Using Linguistic Entrainment to Evaluate Large Language Models for Use in Cognitive Behavioral Therapy

Mina Kian[1,*], Kaleen Shrestha[1,*], Katrin Fischer[2], Xiaoyuan Zhu[1], Jonathan Ong[1], Aryan Trehan[1], Jessica Wang[1], Gloria Chang[1], Sébastien M. R. Arnold[1], Maja Matarić[1]

[1]Department of Computer Science, University of Southern California,
[2]Annenberg School for Communication and Journalism, University of Southern California
{kian, kshresth, katrinfi, xzhu9839, ongjd, atrehan, jmwang, changglo, seb.arnold, mataric}@usc.edu

Abstract

Entrainment, 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 entrainment 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 entrainment of an LLM (ChatGPT 3.5-turbo) in a mental health dialog setting. We first validate computational measures of linguistic entrainment with two measures of the quality of client self-disclosures: intimacy and engagement ($p < 0.05$). We then compare the linguistic entrainment 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 entrainment ($p < 0.001$). These results support the need to be cautious in using LLMs out-of-the-box for mental health applications.

1 Introduction

Entrainment 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). Entrainment 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* entrainment 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 such contexts (Chiu et al., 2024; Cho et al., 2023). To the best of our knowledge, entrainment has not yet been evaluated as a performance indicator, in spite of its critical role (Kejriwal and Benus, 2024) in developing a strong therapist-client relationship in mental health therapy. Therefore, in this work we measure the performance of an LLM (GPT-3.5-Turbo) in a mental health setting with respect to linguistic entrainment. We demonstrate that there is a significant relationship between linguistic entrainment and two measures of the quality of client self-disclosures: intimacy and engagement. We operationalize linguistic entrainment through Linguistic Inquiry and Word Count (LIWC) (Gonzales et al., 2010; Ireland and Pennebaker, 2010) and normalized Conversational Linguistic Distance (nCLiD) (Nasir et al., 2019). We then compare the LLM performance to trained therapists and non-expert online peer supporters in a cognitive behavioral therapy (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 that linguistic entrainment may shed light on the evaluation of LLMs intended for use in mental health contexts.

2 Background

We overview cognitive behavioral therapy, measures of therapy effectiveness, and the key role of linguistic entrainment 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

[*] Equal contribution

Findings of the Association for Computational Linguistics: NAACL 2025, pages 7739–7758
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

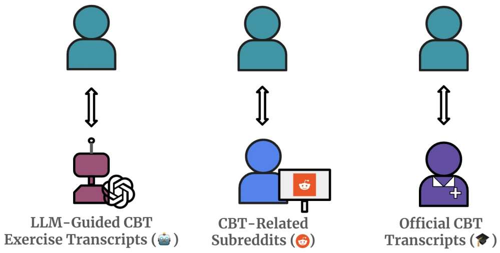

Figure 1: In our assessments, we compare the linguistic entrainment of CBT-related dialogues with an LLM (left), online peer supporters on Reddit (middle), and trained CBT therapists (right). More information about the data collection can be found in Sections 4 and 5.

adaptive behaviors (Wenzel, 2017). A core principle of CBT is that psychological disorders and their symptoms arise from unhelpful patterns of thought and behavior and that directly addressing these patterns can relieve symptoms.

CBT homework exercises are a critical aspect of CBT treatment that allows patients to practice in a natural environment what they learn in therapy sessions (Prasko et al., 2022). Common examples of CBT homework exercises include scheduling specific activities, such as 10-minute breaks, and self-monitoring by recording the frequency of a behavior in a journal (Kazantzis et al., 2007). CBT homework exercises are a valuable at-home component of CBT therapy that complements therapy sessions and leads to better patient outcomes (Kazantzis et al., 2010). Our work analyzes a dataset from a study that deployed an at-home LLM-guided CBT exercise activity (Section 4). In that context, LLMs can potentially be useful in administering CBT homework exercises.

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 the patient 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 Entrainment in Therapy

It is well established that the quality of the relationship between a therapist and their client plays a significant role in how effective therapy is for that 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 the two form a strong emotional bond (Horvath and Greenberg, 1986).

The concept of *interpersonal entrainment*[*] describes the participants of an interaction adapting and converging on each other's behaviors over time. Higher levels of entrainment 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; Doré and Morris, 2018). Vail et al. (2022) investigate the relationship between language and working alliance, focusing on *linguistic entrainment*, which refers to the interlocutors' similarity in semantics, syntax, or style. They find that the therapist's linguistic entrainment strongly impacts the client's perception of the working alliance. Nasir et al. (2019) introduce the normalized Conversational Linguistic Distance (nCLiD), a metric of linguistic entrainment between two speakers. In their study, they show that nCLiD is associated with positive therapeutic measures, specifically with the therapist's level of empathy towards the client, and with affective behaviors of couples in therapy. Gonzales et al. (2010); Ireland and Pennebaker (2010) use a metric called language style matching (LSM) that aims to capture language style similarity between two interlocutors by measuring the similarity in use of function words (e.g., articles, prepositions, and conjunctions). We use nCLiD and LSM to evaluate the linguistic entrainment in CBT interactions.

[*] The 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 among participants in an interaction. We use the term entrainment in this paper.

3 LLM Use in Mental Health Therapy

The prevalence of mental health conditions and the lack of accessible care has created a void that many have attempted to address with assistive therapeutic technologies powered by LLMs (Youper; Choudhury et al., 2023). 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 their impressive text-generating capabilities, LLMs can cause serious repercussions in sensitive tasks, such as by propagating harmful stereotypes and biases (Bender et al., 2021) and encouraging suicide (Marcus, 2022). For example, LLMs suffer from hallucinations and produce harmful or factually incorrect outputs (Zhang et al., 2023; Ganguli et al., 2022; Maynez et al., 2020), demanding research into techniques that mitigate those 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 (Choudhury et al., 2023; Laestadius et al., 2022; Youper). People have turned to LLMs when facing mental health problems, and reported feeling heard and supported, comparing the experience to that of interacting with a human therapist (Reardon, 2023; Al-Sibai, 2023; Reddit, 2022). Cho et al. (2023) tested an LLM 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 mental health use cases, 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 after it gave misguided medical advice (Jargon, 2023). Replika was implicated in a UK criminal case for encouraging a man to attempt to assassinate the Queen and then commit suicide (Weaver, 2023). Despite many serious issues, 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 entrainment (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). In a study by Cho et al. (2023), clinical psychologists and psychiatrists evaluated an LLM with respect to empathy, communication skills, adaptability, engagement, and ability to establish 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 to high-quality and low-quality human therapist transcripts (Pérez-Rosas et al., 2019; Malhotra et al., 2022). 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 use in therapy. The LLM-participant dataset used in our work comes from an IRB-approved study by Kian et al. (2024) that deployed an LLM in an interactive CBT homework context with university students (Section 4); therefore, our work provides a step toward a more realistic evaluation of LLMs used in therapy. Additionally, given the importance of entrainment (Section 2.3), we introduce the use of linguistic entrainment to evaluate an LLM-powered mental health dialog system (Section 4).

