

Using Linguistic Synchrony to Evaluate Large Language Models for Cognitive Behavioral Therapy

Anonymous ACL submission

Abstract

Synchrony, the responsive communication between interacting individuals, is a crucial process in building a strong relationship between a mental health therapist and their client, leading to positive therapeutic outcomes. However, so far synchrony has not been investigated as a measure of efficacy of large language models (LLMs) delivering mental health therapy. In this work, we evaluate the linguistic synchrony of an LLM (ChatGPT 3.5-turbo) in a mental health dialog setting by first validating a computational measure of linguistic synchrony with two measures of the quality of client self-disclosures—intimacy and engagement ($p < 0.05$). We then compare the linguistic synchrony of the LLM to trained therapists and non-expert online peer supporters in a Cognitive Behavioral Therapy (CBT) setting. We show that the LLM is outperformed by humans with respect to linguistic synchrony ($p < 0.001$). These results support the need to be cautious in using LLMs in mental health applications.

1 Introduction

Synchrony describes responsive communication between individuals and is known to be important in building social relationships and supporting mental health outcomes (Delaherche et al., 2012; Klein, 2023). The phenomenon manifests through various modalities, including physical body movements (mirrored body language) (Ramseyer and Tschacher, 2011), vocals (pitch matching) (Imel et al., 2014), and language (linguistic style matching) (Niederhoffer and Pennebaker, 2002), across a variety of contexts (Kidby et al., 2023; Bonny and Jones, 2023). Synchrony is associated with building a sense of affiliation and improving cooperation and rapport (Vail et al., 2022); it is critical in therapist-client relationships (Colton, 2022). In this work, we focus on *linguistic* synchrony in the context of mental health therapy.

LLMs are increasingly used in dialogue systems for mental health, leading to the investigation of their efficacy in that application (Chiu et al., 2024; Cho et al., 2023). To the best of our knowledge, synchrony has not yet been evaluated as a performance indicator, in spite of its critical role (Kerjwal and Benus, 2024) in developing a strong therapist-client relationship. Therefore, in this work we measure the performance of an LLM (GPT-3.5-Turbo) in a mental health setting with respect to linguistic synchrony. We demonstrate that there is a significant relationship between linguistic synchrony, which we operationalize through the normalized Conversational Linguistic Distance (nCLiD) (Nasir et al., 2019), and two measures of the quality of client self-disclosures - intimacy and engagement. Then we compare the performance of the LLM to trained therapists and non-expert online peer supporters in a CBT setting (Figure 1). We show that the LLM is outperformed by both groups. This indicates that LLMs are not yet at the level of humans in generating high-quality therapeutic responses, and we suggest that synchrony can serve as an evaluation criterion for LLMs in mental health contexts.

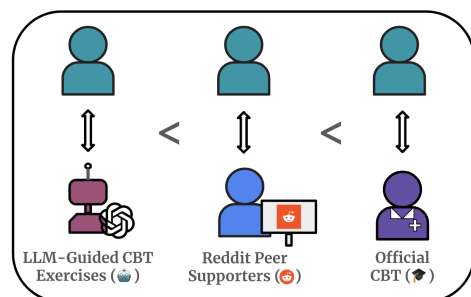


Figure 1: We find that linguistic synchrony is highest for trained CBT therapists (right), with online peer supporters following in second highest (middle), and lowest for LLMs (left).

2 Background

We explore CBT as the application domain for this work, and present background on important measures of therapy effectiveness and the key role of linguistic synchrony in the quality of the therapist-patient relationship.

2.1 Cognitive Behavioral Therapy (CBT)

CBT is a psychotherapeutic intervention with the goal of reducing emotional distress and increasing adaptive behaviors (Wenzel, 2017). A core principle of CBT is that psychological disorders and their symptoms arise from unhelpful patterns of thought and behavior. Directly addressing these patterns can relieve symptoms. CBT homework exercises help individuals practice what they learn in therapy sessions in their daily life (Prasko et al., 2022).

2.2 Intimacy and Engagement in Self-Disclosure

Patient self-disclosure is an essential component of psychotherapy and is associated with positive treatment outcomes (Farber, 2003; Farber et al., 2006). Self-disclosure relies on establishing a trusting relationship between a patient and therapist to allow patients to disclose their problems and achieve constructive change (Newman, 2002). Intimacy within dyadic relationships in a healthcare setting is an important predictor of positive health outcomes (Kadner, 1994). Morton (1978) defined intimacy as having two dimensions: descriptive and evaluative. *Descriptive intimacy* involves the disclosure of private facts, while *evaluative intimacy* involves the disclosure of personal opinions and information. *Engagement* is the extent to which a patient actively participates in the therapeutic process beyond simply being present (Tetley et al., 2011; O’Brien et al., 2009). It can be defined as active or passive (Nguyen et al., 2018). Engagement, shown by involvement in therapy and earnest self-disclosure, also leads to positive therapeutic outcomes (Farber, 2003).

2.3 Linguistic Synchrony in Therapy

It is well known that the quality of the relationship between a therapist and their client plays a significant role in how effective therapy is for the client (Horvath and Symonds, 1991; Martin et al., 2000). The term *working alliance* captures the collaborative aspect of the therapist-client relationship, when the goals of the therapist and client align and

they have a strong emotional bond (Horvath and Greenberg, 1986).

The concept of *interpersonal synchrony*¹ describes when the participants of an interaction adapt and converge on each other’s behaviors over time. Higher levels of synchrony between participants are associated with positive interpersonal outcomes such as better collaboration, increased rapport, and a sense of affiliation (Rennung and Göritz, 2016; Hove and Risen, 2009). Vail et al. (2022) investigate the relationship between language and working alliance, specifically focusing on *linguistic synchrony*, which refers to the similarity between interlocutors in semantics, syntax, or style. They find that the therapist’s linguistic synchrony strongly impacts the client’s perception of the working alliance. Nasir et al. (2019) introduce the normalized Conversational Linguistic Distance (nCLiD), a metric that measures linguistic synchrony between two speakers. nCLiD was found to be associated with the therapist’s level of empathy towards the client, and affective behaviors of couples in therapy. We use nCLiD to evaluate the linguistic synchrony in CBT interactions.

3 LLMs in Mental Health

The prevalence of mental health conditions and lack of accessible care leaves a void that many have attempted to address (Youper; Choudhury et al., 2023; Fang et al., 2022) with assistive therapeutic technologies powered by LLMs. We present background on LLMs used in mental health, and how they have been evaluated in this context.

3.1 Applications and Challenges

The promising capabilities of LLMs such as the OpenAI GPT series (Radford et al., 2018; Brown et al., 2020) have catalyzed the development of various general-purpose LLMs (Touvron et al., 2023; Anil et al., 2023; Jiang et al., 2023) and domain-specific LLMs (Liu et al., 2023; Chen et al., 2021; Ganguli et al., 2022; Yang et al., 2023; Taylor et al., 2022). Despite the impressively human-like text that LLMs can generate, there are many issues with LLMs that can have serious repercussions in sensitive downstream tasks, such as propagating

¹The terms entrainment, synchrony, and coordination are used interchangeably across psychology, computer science, and social-behavior literature (Wynn and Borrie, 2022) to mean the convergence of a type of behavior amongst participants in an interaction. We use the term synchrony in this paper.

harmful stereotypes and biases (Bender et al., 2021) and encouraging suicide (Marcus, 2022), which is exacerbated by the fact that the text generated by LLMs seems coherent. For example, LLMs are known to suffer from hallucinations and produce harmful or factually incorrect outputs (Zhang et al., 2023; Ganguli et al., 2022; Maynez et al., 2020), resulting in a surge of research in techniques to mitigate these issues (Stiennon et al., 2020; Bai et al., 2022; Wei et al., 2022).

