upon the previous utterance. Demszky et al. (2021) formalize this concept (in the context of student-teacher interactions) as the dependence of the reply on the utterance via point-wise Jensen-Shannon divergence, which uses next-sentence *prediction* to quantify the extent to which a given reply is a *probable* response.

While capturing redirection realization better than similarity, uptake does not account for the redirection intent. For instance, two examples in Figure 1 (top left branch vs. bottom) have high uptake, with the patient replies being consistent with the respective therapist's utterance. They, however, show different levels of redirection, as only in the first example does the therapist intentionally shift the focus of conversation.

Our approach: Redirection. We combine the insights of these existing measures to develop a new redirection measure that captures both the intention and realization components. In particular, our analysis of the previous measures points toward the need to consider the actual reply the utterance receives, to use a predictive component to capture the extent to which the reply naturally follows the utterance, and to use a point of reference to capture a change in direction.

More concretely, to quantify the redirecting effect of an utterance, we first consider the likelihood of the patient's reply given the previous context:

$$\mathcal{P}_k(C_{k+1}) \triangleq P(C_{k+1}|C_k, T_k).$$

We condition on the most recent utterances from both speakers, C_k and T_k, to capture both of their prior conversational context.

We use as point of reference the extreme scenario wherein the therapist simply repeats their previous utterance T_{k-1} as T_k, demonstrating absolutely zero intent to redirect:[4]

$$\mathcal{Q}_k(C_{k+1}) \triangleq P(C_{k+1}|C_k, T_{\mathbf{k-1}}).$$

We then formally define the *redirection* of T_k as the log-odds ratio of these two probabilities:

$$R(T_k) \triangleq \log\left(\frac{\mathcal{P}_k(C_{k+1})}{1-\mathcal{P}_k(C_{k+1})} \Big/ \frac{\mathcal{Q}_k(C_{k+1})}{1-\mathcal{Q}_k(C_{k+1})}\right) \quad (1)$$

When redirection is high, $\mathcal{P}_k(C_{k+1}) \gg \mathcal{Q}_k(C_{k+1})$, T_k alters the direction of the conversation, in the sense that the patient's reply C_{k+1} is likely as a reply to this utterance, but not as a continuation of the previous direction of the conversation. Conversely, in cases of low redirection, the patient's response is less affected by the therapist's utterance.

We compute the redirection of a patient's utterance in the corresponding way.

Operationalization. To compute the probabilities for our redirection measure, we fine-tuned the Gemma-2B model (Mesnard et al., 2024) with 4-bit QLoRA (Dettmers et al., 2023) using a held-out dataset of $8,000$ therapy conversations.[5]

The operationalization for the related measures discussed above is included in Appendix B.1.

3.2 Session-level aggregation

To apply and compare our utterance-level measures across therapy sessions, we use two aggregation methods. The first method averages the utterance-level measures per session for each speaker. A speaker's average redirection in a session is computed as the mean redirection of their utterances in that session, which we denote with T_{avg} and C_{avg} for the therapist and patient respectively.

The second method examines the balance of redirection between speakers to determine who is redirecting the conversation more in a given session. Analyzing the balance of redirection can reveal how the control of the discussion is shared between

[4] One naive alternative might be to calculate $\mathcal{Q}_k(C_{k+1})$ without conditioning on the previous therapist utterance T_{k-1}, to simulate the "absence" of any therapist utterance. This alternative, however, creates an unnatural situation with the patient replying to themselves (which would artificially have a low probability). Furthermore, such an alternative would not be directly comparable with $\mathcal{P}_k(C_{k+1})$, which is conditioned on the previous utterances of both speakers. Therefore, we employ a reference point with equivalent conditioning.

[5] The models were trained on internal GPU servers in consideration of the sensitive nature of the data. Training details are included in Appendix B.1.

the patient and therapist. Equation 2 formalizes relative redirection from the therapist's perspective:

$$T_{rel} = \frac{\exp(T_{avg})}{\exp(T_{avg}) + \exp(C_{avg})}. \quad (2)$$

The relative value for the patient C_{rel} is computed the corresponding fashion. Note that mathematically, $T_{rel} + C_{rel} = 1$.

4 Redirection in Online Therapy

We employ our formalism to study redirection in online text-based therapies. We first validate that our redirection measure corresponds to human intuition in the therapy setting. We then explore how redirection relates to the development of the therapeutic relationship. We move on to study the relation between redirection and the patient's perception of the quality of the therapy. Finally, we apply the related measures discussed in Section 3 to gain additional insight into which interactional dynamics might account for the observed trends.

4.1 Validation studies

Human validation. To check how aligned our measure is with our intuitive understanding of redirection in this particular online therapy setting, we designed a redirection identification experiment. A participant is shown a pair of two short interactions, each consisting of 4 utterances (like the examples in Figure 1). Their task is to pick which one of them has a higher redirection effect in the second to last utterance (e.g., T_k in Figure 1).

Each pair for this experiment is constructed by first picking a random therapy, selecting the utterance with the highest and lowest redirection effect according to our measure, and considering the interaction surrounding those utterances. We select 10 such pairs using our redirection measure, and for comparison, we select 10 more pairs for each of the other measures discussed in Section 3. If a measure properly captures the redirection effect, we expect it to match the human rankings.

To respect the privacy of the data and adhere to IRB protocol, the task was performed by one of the authors who was authorized to read the therapy text. To avoid author bias, we administered the task in a blind fashion, with the participant not knowing which pairs were selected using which measure.

