Psycholinguistic Profiles of Cognitive Distortions in Reddit Data

Neha Sharma
University of Tartu, Estonia
neha.sharma@ut.ee

Navneet Agarwal
University of Tartu, Estonia
navneet.agarwal@ut.ee

Kairit Sirts
University of Tartu, Estonia
kairit.sirts@ut.ee

Abstract

Cognitive distortions (CDs) are systematically biased patterns of thinking associated with the onset and maintenance of mental health conditions such as depression and anxiety. Computational research on CDs has primarily focused on detection and classification, while the linguistic characterization of distorted language; what psycholinguistic features distinguish distorted from non-distorted text, and whether individual distortion types carry distinct language patterns, remains largely unexplored. Using a Reddit dataset, we apply a Generalized Linear Model (GLM) with bootstrap sampling to LIWC-derived features and find that CD language is psycholinguistically distinct from non-distorted language. We further characterize type-specific psycholinguistic profiles for each CD, and through hierarchical clustering show that CD types are not fully separable, with certain distortions sharing stable linguistic signatures. Together, these findings contribute to the linguistic characterization of CDs, offering an empirically grounded account of the psycholinguistic properties that distinguish distorted language at the level of CDs as a whole and across specific distortion types.

1 Introduction

Cognitive distortions (CDs) are systematic patterns of negatively biased thinking that distort how individuals interpret events and reality (Beck, 1963; Beck et al., 1979). Distorted thinking patterns have been associated with depression, anxiety, and a range of psychological difficulties, making their study important for both clinical practice and mental health research (Joormann and Stanton, 2016; Ouhmad et al., 2024). Crucially, CDs are not purely internal cognitive events; they can be reflected in language. Beck (1963)'s work identified distortions through the verbal content of patient speech, and subsequent theoretical work has described each distortion type in terms of characteristic patterns of expression (Burns, 1980). This makes language a natural and accessible medium through which distorted thinking can be studied.

Computational research on CDs has approached the phenomenon as a detection and classification problem: given a text segment, determining whether it contains a distortion (binary classification) and, if so, which distortion type(s) (multi-label, multi-class classification) (Sage et al., 2025). A range of methods have been applied to these tasks, from rule-based systems to fine-tuned language models (Sage et al., 2025), which typically enable inferences at the instance level.

A complementary but distinct direction is linguistic profiling, which characterizes the psycholinguistic[1] properties of distorted language and examines whether individual distortion types leave distinguishable traces in language use. Rather than classifying individual texts, this approach aims to make inferences at the population or phenomenological level, describing what distorted language looks like as a broader linguistic phenomenon. Despite its potential, this direction has so far received limited attention.

This approach is well-established in the broader mental health NLP literature, where linguistic markers have been used to characterize conditions such as depression, anxiety, PTSD, and suicidality, yielding interpretable, theoretically grounded insights into the language of these conditions (Homan et al., 2022; Khuon et al., 2026). These studies have consistently demonstrated that different psychological states leave systematic, identifiable traces in language; traces that can be characterized in ways that are meaningful to both clinicians and researchers.

The groundwork for linguistic profiling of CDs

[1] In this work, "psycholinguistic" refers to the characterization of text through lexical features that capture both linguistic structure (e.g., word choice, syntactic markers, temporal orientation) and psychological processes (e.g., affective, cognitive, and social dimensions), as operationalized by LIWC-22

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 306–323

already exists. Burns (1980) described characteristic linguistic patterns for individual distortion types, for example "Overgeneralization" defined as drawing broad conclusions from a single negative event, expressed through words such as "always" or "never." Bathina et al. (2021) subsequently demonstrated that these descriptions can be translated into identifiable linguistic units detectable in text. Despite growing computational interest in CDs, systematic linguistic characterization of individual distortion types remains limited. The psycholinguistic profiles that would make distortion-specific language patterns interpretable and comparable have not yet been constructed. This study tries to address this research gap guided by the following research questions:

RQ1: Do texts containing cognitive distortions exhibit distinct psycholinguistic characteristics compared to texts without cognitive distortions, as measured by LIWC-derived features?

RQ2: Can individual cognitive distortion types be characterized by distinct psycholinguistic profiles, and to what extent are these profiles linguistically distinguishable from one another?

Guided by these research questions, we first establish statistical differences in language between texts containing CDs v/s those without. We further construct psycholinguistic profiles for individual distortion types considered in this study, and explore similarities between them.

2 Related Work

CD detection has followed a familiar methodological progression in NLP, from lexicon-driven approaches such as curated n-gram schemata (Bathina et al., 2021), to feature-engineered representations (e.g., n-grams/TF–IDF and psycholinguistic features) paired with classical machine learning (Simms et al., 2017; Shickel et al., 2019; Shreevastava and Foltz, 2021), and more recently to neural and transformer-based modeling (Rojas-Barahona et al., 2018; Mostafa et al., 2021; Lybarger et al., 2022), followed by LLM-based approaches (Chen et al., 2023; Lim et al., 2024; **?**).

While predictive modeling of CDs has been extensively explored, the linguistic characterization of distorted language remains comparatively understudied. Research in mental health NLP has long used linguistic markers to study conditions such as depression (Rude et al., 2004; De Choudhury et al., 2021; Eichstaedt et al., 2018), anxiety (Al-Mosaiwi and Johnstone, 2018; Abutara et al., 2025), PTSD (Coppersmith et al., 2014; Quillivic et al., 2025), and suicidality (Dobbs et al., 2023; Agurto et al., 2018), showing that language patterns can provide interpretable insights into underlying psychological states. Linguistic Inquiry Word Count (LIWC) framework (Boyd et al., 2022; Tausczik and Pennebaker, 2010) is one such tool which has consistently revealed systematic differences in language use across clinical populations. For example, depressed individuals tend to use more negative emotion words, first-person singular pronouns, and absolutist language (Tackman et al., 2019), while anxious language is associated with elevated uncertainty and threat-related vocabulary (Al-Mosaiwi and Johnstone, 2018).

Several studies have also incorporated LIWC features into CD detection models, improving classification performance (Shreevastava and Foltz, 2021; Bollen et al., 2021; Simms et al., 2017; Wiemer-Hastings et al., 2004). However, these studies treat LIWC features as model inputs rather than objects of analysis, leaving the linguistic characteristics of individual distortion types largely unexamined.

Bathina et al. (2021) developed 241 n-grams (cognitive distortion schemata, CDS) to capture the linguistic characteristics of 12 CD types, validated with CBT experts and computational linguists. The same lexicon was subsequently applied to historical book corpora (Bollen et al., 2021) and psychotherapy transcripts (Lalk et al., 2024), demonstrating that CD-specific language is detectable across diverse text sources. However, these operationalizations remain anchored to surface n-gram matching and have been used primarily for detection and prevalence tracking rather than systematic linguistic characterization of individual distortion types.