The risks are especially serious for applications in sensitive domains such as mental health, where LLMs are becoming increasingly popular and promising for mental health support (Choudhury et al., 2023; Laestadius et al., 2022; Youper). People have turned to LLMs when facing mental health problems, feeling heard and supported, with many comparing the experience to that of interacting with a real therapist (Reardon, 2023; Al-Sibai, 2023; Reddit, 2022). Also, Cho et al. (2023) tested an LLM as a therapist in interactive language therapy for autistic adolescents, showing significant strengths in empathetic engagement and adaptability. However, numerous cases have shown that LLMs pose substantial risks in this use case, such as racial and gender biases (Zack et al., 2023; Omiye et al., 2023), raising serious concerns among interdisciplinary mental-health experts (Stade et al., 2024; Choudhury et al., 2023; Li et al., 2020). These risks have already resulted in real-life consequences. For example, the National Eating Disorder Association shut down their chatbot for giving misguided medical advice (Jargon, 2023). Replika was implicated in a UK criminal case for encouraging a man to assassinate the Queen and then commit suicide (Weaver, 2023). Yet despite the potential problems, the popularity of LLM-powered mental health services continues to rise (van Heerden et al., 2023).

3.2 Evaluation Methods

Computational methods have been developed to assess the performance of human therapist responses in therapeutic dialog with respect to various psychotherapy criteria such as empathy (Sharma et al., 2020), warmth (Zech et al., 2022), and linguistic synchrony (Nasir et al., 2019; Shapira et al., 2022). With LLMs being increasingly explored in mental health dialog systems, some of these evaluation methods have been applied to LLMs as well (Cho et al., 2023; Chiu et al., 2024). Cho et al. (2023) had clinical psychologists and psychiatrists

evaluate an LLM with respect to empathy, communication skills, adaptability, engagement, and ability to establish a therapeutic alliance. Recently, Chiu et al. (2024) proposed a computational framework to evaluate LLMs with respect to reflections, questions, solutions, normalizing, and psychoeducation by comparing them against high-quality and low-quality human therapist transcripts. Both Cho et al. (2023) and (Chiu et al., 2024) simulate the client side of the LLM-client conversation due to ethical concerns of having an LLM advise vulnerable populations. However, this prevents a realistic evaluation of LLMs for therapy. The LLM-participant dataset used in this work comes from an IRB-approved study (Kian et al., 2024) in which they deployed LLMs in an interactive CBT homework context with students (Section 4), which provides a step towards more realistic evaluations of LLMs in therapy. Additionally, linguistic synchrony has thus not been used to evaluate LLM-powered mental health dialog systems. Given its importance (Section 2.3), we investigate linguistic synchrony in a LLM-powered mental health dialog system (Section 4).

4 Study 1: Evaluation of Linguistic Synchrony in Therapy

In this study, we aim to evaluate linguistic synchrony, operationalized by nCLiD, as a measure of therapist quality by demonstrating that it is associated with indicators of positive therapeutic outcomes, specifically engagement, and intimacy. Note: higher linguistic synchrony is operationalized through a lower nCLiD score. We put forth the following hypotheses:

H1a: There will be a positive significant relationship between high evaluative intimacy and linguistic synchrony.

H1b: There will be a positive significant relationship between high descriptive intimacy and linguistic synchrony.

H1c: There will be a positive significant relationship between active engagement and linguistic synchrony.

4.1 Methodology

We run our analysis on English language transcripts of LLM-guided CBT exercises annotated for intimacy and engagement and calculate nCLiD scores for these transcripts. We conduct linear regressions to analyze the relationship between nCLiD and the

259 annotations.

260 4.1.1 Participants and Procedure

261 The analysis in our research utilized transcripts de-
262 rived from LLM-powered SAR and LLM-powered
263 chatbot CBT homework interactions with univer-
264 sity students (Kian et al., 2024). Students were
265 screened to be over 18 years of age, proficient in
266 English, have normal or corrected-to-normal vision
267 and hearing, and live near campus. The Patient
268 Health Questionnaire-9 (PHQ-9) (Kroenke et al.,
269 2001) was used as a screening tool, and individuals
270 with a score of 15 or higher, indicating moderately
271 severe to severe depression, were excluded as a
272 safety measure. All individuals who filled out the
273 screening materials were shown a page with univer-
274 sity mental health resources. A total of 26 students
275 participated in the study conditions we assess in
276 this work. Before the start of the study, all par-
277 ticipants had an informed consent meeting with
278 a member of the research team. This study was
279 approved by their university’s IRB, and all partici-
280 pants were compensated with a US \$150.00 Ama-
281 zon gift card. This amount was calculated based on
282 expected hours spent on the study and local mini-
283 mum wage. The study duration was 15 days; for
284 the first 8 days, the CBT homework sessions were
285 compulsory, but during the last 7 sessions, they
286 were no longer compulsory. Each day, the partici-
287 pants log into a secure portal and either complete
288 an LLM-powered robot- or chatbot-guided CBT
289 exercise. They could select from two CBT exercise
290 options: Cognitive Restructuring (Clark, 2013) or
291 Coping Strategies (Association and of Clinical So-
292 ciety, 2017). The LLM used was GPT-3.5-turbo²,
293 where the LLM was prompted to utilize the cho-
294 sen strategy while acting as a therapy guide (see
295 Appendix A for the prompts used). All identifi-
296 able data for this study were securely stored on
297 IRB-approved secure cloud storage. Only IRB-
298 approved researchers with the appropriate training
299 were allowed access to the data. We will refer to
300 this as the LLM-guided CBT Exercises Dataset.

301 4.1.2 Measures

302 Descriptive and evaluative intimacy were assessed
303 according to the Morton (1978) framework. *De-*
304 *scriptive intimacy* involves the disclosure of private
305 facts, while *evaluative intimacy* involves the dis-
306 closure of personal opinions and feelings. Each

²<https://platform.openai.com/docs/models/gpt-3-5-turbo>

307 dimension was dichotomized into high and low dis-
308 closure levels, as recommended by Tolstedt and
309 Stokes (1984). Secondly, we assess engagement,
310 which measures how much a participant actively
311 participates in the sessions. Engagement was anno-
312 tated to be active or passive according to Nguyen
313 et al. (2018). Finally, to operationalize the linguis-
314 tic coordination between the participants and the
315 LLM, we utilize the normalized Conversational
316 Linguistic Distance (nCLiD) by Nasir et al. (2019).

317 4.1.3 Annotation Process

318 The CBT exercise transcripts were annotated for
319 three variables: descriptive intimacy, evaluative in-
320 timacy, and engagement. Four undergraduate anno-
321 tators (two female, two male) were trained through
322 workshops led by graduate student instructors for
323 two weeks to annotate the data for the selected
324 variables. Each participant’s turn in response to
325 the LLM was annotated, resulting in an average of
326 10-15 annotations per participant per day. The Inter-
327 Coder Reliability (ICR) was measured using 10%
328 of the dataset, resulting in 83.539% and Cohen’s
329 average kappa score of $\kappa = 0.602283$. Finally, an-
330 notations were aggregated to yield percentages of
331 active engagement, high descriptive intimacy, and
332 high evaluative intimacy averaged across all study
333 days per participant, which we use in subsequent
334 analyses.

335 4.1.4 nCLiD Algorithm

336 The Conversational Linguistic Distance (CLiD)
337 (Nasir et al., 2019) is an asymmetric distance met-
338 ric that quantifies the interpersonal linguistic syn-
339 chrony between two speakers. Therefore, higher
340 linguistic synchrony is described by lower CLiD
341 scores and vice versa. Nasir et al. (2019) demon-
342 strated that nCLiD correlates with ratings of a ther-
343 apist’s empathy towards their patient (CLiD is lower
344 for a higher therapist empathy rating) and affective
345 behaviors in Couples Therapy (CLiD is lower for
346 lower negative affect and higher for lower positive
347 affect).

348 For a therapy session text record D be-
349 tween a therapist T and a patient P consist-
350 ing of N turns of interleaving utterances with
351 $D = [t_1, p_1, t_2, p_2, \dots, t_N, p_N]$, let us consider one
352 speaker as the anchor A , and another as the coordi-
353 nator C . For each anchor utterance a_i , we compute
354 the minimum distance $d_i^{C \rightarrow A}$ for the minimum dis-
355 tance between the sequences of *word2vec* (Mikolov
356 et al., 2013) embeddings of a_i and the following c_j

with a context length k , and we use Word Mover’s Distance (WMD) (Kusner et al., 2015) to measure linguistic difference between two utterances:

$$d_i^{C \rightarrow A} = \min_{i \leq j \leq i+k-1 \leq N} WMD(a_i, c_j) \quad (1)$$

The context length, k , accounts for the observation that local coordination may not occur only in the immediate turn, but may occur a few turns later.