4 Study 1: Evaluation of Linguistic Entrainment in Therapy

In this study, we aim to evaluate linguistic entrainment, operationalized by LSM and nCLiD, as a measure of therapist response quality, by demonstrating that these entrainment measures are associated with indicators of positive therapeutic outcomes, specifically engagement and intimacy

(Note: higher linguistic entrainment is operationalized through a *higher* LSM score and *lower* nCLiD score). We present the following hypotheses:

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

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

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

4.1 Methodology

We performed our analysis on English language transcripts of LLM-guided CBT homework exercises annotated for intimacy and engagement and calculated LSM and nCLiD scores for all transcripts. We conducted linear regressions to analyze the relationship between LSM and nCLiD with the therapeutic measures.

4.1.1 Participants and Procedure

We analyzed transcripts derived from a dataset by Kian et al. (2024) of LLM-powered robot and LLM-powered chatbot CBT homework exercise interactions with university students. Study participants were screened to be over 18 years of age, proficient in English, have normal or corrected-to-normal vision and hearing, and live near campus. The Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001) was used as a screening tool, and individuals with a score of 15 or higher, indicating moderately severe to severe depression, were excluded as a safety measure. All individuals who filled out the screening materials were shown information about university mental health resources. A total of 26 students participated in the study conditions we assess in this work (we excluded the condition not using an LLM). Before the start of the study, all participants had an informed consent meeting with a member of the research team. This study was approved by their university's Institutional Review Board (IRB), and all participants were compensated with a US $150 Amazon gift card. This amount was calculated based on expected hours spent on the study and local minimum wage. The study duration was 15 days; in the first 8 days, the CBT homework sessions were required, while during the last 7 sessions, they were optional. Each day, the participants logged into a secure portal and completed an LLM-powered robot- or chatbot-guided CBT exercise (depending on the study condition). They selected from two CBT exercise options: Cognitive Restructuring (Clark, 2013) or Coping Strategies (Association and of Clinical Society, 2017). The LLM was GPT-3.5-turbo[*], prompted to use the participant-selected strategy while acting as a therapy guide (see Appendix A for the prompts used). All identifiable data for this study were securely stored on IRB-approved secure cloud storage. Only IRB-approved researchers had access to the data. The study produced the LLM-Guided CBT Exercises Dataset used in this work. This study was approved by our university's IRB under UP-22-01080.

4.1.2 Measures

Descriptive and evaluative intimacy were assessed according to the Morton (1978) framework. *Descriptive intimacy* involves the disclosure of private facts, while *evaluative intimacy* involves the disclosure of personal opinions and feelings. Each dimension was dichotomized into high and low disclosure levels, as recommended by Tolstedt and Stokes (1984). Next, we assessed engagement, which measures how much a participant actively participated in the sessions. Engagement was annotated to be active or passive according to Nguyen et al. (2018). Finally, to operationalize the linguistic coordination between the participants and the LLM, we used the normalized Conversational Linguistic Distance (nCLiD) by Nasir et al. (2019).

4.1.3 Annotation Process

In prior work, the CBT exercise transcripts were annotated for three variables: descriptive intimacy (DI), evaluative intimacy (EI), and engagement (Eng). In that work, four undergraduate student annotators (two female, two male) were trained through workshops led by PhD student instructors for two weeks to annotate the data for the selected variables. Each participant's turn in response to the LLM was annotated, resulting in an average of 10-15 annotations per participant per day. The Inter-Coder Reliability (ICR) was measured using 10% of the dataset, resulting in 83.5% and Cohen's average kappa score of $\kappa = 0.602$. Finally, annotations were aggregated to yield percentages of active engagement, high descriptive intimacy, and high evaluative intimacy averaged across all study days per participant, which we use in subsequent

[*] `https://platform.openai.com/docs/models/gpt-3-5-turbo`

analyses. For additional details on the annotation process, see Appendix B.

4.1.4 Language Style Matching

LSM[*] quantifies language style similarity between two interlocuters by measuring similarity in usage of function words (e.g., articles). Unlike content words (e.g., nouns), function words capture the speaker's style because they do not have meaning on their own (Ireland and Pennebaker, 2010). Function words are connected to social, situational, and individual processes and have been studied with respect to linguistic entrainment in conversational settings (Ireland and Pennebaker, 2010; Gonzales et al., 2010).

4.1.5 nCLiD Algorithm

The Conversational Linguistic Distance (CLiD) (Nasir et al., 2019) is an asymmetric distance metric that quantifies the interpersonal linguistic entrainment between two speakers. Higher linguistic entrainment is described by lower CLiD scores, and vice versa. Nasir et al. (2019) demonstrated that nCLiD correlates with ratings of a therapist's empathy toward their patient (CLiD is lower for a higher therapist empathy rating) and with affective behaviors in Couples Therapy (CLiD is proportional to negative affect, and inversely proportional to positive affect).

For a therapy session text record D between a therapist T and a patient P, consisting of N turns of interleaving utterances with $D = [t_1, p_1, t_2, p_2, ..., t_N, p_N]$, let us consider one speaker as the anchor A, and the other as the coordinator C. For each anchor utterance a_i, we compute $d_i^{C \to A}$ for the minimum distance between the sequences of *word2vec* (Mikolov 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 the linguistic difference between the two utterances:

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

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, below). 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.

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

The normalization factor α accounts for spurious coordination by accounting for potential coordination within A and B, and between A and B. Appendix C includes the full equation for α.

We performed a swap experiment to determine if nCLiD scores reflect entrainment, on the LLM-Guided CBT Exercises Dataset. For each conversation D in the LLM-Guided CBT Exercises Dataset, we selected the first five rounds of conversations between the LLM T and participants P, forming $D_a = [t_1, p_1, ..., t_5, p_5]$. For each round in D_a, we swap the LLM responses with the LLM responses from another conversation D_b to form $D_{swapped_{ab}} = [t_1^b, p_1^a, ..., t_5^b, p_5^a]$. $D_{swapped_{ab}}$ represents a conversation with low entrainment since we swapped out the original LLM responses with unrelated responses from a different dialogue. We then calculate the nCLiD scores for the rounds in D_a and mismatched rounds $D_{swapped_{ab}}$. The results are shown in Figure 2. The mean and standard deviation for the selected rounds D_a are $\mu_A = 0.335$ and $\sigma_A = 0.015$, while for the mismatched rounds $D_{swapped_{ab}}$, they are $\mu_{AB} = 0.346$ and $\sigma_{swapped_{ab}} = 0.011$. We conduct a Welch's t-test and find a significant difference between the nCLiD scores for D_a and $D_{swapped_{ab}}$: $t(539.95) = -12.82, p < 0.001$. The lower distribution of D_a nCLiD scores supports the use of nCLiD as a measure of entrainment. For further illustration of the mechanics of WMD with respect to entrainment, see Appendix D.

