Linking Language-based Distortion Detection to Mental Health Outcomes

Vasudha Varadarajan[1], **Allison Lahnala**[2], **Sujeeth Vankudari**[1] **Akshay Raghavan**[1], **Scott Feltman**[1], **Syeda Mahwish**[1], **Camilo Ruggero**[1], **Roman Kotov**[1], **H. Andrew Schwartz**[1]
[1]Stony Brook University, [2]McMaster University
{vvaradarajan,has}@cs.stonybrook.edu

Abstract

Recent work has suggested detection of cognitive distortions as an impactful task for NLP in the clinical space, but the connection between language-detected distortions and validated mental health outcomes has been elusive. In this work, we evaluate the co-occurrence of (a) 10 distortions derived from language-based detectors trained over two common distortion datasets with (b) 12 mental health outcomes contained within two new language-to-mental-health datasets: DS4UD and iHiTOP. We find higher rates of distortions for those with greater mental health condition severity (ranging from $r = 0.16$ for thought disorders to $r = 0.46$ for depressed mood), and that the specific distortions of *should statements* and *fortune telling* were associated with a depressed mood and being emotionally drained, respectively. This suggested that language-based assessments of cognitive distortion could play a significant role in detection and monitoring of mental health conditions.

1 Introduction

Cognitive distortions–systematic thinking patterns that cause inaccurate perceptions of reality–contribute to maintaining or worsening mental health conditions, such as depression and anxiety (Beck, 1963). The practice of recognizing one's own cognitive distortions is a core component of cognitive behavioral therapy (CBT), one of the most effective non-medicinal therapies for depression (Hofmann et al., 2012). Recent advances in natural language processing (NLP) have opened new avenues for automatically detecting distortions as well as generating text to reframe the distortions (de Toledo Rodriguez et al., 2021; Lim et al., 2024), potentially extending accessibility to therapeutic practices like CBT. Reliable detection of cognitive distortions in text holds promise for scalable mental health assessments to increase their efficacy and adds a layer of explainability. However, a key step in this vision is to validate that language-detected distortions do in fact have associations with validated mental health outcomes.

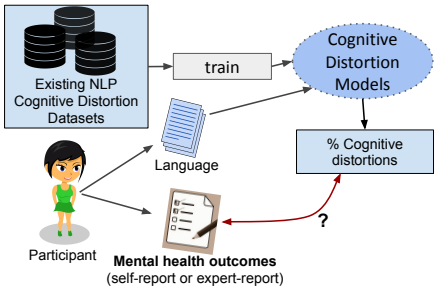

Figure 1: We train distortion detection models on existing cognitive distortion datasets, apply them to identify cognitive distortions in language, and evaluate their relationship with HiTOP (Kotov et al., 2022) and DS4UD (Nilsson et al., 2024) mental health outcomes over two new datasets.

This study seeks to empirically evaluate whether language-detected distortions do in fact show connections to expected mental health outcomes over both clinical interviews as well as standard self-report assessments. Our paper highlights two key findings: (1) our analyses validate the co-occurrence of cognitive distortions with mental health conditions, demonstrating that higher rates of distorted thinking patterns generally correspond to greater severity of mental health symptoms; (2) we identify specific distortion types that exhibit stronger correlations with certain mental health indicators, suggesting they may be useful language markers of particular health indicators. We also identify where better detection performance and connections to mental health outcomes could be stronger, motivating directions for future work.

The established links between distortions and mental health conditions have motivated language analysis on social networks for early detection of depression markers of depression in social media posts (Ophir et al., 2017; Bathina et al., 2021;

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 62–68

A. Rutter et al., 2025). Our study underscores that NLP models of cognitive distortions effectively align language with actual mental health conditions, and contributes to real-world monitoring or intervention strategies through advanced detection capabilities.