To our knowledge, no study has examined the broader linguistic profiles of individual CD types across a common feature space. This work addresses this gap by constructing per-CD linguistic profiles using psycholinguistic features and examining the differences and similarities between linguistic profiles of different CDs.

3 Data and Annotations

This section describes the dataset, CD labeling process, and LIWC features used in this study.

Stage	Users		Posts		Avg.	
	C	D	C	D	C	D
raw	35,753	3,070	∼32.0M	∼5.34M	895	1,739
proc.	35,697	3,069	∼19.71M	∼3.22M	552	1,048
balanced	3,069	3,069	∼1.69M	∼3.22M	551	1,048
final	3,069	3,069	∼1.53M	∼3.20M	498	1,041

Table 1: RSDD statistics. proc.: preprocessed. C/D denote Control/Depression. final: corpus after CD labeling (§3.2). Avg.: average number of posts per user.

3.1 RSDD Dataset

In this study we use the Reddit Self-Reported Depression Diagnosis (RSDD) dataset (Yates et al., 2017), a large corpus of Reddit posts from users with self-reported depression and a matched control cohort; for details about the original dataset refer to the original study. To this dataset, we first applied basic pre-processing involving removal of empty strings, posts without user labels, and non-English text. We then balanced the dataset by taking an equal number of users in both cohorts. Refer to Table 1 and Appendix A for additional details.

3.2 Cognitive Distortion Labeling

The RSDD corpus does not contain cognitive distortion labels. To enable our linguistic analyses, we therefore need to construct CD labels for RSDD posts. The publicly available Therapist Q&A dataset (Shreevastava and Foltz, 2021) provides CD annotations on user inputs across 2,530 user-therapist exchanges and has supported prior work on CD detection and classification.

CD detection is, however, an inherently subjective task. Sharma et al. (2026) argue that in such domains, where objective ground truth is unattainable, label reliability provides the practical way forward: when the same label is consistently produced across independent annotation passes, this consistency indicates that the underlying text contains genuine signals rather than annotation-specific noise. This consideration is directly relevant to Therapist Q&A: its human inter-annotator agreement of 33.7% reflects substantial annotator disagreement, indicating that any single annotation set on this corpus cannot be treated as reliable ground truth. We therefore do not use Therapist Q&A directly for our analyses, and instead build a labeling pipeline that applies the reliability principle to generate labels for RSDD using Therapist Q&A annotations as input.

Sharma et al. (2026) operationalize this principle through multiple LLM annotation runs, retaining labels that appear consistently across runs as reliable. We adapt the same principle to operate across two independently produced annotation sets of the Therapist Q&A dataset: the human annotations of Shreevastava and Foltz (2021) (IAA 33.7%) and the LLM-based annotations of Sharma et al. (2026) (Fleiss's κ 0.78).

We train two separate multi-label mentalRoBERTa-based classifiers (Ji et al., 2022) with default parameters, one on each annotation set. Both classifiers follow the architecture, hyperparameters, and data splits of Sharma et al. (2026), and reproduce classification performance comparable to that reported in the original study. Both classifiers are then retrained on the full Therapist Q&A dataset, leveraging all available annotated examples to maximize the training signal for label transfer, and applied to the RSDD corpus, producing two independent sets of predicted CD labels per post.

Since our goal is linguistic profiling rather than classification, label reliability is paramount. We adopt a conservative agreement-based filtering strategy; we retain only RSDD posts where the two label sets agree in one of two forms. Complete agreement indicates that both classifiers assign identical CD label sets to a post. Partial agreement indicates that the two label sets share at least one common CD label but are not identical; in this case, we take the shared label(s) as the final assignment. Posts with entirely disjoint predictions are discarded. The rationale is that labels that survive this cross-source agreement are more likely to reflect genuine distortion-relevant linguistic signals than artifacts of any specific annotation source. The retained labels form the basis for all downstream linguistic analyses. Table 1 (final row) reports the resulting labeled corpus, and Table 2 lists the CD types considered. Full details of the labeling pipeline and classifier training are provided in Appendix B.

3.3 LIWC

Psycholinguistic features were extracted for the dataset using LIWC-22[2] (Boyd et al., 2022), which is a dictionary-based tool that maps text onto interpretable linguistic and psychological dimensions. From the full LIWC feature set (120+ categories), we selected a subset capturing core linguistic struc-

[2]LIWC-22 was accessed under a research license

No.	Cognitive Distortion	Description and Example
1.	Emotional Reasoning (ER)	Assuming emotions reflect reality. *Example:* "I feel worthless, so I must be a failure."
2.	Overgeneralization (Over)	Drawing broad conclusions from limited experiences. *Example:* "I failed this interview, I'll never get a job."
3.	Should Statements (SS)	Holding rigid expectations for oneself or others. *Example:* "I should always be calm and never get upset."
4.	All-or-Nothing Thinking (AoN)	Viewing situations in extremes. *Example:* "If I'm not the best, I'm a total failure."
5.	Mind Reading (MR)	Presuming negative judgments from others. *Example:* "She didn't say hi, she must think I'm annoying."
6.	Fortune Telling (FT)	Predicting negative outcomes with certainty. *Example:* "There's no point in applying, I know I won't get accepted."
7.	Magnification (Mag)	Exaggerating potential problems. *Example:* "If I mess up this report, I'll lose my job and never recover."
8.	Personalization (Per)	Taking undue responsibility for external events. *Example:* "My friend is upset, it must be something I did wrong."
9.	Labeling (La)	Defining oneself or others by single traits. *Example:* "I missed a deadline, I'm so incompetent."
10.	Mental Filter (MF)	Focusing only on negative aspects. *Example:* "Everyone said my presentation was good, but one person criticized it, so it must have been terrible."

Table 2: List of CDs considered in this study based on Shreevastava and Foltz (2021). Mental Filter is excluded from our analysis due to negligible presence (26 posts out of 4.73M) in the RSDD predictions.

ture, cognitive processing, temporal orientation, and affective dimensions, excluding hierarchically redundant or task-irrelevant categories. Appendix C provides more details on selected features.

4 Methodology

This section details our analysis process with the following subsections explaining task formulation, statistical modeling, and sampling strategies used within the process.

4.1 Task Formulation

To answer our research questions, we define three complementary tasks.

1. **TASK(CD/ND):** Posts containing at least one CD (positive class) are compared against No Distortion (ND) posts (negative class), capturing broad linguistic differences between distorted and non-distorted language.

2. **TASK(XCD/ND):** Identifying CD-specific features relative to No distortion posts.

3. **TASK(XCD/¬XCDS):** Each CD type (positive class) is compared against all remaining CD types (negative class) (one-vs-rest), isolating what linguistically distinguishes it from other CDs.