We implemented nCLiD using the WMD algorithm from the gensim 4.3.2 library (Řehůřek and Sojka, 2010) 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 use of stop words.

[*] We use the LSM implementation provided by the Linguistic Inquiry and Word Count (LIWC) software under a purchased academic license (Boyd et al., 2022)

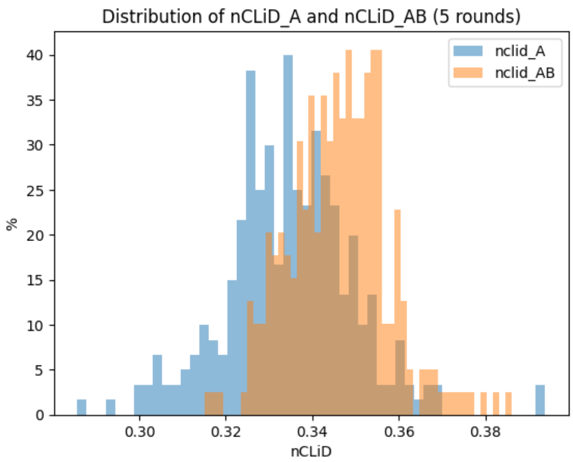

Figure 2: nCLiD distributions for selected rounds D_a and mismatched rounds D_{ab}.

4.1.6 Data Analysis

We implemented tests in R version 4.3.2; a list of all R packages and their versions is available in Appendix E. We ran linear regression tests to assess the relationship between the entrainment scores and therapeutic measures. We used the Durbin-Watson test of autocorrelation 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 failed the Breusch-Pagan test, we applied Huber-White standard errors.

4.2 Results

We performed linear regression of the transcripts' intimacy and engagement onto their entrainment (nCLiD and LSM) scores. The results are in Table 1, with visualizations in Figure 3. As seen in the table, nCLiD and LSM were both significant predictors of intimacy (DI and EI) and engagement, respectively.

We find that nCLiD is inversely related to all three measures, as indicated by the negative value for the main effect of each of the relationships. Similarly, we found that LSM is positively related to all three measures, as indicated by the positive value for the main effect of each of the relationships. This indicates that higher entrainment is associated with higher intimacy and active engagement (*supporting H1a, H1b, and H1c*). For additional assessments evaluating the disaggregated LSM function word categories, see Appendix F.

5 Study 2: LLM vs. Human Comparison

In this study, we compare the linguistic entrainment of the LLM against trained mental health therapists and non-expert online peer supporters. To do so, we obtained two more datasets: a subset of the Alexander Street Press Counseling and Psychotherapy Transcripts (ASPCPT) with trained expert human therapists (Official CBT Dataset), and a dataset developed from CBT-like conversations on Reddit with online non-expert peer supporters (Reddit Dataset). We put forth the following hypotheses.

Transcripts from trained CBT therapists will have higher linguistic entrainment than the LLM-guided exercises, which will, in turn, have higher linguistic entrainment than non-expert online peer supporters (note: higher linguistic entrainment is operationalized through a *higher* LSM score and *lower* nCLiD score). Specifically:

H2a: Linguistic entrainment 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 entrainment 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 entrainment 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 conducted a one-way Analysis of Variance (ANOVA) test across entrainment scores for the three datasets to determine how they perform relative to one another.

5.1.1 Added Datasets

Reddit Dataset We extracted a collection of 30 English CBT-related dyadic conversations from Reddit[*] posts in Online Mental Health Communities (OMHCs) (Sharma and De Choudhury, 2018) that included indicators of coping strategy (Courtney E. Ackerman, 2017) or cognitive restructuring (TherapistAid, 2017; Clark and Egan, 2015) exercises. We chose these exercises because they were the ones used in the LLM-guided CBT transcripts

[*] The Reddit Dataset can be found on ConvoKit (Chang et al., 2020): `https://convokit.cornell.edu/documentation/subreddit.html`

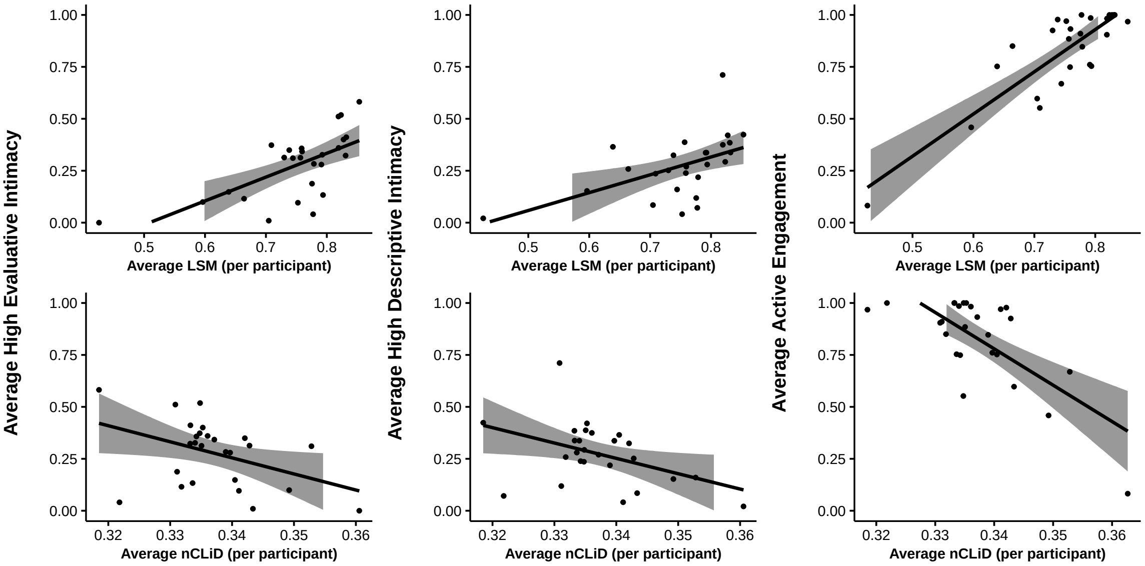

Figure 3: Scatterplots of average high descriptive intimacy, average high evaluative intimacy, and average active engagement vs. average LSM (top) and average nCLiD (bottom) scores per participant with a regression line of best fit.

	nCLiD			**LSM**		
	DI	**EI**	**Eng**	**DI**	**EI**	**Eng**
Intercept	2.765*	2.881*	6.760 ***	-0.371	-0.583 **	-0.696 **
	(1.129)	(1.145)	(1.236)	(0.214)	(0.201)	(0.191)
Entrainment	-7.389*	-7.724 *	-17.592 ***	0.859 **	1.147 ***	2.032 ***
	(3.304)	(3.394)	(3.367)	(0.283)	(0.266)	(0.254)
Test Statistic	χ^2(1, 24) = 5.31	F (1, 24) = 5.18	F(1, 24) = 23.05	F (1, 24) = 9.183 **	F (1, 24) = 18.54 ***	F (1, 24) = 64.24 ***
Adj. R^2	0.15	0.143	0.47	0.247	0.412	0.717
		a			**b**	

Table 1: This table demonstrates the linear regression results for **(a)** nCLiD and **(b)** LSM. We performed a simple linear regression for each measure: Descriptive Intimacy (DI), Evaluative Intimacy (EI), and Engagement (Eng). As seen in the second row, all main effect results are significant with at least $p < 0.05$. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

described in Section 4. The aim of creating the Reddit Dataset was to capture linguistic characteristics of individuals untrained in therapy (non-expert peer supporters) engaging in conversations that parallel guided CBT exercises. This dataset enabled us to establish a baseline to compare against the performance of the LLM. Appendix G provides complete details on data collection and cleaning.