2 Background

Cognitive distortions are systematic patterns of biased thinking and false self-beliefs that can lead to negative moods and behaviors, playing a role mental health conditions like depression (Beck, 1963). Therapies like cognitive behavioral therapy (CBT) involve the practice of identifying and reframing distortions to support individuals in adopting healthier thinking patterns (Rupke et al., 2006). There is strong evidence that this form of therapy is effective for managing conditions like anxiety and depression (Hofmann et al., 2012). Since the COVID-19 pandemic, therapy has increasingly transitioned into the telehealth space (Leroy et al., 2025), highlighting a need for automated detection tools in conversations which would allow therapists in recognizing distorted thinking within the vast amount of information they process during a therapy session. This shift has encouraged the exploration of various ways in which agents within telehealth sessions can assist therapists in recognizing patterns, creating opportunities for the application of distortion detection tools. Tools such as this can enable timely interventions, helping patients recognize and work through cognitive distortions. Additionally, distortion detection tools can be integrated into developing assistive agents for therapy homework after CBT, bringing a more patient-facing support by flagging cognitive distortions and prompting the need for reappraisal (Stade et al., 2024).

As distortion reframing occurs through language, recent research has explored NLP-based approaches for cognitive distortion detection, reframing, and positive reformulation. Various efforts have been dedicated toward cognitive distortion detection and classification models (Shreevastava and Foltz, 2021; Chen et al., 2023; Lim et al., 2024). Datasets of situations, thoughts, and reframes have been created to train generative models (Sharma et al., 2023; Maddela et al., 2023). This research has focused on models that perform positive reformulation of distorted thoughts to more constructive ones (de Toledo Rodriguez et al., 2021), by adopting, for instance, strategies from positive psychology (Ziems et al., 2022). Others have also aimed to build chat systems that guide users through cognitive restructuring (Sharma et al., 2024). These studies highlight the potential of NLP-driven interventions in fostering cognitive shifts and improving mental health, yet there is limited work into how automatically detected distortions correspond to existing mental health conditions.

3 Data

3.1 Mental Health Outcomes Datasets

iHiTOP The iHiTOP dataset contains transcribed clinical interviews with psychiatric outpatients, aligned with the HiTOP taxonomy—a modern mental health taxonomy mappable to DSM-V (Kotov et al., 2022; Regier et al., 2009). These semi-structured interviews, lasting 45-90 minutes, were diarized and transcribed using NVIDIA NeMo and openai/whisper-large-v2.[1] We use the transcribed text of the interviewee in our analysis.

Dataset	Num participants	Num messages	Num spans	Mental Health Outcomes	Report
iHiTOP	536	536	568989	Internalizing Mania Ankankastia Thought disorder Detachment Disinhibition Antagonism	Expert
DS4UD	587	32773	58103	Depressed Mood Daily Stress Daily Drain Wave Anxiety Wave Depression	Self

Table 1: Descriptions of Mental Health datasets.

The dataset includes patient scores across seven "spectra": *internalizing*, *mania*, *anankastia*, *thought disorder*, *detachment*, *disinhibition*, and *antagonism*. After filtering segments shorter than 4 words, the average segment length was 12 words, with interviews averaging 8,217 words per patient.

DS4UD The Data Science for Unhealthy Drinking Study (DS4UD) dataset (Nilsson et al., 2024) comprises mental health assessments and language data collected from U.S. service industry workers over two years. We focus on daily diary language from Ecological Momentary Assessments (EMAs). Participants provided three daily EMA responses across six 14-day waves, responding to: *"Please describe in 2 to 3 sentences how you are currently*

[1] 'iHiTOP' is also the name of the instrument used to assess HiTOP mental health scores. This is the first dataset to use it so it is named the same.

feeling." With responses averaging 50 words (average 11 words per sentence), each participant could contribute up to 252 responses. The dataset includes daily metrics (affect, stress, alcohol consumption, and cravings) and WAVE measurements of anxiety and depression.