The participant ranking perfectly matched that of our redirection measure, compared to 8 out of 10 matches for the similarity difference, 5 out of 10 for orientation, and 4 out of 10 for uptake. While limited in scale due to the privacy restrictions on the data, these results suggest that our measure does indeed capture an intuitive notion of redirection which the other measures do not. The experiment also inspired a qualitative analysis comparing examples in which the metrics disagree (Section 5).

Shuffle check. Given that redirection is inherently an aspect of conversation flow, any proposed measure of it should be sensitive to utterance order. To check for this intuitive property, we can *shuffle* a session's utterances: we would expect no redirection to occur in this shuffled setting.

Indeed, while the average redirection across all (non-shuffled) sessions is 6.40, in the shuffled case it is near zero (-0.025; $p < 0.0001$ for the difference, Wilcoxon signed-rank test).

In our main analysis below, we use the shuffle check to ensure that the trends we observe are actually related to conversational processes, as opposed to spurious non-conversational phenomena such as changes in session or utterance length.

4.2 Evolution of the therapeutic relationship

We now present our main analysis: connecting redirection to the development and quality of the therapeutic relationship. We start with examining the change in redirection as the therapist-patient relation develops, focusing on sustained therapies having at least 10 sessions (3,764 in total).

Figure 2 shows the average redirection measure T_{avg} and C_{avg} in each session for the first 5 and last 5 sessions. For both the patient and therapist, there is a downward trend in local redirection as therapy progresses ($p < 0.0001$ according to a

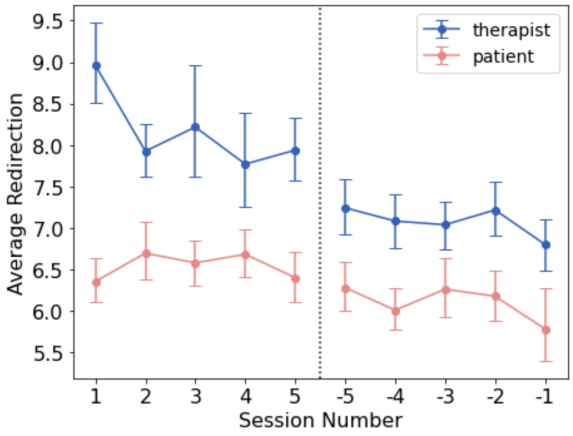

Figure 2: Average redirection across the first 5 and last 5 sessions. Throughout, error bars indicate 95% confidence intervals estimated through bootstrap sampling.

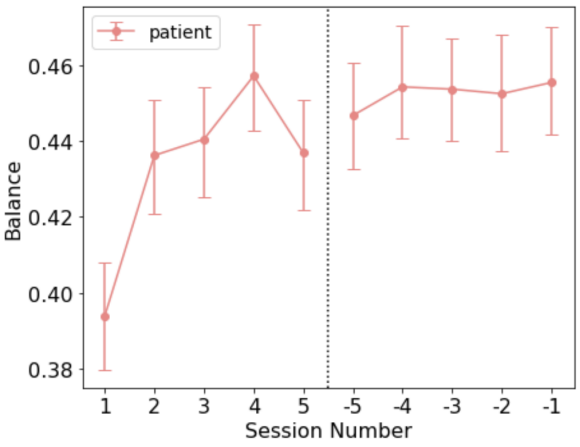

Figure 3: Patient redirection relative to that of the therapist across the first 5 and last 5 sessions.

Wilcoxon signed-rank test comparing the first and last 5 sessions). Thus, as the therapy progresses, redirection in conversation occurs less from both speakers, perhaps suggesting a smoother flow of the conversation with more stable conversational goals (see Section 5 for a qualitative analysis).

We now switch from considering each speaker's redirection separately to examining how the *balance* between the two evolves over time. Figure 3 shows the patient's relative redirection C_{rel} across the first 5 vs. last 5 sessions. As the relationship progresses, the patient share of redirects increases ($p < 0.0001$), suggesting that the patients gain more relative control over the flow of the conversation.

We re-analyze the trends after shuffling the utterances within a session. After the shuffle, all observed trends in the original data disappear (no statistical difference), suggesting that they are indeed tied to the conversation dynamics.

4.3 Unsuccessful relationships

We now investigate how redirection is related to the quality of the therapeutic relationship. For this purpose, we consider a therapeutic relation to be "unsuccessful" if the patient eventually abandons it via a request to switch therapists or to cancel the subscription *and* provides a reason that explicitly cites dissatisfaction with the therapist or with the relation (see Appendix A.3 for a list of such reasons). After filtering out therapies that had fewer than 3 sessions (to ensure that the patient actually "gave it a try"), we are left with 817 such unsuccessful relations. For a meaningful comparison, we consider a control group with the same number of therapies where the patient did not request to cancel or switch the therapist, and that also has at least 3 sessions. To discard effects due to the duration of the therapy and focus on signals that could be perceived early in the therapeutic process, we only consider the first 3 sessions in this analysis. Table 1 shows that in unsuccessful relationships, patients redirect the conversations less than in the control group ($p = 0.018$; Mann-Whitney U test), while we do not find any difference for therapists ($p = 0.5$). We note these differences disappear after shuffling the order of the utterances in the session ($p = 0.17$), thus they reflect differences tied to conversation dynamics.