Official CBT Dataset The Alexander Street Press Counseling and Psychotherapy Transcripts (ASPCPT) Dataset is a therapy and counseling dataset (Alexander Street Press, 2023)[*]. The ASPCPT Dataset was created by transcribing sessions featuring expert therapists working with individual clients or families. We used a subset of the ASPCPT Dataset that falls under the CBT therapy category, which we refer to as the Official CBT Dataset. This subset excludes transcripts of interviews and family therapies because they are not dyadic conversations between a therapist and one client. Clients are anonymized using unique participant IDs. 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 E. The assumption of a normal distribution was assessed via the Shapiro-Wilk test and homogeneity of variance was evaluated with Levene's test. Unequal variances were addressed by employing a Welch's ANOVA, which accounts for the differences in variations between the LLM-Guided CBT Exercises Dataset, Official CBT Dataset, and Reddit Dataset. Non-normal distributions were

[*] Accessed through authors' institution subscription: `https://search.alexanderstreet.com/psyc` (Volume 1) and `https://search.alexanderstreet.com/ctrn` (Volume 2).

addressed by employing a non-parametric Kruskal-Wallis test.

5.2 Results

A Welch's ANOVA was conducted to compare the nCLiD scores between the LLM-Guided CBT Exercises ($M = 0.34, SD = 0.01$), 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 4). 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-Guided CBT Exercises Dataset than in the Official CBT and Reddit Datasets. Additionally, the nCLiD scores for the Reddit Dataset were significantly higher than the Official CBT Dataset.

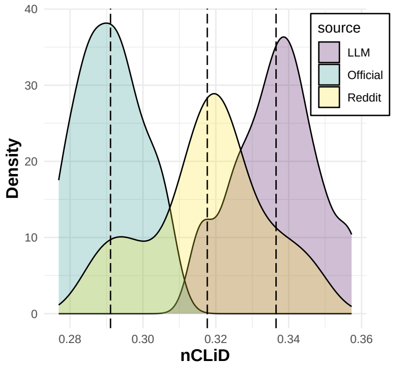

Figure 4: Distributions of nCLiD scores for LLM-Guided CBT Exercises, Official CBT, and Reddit Datasets.

A Kruskal–Wallis test was conducted to compare the LSM scores between the LLM-Guided CBT Exercises ($M = 0.76, SD = 0.09$), Official CBT ($M = 0.92, SD = 0.02$), and Reddit ($M = 0.87, SD = 0.06$) Datasets. The ANOVA was significant at the $p < 0.001$ level, $df = 2$, $\chi^2 = 131.02$ (Figure 5). A post-hoc Dunn Howell test indicated that the LSM scores were significantly different among the LLM-Guided CBT and Official CBT Datasets as well as the LLM-Guided CBT and Reddit Datasets, at the $p < 0.001$ level. In particular, LSM scores were significantly lower in the LLM-Guided CBT Dataset than in the Official CBT and Reddit Datasets. There was no significant difference between the distribution of LSM scores for the Reddit and Official CBT Datasets.

The results of the ANOVAs indicate that the distributions of entrainment scores among the LLM-Guided CBT Exercises, Official CBT, and Reddit Datasets were significantly different. The CBT practitioners in the Official CBT Dataset had significantly higher linguistic entrainment than the LLM (*supporting H2a*) as well as those from non-expert online peer supporters (*supporting H2b*, although this was only captured by nCLiD). Interestingly, non-expert online peer supporters had significantly higher linguistic entrainment than the LLM (*H2c not supported*). For details about the distribution of entrainment scores for all three datasets, see Appendix H.

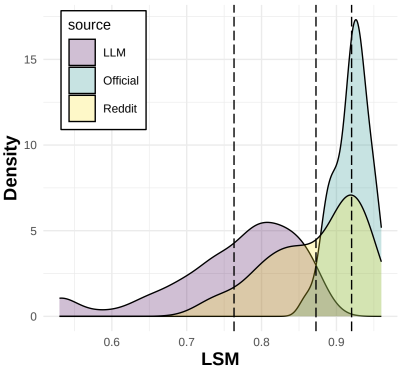

Figure 5: Distributions of LSM scores for LLM-Guided CBT Exercises, Official CBT, and Reddit Datasets.

6 Discussion

We hypothesized that in a therapeutic setting, a therapist's linguistic entrainment with the client encourages greater self-disclosures and, subsequently, higher levels of intimacy and engagement. This relationship was supported by Beňuš (2014) in their review, where they found a connection between entrainment and social distance and suggest that entrainment can help a medical professional develop closeness and trust with clients, which is critical for encouraging greater self-disclosure (Newman, 2002) leading to improved therapeutic outcomes (Farber, 2003; Farber et al., 2006). Colton (2022) also found that linguistic entrainment "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 entrainment, and intimacy and engagement, two measures of patient self-disclosures, which is also supported by our results from the first experiment.

The significant relationship we found between linguistic entrainment and descriptive and evaluative intimacy and engagement indicates that LSM and nCLiD show promise as measures of therapeutic outcomes. This creates an approach to quantitative analysis of dyadic therapeutic interactions without costly annotations for therapeutic measures.

The results from our second experiment align with our hypotheses that the CBT therapists would have the highest linguistic entrainment. These results make sense since CBT therapists undergo years of training to offer patients a high-quality therapeutic experience, which will naturally outperform LLMs. Interestingly, there was not a significant difference between the distribution of LSM scores for the two human datasets, but nCLiD was able to detect the difference between these two distributions. We attribute this to nCLiD being a more sophisticated measure of linguistic entrainment that can capture the difference in performance between experts and non-experts.

Notably, non-expert online peer supporters had significantly higher linguistic entrainment than the LLM. Initially, we hypothesized that the LLM would have higher linguistic entrainment 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, even non-expert humans had higher linguistic entrainment than a prompted LLM. 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 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 analysis of the LLM-Guided CBT Exercises Dataset, we also observed patterns of the LLM repeatedly using the same response frame. These tendencies of the LLM to be less varied in its responses may, therefore, lead to lower linguistic entrainment. 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 more familiar with therapy and are better able to mimic therapeutic dialogue.