3.2 Cognitive Distortions Training Data

Patient Queries dataset (PQ) Shreevastava and Foltz (2021) contains patient queries to therapists, which include questions, concerns, descriptions of circumstances, and symptoms, among other topics. Each example is labeled with 1-2 dominant cognitive distortions, from 10 common types – All-or-Nothing Thinking, Overgeneralizing, Labeling, Fortune Telling, Mind Reading, Emotional Reasoning, Should Statements, Personalization, Mental Filter, and Magnification. There are 1597 instances of distorted spans (average length: 36 words) annotated with one of the ten types, from a total of 2530 messages (average length: 166 words).

Thinking Traps dataset (TT) Sharma et al. (2023) covers a set of 13 cognitive distortions: *All-or-Nothing Thinking, Overgeneralizing, Labeling, Fortune Telling, Mind Reading, Emotional Reasoning, Should Statements, Personalization, Disqualifying the Positive (Mental Filter), Catastrophizing (Magnification), Comparing and Despairing, Blaming, Negative Feeling or Emotion*. We drop the classes *Blaming, Comparing* and *Negative Emotion* due to the lack of enough examples in the dataset, and to maintain the same set of distortions in both the datasets. Our final dataset contains 1011 spans (average length: 21 words) that describe a situation and lead to a distorted thought leading from the situation.

4 Methods

We develop models to detect cognitive distortions in text as a means to study their relationship with mental health outcomes. Following established approaches in mental health NLP (Ganesan et al., 2021), we utilize transformer-based language models (LMs) and their contextual embeddings rather than pursuing incremental architectural improvements. These models enable us to quantify distortion rates per participant and examine their associations with mental health measures, addressing our primary research question.

Task 1: Distortion Detection This binary classification task assessed the models' ability to distinguish between messages containing cognitive distortions and those without. The objective was to make a fundamental present/absent determination for distorted thinking patterns.

	Detection		Classification	
Model	**F1**	**AUC**	**F1**	**AUC**
TT	.597	.813	.276	.755
PQ (span)	.823	.917	**.369**	**.876**
PQ (full)	.693	.766	-	-
TT + PQ (span)	**.833**	**.921**	.366	.847

Table 2: Cross-validation metrics for distortion detection and 11-way classification models. Note that PQ (full) contains full passages and could contain many distortions, so it wasn't used for the classification task.

Model	**F1**	**AUC**
All-or-Nothing Thinking	.506	.768
Overgeneralizing	.581	.735
Labeling	**.607**	**.853**
Fortune Telling	**.612**	**.878**
Mind Reading	**.675**	**.871**
Emotional Reasoning	.525	.753
Should Statements	**.696**	**.874**
Personalization	.554	.797
Mental Filter	.526	.783
Catastrophizing	.522	.706

Table 3: Cross-validation metrics for one-vs-rest distortion classification models. We pick the models with F1 $>$ 0.6 (bolded) for validation on the mental health datasets.

Task 2: Distortion Classification We formulate this in two ways: a multi-class task (Table 2) and a one-vs-rest task (Table 3). The multi-class classification task required models to categorize messages according to specific distortion types identified in the training data. Notably, we included "No Distortion" as a distinct category by augmenting with sentences from the PQ dataset that were not annotated with a distortion, representing the absence of any recognized cognitive distortion patterns. This approach allowed for a more nuanced analysis of distinct cognitive distortion types in relation to mental health outcomes. Further, one-vs-rest task was used to build distortion type-specific classifiers with positive class being a distortion type and the rest of the examples from all the other classes (including No Distortion).

We report the F1 and AUC scores for comparing the performance of the models [2].