The transcript-level unnormalized Conversational Linguistic Distance (uCLiD) is a simple average of local linguistic distance d_i over the whole session (numerator in equation 2). The normalized Conversational Linguistic Distance (nCLiD) normalizes uCLiD to account for the other reasons that may result in spurious coordination, such as a structured conversation on a pre-decided topic or similar language due to coordination of each speaker to their own language, etc.

$$nCLiD = \frac{uCLiD}{\alpha} = \frac{\frac{1}{N} \sum_{i=1}^N d_i^{C \rightarrow A}}{\alpha} \quad (2)$$

The normalization factor α accounts for spurious coordinations by accounting for potential coordination within A and B, and between A and B. The full equation is available in Appendix F.

We implement nCLiD using the WMD algorithm from the gensim 4.3.2 library³ with Python 3.8, using 300-dimensional *word2vec* word embeddings trained on the Google News corpus provided by gensim. The text is tokenized by whitespace, and stop words were not removed, following the example of Nasir et al. (2019) to account for possible linguistic similarity associated with similar usage of stop words.

4.1.5 Data Analysis

We implement tests in R version 4.3.2; a list of all R packages and their versions are available in Appendix B. We run linear regression tests to assess the relationship between nCLiD scores and derived metrics for the LLM transcripts. The Durbin-Watson test of autocorrelation is used to test the assumption of independence, and the Shapiro-Wilk and Breusch-Pagan tests to assess normality of residuals and homoscedasticity, respectively. If a model’s residuals fail the Breusch-Pagan test, we apply Huber-White standard errors.

³<https://radimrehurek.com/gensim/index.html>

4.2 Results

A simple linear regression was performed regressing the transcripts’ descriptive intimacy onto their nCLiD scores. Heteroscedasticity was addressed by employing Huber-White standard errors and the model was significant ($\chi^2(1, 24) = 5.31, p < 0.05, Adj.R^2 = 0.15$). Residuals were normally distributed. There was a significant main effect of nCLiD on descriptive intimacy score, $b = -7.40, t = -2.24, p < 0.05$ (Figure 2).

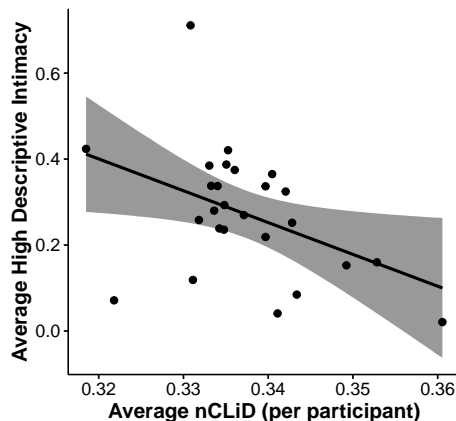


Figure 2: Scatterplot of average high descriptive intimacy vs. average nCLiD score per participant with a regression line of best fit.

A simple linear regression was performed regressing the transcripts’ evaluative intimacy onto their nCLiD scores and the model was significant ($F(1, 24) = 5.18, p < 0.05, Adj.R^2 = 0.14$). The residuals of this model were normally distributed as indicated by the Shapiro-Wilk test ($p > 0.05$) and homoscedastic as indicated by a Breusch-Pagan test ($p > 0.05$). There was a significant main effect of nCLiD on evaluative intimacy score, $b = -7.72, t = -2.28, p < 0.05$ (Figure 3).

A simple linear regression was performed regressing the transcripts’ active engagement score onto their nCLiD scores and the model was significant ($F(1, 24) = 23.05, p < 0.001, Adj.R^2 = 0.47$). The residuals were normally distributed and homoscedastic. There was a significant main effect of nCLiD on engagement score, $b = -17.59, t = -4.80, p < 0.001$ (Figure 4).

5 Study 2: LLM vs. Human Comparison

In this study, we compare the linguistic synchrony of the LLM against trained mental health therapists and non-expert online peer supporters. We put forth the following hypotheses:

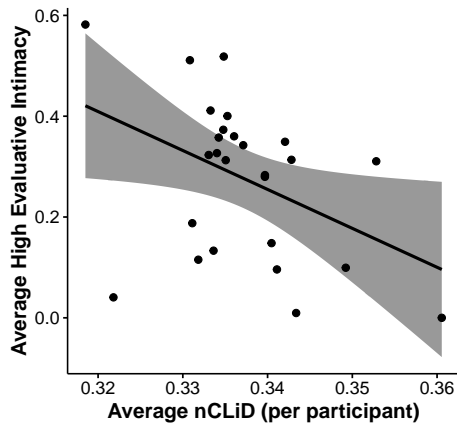


Figure 3: Scatterplot of average high evaluative intimacy vs. average nCLiD score per participant with a regression line of best fit.

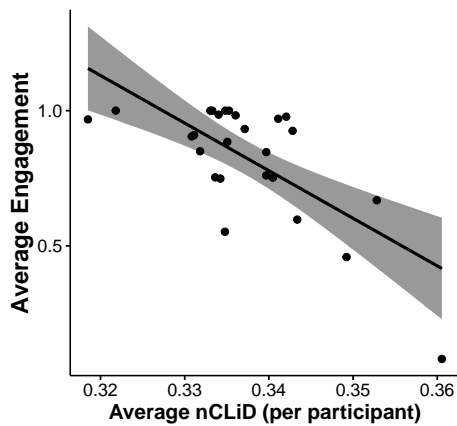


Figure 4: Scatterplot of average engagement vs. average nCLiD score per participant with a regression line of best fit.

Hypotheses: Transcripts from trained CBT therapists will have higher linguistic synchrony than the LLM-guided exercises, which will, in turn, have higher linguistic synchrony than non-expert online peer supporters. Note: higher linguistic synchrony is operationalized through a lower nCLiD score. More specifically,

H2a: Linguistic synchrony will be higher in the therapeutic transcripts from trained CBT therapists (Official CBT Dataset) than from an LLM (LLM-guided CBT Exercises Dataset).

H2b: Linguistic synchrony will be higher in the therapeutic transcripts from trained CBT therapists (Official CBT Dataset) than from non-expert online peer supporters (Reddit Dataset).

H2c: Linguistic synchrony will be higher in the therapeutic transcripts from an LLM (LLM-guided

CBT Exercises Dataset) than from non-expert online peer supporters (Reddit Dataset).

5.1 Methodology

We compare the LLM-guided CBT Exercises Dataset with two other English language datasets. The first is a dataset developed from therapy-like conversations on Reddit with online non-expert peer supporters, and the second is a subset of the Alexander Street Press Counseling (ASPCPT) and Psychotherapy Transcripts with trained expert human therapists. We compare by conducting a one-way Analysis of Variance (ANOVA) test across nCLiD scores for the three datasets to determine how they perform against each other.

5.1.1 Datasets

We introduce two additional datasets to compare against the LLM dataset in Section 4.

Reddit Dataset The Reddit⁴ dataset is a collection of 30 English dyadic conversations we extracted from the Reddit posts of people in Online Mental Health Communities (OMHCs) (Sharma and De Choudhury, 2018) where there are indicators of coping strategy (Courtney E. Ackerman, 2017) or cognitive restructuring (Aid, 2017) exercises. We chose these exercises as they were the ones used in the LLM-guided CBT transcripts described in Section 4 above. The aim of creating the Reddit dataset is to capture the linguistic characteristics of humans, untrained in therapy (non-expert peer supporters), engaging in conversations that parallel guided CBT exercises. This dataset enables us to establish a baseline to compare against the performance of the LLM. Refer to Appendix E for complete details on our data collection procedure.

Official CBT Dataset The Alexander Street Press Counseling and Psychotherapy Transcripts (ASPCPT) dataset is a therapy and counseling dataset (Alexander Street Press, 2023)⁵. The ASPCPT dataset was created by transcribing sessions featuring expert therapists working with a client or a family. We use a subset of the ASPCPT dataset with the “cognitive behavioral therapy” therapy type, which we refer to as the “Official CBT”

⁴<https://www.reddit.com/>

⁵Accessed through authors’ institution subscription: <https://search.alexanderstreet.com/psyc> (Volume 1) and <https://search.alexanderstreet.com/ctrn> (Volume 2)

dataset in this study. This subset excludes transcripts of interviews and family therapy because they are not dyadic conversations between a therapist and one client. Clients are anonymized using unique participant IDs. Transcripts are preprocessed to remove brackets and parentheses of text (e.g., “[00:04:16]”), which are time stamps from the transcription process. The Official CBT dataset contains 39 transcripts in English.