	Unsuccessful	*Control*	**p-Value**
Actual	6.06	6.91	0.018*
Shuffled	0.14	0.043	0.17

Table 1: Patient average redirection is smaller at the start of (eventually) unsuccessful relationships (statistical significance indicated with *; $p < 0.05$). This effect disappears in the shuffled setting, as shown by the lack of statistical significance marker.

4.4 Comparison to other measures

We also explore how our redirection measure compares to the other related measures discussed in Section 3. The results are summarized in Table 2. It is worth noting that none of the related measures show a significant difference between the "unsuccessful" therapies and the control group, and only the similarity difference measure exhibits temporal trends that pass the shuffling test.

We highlight here the results for orientation, since by capturing the redirection intent separately, they provide further context for interpreting the results discussed above. Orientation shows a significant downward trend for therapists; thus the observed decrease in redirection can be at least partially attributed to the decrease in their attempts to change the course of the conversation (rather than to the patient's unwillingness to realize those attempts). Furthermore, there is no difference in patient orientation between the "unsuccessful" and control groups ($p = 0.5$), suggesting that reduced patient redirection is not due to lack of patient redirection bids, but rather because the therapist is not accepting those bids.

5 Qualitative Analysis

We conduct a qualitative analysis of high- and low-redirection examples to explore therapy strategies that are tied to this phenomenon and to further

Analysis	*Orientation*	*Similarity Difference*	*Uptake*	***Redirection***
Start/end difference in average value	- / ↓	↓ / ↓	↑ / ↑	↓ / ↓
Start/end difference in balance	↑ / ↓	- / -	- / -	↑ / ↓
Reflects temporal order (i.e., does shuffling remove the trend)	No	**Yes**	No	**Yes**
Distinguishes unsuccessful relations	No	No	No	**Yes**

Table 2: Comparing redirection with three related measures. Slashes ("/") separate patient from therapist effects. ↑ indicates a higher level in the end than the start; ↓ indicates vice-versa; "-" indicates no significance.

interpret the observed trends. Additionally, we analyze examples where the redirection measure deviates from the related measures to check our intuition about differences in what they capture. The discussion follows examples from Table 3.

Redirection strategies. In our setting, highly redirecting therapist utterances typically involve exploration and surveying. Therapists tend to use open-ended questions to assess their patients' given situations and viewpoints (Example 1: *"Where does this [issue] originate from"*). They may also suggest different perspectives or provide guidance on addressing these issues (Example 2: *"Try sitting with your feelings ..."*). Conversely, less-redirecting utterances tend to echo or validate the patient's statements or empathize with their struggles. They may also reformulate what the patient has said and keep the focus on the immediate problem at hand (Example 4: *"Sounds awful. Is it..."*).

For patients, highly redirecting utterances tend to introduce personal experiences or feelings that shape the immediate course of the conversation. They may initiate subjects they wish to explore or verbalize the specific obstacles they are facing (Example 6: Challenges at work). Low-redirecting patient utterances, on the other hand, extend the current subject of discussion.

These observations are consistent with the decreasing trend in redirection for both speakers. Initially, patients and therapists both focus on introducing the context of the therapy and setting therapeutic goals. Therapists tend to actively discuss the objectives of the therapy and suggest behavioral strategies and cognitive techniques. Patients, likewise, will share their background or expectations as they start out their therapeutic relationship. Later sessions, conversely, usually consist of more concentrated discussions on the identified issues.

Comparison with other metrics. We also examined cases where the redirection measure is at odds with the related measures. As expected, orientation can be high even when redirection is resisted. For instance, in Example 3, the therapist attempts to steer the conversation towards discussing who to talk to, but the patient disregards their suggestions and continues explaining their plans, thus causing redirection to be low.

Importantly, unlike similarity difference, our metric captures utterances with redirection effects even when the reply is not semantically similar. In Example 2, the therapist redirects the focus of the conversation by suggesting a potential solution to the patient's problem, which the patient acknowledges in their reply. While similarity difference is low since acknowledging the suggestion is not semantically similar, the patient does reflect on the therapist's suggestion, which deviates from their prior discussion focused on the problem.

Our measure also differs from uptake in that it uses a reference point: the previous conversational context. Example 7 illustrates a segment where the patient reflects on having someone to vent to. The therapist addresses the patient's utterance by continuing to encourage the patient to vent, exhibiting high uptake. However, the focus of the conversation does not change, indicating low redirection.

6 Further Related Work

Our work relates to prior research on conversational dynamics, analyzing the development and quality of therapeutic relationships, and exploring language on mental-health platforms.

Development and quality of therapeutic relationships. The development of the patient-therapist relationship and its impact on therapy outcomes have been extensively studied (Gelso and Carter,