[2]F1 is a metric that is calculated as the harmonic mean

Model	iHiTOP								DS4UD					
									EMA			Wave		
	Internalizing	Mania	Anankastia	Thought Disorder	Detachment	Disihibition	Antagonism	Distort Rate (%)	Depressed Mood	Emotionally Drained	Nervous Stress	Anxiety	Depression	Distort Rate (%)
TT + PQ	0.24	0.18	0.16	0.16	0.11	0.29	0.19	9.78	**0.42**	**0.29**	**0.17**	**0.30**	**0.27**	8.47
PQ	**0.34**	**0.21**	**0.22**	**0.16**	**0.24**	**0.35**	**0.21**	11.36	**0.46**	**0.29**	**0.25**	**0.32**	**0.28**	15.03

Table 4: **Between-user correlations** (Pearson r) between overall percentage of distortions and mental health assessment scores in the iHiTOP and DS4UD dataset. Bold indicates statistically significant (p-values < 0.05). Correlations between a behavior (distortion mention) and psychological variables have a modal correlation between 0.1 to 0.4 and those above are considered very large (Roberts et al., 2007).

4.1 Modeling

We implemented two distinct approaches for fine-tuning encoder models:

Span-only (span) In this approach, we utilized only the text spans explicitly annotated as cognitive distortions. These spans were processed through a RoBERTa-base model, from which we derived averaged embeddings to train task-specific linear classifiers. This methodology was employed for both the therapist QA dataset and the thinking traps corpus (the latter consisting exclusively of short, distortion-containing sentences).

Full message (full) We expanded the input to encompass the complete message context from which the distortion spans were originally annotated in the therapist QA dataset. The processing pipeline remained consistent with the span-only approach, utilizing the same model architecture and classification framework. Both approaches leveraged the DistilRoBERTa-base architecture as our foundation, with subsequent linear classification layers optimized for each specific task.

We trained and validated our models on stratified train-test splits of the distortion-labeled datasets. We selected the **PQ (span)** and **TT + PQ (span)** to apply to our mental health outcomes data as they were the top performers by F1 score (Table 2) for detection. Some of the one-vs-rest models perform better than the others (Table 3, this could be attributed to PQ dataset: it has 73% examples with more than one distortion type annotated, the models might not pick up on the signals for certain types effectively. We select four one-vs-rest classification models for distortion classification due to their superior performance compared to the other models.

of precision and recall for a class. AUC is short for AUC-ROC, which stands for Area under the Receiving Operating Characteristic curve, a measure of binary classification models' ability to distinguish two classes.

4.2 Predictions on Mental Health Outcomes Dataset

We applied our trained detection and classification models to the DS4UD and iHiTOP texts to quantify the presence of distortions within users' language to analyze in relation to their mental health scores. We then compute the percentage of sentences that contain a detected distortion. For each of the distortion classes, we likewise compute the percent of sentences where the distortion was detected.

5 Results

We examine correlations between detected cognitive distortions and mental health outcomes at two levels: between users (in both DS4UD and iHiTOP datasets) and within users over time (in DS4UD). Using Pearson's r, we analyze how distortion rates correlate with mental health indicators across users, as well as how individual-level fluctuations in distortion rates relate to changes in mental health states.

Finding: Cognitive distortions detected in language are linked to mental health outcomes. The results of the between-user correlation analysis are shown in Table 4. Overall, rates of detected distortions are positively correlated with the mental health outcome scores, reflecting increased severity of mental health conditions as associated with elevated patterns of distorted thinking.

In the iHiTOP dataset, the overall rate of detected distortions by the PQ model correlates significantly positively with the spectra. Significant correlations were observed between all indicators and the rate of distortions detected by the PQ model.

In the DS4UD data, both models identified distortions at rates that correlate significantly with levels of user depression, being emotionally drained, and having nervous stress in the EMAs, and anxiety and depression scores measured by WAVE.

Model	DS4UD				
	EMA			Wave	
	Depressed Mood	Emot Drained	Nervous Stress	Anx	Dep
Should Statements	**0.32**	**0.12**	0.04	0.10	**0.18**
Fortune Telling	0.09	**0.16**	0.09	0.12	**0.13**
Mind Reading	-0.07	-0.04	-0.03	-0.06	-0.07
Labeling	0.26	-0.07	-0.10	-0.10	-0.05

Table 5: **Between-user correlations** (Pearson r) between overall percentage of distortions and mental health assessment scores in the DS4UD dataset. Bold indicates statistically significant (p-values < 0.05).