5.1.2 Data Analysis

We implemented tests in R version 4.3.2; a list of all R packages and their versions is available in Appendix B. The assumption of a normal distribution was assessed via the Shapiro-Wilk test and homogeneity of variance was evaluated by use of Levene’s test. Unequal variances are addressed by employing a Welch’s ANOVA which accounts for the differences in variations between the LLM, Official CBT, and Reddit datasets.

5.2 Results

A Welch’s ANOVA was conducted to compare the nCLiD scores between the LLM ($M = 0.34, SD = 0.01$), Official CBT ($M = 0.29, SD = 0.01$), and Reddit ($M = 0.32, SD = 0.02$) datasets. The ANOVA was significant at the $p < 0.001$ level, $F(2, 53.69) = 429.95, p < 2.2e - 16$ (Figure 5). A post-hoc Games-Howell test indicated that the nCLiD scores were significantly different among all pairs of datasets (LLM-Official, LLM-Reddit, Official-Reddit) at the $p < 0.001$ level. In particular, nCLiD scores were higher in the LLM dataset than in the Official dataset and Reddit dataset. Additionally, the nCLiD scores for the Reddit dataset were significantly higher than the Official dataset.

6 Discussion

We first validate a computational measure of linguistic synchrony, nCLiD, as a measure of therapeutic effectiveness (Section 4). We demonstrate a statistically significant relationship ($p < 0.05$) between this measure of linguistic synchrony and the intimacy and engagement of the participant’s self-disclosures - two important predictors of positive therapeutic outcomes (Kadner, 1994; Scott and King, 2007). We find that nCLiD is inversely related to these measures, as indicated by the negative b-value for the main effect of each of the relationships. This means that higher synchrony is

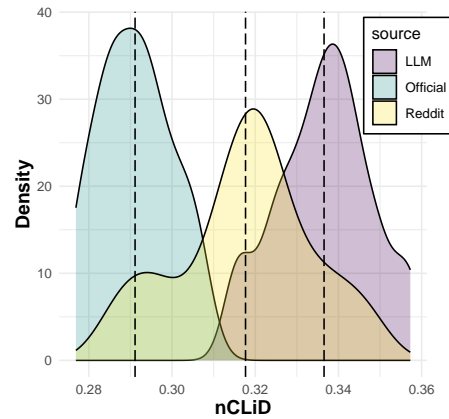


Figure 5: Distributions of nCLiD scores for LLM-Guided CBT exercises, Official CBT, and Reddit datasets.

associated with higher intimacy and active engagement (*Supporting H1a, H1b, and H1c*).

We hypothesize that in a therapeutic setting, a therapist’s linguistic synchrony with the client encourages greater self-disclosures and, subsequently, higher levels of intimacy and engagement. This relationship is supported by Beňuš (2014) in their review, where they find a connection between synchrony and social distance and suggest that synchrony can help a medical professional develop closeness and trust with their clients. This closeness is critical for encouraging greater self-disclosure (Newman, 2002), which leads to improved therapeutic outcomes (Farber, 2003; Farber et al., 2006). Colton (2022) also found that linguistic synchrony “catalyzes” the therapeutic bond, which is further supported by Vail et al. (2022). Therefore, the related literature suggests that there is a relationship between linguistic synchrony, and intimacy and engagement, two measures of patient self-disclosures.

The significant relationship between linguistic synchrony and descriptive intimacy, evaluative intimacy, and engagement indicates that nCLiD shows promise as a measure of therapeutic outcomes. This allows for quantitative analysis of dyadic therapeutic interactions without costly annotations for measures such as intimacy or engagement.

Next, we used linguistic synchrony to evaluate the quality of LLM-administered CBT exercises by comparing them to transcripts from trained CBT therapists and therapy-like dialogue from non-expert online peer supporters (Section 5). The results of the ANOVA indicate that the distributions of nCLiD scores among the LLM, Official,

and Reddit datasets are significantly different. The CBT practitioners in the Official dataset had significantly higher linguistic synchrony than the LLM (*supporting H2a*) as well as those from non-expert online peer supporters (*supporting H2b*). These results are as we hypothesized since CBT therapists undergo years of training to offer patients a high-quality therapeutic experience. Interestingly, non-expert online peer supporters had significantly higher linguistic synchrony than the LLM (*H2c not supported*).

Initially, we hypothesized that the LLM would have higher linguistic synchrony than the online peer supporters because of the demonstrated high level of mental health domain knowledge found in LLMs (Heinz et al., 2023; Lamichhane, 2023) and the increased usage of LLMs in mental health therapy applications (Youper; Reardon, 2023; Al-Sibai, 2023; Reddit, 2022). However, it seems that even non-expert humans have higher linguistic synchrony than a prompted LLM. This suggests that LLMs underperform even untrained people in a therapeutic context. It may be that Reinforcement Learning from Human Feedback, a popular alignment technique employed in LLMs, makes LLMs overly focused on offering advice and problem-solving, as noted by Chiu et al. (2024). This may lead LLMs to have a less varied and nuanced conversational style, making the LLM output more formulaic, aligning with the given instruction, as also seen by Shaikh et al. (2024) in their LLM-based conversational system. In our own analysis of the LLM dataset, we also observe patterns of the LLM repeatedly using the same response frame (see Appendix D for excerpts). These tendencies of the LLM to be less varied in its responses may, therefore, lead to lower linguistic synchrony. It is also important to note that the individuals who self-select to participate in discussions on mental health subreddits and offer support to their peers are not representative of the average social media user. While these individuals are non-experts, it is possible that they are still more familiar with therapy and able to better mimic the expected dialogue.

Our results further confirm the need to be cautious in applying LLMs for therapeutic contexts. While they are able to manage various therapeutic tasks, as found by Cho et al. (2023); Kian et al. (2024), their dialog is ultimately inferior to that of therapists. Thus, researchers must carefully assess each application domain and determine if the LLM can meet the expected threshold of perfor-

mance. Furthermore, suggestions to use LLMs as a replacement instead of augmentative therapeutic technologies should be cautioned, as these results demonstrate that even lay people outperform LLMs in their current stage of development.

7 Conclusion and Future Work

In this work, we investigate the linguistic synchrony of an LLM in an interactive therapy session. We demonstrate that there is a statistically significant relationship between the linguistic synchrony of the LLM and the percentage of high intimacy and active engagement responses from the participants. We next compare the LLM’s linguistic synchrony with that of trained CBT therapists and non-expert online peer supporters. We find that the LLM is outperformed by both experts and non-experts in guiding participants through a CBT interaction.

In the future, we would like to investigate the use of other measures of linguistic synchrony, such as those based on part-of-speech distributions (Shapira et al., 2022) and usage rate of function words (Niederhoffer and Pennebaker, 2002), as a measure of therapeutic effectiveness. We selected nCLiD in this work because of its previous validation as a therapeutic measure (Nasir et al., 2019). However, we acknowledge that the use of static word embeddings in nCLiD does not account for context-aware word representations. We have designed metrics based on nCLiD that utilize contextual word embeddings, like BERT embeddings (Devlin et al., 2018), and in the future look to validate these metrics in a therapeutic context. Additionally, we would like to extend our analysis of the association of linguistic synchrony with other therapeutic measures beyond intimacy and engagement.

The LLM dataset used in this study was collected in interactions between GPT-3.5-turbo and participants. In future iterations of this project, we would like to test this with newer LLMs with an expanded token limit that would allow a longer interaction that parallels the interaction length of a full CBT session. In the LLM dataset, the responses were generated by a prompted model, and in the future, we would like to prompt an LLM that has been fine-tuned on therapy data to see if this improves the generated responses.