Our results reinforce the need to be cautious in applying LLMs out-of-the-box in therapeutic contexts. While they are able to manage various therapeutic tasks (Cho et al., 2023; Kian et al., 2024), LLM dialog is inferior to that of therapists. Thus, researchers must carefully assess each application domain and determine if the LLM can meet the expected threshold of performance. Furthermore, suggestions to use LLMs as a replacement instead of augmentative therapeutic technologies should be cautioned, as our results demonstrate that even untrained people outperform LLMs in their current stage of development.

7 Conclusion and Future Work

This work investigated the linguistic entrainment of an LLM in an interactive therapy session. We demonstrated that there is a statistically significant relationship between the linguistic entrainment of the LLM and the percentage of high intimacy and active engagement responses from the users. We next compared the LLM's linguistic entrainment with that of trained CBT therapists and non-expert online peer supporters and found, with respect to linguistic entrainment, 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 entrainment, such as those based on part-of-speech distributions (Shapira et al., 2022), 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 aim to validate these metrics in the future. We would also like to extend our analysis to additional therapeutic measures.

The LLM-Guided CBT Exercises Dataset was collected from interactions between GPT-3.5-turbo and participants. It would be interesting to evaluate LLMs with an expanded token limit that would allow for longer interactions. In the LLM-Guided CBT Exercises Dataset, the responses were generated by a prompted model; evaluating an LLM that has been fine-tuned on therapy data may further improve the generated responses.

Limitations

The data in the three datasets we analyzed, Official, LLM-Guided CBT Exercises, and Reddit Datasets (Section 5), come from inherently different channels of communication: the Official Dataset contains transcriptions of real-time human-to-human spoken conversations between a therapist and client, while the Reddit Dataset contains asynchronous, online-typed conversations, and the LLM-Guided CBT Exercises Dataset includes real-time typed conversations between a human and either a robot or chatbot. The differences in modality can lead to differences in the nature of conversations and introduce confounding variables in linguistic entrainment. Additionally, although we worked to find consistent data from the CBT-related mental health domain for all three datasets, the premise in each dataset we used is different. The Official Dataset comprises of full CBT sessions, while the LLM-Guided CBT Exercises Dataset comprises of CBT exercises for a shorter duration. On Reddit, people responded to posts asynchronously without adhering to specific therapy guidelines. Since the Reddit Dataset tends to follow a short-form interaction instead of the length expected in a full therapy session, its premise is similar to that of the LLM-Guided CBT Exercises Dataset.

We also note that our datasets were quite small, with approximately 30 interactions per dataset. Larger sample sizes, when available, could yield more insightful results.

Another limitation in our work is that the nCLiD algorithm uses word2vec word embeddings, which are static and limit the use of multiple meanings of words depending on the context, unlike newer transformer-based contextual word embeddings such as BERT-based embeddings (Vaswani et al., 2017; Devlin et al., 2018). We chose nCLiD for this work since Nasir et al. (2019) validated this metric in a therapy setting by demonstrating its association with empathy. Additionally, in order for nCLiD to be implemented with contextual word embeddings, the nCLiD algorithm needs to be changed fundamentally since it depends on word frequency counts. This leads to a different metric based on nCLiD and would require additional validation.

We note that nCLiD averages the Word Mover's Distance values over all the turns in the conversation, therefore potentially not capturing temporal shifts in linguistic entrainment. nCLiD captures lexical semantic similarity, a specific aspect of linguistic entrainment. There are other measures of linguistic entrainment that, for example, consider part of speech distributions (Shapira et al., 2022). A fuller picture of linguistic entrainment could be obtained by evaluating our datasets with those metrics as well.

Finally, we used data from interactions with one version of an LLM (GPT-3.5-Turbo). Performance across different LLMs can vary.

Ethical Considerations

The use of LLMs and conversational agents in mental health contexts can be risky. LLMs can hallucinate, make false promises, and encourage inappropriate ideas. While there are many benefits of LLM-based systems, such as enabling frequent, interactive conversations that the mental healthcare system cannot always provide, we caution against their use because of the potential negative impacts. We advocate that LLMs can augment therapists by providing an accessible, interactive version of the at-home exercises, as was done in the LLM-guided CBT exercises study (Kian et al., 2024). We do not support the use of LLMs as a replacement technology for human therapists. Additionally, to ensure safety, measuring the quality of LLMs in the mental health domain is critical. We hope that this work contributes to the growing effort of evaluating LLMs used in mental health domains.

Acknowledgments

We would like to thank Amy O'Connell, Abhyuday Singh, Anna-Maria Velentza, and Mingyu Zong for their support on this project.

This material is based upon work supported in part by the NSF CISE Graduate Fellowship CSGrad4US under Grant No. 2313998 (Award ID G-2A-061). This work is also supported in part by the Center for Undergraduate Research in Viterbi Engineering (CURVE) Fellowship and the University of Southern California Undergraduate Research Associates Program (URAP).

References

Noor Al-Sibai. 2023. Openai employee says she's never tried therapy but chatgpt is pretty much a replacement for it.

Alexander Street Press. 2023. Counseling and psychotherapy transcripts: Volume i. Dataset.

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, Xuezhi 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. *Preprint*, arXiv:2305.10403.

American Psychological Association and Society of Clinical Society. 2017. What is cognitive behavioral therapy?

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. *Preprint*, arXiv:2212.08073.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Štefan Beňuš. 2014. Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6:802–813.

Justin W. Bonny and Anya M. Jones. 2023. Teams moving more synchronously are perceived as socially dominant. *Acta Psychologica*, 237:103952.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

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. *Preprint*, arXiv:2005.14165.

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 the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

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. *Preprint*, arXiv:2107.03374.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

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 autistic adolescent psychological counjaseling. *arXiv preprint arXiv:2311.09243*.

Munmun De Choudhury, Sachin R. Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *Preprint*, arXiv:2311.14693.

David A. Clark. 2013. *Cognitive Restructuring*, chapter 2. John Wiley Sons, Ltd.

Gavin I Clark and Sarah J Egan. 2015. The socratic method in cognitive behavioural therapy: a narrative review. *Cognitive Therapy and Research*, 39:863–879.

Tayler M. S. Colton. 2022. *Linguistic synchrony: indicator or facilitator of therapeutic bond*. Ph.D. thesis, University of British Columbia.

MA. Courtney E. Ackerman. 2017. Cbt techniques: 25 cognitive behavioral therapy worksheets.

Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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 A Farber. 2003. Patient self-disclosure: A review of the research. *Journal of clinical psychology*, 59(5):589–600.

Barry A Farber, Kathryn C Berano, and Joseph A Capobianco. 2006. A temporal model of patient disclosure in psychotherapy. *Psychotherapy Research*, 16(4):463–469.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, 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. *Preprint*, arXiv:2209.07858.

Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19.

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. *Digital Health*, 9:20552076231170499.

Adam O Horvath and Leslie S Greenberg. 1986. The development of the working alliance inventory. *The Psychotherapeutic Process: A Research Handbook.*, pages 529–556.

Adam O Horvath and B Dianne Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139.

Michael J Hove and Jane L Risen. 2009. It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition*, 27(6):949–960.

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. *Journal of counseling psychology*, 61(1):146.

Molly E Ireland and James W Pennebaker. 2010. Language style matching in writing: synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99(3):549.

Julie Jargon. 2023. Wsj news exclusive | how a chatbot went rogue. *WSJ*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Kathleen Kadner. 1994. Therapeutic intimacy in nursing. *Journal of Advanced Nursing*, 19(2):215–218.

Nikolaos Kazantzis, Luciano L'Abate, et al. 2007. *Handbook of homework assignments in psychotherapy*. Springer.

Nikolaos Kazantzis, Craig Whittington, and Frank Dattilio. 2010. Meta-analysis of homework effects in cognitive and behavioral therapy: A replication and extension. *Clinical Psychology: Science and Practice*, 17(2):144.

Jay Kejriwal and Stefan Benus. 2024. Lexical, syntactic, semantic and acoustic entrainment in slovak, spanish, english, and hungarian: A cross-linguistic comparison. *SSRN*.

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. *Preprint*, arXiv:2402.17937.

Sayaka Kidby, Dave Neale, Sam Wass, and Victoria Leong. 2023. Parent–infant affect synchrony during social and solo play. *Philosophical Transactions of the Royal Society B*, 378(1875):20210482.

Lauren Rebecca Klein. 2023. *Modeling Dyadic Synchrony with Heterogeneous Data: Validation in Infant-Mother and Infant-Robot Interactions*. Ph.D. thesis.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, 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. *New Media & Society*, 0(0):14614448221142007.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Ron Li, Steven Asch, and Nigam Shah. 2020. Developing a delivery science for artificial intelligence in healthcare. *npj Digital Medicine*, 3:107.

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. *Preprint*, arXiv:2312.06550.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 735–745, New York, NY, USA. Association for Computing Machinery.

Gary Marcus. 2022. The dark risk of large language models. *Wired Magazine*. Https://archive.ph/Dpdg7.

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. *Journal of consulting and clinical psychology*, 68(3):438.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Preprint*, arXiv:1310.4546.

Teru L Morton. 1978. Intimacy and reciprocity of exchange: A comparison of spouses and strangers. *Journal of Personality and Social Psychology*, 36(1):72.

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. *Interspeech*, 2019:1423–1427.

Cory F Newman. 2002. A cognitive perspective on resistance in psychotherapy. *Journal of clinical psychology*, 58(2):165–174.

Tuan Dinh Nguyen, Marisa Cannata, and Jason Miller. 2018. Understanding student behavioral engagement: Importance of student interaction with peers and teachers. *The journal of educational research*, 111(2):163–174.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 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. *medRxiv*.

Aileen O'Brien, Rana Fahmy, and Swaran P Singh. 2009. Disengagement from mental health services: a literature review. *Social psychiatry and psychiatric epidemiology*, 44:558–568.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Jan Prasko, Ilona Krone, Julius Burkauskas, Jakub Vanek, Marija Abeltina, Alicja Juskiene, Tomas Sollar, Ieva Bite, Milos Slepecky, and Marie Ociskova. 2022. Homework in cognitive behavioral supervision: theoretical background and clinical application. *Psychology Research and Behavior Management*, pages 3809–3824.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Fabian Ramseyer and Wolfgang Tschacher. 2011. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 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.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Miriam Rennung and Anja S Göritz. 2016. Prosocial consequences of interpersonal synchrony. *Zeitschrift für Psychologie*.

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Natalie Shapira, Dana Atzil-Slonim, Rivka Tuval-Mashiach, and Ori Shapira. 2022. Measuring linguistic synchrony in psychotherapy. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 158–176.

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 *EMNLP*.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12.

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. *Advances in Neural Information Processing Systems*, 33:3008–3021.

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. *Preprint*, arXiv:2211.09085.

Amanda Tetley, Mary Jinks, Nick Huband, and Kevin Howells. 2011. A systematic review of measures of therapeutic engagement in psychosocial and psychological treatment. *Journal of Clinical Psychology*, 67(9):927–941.

TherapistAid. 2017. Cognitive restructuring: Socratic questions: Worksheet.

Betsy E Tolstedt and Joseph P Stokes. 1984. Self-disclosure, intimacy, and the depenetration process. *Journal of Personality and Social Psychology*, 46(1):84.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alexandria Vail, Jeffrey Girard, Lauren Bylsma, Jeffrey Cohn, Jay Fournier, Holly Swartz, and Louis-Philippe Morency. 2022. Toward causal understanding of therapist-client relationships: A study of language modality and social entrainment. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 487–494.

Alastair C. van Heerden, Julia R. Pozuelo, and Brandon A. Kohrt. 2023. Global Mental Health Services and the Impact of Artificial Intelligence–Powered Large Language Models. *JAMA Psychiatry*, 80(7):662–664.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Matthew Weaver. 2023. Ai chatbot 'encouraged' man who planned to kill queen, court told. *The Guardian*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Amy Wenzel. 2017. Basic strategies of cognitive behavioral therapy. *Psychiatric Clinics*, 40(4):597–609.

Camille J. Wynn and Stephanie A. Borrie. 2022. Classifying conversational entrainment of speech behavior: An expanded framework and review. *Journal of Phonetics*, 94:101173.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Youper. Mental health gpts.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, Raja-Elie E. Abdulnour, Atul J. Butte, and Emily Alsentzer. 2023. Coding inequity: Assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. *medRxiv*.

James M Zech, Robert Steele, Victoria K Foley, and Thomas D Hull. 2022. Automatic rating of therapist facilitative interpersonal skills in text: A natural language processing application. *Frontiers in Digital Health*, 4:917918.

Muru Zhang, Ofir Press, Will Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. *ArXiv*, abs/2305.13534.