Distortion Type	Depressed	Emotionally Drained	Nervous Stress
TT + PQ	0.12	0.12	0.13
PQ	0.25	0.13	0.16
Should Statements	-0.02	0.03	0.00
Fortune Telling	0.14	0.08	0.07
Mind Reading	0.00	0.03	0.00
Labeling	-0.23	0.05	0.00

Table 6: **Within user correlations** (mean of Pearson r across users) of the overall percentage of distortions with Psychological State Indicators aligned in time. A higher value means that as the distortion increases so too does the reported condition severity where as a negative correlation indicates the severity decreases as the condition increases.

Considering specific classes of distortions, weak but significant positive correlations were observed between the presence of *should statements* with depressed and emotionally drained states from the EMAs, and depression from WAVE, and likewise for *fortune telling* except which does not have a significant relationship with the depressed EMA state. The difference in the rates of various types of distortions has also been observed in other studies with respect to emotional stress, depressive symptoms and anxiety (Jha et al., 2022; Wang et al., 2025), which could indicate distinct thinking patterns for specific mental health conditions. However, the low degree of associations observed should not necessarily mean that some of the cognitive distortion types could be disregarded.

The within-person analysis for DS4UD dataset is discussed in Table 6. Significance is not reported for this analysis since it captures the average Pearson correlations across a user timeline, and the maximum number of user EMAs is 252 (See §3.1). However, we still observe positive r values for the EMA-level outcomes, which indicates that increase in cognitive distortions expressed in language is weakly positively correlated to worsening mental health scores, even at a user-level. We note that we have a small number of repeated measures for these users (six), which limits the scope of observing within-person patterns in the WAVE outcomes. Future research can explore these relationships with more longitudinal data to assess whether models would detect within-user fluctuations in mental health states and thinking patterns.

6 Conclusion

We evaluated the link between distortion models and mental health outcomes for the authors of the language across two language-to-mental-health datasets: DS4UD and iHiTOP. We found automatically detected distortions in language correlated in general with higher anxiety and depression-related outcomes. In particular, we found that *should statements* and *fortune telling* associated with depressed states. Other types of distortions were not as easy to detect, suggesting further development may unlock additional benefits of NLP-based distortion detectors. Our findings establish language-based distortion detection as a promising tool for mental health professionals, offering empirically-validated support for identifying and addressing cognitive distortions in clinical settings. Our work contributes to advancing methods for early detection of mental health conditions like depression that can be integrated in real-world monitoring and intervention strategies.

Limitations

Our study faces methodological and data constraints that warrant consideration. The classification models show variable performance across different types of cognitive distortions, with some categories like mind reading and personalization showing particularly weak correlations with mental health outcomes. We also drop three classes of cognitive distortions from TT in our analyses for ease of combining the datasets. This suggests room for improvement in capturing more nuanced forms of distorted thinking and exploring more complex frameworks. We have limited our analysis to two datasets that could have potential sampling biases. We limit the analysis to English-language. Further, our computational approach faces several challenges. The reliance on automated detection methods may miss contextual nuances that human clinicians typically observe. These methods should

be used as an assistance rather than a replacement for human clinicians.

Ethics Statement

As NLP continues to advance in enhancing human-centered applications such as improving mental health assessments, striking a balance between respecting human privacy and promoting open data sharing becomes increasingly important. In this case, the data was shared with consent solely for academic research and was anonymized. Open sharing would breach the trust with participants and violate agreements with ethical review boards. Ideally, all data should be released while maintaining privacy, however, the limited availability of data underscores the need for those with access to share their work openly within established ethical guidelines, such as the training datasets used in this work.