677 Limitations

678 In our comparison of the Official, LLM-Guided
679 CBT Exercises, and Reddit datasets (Section 5), the
680 data in each of these datasets come from inherently
681 different channels of communication: the Official
682 dataset contains transcriptions of real-time human-
683 to-human spoken conversations between a therapist
684 and client, while the Reddit dataset contains
685 asynchronous, online typed conversations, and the
686 LLM-Guided CBT Exercises dataset has real-time
687 typed conversations between a human and either a
688 robot or chatbot system. The difference in modal-
689 ity can lead to differences in the nature of the con-
690 versations and, therefore, introduce confounding
691 variables in our analysis of linguistic synchrony.
692 Additionally, although we worked to find data from
693 the CBT-related mental-health domain, the premise
694 in each dataset is also different. The Official dataset
695 comprises of full CBT sessions, while the LLM-
696 Guided CBT Exercises dataset comprises of CBT
697 exercises for a shorter duration. In Reddit, peo-
698 ple responded to posts asynchronously without a
699 specific therapy guideline to which they needed
700 to adhere. Since the Reddit dataset tends to fol-
701 low a short-form interaction instead of the length
702 expected in a full therapy session, its premise is
703 similar to that of the LLM-guided CBT exercises.

704 We also note that our datasets were quite small,
705 with approximately 30 interactions per dataset. In
706 the future, a larger sample size would yield more
707 insightful results.

708 Another limitation in our work is that the nCLiD
709 algorithm uses word2vec word embeddings, which
710 are static word embeddings. It is known that static
711 word embeddings limit the usage of multiple mean-
712 ings of words depending on the context, unlike
713 newer transformer-based contextual word embed-
714 dings such as BERT-based embeddings (Vaswani
715 et al., 2017; Devlin et al., 2018). We chose nCLiD
716 for this work since Nasir et al. (2019) validated
717 this metric in a therapy setting by demonstrating
718 its association with empathy. Additionally, in or-
719 der for nCLiD to be implemented with contextual
720 word embeddings, the nCLiD algorithm needs to
721 be changed fundamentally since it depends on word
722 frequency counts. This leads to a different metric
723 based on nCLiD and will therefore require addi-
724 tional validation.

725 Another note is that nCLiD averages the Word
726 Mover’s Distance values over all the turns in the
727 conversation, therefore potentially not capturing

temporal shifts in linguistic synchrony. nCLiD cap-
728 tures a specific aspect of linguistic synchrony; that
729 is, in terms of lexical semantic similarity. There
730 are other measures of linguistic synchrony that
731 look into, for example, part of speech distributions
732 (Shapira et al., 2022) and usage rate of function
733 words (Niederhoffer and Pennebaker, 2002). A
734 fuller picture of linguistic synchrony could be ob-
735 tained by evaluating our datasets with those metrics
736 as well.

737 Finally, we also note that we used data from
738 interactions with one version of an LLM (GPT-
739 3.5-Turbo). It is known that different LLMs have
740 varying levels of performance, and these findings
741 may only pertain to GPT-3.5-Turbo. 742

Ethical Considerations 743

744 The usage of LLMs, especially generative conversa-
745 tional models, in mental health can be risky. LLMs
746 can hallucinate, make false promises, and gener-
747 ate inappropriate ideas. In the LLM dataset, the
748 participants were screened for depression to ex-
749 clude vulnerable students. However, even in this
750 context, there is a risk that the LLM will gener-
751 ate something unhelpful, hurtful, or triggering that
752 negatively impacts the participant. While there are
753 many benefits of these LLM-based systems, such
754 as enabling frequent, interactive conversations that
755 the mental healthcare system cannot always pro-
756 vide, we caution against their use because of the
757 potential negative impacts. We advocate that LLMs
758 augment therapists by providing an accessible, in-
759 teractive version of the at-home exercises as done
760 in the LLM-guided CBT exercises study. Under
761 no circumstances do we support the use of LLMs
762 as a replacement technology, a narrative around
763 the potential use of LLMs that we find concerning.
764 Additionally, to ensure safety for those involved,
765 measuring the quality of LLMs with respect to the
766 mental health domain is critical. We hope that with
767 this work, we will contribute to the growing effort
768 of evaluating LLMs in the mental health domains.

References 769

- 770 Therapist Aid. 2017. *Cognitive restructuring: Socratic*
771 *questions: Worksheet.*
- 772 Noor Al-Sibai. 2023. *Openai employee says she’s never*
773 *tried therapy but chatgpt is pretty much a replacement*
774 *for it.*
- 775 Alexander Street Press. 2023. *Counseling and psy-*
776 *chotherapy transcripts: Volume i. Dataset.*

777	Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report . <i>Preprint</i> , arXiv:2305.10403.	
817	American Psychological Association and Society of Clinical Society. 2017. What is cognitive behavioral therapy?	
820	Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback . <i>Preprint</i> , arXiv:2212.08073.	
838	Emily M. Bender, Timnit Gebru, Angelina McMillan-	
	Major, and Shmargaret Shmittchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . <i>FACCT '21</i> , page 610–623, New York, NY, USA. Association for Computing Machinery.	839 840 841 842
	Štefan Beňuš. 2014. Social aspects of entrainment in spoken interaction. <i>Cognitive Computation</i> , 6:802–813.	843 844 845
	Justin W. Bonny and Anya M. Jones. 2023. Teams moving more synchronously are perceived as socially dominant . <i>Acta Psychologica</i> , 237:103952.	846 847 848
	Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners . <i>Preprint</i> , arXiv:2005.14165.	849 850 851 852 853 854 855 856 857 858 859 860
	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 <i>Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue</i> , pages 57–60, 1st virtual meeting. Association for Computational Linguistics.	861 862 863 864 865 866 867
	Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code . <i>Preprint</i> , arXiv:2107.03374.	868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888
	Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists . <i>Preprint</i> , arXiv:2401.00820.	889 890 891 892
	Yujin Cho, Mingeon Kim, Seojin Kim, OYun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. 2023. Evaluating the efficacy of interactive language therapy based on llm for high-functioning	893 894 895 896

897	autistic adolescent psychological counseling. <i>arXiv preprint arXiv:2311.09243</i> .	Adam O Horvath and Leslie S Greenberg. 1986. The development of the working alliance inventory. <i>The Psychotherapeutic Process: A Research Handbook.</i> , pages 529–556.	953
898			954
899	Munmun De Choudhury, Sachin R. Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health . <i>Preprint</i> , arXiv:2311.14693.	Adam O Horvath and B Dianne Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. <i>Journal of counseling psychology</i> , 38(2):139.	955
900			956
901			957
902			958
903	David A. Clark. 2013. <i>Cognitive Restructuring</i> , chapter 2. John Wiley Sons, Ltd.	Michael J Hove and Jane L Risen. 2009. It’s all in the timing: Interpersonal synchrony increases affiliation. <i>Social cognition</i> , 27(6):949–960.	959
904			960
905	Taylor M. S. Colton. 2022. <i>Linguistic synchrony: indicator or facilitator of therapeutic bond</i> . Ph.D. thesis, University of British Columbia.	Zac E Imel, Jacqueline S Barco, Halley J Brown, Brian R Baucom, John S Baer, John C Kircher, and David C Atkins. 2014. The association of therapist empathy and synchrony in vocally encoded arousal. <i>Journal of counseling psychology</i> , 61(1):146.	961
906			962
907			963
908	MA. Courtney E. Ackerman. 2017. Cbt techniques: 25 cognitive behavioral therapy worksheets .	Julie Jargon. 2023. Wsj news exclusive how a chatbot went rogue . <i>WSJ</i> .	964
909			965
910	Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. <i>IEEE Transactions on Affective Computing</i> , 3(3):349–365.	Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L�lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth�e Lacroix, and William El Sayed. 2023. Mistral 7b . <i>Preprint</i> , arXiv:2310.06825.	966
911			967
912			968
913			969
914			970
915			971
916	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. <i>arXiv preprint arXiv:1810.04805</i> .	Kathleen Kadner. 1994. Therapeutic intimacy in nursing. <i>Journal of Advanced Nursing</i> , 19(2):215–218.	972
917			973
918			974
919			975
920	Ning Fang, Chao Zhang, Supraja Sankaran, and Shaoya Ren. 2022. Role of socially assistive robots in reducing anxiety and preserving autonomy in children. In <i>2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)</i> , pages 754–759. IEEE.	Jay Kejriwal and Stefan Benus. 2024. Lexical, syntactic, semantic and acoustic entrainment in slovak, spanish, english, and hungarian: A cross-linguistic comparison. <i>SSRN</i> .	976
921			977
922			978
923			979
924			980
925			981
926	Barry A Farber. 2003. Patient self-disclosure: A review of the research. <i>Journal of clinical psychology</i> , 59(5):589–600.	Mina J. Kian, Mingyu Zong, Katrin Fischer, Abhyuday Singh, Anna-Maria Velentza, Pau Sang, Shriya Upadhyay, Anika Gupta, Misha A. Faruki, Wallace Browning, Sebastien M. R. Arnold, Bhaskar Krishnamachari, and Maja J. Mataric. 2024. Can an llm-powered socially assistive robot effectively and safely deliver cognitive behavioral therapy? a study with university students . <i>Preprint</i> , arXiv:2402.17937.	982
927			983
928			984
929	Barry A Farber, Kathryn C Berano, and Joseph A Capobianco. 2006. A temporal model of patient disclosure in psychotherapy. <i>Psychotherapy Research</i> , 16(4):463–469.	Sayaka Kidby, Dave Neale, Sam Wass, and Victoria Leong. 2023. Parent–infant affect synchrony during social and solo play. <i>Philosophical Transactions of the Royal Society B</i> , 378(1875):20210482.	985
930			986
931			987
932			988
933	Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askill, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned . <i>Preprint</i> , arXiv:2209.07858.	Lauren Rebecca Klein. 2023. <i>Modeling Dyadic Synchrony with Heterogeneous Data: Validation in Infant-Mother and Infant-Robot Interactions</i> . Ph.D. thesis.	989
934			990
935			991
936			992
937			993
938			994
939			995
940			996
941			997
942			998
943			999
944			1000
945			1001
946			1002
947	Michael V Heinz, Sukanya Bhattacharya, Brianna Trudeau, Rachel Quist, Seo Ho Song, Camilla M Lee, and Nicholas C Jacobson. 2023. Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. <i>Digital Health</i> , 9:20552076231170499.	Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. <i>Journal of general internal medicine</i> , 16(9):606–613.	1003
948			1004
949			
950			
951			
952			