Appendices

A GPT-3.5 Prompts and Parameters in LLM-Guided CBT Exercises

Prompt for coping strategies exercise:

```
Coping strategy is used to help patients identify problems they encountered and the triggers. When a problem is defined, a therapist will help the patient figure out ways to cope with it. You are a therapist who uses coping strategies to help your patient in this session.
```

Prompt for cognitive restructuring exercise:

```
Cognitive restructuring is a strategy to help the patient identify cognitive distortion and find evidence to challenge the distortion. You are a therapist who uses cognitive restructuring to help your patient in this 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 2: Input parameters for OpenAI's chat completion API.

B Therapeutic Measures Annotations

Each participant utterance in the LLM-Guided CBT Exercises Dataset was evaluated on the following dimensions: descriptive intimacy, evaluative intimacy (Tolstedt and Stokes, 1984; Morton, 1978), and engagement (Nguyen et al., 2018). The levels of the two intimacy measures were evaluated to be *High* or *Low*, while engagement was evaluated to be *Active* or *Passive*. The definitions of these variables and examples are provided in Table 3 (note: these are hypothetical examples for illustrative purposes).

C nCLiD normalization factor α

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

$$nCLiD = \frac{uCLiD}{\alpha} \tag{3}$$

$$\begin{aligned}\alpha = &\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N} WMD(a_i, a_j)\\ &+\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N} WMD(c_i, c_j)\\ &+\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i}^{N} WMD(a_i, c_j)\end{aligned} \tag{4}$$

D Word Mover's Distance Experiment

We first give a simple example demonstrating the performance of WMD. We selected two dialogue rounds from two independent transcripts in which the LLM responds to participants' concerns:

We then perform an experiment where we swap the LLM responses between the two rounds from dialogues **A** and **B**.

Variable	Label	Definition	Example
Engagement	Active	Giving a thorough and thoughtful response	"How can I restructure my thoughts to change the way I perceive guilt?"
	Passive	Giving short responses and not continuing the dialogue	"I guess. . . ."
Evaluative Intimacy	High	Disclosing opinions, judgements, and emotions	“I’m feeling stressed and anxious about my classes this semester. I don’t feel smart enough to do well in them."
	Low	Not disclosing opinions, judgements, and emotions in a lot of detail	“I am taking academically challenging courses this semester.”
Descriptive Intimacy	High	Disclosing facts about oneself	“I have a final exam tomorrow for Physics that could determine whether I pass or fail my class.”
	Low	Not disclosing facts about oneself in a lot of detail	“I have a big event."

Table 3: Definitions and examples for Descriptive Intimacy, Evaluative Intimacy, and Engagement.

Dialogue	Speaker	Content
A	**Participant**	"I’m insecure when recruiters don’t get back to me and feel like they rejected me after I sent them my resume."
	LLM	"Understand it can be frustrating when you don’t hear back from recruiters after submitting your resume. When you feel rejected, what thoughts come to mind?"
B	**Participant**	"Yeah, I’ve recently been really anxious about the idea that I won’t have enough time to study for final exams and that I might do poorly on the exams because of this."
	LLM	"Okay, that sounds like a form of catastrophizing. When you think about this scenario where you don’t have enough time to study and perform poorly on the exams, what evidence do you have that supports or contradicts these thoughts?"

Table 4: Selected rounds from two dialogues.

Dialogue	**Response**	**WMD Score (↓)**
A	**Original**	0.71187
	Swapped LLM response	1.01617 (↑)
B	**Original**	0.69998
	Swapped LLM response	1.08221 (↑)

Table 5: WMD score results for original rounds and swapped rounds.

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 6: R package versions used in our analysis.

As shown in the experiment above in Table 5, swapping the LLM responses between the two dialogues increases the WMD scores, aligning with our hypothesis that the WMD score can reflect the "entrainment" between the speakers.

We also performed an ablation test for the selected dialogue rounds in Table 4. Specifically, for each LLM response, we removed one word at a time and recorded the corresponding change to the WMD scores. Darker colors demonstrate words that contribute the most to a lower WMD, and thus a higher entrainment, with the participant’s turn. The normalized changes for each word removal are shown in the heatmaps below (Figures 6-9). As demonstrated in Figure 6, the key words intuitively make sense as those relevant to the WMD score. The examples shown in the figures have exact word matches in the participant’s and LLM’s turns; however, since WMD uses *word2vec* embeddings, a less exact match in words, such as synonyms, would also contribute to a low WMD (high entrainment). Since the response in Figure 7 is from a different transcript, none of the key words from the original response are present, resulting in an overall worse WMD score. These patterns are replicated in Figures 8 and 9. Since a key part of CBT is reframing, therapists often repeat key aspects of their clients’ disclosures, as also seen in the examples above. They also try to match the language of their clients, both of which result in higher entrainment.

E R Packages and Versions

See Table 6 for the list of R packages used.

F LSM Function Word Assessment

LSM calculates the similarity between two texts with respect to the similarity in frequency of function words used by the two interlocutors. Specifically, LSM uses the following eight function word types: *prepositions* (prep) (e.g. to, of, in, for), *articles* (e.g. a, an, the, alot), *auxiliary verbs* (auxverb)

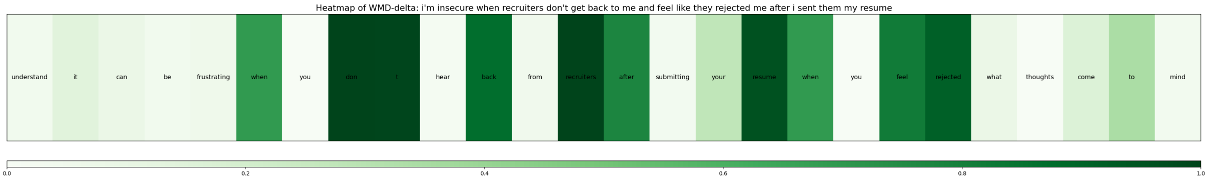

Figure 6: Heat map for an ablation test between a turn from Participant$_{\mathrm{A}}$ and the response from LLM$_{\mathrm{A}}$. As seen in this example: "I'm insecure when recruiters don't get back to me and feel like they rejected me after I send them my resume." The italicized words are most important in the LLM's response "Understand it can be frustrating when you *don't* hear back from *recruiters* after submitting your *resume*. When you feel *rejected*, what thoughts come to mind?"