All data collection, storage, and secondary analyses procedures were approved by an academic ethics institutional review board.

7 Acknowledgements

This work was supported in part by a grant from the NIH-NIAAA (R01 AA028032) awarded to H. Andrew Schwartz at Stony Brook University. The conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, NIH, any other government organization, or the U.S. Government.

References

Lauren A. Rutter, Andy Edinger, Lorenzo Lorenzo-Luaces, Marijn ten Thij, Danny Valdez, and Johan Bollen. 2025. Anxiety and depression are associated with more distorted thinking on social media: A longitudinal multi-method study. *Cognitive Therapy and Research*, pages 1–9.

Krishna C Bathina, Marijn Ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. 2021. Individuals with depression express more distorted thinking on social media. *Nature human behaviour*, 5(4):458–466.

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Ignacio de Toledo Rodriguez, Giancarlo Salton, and Robert Ross. 2021. Formulating automated responses to cognitive distortions for CBT interactions. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 108–116, Trento, Italy. Association for Computational Linguistics.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level nlp: The role of sample size and dimensionality. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4515. NIH Public Access.

Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36:427–440.

Ajeya Jha, Akash Kumar Bhoi, Saibal Kumar Saha, Ankit Singh, Samrat Mukherjee, Bibeth Sharma, and Jayarani. 2022. Impact of select cognitive distortions on emotional stress. *Cognitive Computing for Risk Management*, pages 31–44.

Roman Kotov, David C Cicero, Christopher C Conway, Colin G DeYoung, Alexandre Dombrovski, Nicholas R Eaton, Michael B First, Miriam K Forbes, Steven E Hyman, Katherine G Jonas, et al. 2022. The hierarchical taxonomy of psychopathology (hitop) in psychiatric practice and research. *Psychological medicine*, 52(9):1666–1678.

Tatiana Leroy, David Nicholas Top Jr., Russell J. Bailey, Christina Bartholomew, Julia Toomey, Taylor Baker, Josephine Schwalbe, and Kirra D. Jensen and. 2025. Accessing care: qualitative analysis of counseling center therapists, Âô experiences of transitioning to telehealth. *Cogent Mental Health*, 4(1):1–46.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

August Håkan Nilsson, Hansen Andrew Schwartz, Richard N Rosenthal, James R McKay, Huy Vu, Young-Min Cho, Syeda Mahwish, Adithya V Ganesan, and Lyle Ungar. 2024. Language-based ema assessments help understand problematic alcohol consumption. *Plos one*, 19(3):e0298300.

Yaakov Ophir, Christa SC Asterhan, and Baruch B Schwarz. 2017. Unfolding the notes from the walls: Adolescents' depression manifestations on facebook. *Computers in Human Behavior*, 72:96–107.

Darrel A Regier, William E Narrow, Emily A Kuhl, and David J Kupfer. 2009. The conceptual development of dsm-v. *American Journal of Psychiatry*, 166(6):645–650.

Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345.

Stuart J Rupke, David Blecke, and Marjorie Renfrow. 2006. Cognitive therapy for depression. *American family physician*, 73(1):83–86.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

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.

Xi Wang, Yujia Zhou, and Guangyu Zhou. 2025. Unveiling the cognitive burden: The impact of stigma on distorted thinking among individuals living with hepatitis b. *International Journal of Clinical and Health Psychology*, 25(1):100556.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.

A Within-person Analysis

Significance is not reported for this analysis since it captures the average Pearson correlations across a user timeline, and the maximum number of user EMAs is 252 (See §3.1). However, we still observe positive r values for the EMA-level outcomes, which indicates that increase in cognitive distortions expressed in language is weakly positively correlated to worsening mental health scores, even at a user-level. We note that we have a small number of repeated measures for these users (six), which limits the scope of observing within-person patterns in the WAVE outcomes. Future research can explore these relationships with more longitudinal data to assess whether models would detect within-user fluctuations in mental health states and thinking patterns.