1005	Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances . In <i>International Conference on Machine Learning</i> .	1061
1006		1062
1007		1063
1008		1064
1009	Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčik, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika . <i>New Media & Society</i> , 0(0):14614448221142007.	1065
1010		1066
1011		1067
1012		1068
1013		1069
1014		
1015	Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. <i>arXiv preprint arXiv:2303.15727</i> .	1070
1016		1071
1017		1072
1018	Ron Li, Steven Asch, and Nigam Shah. 2020. Developing a delivery science for artificial intelligence in healthcare . <i>npj Digital Medicine</i> , 3:107.	1073
1019		1074
1020		1075
1021	Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023. Llm360: Towards fully transparent open-source llms . <i>Preprint</i> , arXiv:2312.06550.	1076
1022		1077
1023		1078
1024		1079
1025		1080
1026		1081
1027		1082
1028		1083
1029		1084
1030		
1031	Gary Marcus. 2022. The dark risk of large language models. <i>Wired Magazine</i> . https://archive.ph/Dpdg7 .	1085
1032		1086
1033	Daniel J Martin, John P Garske, and M Katherine Davis. 2000. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. <i>Journal of consulting and clinical psychology</i> , 68(3):438.	1087
1034		1088
1035		1089
1036		1090
1037		1091
1038	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 1906–1919, Online. Association for Computational Linguistics.	1092
1039		1093
1040		1094
1041		1095
1042		1096
1043		
1044	Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality . <i>Preprint</i> , arXiv:1310.4546.	1097
1045		1098
1046		1099
1047		
1048	Teru L Morton. 1978. Intimacy and reciprocity of exchange: A comparison of spouses and strangers. <i>Journal of Personality and Social Psychology</i> , 36(1):72.	1100
1049		1101
1050		1102
1051		1103
1052	Md. Nasir, Sandeep Nallan Chakravarthula, Brian R. Baucom, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2019. Modeling interpersonal linguistic coordination in conversations using word mover’s distance . <i>Interspeech</i> , 2019:1423–1427.	1104
1053		1105
1054		1106
1055		1107
1056		1108
1057		
1058	Cory F Newman. 2002. A cognitive perspective on resistance in psychotherapy. <i>Journal of clinical psychology</i> , 58(2):165–174.	1109
1059		1110
1060		1111
		1112
		1113
	Tuan Dinh Nguyen, Marisa Cannata, and Jason Miller. 2018. Understanding student behavioral engagement: Importance of student interaction with peers and teachers. <i>The journal of educational research</i> , 111(2):163–174.	
	Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. <i>Journal of Language and Social Psychology</i> , 21(4):337–360.	
	Jesutofunmi A. Omiye, Jenna Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Beyond the hype: large language models propagate race-based medicine . <i>medRxiv</i> .	
	Aileen O’Brien, Rana Fahmy, and Swaran P Singh. 2009. Disengagement from mental health services: a literature review. <i>Social psychiatry and psychiatric epidemiology</i> , 44:558–568.	
	Jan Prasko, Iлона Krone, Julius Burkauskas, Jakub Vanek, Marija Abeltina, Alicja Juskiene, Tomas Soliar, Ieva Bite, Milos Slepecky, and Marie Ociskova. 2022. Homework in cognitive behavioral supervision: theoretical background and clinical application. <i>Psychology Research and Behavior Management</i> , pages 3809–3824.	
	Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.	
	Fabian Ramseier and Wolfgang Tschacher. 2011. Non-verbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. <i>Journal of consulting and clinical psychology</i> , 79(3):284.	
	Sara Reardon. 2023. Ai chatbots could help provide therapy, but caution is needed .	
	Reddit. 2022. Chatgpt is better than my therapist, holy shit .	
	Miriam Rennung and Anja S Göritz. 2016. Prosocial consequences of interpersonal synchrony. <i>Zeitschrift für Psychologie</i> .	
	Katreena L Scott and Colin B King. 2007. Resistance, reluctance, and readiness in perpetrators of abuse against women and children. <i>Trauma, Violence, & Abuse</i> , 8(4):401–417.	
	Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In <i>Proceedings of the CHI Conference on Human Factors in Computing Systems</i> , pages 1–20.	
	Natalie Shapira, Dana Atzil-Slonim, Rivka Tuval-Mashiach, and Ori Shapira. 2022. Measuring linguistic synchrony in psychotherapy. In <i>Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology</i> , pages 158–176.	

1114	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 <i>EMNLP</i> .	1170
1115		1171
1116		1172
1117		1173
1118		1174
1118	Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities . In <i>Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems</i> , CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.	1175
1119		1176
1120		
1121		1177
1122		1178
1123		1179
1124	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. <i>npj Mental Health Research</i> , 3(1):12.	1180
1125		1181
1126		1182
1127		1183
1128		
1129		1184
1130		1185
1131	Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. <i>Advances in Neural Information Processing Systems</i> , 33:3008–3021.	1186
1132		1187
1133		
1134		1188
1135		1189
1136		1190
1137		1191
1138		1192
1139	Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science . <i>Preprint</i> , arXiv:2211.09085.	1193
1140		
1141		1194
1142		1195
1143	Amanda Tetley, Mary Jinks, Nick Huband, and Kevin Howells. 2011. A systematic review of measures of therapeutic engagement in psychosocial and psychological treatment. <i>Journal of Clinical Psychology</i> , 67(9):927–941.	1196
1144		1197
1145		1198
1146		1199
1147		1200
1148		
1149	Betsy E Tolstedt and Joseph P Stokes. 1984. Self-disclosure, intimacy, and the depenetration process. <i>Journal of Personality and Social Psychology</i> , 46(1):84.	1201
1150		1202
1151		1203
1152		1204
1153		1205
1154		
1155		1206
1156		1207
1157		1208
1158		
1159		1209
1160		
1161		1210
1162		1211
1163		
1164		1212
1165		1213
1166		1214
1167		1215
1168		1216
1169		1217
		1218
		1219
		1220
		1221
		1222

1223 **Prompt for cognitive restructuring exercise:**

1224 **Prompt:** Cognitive restructuring
 1225 is a strategy to help the patient
 1226 identify cognitive distortion and
 1227 find evidence to challenge the
 1228 distortion. You are a therapist
 1229 who uses cognitive restructuring
 1230 to help your patient in this
 1231 session.