#	*Example*	*R*	*O*	*SD*	*U*
1	T: It seems like this anxiety doesn't arise when you're in a relationship with someone. P: Now that I remember no. I feel more secure. **T: Where does the strong reaction to being alone originate from?** P: Honestly, I think it has to do with my relationship with my family.	High	Low	Low	Mid
2	T: It seems like the issue is your lack of control over being alone. P: Yes, that's interesting; I think that's a part of it. **T: Try sitting with your feelings and exploring them without blocking them, and see what insights come up for you.** P: Okay, I'll meditate on that. Thanks.	High	Mid	Low	Mid
3	T: I believe using conflict resolution methods might help you better communicate with your husband about this issue, rather than withdrawing and shutting down. P: He keeps saying he doesn't want to get involved with her problems. **T: Then, are you planning to talk to your daughter first, or are you waiting to speak with her counselor?** P: I'll just ground her and take things away.	Low	High	Low	Mid
4	T: When you are ready to process the trauma, we can go through it. P: Now, I'm feeling better mentally, but physically my heart is thumping, and I began having panic attacks at work. **T: That sounds awful. Is it a thumping sensation in your chest or feeling in your ears?** P: I wore a heart monitor in my chest. I feel a thump each time before my heart skips beats.	Low	Low	High	High
5	P: I dislike how I get easily attached to a man I like. I had thoughts of being in a relationship just because we were spending time with each other. T: Some people are attachers in relationships, while some people it takes a while for the heart to thaw out. **P: I feel like it may be because of not having my dad present.** T: I think that is a good insight on your part. Your upbringing and personality formation definitely affects attachment styles.	High	Low	Low	Low
6	P: How was your day? T: I'm also waiting for warm weather; I think today's supposed to be pretty nice! How are things going at home for you? **P: It's going okay. I'm nervous about having to hit numbers at work every day though. That is going to be a challenge.** T: I understand productivity requirements. It sounds like you have an amazing work ethic and your trainer is already impressed with you though.	High	High	Mid	Low
7	P: I'm going to take a walk tonight if weather permits. T: The walk sounds like a great idea. It is good to let it all out, so don't be afraid to vent as much as you need to. **P: Thanks. Not used to having someone to vent to.** T: Hope the rest of the day went better and the weather allows you to take a walk. Feel free to vent as things come up.	Low	Low	Low	High
8	P: When I have trouble concentrating, I tend to eat. T: Does that help you concentrate better? **P: I don't know.** T: So what's behind eating when you are having trouble concentrating?	Low	Low	Low	Mid

Table 3: Qualitative examples comparing **R**edirection with related measures **O**rientation, **S**imilarity **D**ifference, and **U**ptake, all applied to the colored utterance (Therapist for the first four examples, Patient for the remaining four). Low/High/Mid: the bottom 25th percentile of the measure values for the respective speaker type, the top 75th percentile,and 25-75th percentile, respectively. Examples are paraphrased to preserve privacy.

1985; Norcross, 2010). Previous work characterized therapeutic alliance (Goldberg et al., 2020) and ruptures (Tsakalidis et al., 2021) through emotional engagement (Christian et al., 2021), sentiment (Syzdek, 2020), linguistic coordination (Nasir et al., 2019), and synchrony (Doré and Morris, 2018). Our work characterizes an additional dimension of the patient-therapist relationship based on the joint act of redirection.

Conversational dynamics. Prior computational work examined different aspects of conversation flow (Zhang et al., 2016), including work on topic segmentation (Eisenstein and Barzilay, 2008; Nguyen et al., 2014; Purver, 2011; Glavas and Somasundaran, 2020; Jiang et al., 2023) and topic shift (Xie et al., 2021). While these conversation-level concepts are related, our probabilistic measure seeks to quantify the immediate effect of a

single utterance, an effect different from semantic similarity and topic coherence. Furthermore, we develop a framework for a *longitudinal* analysis of conversational dynamics in the distinctive domain of online mental-health therapy.

Online mental-health platforms. Prior literature extensively explores linguistic behaviors and conversational choices users make in online mental-health-related platforms, whether in crisis counseling platforms (Althoff et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020), peer-support platforms (Yang et al., 2017; Pruksachatkun et al., 2019; Yang and Jurgens, 2024), or therapy platforms (Malgaroli et al., 2023b). Accompanying growing attention to online mental-health resources and use of technology in psychotherapy (Anthony, 2003; Barak et al., 2008), many studies highlighted the benefits of online social support (De Choudhury and Kıcıman, 2017; Newman et al., 2011). We specifically focus on *long-term* sustained relationships in text-based therapy platforms, offering a novel perspective on development of the patient-therapy relation through longitudinal analyses.

7 Discussion and Conclusion

Redirection is a joint act often performed by therapists and patients. Its realization requires both speakers to understand the initial direction of the conversation, one's attempt to redirect it, and the other's compliance with this attempt. As such, its study has the potential to characterize the patient-therapist dyad, in particular with respect to their ability to negotiate the focus of the conversation.

In this work, we introduce a computational method for quantifying the redirection effect of an utterance and apply it to a mental-health-therapy domain. We find that both the patient and the therapist redirect less as therapy progresses. Our qualitative analysis suggests that after initial exploration and contextualization prompted by both speakers, the relationship matures to a more stable stage with less redirection. Moreover, we reveal that the less patients redirect early on, the more likely they are to eventually express dissatisfaction with the therapist and abandon the relationship.

Connection with psychotherapy literature. Several decades of research in psychotherapy suggest that the "treatment model" approach in which an expert clinician provides curative treatment that the patient passively receives is not well supported by outcomes or engagement data (Bohart, 2000; Duncan and Miller, 2000). Instead, shifting focus from which aspects of treatment are delivered to how dyads collaboratively negotiate therapeutic conversations has provided evidence of improved access, efficiency, and effectiveness of mental-health services for "patient-led" approaches (Carey, 2010; Carey et al., 2013; Huber et al., 2021). Our results offer a computational perspective into one dimension of patient agency in therapeutic relationships.