TASK(CD/ND) answers RQ1 and also provides a reference point for interpreting results for RQ2. TASK(XCD/ND) and TASK(XCD/¬XCDS) together drive our CD-type profiling (RQ2).

4.2 Sampling Strategy

Our corpus contains 6,138 users and ∼4.73M posts, with multiple posts per user. Running inference on the full dataset is computationally expensive and can yield overly small p-values due to the very large n. We therefore use a *user-clustered bootstrap* with a controlled subsampling procedure that (i) preserves user-level dependence and (ii) yields balanced, comparable bootstrap samples. The following procedure generates one bootstrap sample:

1. **Eligible users:** Identify users who have at least one *positive* and one *negative* class post under the task definition (§4.1).

2. **Stratify users:** Sample an equal number of users from the depression and control groups (these group labels are *not* used as predictors; they are used only to avoid dominance of one user population in subsamples).

3. **Balance classes:** Within sampled users, draw posts to obtain equal numbers of positive and negative examples for each user.

4. **Cap size:** Continue sampling users until the total post count first exceeds 10,000 posts. Based on pilot experiments, a cap of 10,000 posts provided a practical stability–compute trade-off.

4.3 Statistical Modeling

We initially considered mixed-effects multivariable logistic regression (Generalized Linear Mixed

Model; GLMM) (Bates et al., 2015) shown in equation 1 for our analysis. This choice was based on the binary nature of our tasks §4.1, multiple LIWC features as predictors (fixed effect), and use of random intercept to account for dependency due to multiple posts per user.

$$y \sim \underbrace{x_1 + x_2 + \cdots + x_K}_{\textit{Fixed Effects}} + \underbrace{(1 \mid \text{user})}_{\textit{Random Intercept}} \quad (1)$$

In preliminary analysis, the estimated variance of the random intercept consistently approached zero (i.e., boundary/singular fits) under our sampling design, indicating that the user-level random effect contributed negligibly once LIWC predictors were included and samples were constructed in a user-aware, balanced manner. In this situation, the GLMM reduces to a fixed-effects-only Generalized Linear Model (GLM) (Nelder and Wedderburn, 1972). For computational efficiency and numerical stability we therefore report results from GLM defined as:

$$\text{logit}\big(\Pr(y_i = 1)\big) \;=\; \beta_0 \;+\; \sum_{k=1}^{K} \beta_k \, x_{k,i}. \quad (2)$$

Where, $y_i \in \{0, 1\}$ denote the task-specific label for i^{th} post, $x_{k,i}$ denote the k-th LIWC feature for i^{th} post, β_k is corresponding fixed-effect coefficient, and β_0 is the intercept.

For each task, we fit a GLM using the selected LIWC features and quantify uncertainty via $B = 500$ bootstrap samples. We use bootstrap median coefficient ($\tilde{\beta}$) as the primary effect estimate, together with 95% confidence interval (CI) and bootstrap standard error ($\widehat{\text{SE}}$_boot) values. Coefficients are reported on the log-odds scale, where the sign indicates whether a feature is associated with a higher ($\tilde{\beta} > 0$) or lower ($\tilde{\beta} < 0$) probability of the positive class. For interpretability, we also report odds ratios $\widetilde{\text{OR}} = \exp(\tilde{\beta})$ and the corresponding percentage change in odds, computed as $(\widetilde{\text{OR}} - 1) \times 100$ (Heyard, 2026). Since predictors are standardized (z-scored), these odds ratios represent the change in odds associated with one standard deviation increase in a given LIWC feature, holding other features constant.

To ensure significance, consistency, and stability of reported features, we use 95% CI and Relative Standard Error (RSE). Excluding features with zero within their 95% CI ensures that retained features are both significant and sign consistent across bootstraps. Feature stability across bootstraps is measured using RSE, defined as:

$$\text{RSE} = \frac{\widehat{\text{SE}}\text{\_boot}}{|\tilde{\beta}|}, \quad (3)$$

RSE provides a scale-free measure of stability: the same absolute bootstrap SE ($\widehat{\text{SE}}$_boot) can imply high stability for a large effect (large $\tilde{\beta}$) but low stability for a small effect (small $\tilde{\beta}$).

5 Results

This section discusses the results from our statistical modeling for 3 tasks in §4.1.

5.1 Distorted vs non-distorted language

Table 3 summarizes the significance and effect of LIWC features for the TASK(CD/ND). Overall, CD text corresponds to a more negative *Tone* ($\tilde{\beta}$=−0.61, −45.49%) and shows highest association with *swear* ($\tilde{\beta}$=0.65, +91.31%). Temporally CD language is shown to be less past-oriented ($\tilde{\beta}$=−0.18, −16.09%) and more present-focused ($\tilde{\beta}$=0.10, 10.91%). Absolutist vocabulary, captured by *allnone*, is strongly elevated in CD posts ($\tilde{\beta}$=0.41, +50.60%), and *negations* are associated with higher CD odds ($\tilde{\beta}$=0.29, +33.58%).

These results provide a clear answer to RQ1: texts containing cognitive distortions differ significantly from texts without cognitive distortions on a range of psycholinguistic features. CD text is consistently negative in emotional language, with elevated negative emotion words *(emo_neg)*, reduced positive emotion words *(emo_pos)*, and an overall negative tone *(Tone)*. It is more absolutist *(allnone)* and has a reduced analytic structure. Linguistically, CD text is more likely to contain moral language *(moral)* and achievement-related language *(achieve)*, and is more interpersonally referenced *(socrefs)* while being less affiliative and less polite.

5.2 Cognitive Distortion Profiles

TASK(xCD/ND) and TASK(xCD/¬xCDs) (§4.1), together help define the psycholinguistic profiles of individual CDs considered in this study. The profiles are defined based on 44 unique LIWC features obtained from the union of the top 15 features (highest absolute $\tilde{\beta}$) for each CD within beforemetioned tasks. Figure 1 plots these profiles. TASK(CD/ND) further represents global CD trends, providing a reference point for analyzing individual CD profiles. For each CD we group