Parameter	Value
model	gpt-3.5-turbo
messages	<complete transcript including the user's responses>
stop	Patient
temperature	1
frequency_penalty	2
presence_penalty	2
n	2
max_tokens	150

Table 1: Input parameters for OpenAI's chat completion API

Package	Version
robustHD	0.8.0
readxl	1.4.3
ggcorrplot	0.1.4.1
rstudioapi	0.15.0
dplyr	1.1.4
tidyr	1.3.0
afex	1.3-0
tidyverse	2.0.0
ggpubr	0.6.0
rstatix	0.7.2
outliers	0.15
pastecs	1.4.2
psych	2.3.12
car	3.1-2
lmtest	0.9-40
moments	0.14.1
gmodels	2.19.1
pgirmess	2.0.3
heplots	1.6.2
Rmisc	1.5.1
ggplot2	3.4.4
jmv	2.4.11
haven	2.5.4
stats	4.3.2
multcomp	1.4-25

Table 2: R Package Versions Used in the Analysis

1232 **B R Packages and Versions**

1233 See Table 2 for the list of R packages used.

1234 **C CBT-Related Terms**

1235 CBT-related terms used as keywords when filtering
 1236 relevant Reddit threads.

cbt	cognitive behavioral therapy
coping mechanisms	negative thinking
emotional regulation	reframing
cognitive reframing	coping
coping strategies	coping strategy
coping skill	coping skills
coping mechanisms	coping mechanism
cognitive restructuring	cognitive distortions
cognitive distortion	distortion
distortions	catastrophize
overthink	overthinking
personalize	overgeneralize
mental filter	discount positives
catastrophize	magnifying negatives
minimizing positives	jumping to conclusions
mind read	fortune tell
emotional reasoning	black-and-white thinking
all-or-nothing thinking	all or nothing
mental filter	personalization
should statements	mental filter
labeling	catastrophizing
awfulizing	mind reading
fortune telling	magnification
minimization	disqualification of positives
overgeneralization	jump to conclusions
jumping to conclusions	overgeneralizing
restructuring	

D Excerpts from datasets

1237

WARNING: This section contains expletives and sensitive content related to suicide and self-harm.

1238

1239

Below are random excerpts from each dataset in order for a sample of the nature of each of the datasets.

1240

1241

1242

In Figures 6, 7, and 8, excerpts from a random conversation from the Official dataset are shown, a few turns from the beginning, middle, and end, respectively.

1243

1244

1245

1246

In Figures 9, 10, and 11, excerpts from a random conversation from the Reddit dataset are shown, a few turns from the beginning, middle, and end, respectively.

1247

1248

1249

1250

In Figures 12, 13, and 14, excerpts from a random conversation from the LLM-Guided CBT Exercises dataset are shown, a few turns from the beginning, middle, and end, respectively.

1251

1252

1253

1254

THERAPIST: So what's happening?	
	CLIENT: It's been a while, hasn't it?
THERAPIST: It's been a week, yes. Because I think I saw you a week from Thursday. I saw you like the sixth.	
	CLIENT: Okay.
THERAPIST: Yes.	
	CLIENT: Yeah. It feels like a lot of water under the bridge.
THERAPIST: Or the thirteenth. I can't remember.	
	CLIENT: I think the thirteenth.
...	

Figure 6: An example from the Official dataset: an excerpt of the beginning of the conversation

...	
CLIENT: But, you know, [NAME] is like, "How can you... How can you..." Whatever she said, the word. I don't know. "...break the hearts of our children," and all that kind of thing. And I'm just feeling like continuing this crap and giving them that for how many more years, you know, it's like...	
THERAPIST: But see even that is a very emotional blackmail.	
	CLIENT: Manipulative.
THERAPIST: "Break the hearts of our children." Well, your children are grown and it doesn't mean they're immune to this." Okay? But it means you will have conversations with your grown children and you will say what's true for you and you will acknowledge how difficult it is for the family. But are you breaking the hearts of your children? You know, I don't...	
	CLIENT: Here's another thing. It's like, "Okay." I get that what she's saying.
THERAPIST: Sure, sure. It's definitely sad and they will have a reaction to that. But that's really... That's a finger point. "You're breaking the hearts of our children."	
	CLIENT: Oh yeah.
...	

Figure 7: An example from the Official dataset: an excerpt of the middle of the conversation

E ConvoKit Dyadic Reddit Thread Extraction

Data Collection and Preprocessing To narrow our search, we looked for subreddits (online topic-based communities on Reddit where people view/make posts) where people discussed mental health-related topics. To do so, we selected the subreddits listed as OMHCs that focused on *Psychosis & Anxiety* and *Coping & Therapy* (Sharma and De Choudhury, 2018) as these are most relevant to cognitive restructuring and coping strategy exercises used in the LLM-Guided CBT dataset. Since we were looking for untrained individuals, we reviewed the subreddit rules and descriptions for each listed subreddit and excluded those communities that were explicitly intended for or had a high presence of therapists or professionals advising people (e.g., r/Therapy, r/askatherapist, etc.). From this process, we identified 40 candidate subreddits for further screening.

The ConvoKit Reddit Corpus (Chang et al.,

2020)⁶ is a corpus of Reddit data containing all posts and comments from an individual subreddit from its inception until October 2018. This corpus can be traversed using ConvoKit's API so that each post can be accessed in a thread/tree-like manner, with the root being the main post and each response being a node connected to the post/comment being replied to. Using the ConvoKit 3.0.0 API, we traversed every post in each of the 40 selected subreddits, extracting conversation threads with alternating utterances of responses between the original poster and a particular commenter (see Figure 15). Since Reddit posts can have multiple reply threads with various people replying at each level, we applied this constraint to ensure we only extracted dyadic conversations. After we extracted a dyadic conversation thread, we only include the thread if the number of utterances in the conversation was greater than equal to a minimum threshold (based on the average number of utterances in the LLM-Guided CBT Exercises dataset), ensuring the thread was of sufficient length. We did not keep the user-

⁶distributed under the MIT license

...
CLIENT: I'm surrounded by people that are doing that. My wife's, you know, all of the friends.
THERAPIST: Yes. But that's not okay with you anymore.
CLIENT: They're doing that. I'm like... I have to live on the island of misfit toys. Right?
THERAPIST: (LAUGHTER) So I will see you next week.
CLIENT: Thank you.
THERAPIST: Okay. Let me know how things go.
CLIENT: I am sorry. I appreciate the time.
THERAPIST: No. Problem at all. Take care [NAME]. See you next week.

Figure 8: An example from the Official dataset: an excerpt of the end of the conversation

SEEKER: I want to die. People try to help only to realize I'm a lost cause. There is no helping me. I'm beyond it. It's ok. Maybe i deserve it. Maybe i was meant to be this way. Who knows. But i know that living out each day is a new struggle. I hate it. I want everything to stop.. especially the thoughts. I just want to go away and disappear.
PEER SUPPORTER: Maybe people try to help and merely realize they don't know what they're doing, or don't know where to start. Maybe believing you're a lost cause or that you deserve to be miserable makes it somehow more likely that you stay (and maybe become even more) miserable. I somehow doubt that the depth of the misery and struggle you're in right now is "deserved" by anyone. The shitty part is, often the people who are 'cursed' with the type of misery you describe are the one who can't see that they don't actually deserve it. ...If that makes any sense. On another note, what kinds of thoughts do you wish would stop?
SEEKER: "Life is pointless. I don't want to do this anymore. I <u>D</u> ont want to wake up tomorrow. Damn it I woke up. Why cant i just have an aneurysm? I want to disappear. "
PEER SUPPORTER: Yeah, that sounds pretty fucking awful and I'm sorry you have to deal with it at all, especially on a daily basis. These are the thoughts of desperation, of wanting to give up, of wanting to escape... what are you trying to escape from? (e.g. is there a deeper layer of thoughts that are causing these thoughts to appear? or is it sadness, loneliness, fear... or something else altogether)
...

Figure 9: An example from the Reddit dataset: an excerpt of the beginning of the conversation

names associated with each Reddit post/comment.

Filtering and Screening To ensure that the selected threads were broadly related to CBT, we included only those threads that contained at least one keyword from a dictionary of keywords identified from conversations gathered by Kian in their study (2024). Refer to Appendix C for the full list of terms used. The dictionary of 53 keywords contained common cognitive distortions, thinking traps, and phrases related to CBT. The dictionary filtering step was conducted right after a candidate dyadic thread was identified in the post, and the thread was only included if it also passed the filtering criteria. After running the extraction, preprocessing, and filtering on the 40 selected subreddits, we extracted 683 dyadic conversations.