Future work. A computational approach to redirection enables us to observe the evolution of therapeutic relationships and could assist therapists in fostering them. A complementary line of work can include a more in-depth exploration of how to foster healthy levels of redirection in therapy and the causal relationship between redirection and the quality of therapeutic relationships. For instance, labeling messages with therapeutic strategies and techniques can provide insight into the effectiveness of each strategy and determine which ones are more applicable for specific contexts (e.g., when patients resist therapists' redirection attempts).

In their current observational form, these results suggest that early conversational patterns can signal the eventual dissatisfaction of the patient. Future work could examine the predictive power of these ties and test the extent to which they might be explained by other (unalterable) factors, such as the characteristics of the patient or of their condition.

Our methodology can extend beyond the mental-health domain and be applied to other conversation-rich domains, such as education, online discussions, or interviews, where discussions are carried out in relatively unstructured ways with only a few general agendas set. Exploration of redirection in these contexts may present a unique perspective into how speakers are able to redirect and control the flow of the discussions.

Finally, understanding how humans redirect the flow of conversations is important for supporting more naturalistic human-AI conversations, where the AI could pick up on humans' redirection attempts and initiate their own.

8 Limitations

In the mental-health context, how to define successful therapeutic relations is always an open question. Our work uses patient-provided reasons for canceling the therapy or switching therapists that explicitly mention dissatisfaction with the therapist as an imperfect indicator for failure of the therapeutic

relationship. A reliable positive signal, on the other hand, is difficult to define, especially in a therapy domain where patients can have hidden agendas, present misleading information, or choose to remain in treatment for a number of reasons (Newman and Strauss, 2003). We thus do not consider the control group to necessarily contain "successful" relationships, and thus the interpretation of the results of the comparison should be interpreted accordingly.

Furthermore, these results are purely observational, and future work would be needed to establish if intervening to change the amount of realized redirection (e.g., through therapist training) would have an effect on the quality of the therapeutic relationship.

The notion of session we employed captures the structure of therapy that arises as patients and therapists maintain a long term relationship spanning multiple conversations, rather than just a single encounter. However, sessions' characteristics change as time progresses. We observed that the session length and the token count of utterances in sessions both decrease. While our shuffling experiments suggest that our observations are not merely a result of this variability and are tied to the actual dynamics of the conversation, further work is needed to fully account for this variability.

Our work focuses on text-based conversations. While these are a substantial component of online therapy, they do not capture the occasional video conversations that patients and therapists might have. Future work could explore video conversations and the role they have in the development of relationships (although additional de-identification challenges arise when working with video data).

As our redirection measures reflect change in the subject of the conversation, they require the reply of the utterance to be able to be calculated. This constraint can be restrictive in practice, as an automated system would be unable to analyze an ongoing conversation without a reply it hasn't received yet. Exploring how we can predict whether the upcoming reply will cooperate with one's attempt to redirect remains an interesting direction for future work.

Ethical Concerns. As our work involves highly sensitive data in the form of patient chat-therapy history and surveys, all personally identifiable information was removed. Access to the data was strictly limited to the authors of the paper, and data was processed on restricted servers and will be removed upon publication. All large language models trained on the data were strictly internal and discarded after the analysis.

This work is done in close collaboration with the therapy platform Talkspace. All participants consent to the use of their data in aggregate and de-identified format for research purposes as part of the Terms of Service they review during onboarding (in the case of the patients) or as part of information reviews done during the hiring process (in the case of the therapists). All individuals may opt out of the use of their data for research purposes at any time without penalty.

Given the sensitivity of the domain and the limitations discussed above, extensive work is still required before insights from our results could be used in patient-facing systems.

Acknowledgements

We thank Liye Fu, Emily Tseng, Khonzoda Umarova, and Tony Wang for initial pre-processing of the data. We are grateful for insightful discussions with Team Zissou—including Jonathan P. Chang, Yash Chatha, Nicholas Chernogor, Tushaar Gangavarapu, Kassandra Jordan, Seoyeon Julie Jeong, Ethan Xia, and Sean Zhang—and for the feedback we received from the anonymous reviewers. Finally, we would like to express our gratitude to the patients and therapists on Talkspace who generously shared their experiences and therapy data for research purposes. This material is based upon work supported in part by the U.S. National Science Foundation under Grant No. IIS-1750615 (CAREER). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Cornell University, the National Science Foundation, or Talkspace.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Kate Anthony. 2003. The use and role of technology in counselling and psychotherapy. In Stephen Goss and Kate Anthony, editors, *Technology in Counselling and Psychotherapy: A Practitioner's Guide*, pages 13–35. Macmillan Education UK, London.

Azy Barak, Liat Hen, Meyran Boniel-Nissim, and Na'ama Shapira. 2008. A Comprehensive Review

and a Meta-Analysis of the Effectiveness of Internet-Based Psychotherapeutic Interventions. *Journal of Technology in Human Services*, 26(2-4):109–160.

Arthur C. Bohart. 2000. The client is the most important common factor: Clients' self-healing capacities and psychotherapy. *Journal of Psychotherapy Integration*, 10(2):127–149.

Timothy A. Carey. 2010. Will you follow while they lead? Introducing a patient-led approach to low intensity CBT interventions. In *Oxford guide to low intensity CBT interventions*, Oxford guides in cognitive behavioural therapy, pages 331–338. Oxford University Press, New York, NY, US.

Timothy A. Carey, Sara J. Tai, and William B. Stiles. 2013. Effective and efficient: Using patient-led appointment scheduling in routine mental health practice in remote Australia. *Professional Psychology: Research and Practice*, 44(6):405–414.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of SIGDIAL*.