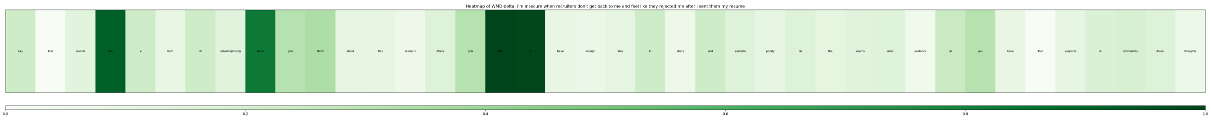

Figure 7: Heatmap for an ablation test between a turn from Participant$_{\mathrm{A}}$ and the response from LLM$_{\mathrm{B}}$. "I'm insecure when recruiters don't get back to me and feel like they rejected me after I send them my resume." The italicized words are most important in the LLM's response "Okay, that sounds *like* a form of catastrophizing. *When* you think about this scenario where you *don't* have enough time to study and perform poorly on the exams, what evidence do you have that supports or contradicts these thoughts?"

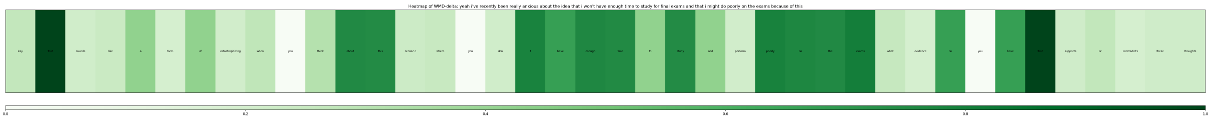

Figure 8: Heatmap for an ablation test between a turn Participant$_{\mathrm{B}}$ and the response from LLM$_{\mathrm{B}}$. "Yeah, I've recently been really anxious about the idea that I won't have enough time to study for final exams and that I might do poorly on the exams because of this." The italicized words are most important in the LLM's response "Okay, *that* sounds like a form of catastrophizing. When you think *about this* scenario where you don't have *enough time* to study and perform *poorly* on the *exams*, what evidence do you have *that* supports or contradicts these thoughts?"

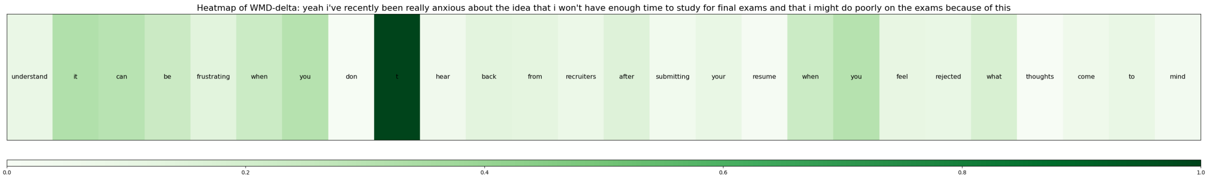

Figure 9: Heatmap for an ablation test between Participant$_{\mathrm{B}}$ and LLM$_{\mathrm{A}}$."Yeah, I've recently been really anxious about the idea that I won't have enough time to study for final exams and that I might do poorly on the exams because of this." The italicized words are most important in the LLM's response "Understand it can be frustrating when you *don't* hear back from recruiters after submitting your resume. When you feel rejected, what thoughts come to mind?"

(e.g. is, was, be, have), *adverbs* (e.g. so, just, about, there), *conjunctions* (conj) (e.g. and, but, so, as), *personal pronouns* (ppron) (e.g. I, you, my, me), *impersonal pronouns* (ipron) (e.g. that, it, this, what), and *negations* (e.g. not, no, never, nothing). The LSM score (Boyd et al., 2022) for each function word category is calculated as follows:

$$LSM_{\text{prep}} = 1 - \frac{|prep_1 - prep_2|}{prep_1 + prep_2 + 0.0001} \tag{5}$$

The final LSM score is an aggregate of the LSM scores for each function word category:

$$\begin{aligned} LSM = average(&LSM_{\text{prep}} + LSM_{\text{article}} + \\ &LSM_{\text{auxverb}} + LSM_{\text{adverb}} + \\ &LSM_{\text{conj}} + LSM_{\text{ppron}} + \\ &LSM_{\text{ipron}} + LSM_{\text{negate}}) \end{aligned} \tag{6}$$

The correlation results between the disaggregated function word LSM scores and the percentage of high intimacy or active engagement are demonstrated in Table 7.

G ConvoKit Dyadic Reddit Thread Extraction

G.1 Data Collection and Preprocessing

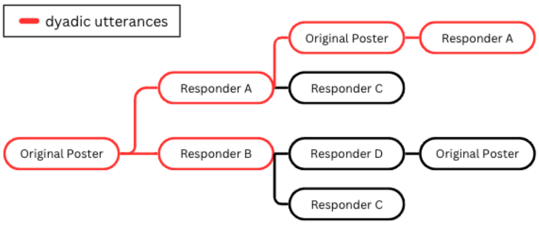

Figure 10: Extraction process of dyadic Reddit comment threads between the original poster and a specific responder.

To narrow our search, we looked for subreddits (online topic-based communities on Reddit) 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 Exercises 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)* 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 10). 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 usernames associated with each Reddit post/comment.

G.2 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 et al. (2024) in their study. For the full list of CBT-related terms used as keywords when filtering relevant Reddit threads, see Table 8. 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

*distributed under the MIT license

	Spearman's Rank Correlation Coefficient r(26)		
	% High Evaluative Intimacy	**% High Descriptive Intimacy**	**% Active Engagement**
$\mathbf{LSM_{prep}}$	0.4537^{*}	X	0.5561^{**}
$\mathbf{LSM_{article}}$	X	X	X
$\mathbf{LSM_{auxverb}}$	0.5385^{**}	0.5501^{**}	0.5479^{**}
$\mathbf{LSM_{adverb}}$	0.5275^{**}	0.4373^{*}	0.5698^{**}
$\mathbf{LSM_{conj}}$	0.6903^{***}	0.6109^{***}	0.659^{***}
$\mathbf{LSM_{ppron}}$	X	X	X
$\mathbf{LSM_{ipron}}$	0.5508^{**}	0.4373^{*}	0.6453^{***}
$\mathbf{LSM_{negate}}$	X	X	0.6672^{***}

Table 7: The rows indicate the LSM function words, while the columns indicate the therapeutic outcome variables. The values demonstrate the Spearman's rank correlation coefficients. The significance thresholds are represented by $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, an X represents a non-significant result.

CBT-Related Keywords	
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	

Table 8: CBT keywords used to filter Reddit threads.

on cognitive restructuring and coping strategy exercises from the LLM-Guided Exercises Dataset, screen conversations to make sure that they paralleled 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 that were 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 in different linguistic characteristics than those captured in real-time conversations. However, we observed that in most OMHC subreddits, posters often seek advice and engage with commenters in a timely manner.

H Synchrony Score Distributions

The distribution of the nCLiD and LSM scores for all three datasets are provided in Table 9. We can consider the entrainment scores of the trained therapist (Official Dataset) a proxy for the standard values for entrainment that a therapist-in-training or at-home CBT dialogue system should aim for. Looking at the distribution of nCLiD and LSM scores for the dialogues with the trained therapists, we see that the average nCLiD score was 0.29, which is the highest entrainment score compared to the Reddit and LLM dialogues. Analogously, we can see that the trained therapist has the highest entrainment score through LSM with a score of 0.92. As a proxy for optimal entrainment in a therapeutic context, we can use these values to compare entrainment for other CBT therapy dialogues.

Dataset	**Min**	**Max**	**Mean**	**Std**
		nCLiD		
LLM	0.2944	0.3761	0.3364	0.0125
Official	0.2631	0.3237	0.2912	0.0111
Reddit	0.2722	0.3526	0.3136	0.0131
		LSM		
LLM	0.2300	0.9400	0.7575	0.1134
Official	0.8600	0.9600	0.9203	0.0236
Reddit	0.0000	0.9700	0.8498	0.1647

Table 9: Statistical summary of the nCLiD (top) and LSM (bottom) score distributions for each of the three datasets.