Feature	$\tilde{\beta}$	$\widetilde{\mathrm{OR}}$	%Δ Odds	RSE	95% CI
swear	0.65	1.91	91.31	••	[0.50, 0.82]
Tone	-0.61	0.55	-45.49	•••	[-0.70, -0.51]
moral	0.44	1.55	55.01	••	[0.31, 0.59]
WC	0.44	1.55	54.75	••	[0.28, 0.65]
allnone	0.41	1.51	50.60	••	[0.31, 0.51]
emo_neg	0.33	1.40	39.68	••	[0.18, 0.51]
negate	0.29	1.34	33.58	••	[0.19, 0.39]
socrefs	0.24	1.27	26.90	•	[0.11, 0.37]
polite	-0.18	0.84	-16.49	•	[-0.34, -0.06]
achieve	0.18	1.19	19.33	••	[0.11, 0.25]
focuspast	-0.18	0.84	-16.09	••	[-0.26, -0.09]
emo_pos	-0.17	0.84	-15.97	•	[-0.27, -0.08]
affiliation	-0.16	0.85	-14.70	•	[-0.25, -0.07]
Analytic	-0.16	0.85	-14.59	•	[-0.25, -0.07]
insight	0.15	1.16	16.24	••	[0.08, 0.21]
conflict	-0.13	0.88	-12.00	•	[-0.21, -0.04]
power	0.12	1.12	12.33	•	[0.04, 0.19]
focuspresent	0.10	1.11	10.91	•	[0.02, 0.18]
certitude	0.09	1.10	9.78	•	[0.04, 0.15]
discrep	-0.08	0.92	-7.91	•	[-0.15, -0.01]
cause	0.08	1.08	8.35	•	[0.02, 0.14]
curiosity	-0.08	0.92	-7.60	•	[-0.16, -0.01]
lack	0.07	1.07	7.29	•	[0.02, 0.15]

Table 3: TASK(CD/ND): GLM feature effects (median across bootstraps) with uncertainty and stability. $\tilde{\beta}$ and $\widetilde{\mathrm{OR}} = \exp(\tilde{\beta})$ denote bootstrap medians. Relative Standard Error (RSE) encodes stability categories: ••• = extremely stable (RSE ≤ 0.10), •• = stable (0.10–0.25], • = moderately stable (0.25–0.50]. %Δ Odds is computed as $(\widetilde{\mathrm{OR}} - 1) \times 100$. All reported effects are extremely sign-consistent across bootstraps (consistency $> 99\%$).

the features into two categories: Core markers and Global markers. Core markers are features where individual CD values deviate considerably from global trends i.e. substantial $\tilde{\beta}$ values for TASK(XCD/ND) and TASK(XCD/¬XCDS), while Global markers are features where CD closely follows global distortion values. For example, in All-or-Nothing Thinking (AoN), *Authentic & socrefs* are Core markers whereas *Tone & i* are Global markers. Appendix D Table 8 presents the complete list of Core and Global markers for individual CDs.

Figure 1 highlights patterns that are characteristic of individual CDs and can help differentiate them from other CDs as well as non-distorted text. Anxiety-related words (*emo_anx*) are characteristic of Emotional Reasoning (ER) and Fortune Telling (FT), while social references (*socrefs*) are characteristic of Mind Reading (MR) and Magnification (Mag). These examples illustrate that individual CDs can be characterized by distinct psycholinguistic patterns in LIWC-derived features, positively answering our second research question.

6 CD Profile Clustering

While constructing individual CD profiles, we observed that many psycholinguistic features appeared across multiple CD types, suggesting substantial overlap in their linguistic signatures. This indicates that CD types are not fully distinct at the linguistic level, and the overlap in their profiles is worth exploring. To this end, we apply hierarchical clustering (Ward, 1963) using the GLM median coefficient ($\tilde{\beta}$) values of the CD profiles (see Appendix E for the full tables of feature and $\tilde{\beta}$).

Figure 2 presents the clustering results for both TASK(XCD/ND) and TASK(XCD/¬XCDS). Although some differences exist, certain groupings remain consistent: Fortune Telling, Emotional Reasoning, and Magnification cluster together, and this grouping is stable across both tasks, appearing with low within-cluster correlation distance, as do Personalization and Labeling. The top 10 features driving each cluster are reported in Table 4.

7 Discussion

Our findings confirm that CD language is psycholinguistically distinct from non-distorted language. Our results show a positive association between CD and LIWC features representing negative emotions *(emo_neg, swear)*, absolutist vocabulary *(allnone, certitude)*, and negation language *(negate)*. These findings align with Burns (1980)'s characterization of CDs as patterns of systemati-