Christopher Christian, Eran Barzilai, Jacob Nyman, and Attà Negri. 2021. Assessing Key Linguistic Dimensions of Ruptures in the Therapeutic Alliance. *Journal of Psycholinguistic Research*, 50(1):143–153.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of WWW*.

Munmun De Choudhury and Emre Kıcıman. 2017. The Language of Social Support in Social Media and its Effect on Suicidal Ideation Risk. *Proceedings of ICSWSM*, 2017:32–41.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B. Hashimoto. 2021. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1638–1653.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs.

Bruce P. Doré and Robert R. Morris. 2018. Linguistic Synchrony Predicts the Immediate and Lasting Impact of Text-Based Emotional Support. *Psychological Science*, 29(10):1716–1723.

Barry L. Duncan and Scott D. Miller. 2000. The client's theory of change: Consulting the client in the integrative process. *Journal of Psychotherapy Integration*, 10(2):169–187.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian Unsupervised Topic Segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.

Charles Gelso and Jean Carter. 1985. The Relationship in Counseling and Psychotherapy: Components, Consequences, and Theoretical Antecedents. *The Counseling Psychologist*, 13(2):155–243.

Goran Glavas and Swapna Somasundaran. 2020. Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7797–7804.

Simon B. Goldberg, Nikolaos Flemotomos, Victor R. Martinez, Michael J. Tanana, Patty B. Kuo, Brian T. Pace, Jennifer L. Villatte, Panayiotis G. Georgiou, Jake Van Epps, Zac E. Imel, Shrikanth S. Narayanan, and David C. Atkins. 2020. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling Psychology*, 67(4):438–448.

Julia Huber, Simone Jennissen, Christoph Nikendei, Henning Schauenburg, and Ulrike Dinger. 2021. Agency and alliance as change factors in psychotherapy. *Journal of Consulting and Clinical Psychology*, 89(3):214–226.

Zac E. Imel, Michael J. Tanana, Christina S. Soma, Thomas D. Hull, Brian T. Pace, Sarah C. Stanco, Torrey A. Creed, Theresa B. Moyers, and David C. Atkins. 2024. Outcomes in Mental Health Counseling From Conversational Content With Transformer-Based Machine Learning. *JAMA Network Open*, 7(1):e2352590.

Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. 2023. SuperDialseg: A Large-scale Dataset for Supervised Dialogue Segmentation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4086–4101, Singapore. Association for Computational Linguistics.

Taisa Kushner and Amit Sharma. 2020. Bursts of Activity: Temporal Patterns of Help-Seeking and Support in Online Mental Health Forums. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2906–2912, Taipei, Taiwan. Association for Computing Machinery.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. *eprint arXiv:1711.05101*.

Matteo Malgaroli, Thomas D. Hull, James M. Zech, and Tim Althoff. 2023a. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):1–17.

Matteo Malgaroli, Emily Tseng, Thomas D. Hull, Emma Jennings, Tanzeem K. Choudhury, and Naomi M. Simon. 2023b. Association of Health Care Work With Anxiety and Depression During the COVID-19 Pandemic: Structural Topic Modeling Study. *JMIR AI*, 2(1):e47223.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4):372–403.

Md Nasir, Sandeep Nallan Chakravarthula, Brian Baucom, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Modeling Interpersonal Linguistic Coordination in Conversations using Word Mover's Distance. *Interspeech*, 2019:1423–1427.

Cory Newman and Jennifer Strauss. 2003. When Clients Are Untruthful: Implications for the Therapeutic Alliance, Case Conceptualization, and Intervention. *Journal of Cognitive Psychotherapy*, 17:241–252.

Mark W. Newman, Debra Lauterbach, Sean A. Munson, Paul Resnick, and Margaret E. Morris. 2011. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 341–350, New York, NY, USA. Association for Computing Machinery.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421.

John C. Norcross. 2010. The therapeutic relationship. In *The heart and soul of change: Delivering what works in therapy, 2nd ed*, pages 113–141. American Psychological Association, Washington, DC, US.

Sungjoon Park, Donghyun Kim, and Alice Oh. 2019. Conversation Model Fine-Tuning for Classifying Client Utterances in Counseling Dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1448–1459, Minneapolis, Minnesota. Association for Computational Linguistics.

Yada Pruksachatkun, Sachin R. Pendse, and Amit Sharma. 2019. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. In *Proceedings of CHI*, page 13.

Matthew Purver. 2011. Topic Segmentation. In *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Chapter 11.

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 *Proceeding of EMNLP*.

Brian Syzdek. 2020. Client and Therapist Psychotherapy Sentiment Interaction Throughout Therapy. *Psychological Studies*, 65:1–11.

Adam Tsakalidis, Dana Atzil-Slonim, Asaf Polakovski, Natalie Shapira, Rivka Tuval-Mashiach, and Maria Liakata. 2021. Automatic Identification of Ruptures in Transcribed Psychotherapy Sessions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 122–128, Online. Association for Computational Linguistics.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. 2021. TIAGE: A Benchmark for Topic-Shift Aware Dialog Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1684–1690, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-Disclosure and Channel Difference in Online Health Support Groups. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):704–707.

Jiamin Yang and David Jurgens. 2024. Modeling Empathetic Alignment in Conversation. In *Proceedings of NAACL*.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proceedings of ACL*, pages 5276–5289, Online.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of NAACL*.