Lastly, to exclude erroneous conversations that may have evaded the filtering process, we had 3 reviewers who were well-versed in conversations on cognitive restructuring and coping strategy exercises from the LLM-Guided Exercises dataset, screen conversations to make sure that they par-

alleled guided CBT exercises. The screeners included conversations in the final dataset if they noticed indicators of cognitive restructuring or coping strategies being discussed. The screeners identified cognitive restructuring in the conversation if the responder sought evidence in the original poster's claims, provided counterarguments to the original poster's beliefs, followed Socratic questioning techniques, or named cognitive distortions they believed the original poster exhibited. Subsequently, if there was any discussion between the original poster and the responder about activities and strategies used by either individual to deal with their emotions (regardless of efficacy), there was evidence for coping strategies being explored. The screeners also excluded conversations if the conversation seemed off topic or if the CBT exercise was only a small part of the conversation. The reviewers screened a subset of the 683 extracted conversations and identified 30 conversations that strongly paralleled CBT exercises and were included in the Reddit dataset.

Since Reddit allows for people to respond to posts and replies asynchronously, this may result

1298
1299

1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320

1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344

...
SEEKER: When all you've known is darkness, how do you even begin to see the light?
PEER SUPPORTER: It makes it much harder, admittedly. But not impossible. I don't know if you're into philosophy or ever read Plato, even in school. But there's a piece he wrote called Allegory of the Cave that described this phenomenon. Basically a bunch of people in a cave away from any light chained up so they can't escape the cave. And how when someone manages to make it out into the sunlight and return to tell everyone about it, he is seen as crazy or whatever. In the end it took a bit of a leap of faith on the part of the chained prisoners, but once they made the leap to believe there was at least a possibility of something other than darkness, the "crazy" man would guide them out of the cave and into light. So, a leap of faith and someone that can guide you may be start?
SEEKER: So what are you getting at?
PEER SUPPORTER: Sorry, my point, I think, got lost in the analogy. When all you've known is darkness the leap of faith is trusting that there is some light out there in addition to the darkness when others tell you there is despite never seeing it yourself. The "someone to guide you" could be a therapist or friend or SO or family member...anyone, though professionals have the training if you're willing to go that route.
...

Figure 10: An example from the Reddit dataset: an excerpt of the middle of the conversation

...
SEEKER: I've had 4. They all say the same thing. Plus the fact that they're getting paid bothers me
PEER SUPPORTER: What do they all say?
SEEKER: It'll get better. Accept yourself. Get on antidepressants. That's the jist of it
PEER SUPPORTER: Sorry for the delayed reply - life got a bit chaotic, as it tends to from time to time. This doesn't seem like solid advice...or much of advice at all. All of the statements you mentioned are useless unless you know what you can DO to get better, or HOW to start accepting yourself, etc. Getting on antidepressants could help, but a therapist's job is not just to be the middle man to get people into a Dr. for meds. They are supposed to help in another way. I hate to say you've been unlucky all these times, and shake my head at the quality of therapists you've ended up with. There are better ones out there, I promise. I know therapists use different types of therapy - psychoanalysis, cognitive behavioral, family systems, attachment based, etc etc etc. MY best luck in finding solutions rather than the same old questions has been with CBT (cognitive behavioral) or ACT (acceptance and commitment therapy). CBT is a bit more common I think but it may be worth trying one more shot for a therapist/psychologist that specializes in CBT in particular because CBT is so behavior and action-based. Meaning you're continually given things to actually do about your depression, anxiety, pain, etc.

Figure 11: An example from the Reddit dataset: an excerpt of the end of the conversation

1345 in different linguistic characteristics than those cap-
1346 tured in real-time conversations. However, we ob-
1347 served that in most OMHC subreddits, posters of-
1348 ten seek advice and engage with commenters in a
1349 timely manner.

1350 **F nCLiD normalization factor α**

1351 The full equation for the normalization factor
1352 *alpha* in:

$$1353 \quad nCLiD = \frac{uCLiD}{\alpha} \quad (3)$$

$$1354 \quad \alpha = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N WMD(a_i, a_j) \\
+ \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N WMD(c_i, c_j) \quad (4) \\
+ \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i}^N WMD(a_i, c_j)$$

PARTICIPANT: Hello. It's going better than yesterday!
LLM: That's great to hear. Can you tell me more about what made yesterday difficult for you?
PARTICIPANT: Yesterday, even after my discussion with you, I couldn't will myself to do the work I wanted to do because I was so tired but I was too stubborn to go to sleep so I got less sleep and less work done than I should have.
LLM: I see. So when you were feeling tired, what thoughts came to your mind that made it difficult for you to go to sleep? ...

Figure 12: An example from the LLM-Guided CBT Exercises dataset: an excerpt of the beginning of the conversation

...
PARTICIPANT: Oh I never thought about putting well-being tasks on my to-do list. I suppose not all things on the list are equally valuable. I have started dividing my to-do list into three sections of urgent, should-work on, and "only if you have extra time" to make myself less overwhelmed. I think I have started putting more items in the urgent to-do list than I realistically should because of my ideal instead of factoring in my well-being or giving myself wiggle room for when I am sick or feeling tired
LLM: That's a great step you have taken by dividing your to-do list into different categories. It shows that you are aware of the importance of setting priorities for yourself and managing your time well. It might also be helpful to schedule in self-care activities such as exercise, meditation or simply taking breaks throughout the day where you can recharge yourself mentally and physically. Remember, it is okay not to check off everything on our lists every single day - what matters most is making sure we take care of ourselves along the way!
PARTICIPANT: I have a question about meditation. What if I have tried and failed having a steady meditation practice several times in the past few years? Is it possible that meditation just doesn't work for me?
LLM: That's a great question! It is important to remember that meditation doesn't have to look one specific way, and what works for someone else might not work as well for you. There are many different types of mindfulness practices out there - perhaps exploring some new techniques or trying guided meditations could be helpful. It might also be beneficial to explore whether any cognitive distortions are impacting your experience with meditation - such as feeling like you need it to perfect every time or being overly critical of yourself when things don't go exactly right- which may affect how open-minded and accepting we can allow ourselves in our practice. Ultimately, the goal is not perfection but finding ways that help us feel more grounded, present on mindful day-to-day experiences & emotions. ...

Figure 13: An example from the LLM-Guided CBT Exercises dataset: an excerpt of the middle of the conversation

...
PARTICIPANT: Would it be fair to say that meditation is kind of like the gym for the mind?
LLM: Yes, that's a great comparison! Just like how the gym helps to strengthen and maintain our physical health, meditation can help us work on strengthening our mental wellbeing by training ourselves to approach challenging situations with more mindfulness and compassion. It takes practice just like going regularly at the gym- so it's important for you not judge yourself too harshly or expect perfection when meditating as well. Thank you for being open about your struggles today - let's revisit next session & continue working together towards these cultivating new goals and building self-confidence in achieving them.
PARTICIPANT: Thank you so much! So just to summarize for myself, not all tasks are equal and taking care of yourself (like getting enough sleep, self-care exercises like exercise or meditation or breaks) is a high priority task that I can potentially add onto my list. Next, meditation is like a gym for the mind and it is not supposed to really look any certain way plus I can try out different techniques
LLM: Absolutely! You've summarized it very well. Remember to take small steps and be kind to yourself as you work towards these goals - progress is not always visible overnight, but with consistent effort and reflection we can cultivate long-term habits that enhance our mental wellbeing. See you next session!

Figure 14: An example from the LLM-Guided CBT Exercises dataset: an excerpt of the end of the conversation

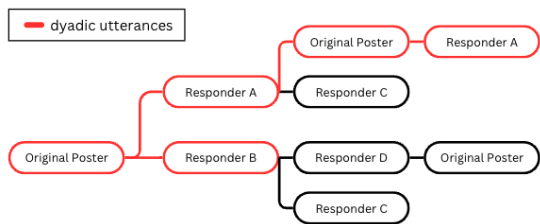


Figure 15: Extraction process of dyadic Reddit comment threads between the original poster and a specific responder.