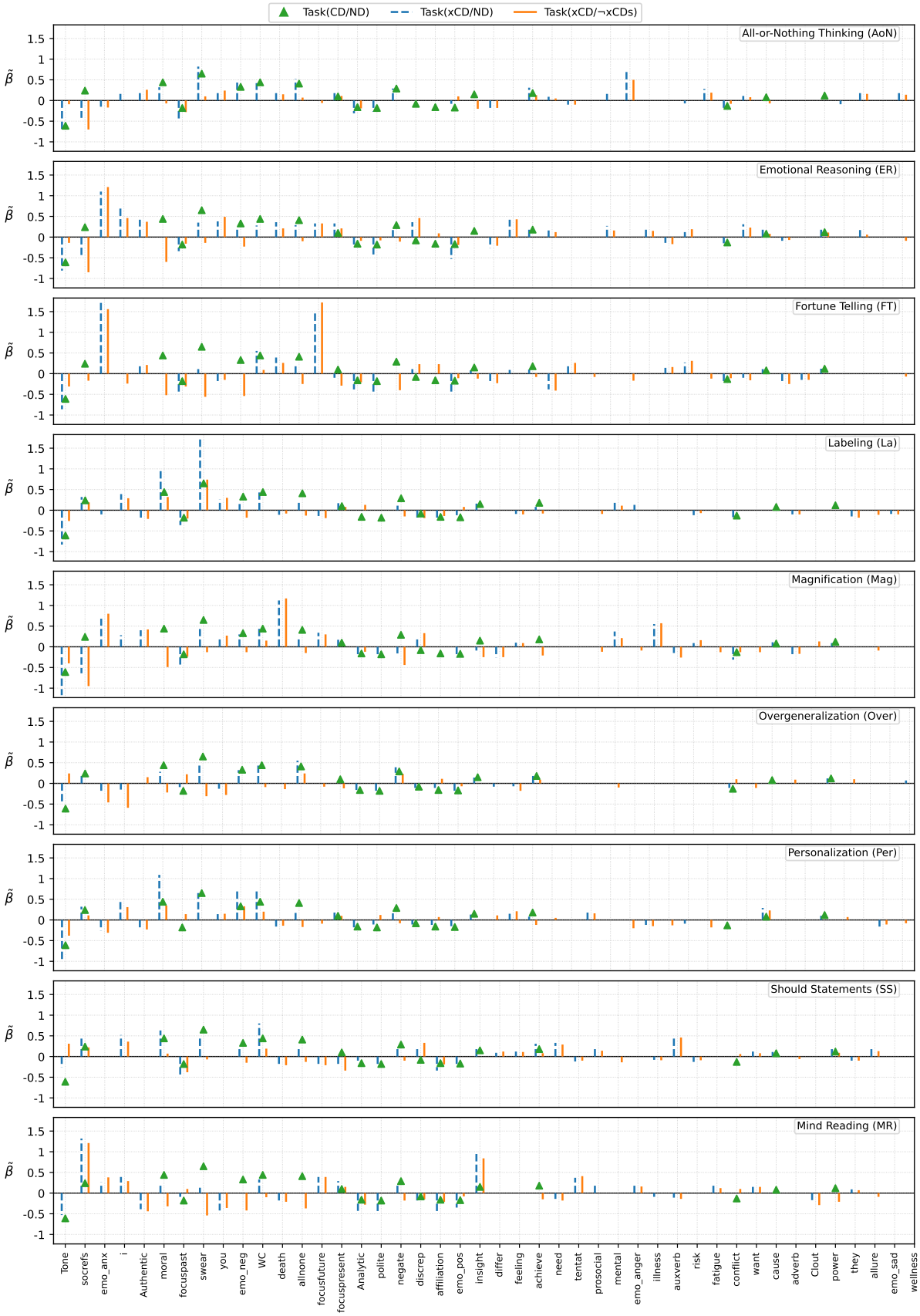

Figure 1: Psycholinguistic profiles of 9 CDs from TASK(CD/ND): ▲, TASK(xCD/ND): −−, and textscTask(xCD/¬xCDs): −. X-axis plots the 44 union LIWC features considered and Y-axis plots the bootstrap median coefficient ($\tilde{\beta}$).

Task	Cluster CDs	Top 10 driver features
TASK(CD/ND)	ER, FT, Mag	+emo_anx, −socrefs, +focusfuture, +death, +Authentic, +swear, −discrep, −Tone, +illness, −emo_pos
	AoN, Over	−emo_anx, +i, +allnone, +emo_anger, −socrefs, +negate, −emo_pos, −Tone, +emo_neg, +achieve
	La, Per, SS	+moral, −emo_anx, +swear, −focusfuture, −death, +socrefs, −Authentic, +WC, +i, +auxverb
	MR	+socrefs, +insight, −you, −swear, −Authentic, −death, +tentat, −affiliation, −discrep, −focuspast
TASK(XCD/¬XCDS)	AoN, ER, FT, Mag	−socrefs, +emo_anx, +focusfuture, +death, +Authentic, −moral, −insight, −differ, +discrep, +you
	La, Per	+moral, −emo_anx, +swear, −focusfuture, −Authentic, +socrefs, −death, +emo_neg, −discrep, −Tone
	MR, SS	+socrefs, +insight, −death, −Authentic, +Tone, −you, +i, −swear, −affiliation, +auxverb
	Over	−emo_anx, −i, +negate, +Tone, −you, +allnone, +focuspast, −focusfuture, −death, −feeling

Table 4: Top 10 driver features per cluster for TASK(CD/ND) and TASK(XCD/¬XCDS). These features indicate which shared positive and negative patterns underlie the observed clustering.

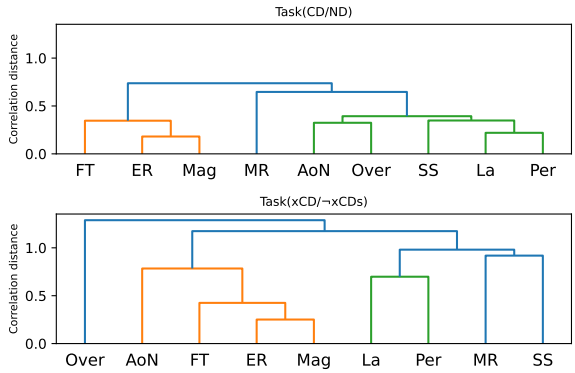

Figure 2: Hierarchal clustering of CDs in TASK(XCD/ND) and TASK(XCD/¬XCDS) based on their lingustic profiles and corresponding $\tilde{\beta}$ values.

cally negative and absolutist thinking. Furthermore, these inferences align with findings from previous studies (Simms et al., 2017; Al-Mosaiwi and Johnstone, 2018), that also show higher use of negative emotions and absolutist language in CD text as compared to control.

Additionally, our results also show that individual CDs carry distinct linguistic cues, not only enabling the characterization of type-specific psycholinguistic profiles, but also allowing differentiation between them. For example, future orientation *(focusfuture)* and anxiety language *(emo_anx)*, combined with elevated tentative language *(tentat)*, are characteristic features of Fortune Telling profile, which directly reflect its clinical definition of predicting negative outcomes as facts without evidence (Burns, 1980, p. 37). Similarly Emotional Reasoning profile shows strong anxiety *(emo_anx)* and feeling *(feeling)* and highly self-referential language *(i)*, which is consistent with its clinical definition of treating emotional states as evidence (Burns, 1980, p. 38). These patterns are further supported by Shreevastava and Foltz (2021), who also found elevated future-focused language for Fortune Telling and elevated feel language for Emotional Reasoning. Similar patterns were observed across all remaining CD types, with each profile reflecting its corresponding clinical definition from Burns (1980, p. 32–43).

The psycholinguistic profiles identified in this study contribute to the linguistic characterization gap in prior work. Rather than treating CD as an opaque categorical label, these profiles specify which linguistic dimensions are elevated or suppressed in distorted text, and how strongly, enabling inferences at the population and phenomenological level. Furthermore, the estimated $\tilde{\beta}$ values provide a quantifiable, and continuous measure of each LIWC-feature and its association with a given CD. This can be used to estimate the severity of predicted distortion by comparing the linguistic profile of a given text with reference to the profile of the corresponding CD. This can also enable longitudinal tracking of changes in a person's cognitive distortion patterns, offering more insights into a person's mental health. This kind of evidence-based linguistic profiling is particularly valuable in clinical and research contexts, where understanding the psycholinguistic nature of distorted thinking at the population level complements downstream applications such as detection and classification.

Beyond individual CD profiles, clustering analysis revealed linguistic similarity across CD types. The consistent grouping of Fortune Telling, Emotional Reasoning, and Magnification across both tasks is theoretically coherent: following Burns (1980, p. 32–43), these distortions share a negative future outlook, reflected in their driver features; anxiety language *(emo_anx)*, future orientation *(focusfuture)*, and death/illness vocabulary *(death, illness)*. This aligns with Shickel et al. (2019), who similarly found natural groupings among overlapping CD types, reinforcing that CD categories are not fully separable linguistically, however, their specific cluster compositions differ from our find-

ings, likely due to the use of different datasets and CD taxonomies.

8 Future Work

In the future, we plan to leverage the temporal structure of the RSDD dataset alongside CD psycholinguistic profiles to track how distortion patterns evolve over time. This would enable estimation of distortion severity and longitudinal monitoring. Additionally, integrating CD profiles with mental health measures, such as self-reported depression, could reveal how specific distortions cooccur with these conditions. This may help identify early indicators or risk factors, providing a more nuanced understanding of the relationship between cognitive distortions and mental health.