A Additional Data Details

A.1 Session Split

Adapting the methodology in Kushner and Sharma (2020), we chose a value of N for the session split based on the number of sessions it produces for $10,000$ random conversations. The median, mean, and standard deviation all plateau starting from $N = 50$ and level off around $N = 100$. Setting N at 100 will provide the reliability of the measure in relation to the selection of its value.

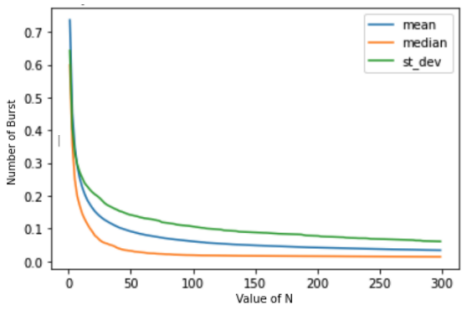

Figure 4: Number of sessions per N from 10,000 random conversations.

A.2 Switch / Cancel Surveys

During therapy, patients may choose to switch therapists or cancel their plan by filling out surveys. In the survey, they are asked to provide a reason for the switch/cancel, which they either select from a fixed list, or enter their own in free text. The provided switch reasons include:

s1 I could not find a time to meet with my provider.

s2 I couldn't form a strong connection with my provider.

s3 I don't feel like my provider was responsive enough.

s4 I just want to try someone new.

s5 I want to select a provider with a different gender.

s6 I was unsatisfied with the quality of care.

s7 Disabled.

s8 Dissatisfied with app.

s9 Dissatisfied with provider.

s10 Expensive.

The provided cancel reasons include:

c1 I met my goal / I feel better.

c2 The cost doesn't fit my budget.

c3 The treatment provided by my therapist was not helpful.

c4 I had technical issues.

c5 My therapist was not responsive to my messages.

A.3 Defining Unsuccessful Relations

The dataset of "unsuccessful" therapies with the patient abandoning the relationship includes therapies where the patient requested either a switch with reasons s2, s3, s4, s6 or a cancellation with reason c3.

B Operationalization

B.1 Implementation Details

For our redirection model, we fine-tuned Gemma-2B (Mesnard et al., 2024) with 4-bit QLoRA (Dettmers et al., 2023) using the huggingface library. We use a 90/10 split for training and validation, and trained for 2 epochs, with LoRA rank = 16 and dropout = 0.05, context length 4096, batch size 2, learning rate 2e-4, AdamW optimizer (Loshchilov and Hutter, 2017), achieving a validation perplexity of 10.97. The total train time is approximately 43 hours on 2 NVIDIA RTX A6000 GPUs. For the model experiments, we conducted a hyperparameter search over learning rates [2e-5, 2e-4] and LoRA rank = [8, 16, 64], and used fixed values for the rest of the parameters.

We used the ConvoKit Python package (Chang et al., 2020) to calculate orientation to employ the same methodology outlined in Zhang and Danescu-Niculescu-Mizil (2020). We trained two separate

models for therapist and patient orientation, using dependency-parse arcs representations for both speakers and 12 SVD dimensions.

To calculate uptake, we fine-tuned a pre-trained BERT-base model for next utterance classification. Two separate models were trained: one for predicting patient's utterance after a therapist and the other for vice-versa. Our training setup follows the original paper (Demszky et al., 2021). We fine-tune our model for 1 epoch with a batch size of 16, max length of 512 tokens for patient's and therapist's utterance each. The learning rate is set at 6.24e-5, with linear decay and AdamW optimizer (Loshchilov and Hutter, 2017) The total train time is approximately 12 hours for each model on 2 NVIDIA RTX A6000 GPUs. We used the two models with the original source code from (Demszky et al., 2021) to calculate uptake.

For similarity difference, we used a pre-trained sentence BERT model 'multi-qa-MiniLM-L6-cos-v1' to map utterances into a 384 dimensional dense vector space. We calculated the similarity between two utterances using cosine similarity of their embeddings.

B.2 Used Artifacts

We list the following artifacts and their licenses that are used in the work.

- ConvoKit 2.5.3: `https://convokit.cornell.edu/`, MIT License
- PyTorch 2.2.1: `https://pytorch.org`, BSD-3 License
- Sentence Transformers 3.0.0: `https://github.com/UKPLab/sentence-transformers`, Apache License 2.0
- Transformers 4.38.2: `https://github.com/huggingface/transformers`, Apache License 2.0
- Conversational Uptake Source Code: `https://github.com/ddemszky/conversational-uptake`, MIT License

C Additional Application to US Supreme Court Oral Arguments

We also examine how our redirection framework can be applied in other domains in addition to mental health. In particular, we apply our method to a publicly available dataset of U.S. Supreme Court oral arguments (Danescu-Niculescu-Mizil et al., 2012; Chang et al., 2020). Although court proceedings differ from therapy in terms of topics, goals, and interaction styles, their relatively unstructured and dynamic nature enables an initial exploration of how such discussions are redirected.

In this setting, we focus on the interactions between justices and lawyers. The power dynamics between these distinct roles reflect the asymmetric relationship between therapists and patients in mental-health domains, where one party generally holds more influence over the conversation.

As expected, our analysis reveals that justices redirect the conversation significantly more than lawyers ($p < 0.001$, according to a Wilcoxon signed-rank test). Our findings suggest that justices exercise more control over the flow of the discussion, steering it towards issues they consider critical to the case.