9 Conclusion

This study investigated the psycholinguistic properties of cognitive distortions, motivated by a gap in prior research: while detection systems are well-studied, the linguistic cues underlying distorted thinking remain relatively unexplored. We proposed that psycholinguistic profiling via LIWC-derived features offers a viable path toward addressing this. Our results confirm that CD language is psycholinguistically distinct from non-distorted language. Beyond the global distinction, individual CD types exhibit meaningful psycholinguistic profiles that broadly mirror their clinical definitions, demonstrating that linguistic markers can characterize not only whether a text is distorted but also what kind of distortion it reflects. Hierarchical clustering further reveals that CD types are not fully separable at the linguistic level, certain distortions share stable psycholinguistic signatures, reflecting their linguistic overlap.

Together, these findings suggest that psycholinguistic profiling provides an empirically grounded and human-interpretable characterization of cognitive distortions at the population level by revealing how cognitive distortions manifest linguistically, a particularly valuable property in the mental health domain where understanding the nature of the phenomenon is as important as its detection.

Limitations

The RSDD dataset consists of Reddit posts, which differ substantially from clinical text; findings may therefore not generalize to clinical populations or settings. Furthermore, although depression and control labels are not used as predictors, they guide sampling, meaning results may be implicitly shaped by this divide. Additionally, since cognitive distortions are inherently context-dependent, the variable length and informal nature of Reddit posts may also introduce noise into the profiles.

The CD labels carry additional uncertainty given the subjective nature of the task. Although we try to mitigate this by combining two different labeling sources for our CD predictions, the final CD labels can still carry noisy predictions arising from original training labels.

As a lexicon-based framework, LIWC lacks sensitivity to meaning and context, meaning psychologically distinct expressions may receive identical feature representation. Furthermore, the initial LIWC feature selection was guided by the researchers' domain knowledge and subjective judgment, and may therefore vary across research teams, potentially yielding different feature sets.

Finally, while the GLMM random intercept collapsing to zero justified the simplification to GLM, this may partly reflect an artifact of the sampling procedure itself, equal per-user balancing may have removed the very user-level variance the random intercept was intended to capture, rather than indicating a genuine absence of user dependency in data.

Acknowledgments

This research was supported by the Estonian Centre of Excellence in AI (EXAI) and by the Estonian Research Council Grant PRG3182.

Ethics Statement

This study uses two datasets. The Therapist Q&A dataset is publicly available and contains anonymized text with no personally identifiable information; all use complies with its terms of use. The RSDD dataset was obtained directly from the original authors under a formal data agreement. No personally identifiable information was accessed, and no attempts were made to identify or contact any individuals in the dataset.

All analyses rely on publicly available/with agreement access data, together with open-source tools and the licensed LIWC-22 framework (accessed under a research license). The broader goal of this work is the linguistic characterization of cognitive distortions to support mental health research. As the data originates from social media,

findings should be interpreted with caution and are not intended for clinical diagnosis or therapeutic decision-making. We encourage responsible use within appropriate research and ethical boundaries.

Data and Code Availability

The Therapist Q&A dataset and its annotations are available through Shreevastava and Foltz (2021) and Sharma et al. (2026) respectively. Access to the RSDD dataset may be requested directly from its original authors. Code to reproduce the analyses presented in this work is available on Github[3].

References

Ana Abutara, Aline Kissimoto, Felipe Oliveira de Aguiar, Victor Otani, Ricardo Riyoiti Uchida, and Lucas Murrins Marques. 2025. Beyond words: understanding anxiety and depression in college applicants through liwc analysis of textual features. *Frontiers in Psychology*, Volume 16 - 2025.

Carla Agurto, Pat Pataranutaporn, Elif K. Eyigoz, Gustavo Stolovitzky, and Guillermo Cecchi. 2018. Predictive linguistic markers of suicidality in poets. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 282–285.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical psychological science*, 6(4):529–542.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

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*, 9:324–333.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*. Guilford Press, New York.

Johan Bollen, Marijn ten Thij, Fritz Breithaupt, Alexander T. J. Barron, Lauren A. Rutter, Lorenzo Lorenzo-Luaces, and Marten Scheffer. 2021. Historical language records reveal a surge of cognitive distortions in recent decades. *Proceedings of the National Academy of Sciences*, 118(30).

[3]https://github.com/nehasharma666/Psycholinguistic-Profiles-of-Cognitive-Distortions-in-Reddit-Data

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(1-47):6.

David D. Burns. 1980. *Feeling Good: The New Mood Therapy*. Signet, New York.

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.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):579–582.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Matthew F. Dobbs, Alessia McGowan, Alexandria Selloni, Zarina Bilgrami, Cansu Sarac, Matthew Cotter, Shaynna N. Herrera, Guillermo A. Cecchi, Marianne Goodman, Cheryl M. Corcoran, and Agrima Srivastava. 2023. Linguistic correlates of suicidal ideation in youth at clinical high-risk for psychosis. *Schizophrenia Research*, 259:20–27. Language and Speech Analysis in Schizophrenia and Related Psychoses.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Rachel Heyard. 2026. Introduction to regression methods for public health using r. *The American Statistician*, 80(1):188–190.

Stephanie Homan, Marion Gabi, Nina Klee, Sandro Bachmann, Ann-Marie Moser, Martina Duri', Sofia Michel, Anna-Marie Bertram, Anke Maatz, Guido Seiler, Elisabeth Stark, and Birgit Kleim. 2022. Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical Psychology Review*, 95:102161.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Jutta Joormann and Colin H Stanton. 2016. Examining emotion regulation in depression: A review and future directions. *Behaviour research and therapy*, 86:35–49.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Clara Khuon, Gabriel Tillman, George Van Doorn, Jacob Dye, Kimberley A. McFarlane, Bridianne O'Dea, and Taylor A. Braund. 2026. Identifying structural linguistic markers of depression in written text: A narrative review of language analysis methods. *Journal of Affective Disorders Reports*, 23:101022.

Christopher Lalk, Lauren A. Rutter, Lorenzo Lorenzo-Luaces, and Johan Bollen. 2024. Depression symptoms are associated with frequency of cognitive distortions in psychotherapy transcripts. *Cognitive Therapy and Research*.

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.

Kevin Lybarger, Justin Tauscher, Xiruo Ding, Dror Benzeev, and Trevor Cohen. 2022. Identifying distorted thinking in patient-therapist text message exchanges by leveraging dynamic multi-turn context. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136, Seattle, USA. Association for Computational Linguistics.

Mai Mostafa, Alia El Bolock, and Slim Abdennadher. 2021. Automatic detection and classification of cognitive distortions in journaling text. In *WEBIST*, pages 444–452.

J. A. Nelder and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nawal Ouhmad, Romain Deperrois, Wissam El Hage, and Nicolas Combalbert. 2024. Cognitive distortions, anxiety, and depression in individuals suffering from ptsd. *International Journal of Mental Health*, 53(4):336–352.

Robin Quillivic, Yann Auxéméry, Frédérique Gayraud, Jacques Dayan, and Salma Mesmoudi. 2025. Linguistic markers for identifying post-traumatic stress disorder and associated symptoms: a systematic literature review. *Journal of the American Medical Informatics Association*, 32(8):1350–1363.

Lina Maria Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. *CoRR*, abs/1809.00640.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion - COGNITION EMOTION*, 18:1121–1133.

Archie Sage, Jeroen Keppens, and Helen Yannakoudakis. 2025. A survey of cognitive distortion detection and classification in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14884–14899, Suzhou, China. Association for Computational Linguistics.

Neha Sharma, Navneet Agarwal, and Kairit Sirts. 2026. Towards consistent detection of cognitive distortions: Llm-based annotation and dataset-agnostic evaluation. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10866–10882, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2019. Automatic detection and classification of cognitive distortions in mental health text. *CoRR*, abs/1909.07502.

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.

Taetem Simms, Clayton Ramstedt, Megan Rich, Michael Richards, Tony R. Martinez, and Christophe G. Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 508–512.

Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To'Meisha S Edwards, James W Pennebaker, and Matthias R Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*, 116(5):817.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Katja Wiemer-Hastings, Adrian S Janit, Peter M Wiemer-Hastings, Steve Cromer, and Jennifer Kinser. 2004. Automatic classification of dysfunctional thoughts: a feasibility test. *Behavior Research Methods, Instruments, & Computers*, 36(2):203–212.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *CoRR*, abs/1709.01848.

Appendix

A Dataset

We use RSDD dataset (Yates et al., 2017) for two reasons. First, it is one of the most widely used datasets in mental health NLP, and was available to us under a data use agreement. Second, its post-level timestamps make it suitable for longitudinal study, where tracking distortion patterns over time is of future interest to us.

The RSDD dataset consists of Reddit posts from users assigned to a depression cohort, identified through high-precision diagnosis patterns (e.g., "I was diagnosed with depression") with crowd-sourced verification, and a matched control cohort. Controls were selected from users with no mental-health-related posts and matched to the diagnosed cohort based on subreddit-usage similarity. All mental-health-related posts were removed from the diagnosed cohort during dataset construction. We refer readers to Yates et al. (2017) for full details of cohort construction.

The raw training set contains ∼40K users and ∼38.15M posts, with strong class imbalance between cohorts. We applied the following cleaning steps:

1. removed posts with empty user labels or empty text,
2. detected language using fastText (Joulin et al., 2017) and retained English-only posts,
3. performed standard text cleaning including removal of URLs and non-linguistic characters,
4. discarded texts with fewer than 3 tokens, as shorter texts lack sufficient linguistic content for meaningful feature extraction and are more likely to represent noise than genuine linguistic expression.
5. To address user-level class imbalance, we randomly downsampled control users to match the size of the depression cohort.

Cleaned corpus contains ∼4.91M posts for 6,138 users. Full corpus statistics at each filtering stage are reported in §3.1, Table 1.

B CD Annotations

This appendix provides further details on the CD labeling pipeline introduced in §3.2.

Annotation Sources: Both Shreevastava and Foltz (2021) and Sharma et al. (2026) use the publicly available Therapist Q&A dataset[4]. The dataset consists of 2,530 user/therapist pairs, with CD annotations applied to the user inputs only. Two independent annotation sets exist for this data: (1) **human annotations** from Shreevastava and Foltz (2021), with a reported inter-annotator agreement of 33.7%, and (2) **LLM-based annotations** from Sharma et al. (2026), obtained via five repeated GPT-4 passes (temperature 0.5), selecting the most recurrent label across runs, achieving a Fleiss' kappa of 0.78. Both annotation sets follow the same label schema: one dominant CD and one secondary CD, where applicable (See Table 2 for CD classes).

Classifier Training: To transfer CD labels to the RSDD corpus, we train two separate multi-label mentalRoBERTa-based classifiers (Ji et al., 2022), one on each annotation set. We follow the same experimental setup as Sharma et al. (2026), using their data splits and default hyperparameters, and reproduce comparable classification performance on their test set (refer to original study for more details and classification results). Both classifiers are then retrained on the full Therapist Q&A dataset and applied to the RSDD corpus, producing two independent sets of predicted CD labels per post.

Agreement-Based Filtering: We adopt a conservative agreement-based filtering strategy: we retain labels where the two models agree completely (identical label sets) or partially (at least one shared CD label), and discard posts with entirely disjoint predictions. For example, if one classifier predicts *Labeling, Personalization* and the other predicts *Labeling, Overgeneralization*, the post is retained with the single shared label *Labeling*. The rationale mirrors the reliability principle underlying multi-pass LLM annotation (Sharma et al., 2026): labels that survive both independently trained models are more likely grounded in consistent textual signals rather than model-specific noise. Table 5 reports the distribution of the resulting CD labels used in this study, overall and by cohort. Since posts may carry multiple CD labels, percentages reflect post-based prevalence and do not necessarily sum to 100%.

[4] https://www.kaggle.com/datasets/arnmaud/therapist-qa/data

Cognitive Distortion	Overall	Control	Depression
Overgeneralization	7.02	5.94	7.54
Labeling	2.04	1.16	2.46
Personalization	1.78	1.17	2.07
All-or-Nothing	0.95	0.51	1.16
Should Statements	0.83	0.52	0.97
Emotional Reasoning	0.48	0.22	0.60
Mind Reading	0.44	0.30	0.50
Magnification	0.41	0.17	0.53
Fortune Telling	0.29	0.22	0.33
Mental Filter	0.00	0.00	0.00
No Distortion	87.21	90.69	85.55

Table 5: CD label distribution (%). Percentages do not sum to 100% as posts may contain multiple CD labels. Mental Filter is excluded from our analysis due to its negligible presence in the RSDD predictions.

C LIWC

LIWC-22 computes feature scores by counting the proportion of words in a text that match each predefined dictionary category, normalizing by total word count to produce comparable scores across texts of varying length (Boyd et al., 2022). From the full set of 120+ categories, we selected features relevant to cognitive distortion profiling across four dimensions: affective processes (e.g., emotional tone, positive and negative emotion, anxiety, anger, etc.), cognitive processes (e.g., analytical thinking, causation, insight, certainty, etc.), temporal orientation (past, present, and future focus), and social and interpersonal language (e.g., social references, affiliation, prosocial behavior, first person pronoun etc.). We excluded categories that were hierarchically redundant, where a broad category and its subcategories would introduce collinearity, as well as highly specific categories with limited relevance to our research questions. The full list of selected features with descriptions is provided in Table 6.