To further examine how redirection unfolds in these exchanges, we use a Bayesian distinguishing word analysis, "Fightin' Words" (Monroe et al., 2017) to compare high and low redirection phrases from both speakers. For justices, highly redirecting utterances frequently involve assertive questioning ("may ask", "ask you", "if he", "is this") or references that draw the court's attention to specific matters ("the court", "court of"). In contrast, low redirecting utterances from justices tend to include responses ("all right", "that right", "no no") or clarifications ("you mean", "mean that") of subjects raised by other parties.

Conversely, highly redirecting utterances by lawyers often direct the court's ("the federal", "this court", "this state") focus to new arguments or highlight their perspectives on the case ("it seems", "to me", "that the"). Low lawyer redirecting utterances, however, tend to affirm and acknowledge the justice's statements ("yes sir", "your honor", "oh yes") or continue discussing the current issue at hand ("it was", "it for").

For reference, we provide additional examples of high and low redirection from both speakers in Table 4.

#	*Example*	*Redirection*
1	J: That's on the outer, outer belt. L: On the outer, outer belt. Now there's no dispute about it. It was admitted by the President of Santa Fe so that the evidence is here but the Commission simply ignored it. **J: Let me ask you [...] Are the gateways of Danville, Decatur Springfield along the Wabash which is a wholly owned subsidiary of the Pennsylvania, are they of any consequence?** L: The Decatur gateway is because as Your Honor would see that leads through the Wabash to the Hannibal bridge crossing the Mississippi and then on to Kansas City. [...] The Springfield gateway is an important gateway but of lesser importance [...].	High
2	J: You happen to know what was your practice, if you had a practice of having you're statements to the grand jury in summary or explanation of what the evidence disclosed or manifest, was that taken down by a stenographer? L: No, I don't recall it ever being done [...]. **J: Did Heras adjure the conspiracy or did you just not have enough – have recent acts on his part?** L: Well, in all frankness, I don't think that Heras have adjured the conspiracy [...].	High
3	J: [...] This protection by the McCarran Act offer the individual State, protection to the State from the paramount federal power is difficult to reconcile with the theory after making one State subject to the laws of another State, in which laws they have no part in making. Do you subscribe to that language? L: Well, the – latter part would raise a question of possible [...]. **J: Well it didn't make any sense to me at all.** L: Well – I said the – the latter part may raise the question as to, the possible conflict were Nebraska interprets deceptive practice in one way – Nebraska corporation [...].	Low
4	J: [...] Do you think it would be a permissible reading of the Act to say that as far as conspiracy versus substantive counts, yes, Congress must be taken to have intended cumulative punishments, authorized cumulative punishments, but with respect to the two substantive accounts, it cannot be so taken to have intended? L: Well, I can't say that you can't read it that way [...] **J: I think your view is to whether there's a difference between the approach that the Court should take in a case of this kind, if there is any difference.** L: Well, I think the thing that bothers me is that in the context in which these cases arise, the question of power goes to the offense [...].	Low
5	L: [...] I just wanted to say I thought that distinguished this case from many others which might be put. J: The question I wanted to ask you was, is the restaurant itself, either on its menus or its advertising literature, carry any notation that it's identified in any way with the Delaware [...] **L: [...] The third point to emphasize is that, [...] the entire enterprise [is] bound up financially into a single project. The Supreme Court of Delaware, [...] ruled that the leases could be permitted only to the extent that such leasing is necessary and feasible to enable the Authority to finance the project [...].** J: [...] That third fact would apply to every leasing because whether the State leases in order to derive money from the leasing or part of the money for maintaining the state enterprise [...] it seems to be immaterial [...].	High
6	L: [...] that I would doubt whether it was a deep-seated. J: Look at the alibi of Mapp for a good illustration. **L: [...] one of the parts that bothers me here, Mr. Justice, it's the converse in a sense of what I said earlier. I spoke of the affirmative effect of a decree pointing out the constitutional duty of the legislature. [...] you said since the Supreme Court of Tennessee refused to act, that established that there was no violation of the Tennessee Constitution.** J: That isn't what I said. I said that that decision isn't done nothing [...].	High
7	L: [...] There would be no reason for him to have been walking out along that ledge of that barge [...]. J: If he had been ordered to go to do the work? **L: And he had been ordered to go do that work on the raft, that's correct sir.** J: And what you're saying is as a matter of law, it has to defend as a matter – you have to defend it, right, it has been found as a matter of law. And when he walked from there – from the Pfeifer or to the Winisook catwalk, and before he fell on it, he was at that time as a matter of law, no longer engaged in any duty as mate.	Low
8	L: They are outlined in our briefs and needless to say, I do not have the time and shouldn't take the time to outline them all [...] let's make a further argument or conclusion or state my own opinion and then I shall develop the facts [...]. J: If you don't mind I suggest to you, I think you help us a lot more if you could guide us through what you consider to be the controlling thing instead of characterizing them all. **L: Your Honor it's quite, I'm sure Your Honor is quite right.** J: [...] Now I hope you are going to discuss them separately [...].	Low

Table 4: Examples comparing high and low redirection from justices and lawyers in Supreme Court oral arguments. The measure is applied to the colored utterance (Justice for the first 4 examples and Lawyer for the remaining 4). Low indicates the bottom 25th percentile of the measure values for the respective speaker type; High indicates the top 75th percentile.