C.1 Descriptive Statistics of LIWC Features

Table 7 reports descriptive statistics for the LIWC features used in this study, computed over the final RSDD corpus (n $\approx$ 4.73M posts). For each feature, we report the mean, median, standard deviation, minimum, maximum, and the percentage of posts with a zero value (% Zero) and non-zero value (% Non-zero).

Several categories exhibit a high proportion of zero values, with the modal post containing no words from the corresponding LIWC dictionary. For example, *moral* (91.5% zero), *death* (95.7% zero), and *emo_anx* (96.7% zero) appear in fewer than 10% of posts each. We note that a zero value here is not a missing or invalid observation; it is a valid measurement indicating that the post simply does not contain any words from that category. Such sparsity is expected for content-specific LIWC categories, particularly in short informal text such as Reddit posts.

Given the corpus size, even sparse categories contain substantial absolute numbers of non-zero observations (e.g., *moral* appears in $\sim$404K posts, *emo_anx* in $\sim$154K), providing sufficient observations for stable coefficient estimation under bootstrap resampling. Effects involving such categories should be interpreted as reflecting differences in feature *prevalence* across CD and ND posts rather than continuous variation in usage intensity. In contrast, features with broader non-zero variation (e.g., *socrefs*, *i*, *focuspresent*, *auxverb*) support a continuous-gradient interpretation.

D Cognitive Distortion Linguistic Profiles (Core and Global Markers)

Table 8 shows the Core and Global markers for each CD, which constructs their Linguistic profiles. Core markers are the most CD-specific features, while global-like markers capture features that are relevant for the CD but align more closely with broader distorted-language patterns. Features within each cell are listed in descending order of profile strength, based on the median coefficient ($\tilde{\beta}$). The sign indicates the direction of the effect; when two signs are shown (e.g., $(-, +)$ or $(+, -)$), the first corresponds to the TASK(XCD/ND) and the second to the TASK(XCD/$\neg$XCDS).

E GLM Feature Effects for Tasks

Table 9 and Table 10 report the GLM feature effects used in the CD profiling and clustering analyses. The tables present the union of linguistic markers identified during the CD profiling stage. Specifically, the top 15 features from each CD were combined, resulting in a set of 44 unique features. For each feature, the tables report the median coefficient ($\tilde{\beta}$) obtained from the GLM models along with the corresponding Relative Standard Error (RSE) stability indicators. Results are shown separately for TASK(XCD/ND) in Table 9 and TASK(XCD/$\neg$XCDS) in Table 10.

Category	Description	Category	Description
Summary Variables		**Social Processes**	
Word count	Total word count	Prosocial behavior	care, help, thank, please
Analytical thinking	Metric of logical, formal thinking	Politeness	thank, please, thanks, good morning
Clout	Language of leadership, status	Interpersonal conflict	fight, kill, killed, attack
Authentic	Perceived honesty, genuineness	Moralization	wrong, honor*, deserv*, judge
Emotional tone	Degree of positive (negative) tone	Social referents	you, we, he, she
Linguistic Dimensions		**Expanded Dictionary**	
1st person singular	I, me, my, myself	Illness	hospital*, cancer*, sick, pain
1st person plural	we, our, us, lets	Wellness	healthy, gym*, supported, diet
2nd person	you, your, u, yourself	Mental health	mental health, depressed, suicid*
3rd person plural	they, their, them, themsel*	Death	death*, dead, die, kill
Auxiliary verbs	is, was, be, have	**States**	
Adverbs	so, just, about, there	Need	have to, need, had to, must
Conjunctions	and, but, so, as	Want	want, hope, wanted, wish
Negations	not, no, never, nothing	Acquire	get, got, take, getting
		Lack	don't have, didn't have, *less, hungry
		Fulfilled	enough, full, complete, extra
		Fatigue	tired, bored, don't care, boring
Psychological Processes		**Motives**	
Affiliation	we, our, us, help	Reward	opportun*, win, gain*, benefit*
Achievement	work, better, best, working	Risk	secur*, protect*, pain, risk*
Power	own, order, allow, power	Curiosity	scien*, look* for, research*, wonder
All-or-none	all, no, never, always	Allure	have, like, out, know
Insight	know, how, think, feel	**Perception**	
Causation	how, because, make, why	Attention	look, look* for, watch, check
Discrepancy	would, can, want, could	Feeling	feel, hard, cool, felt
Tentative	if, or, any, something	**Time Orientation**	
Certitude	really, actually, of course, real	Past focus	was, had, were, been
Differentiation	but, not, if, or	Present focus	is, are, I'm, can
Positive emotion	good, love, happy, hope	Future focus	will, going to, have to, may
Negative emotion	bad, hate, hurt, tired		
Anxiety	worry, fear, afraid, nervous		
Anger	hate, mad, angry, frustr*		
Sadness	:(, sad, disappoint*, cry		
Swear words	shit, fuckin*, fuck, damn		

Table 6: Selected LIWC-22 features used in this study. Example words illustrate dictionary entries; * denotes wildcard matching.

Feature	Mean	Median	SD	Min	Max	% Zero	% Non-zero
WC	41.03	23.00	61.58	3.0	5934.00	0.00	100.00
Analytic	38.30	30.52	32.92	1.0	99.00	0.00	99.12
Clout	39.33	25.50	37.98	1.0	99.00	0.00	91.70
Authentic	55.45	63.35	37.00	1.0	99.00	0.00	93.05
Tone	59.75	74.09	39.20	1.0	99.00	0.00	69.30
i	5.09	3.45	5.91	0.0	75.00	38.24	61.76
we	0.43	0.00	1.85	0.0	66.67	89.53	10.47
you	2.59	0.00	4.86	0.0	75.00	61.84	38.16
they	0.91	0.00	2.53	0.0	66.67	78.46	21.54
auxverb	10.22	10.00	7.02	0.0	100.00	14.52	85.48
adverb	6.68	5.88	6.55	0.0	100.00	26.37	73.63
conj	6.00	5.95	5.32	0.0	100.00	29.09	70.91
negate	2.07	0.00	3.76	0.0	80.00	56.89	43.11
affiliation	1.32	0.00	3.42	0.0	100.00	73.75	26.25
achieve	1.07	0.00	2.74	0.0	100.00	73.15	26.85
power	0.95	0.00	2.69	0.0	100.00	77.26	22.74
allnone	1.36	0.00	3.37	0.0	100.00	69.81	30.19
insight	2.57	0.00	4.21	0.0	100.00	53.34	46.66
cause	1.58	0.00	3.15	0.0	100.00	63.82	36.18
discrep	2.12	0.00	3.70	0.0	75.00	57.08	42.92
tentat	3.05	0.96	4.56	0.0	100.00	49.36	50.64
certitude	0.99	0.00	2.96	0.0	100.00	76.69	23.31
differ	3.96	2.99	4.83	0.0	100.00	40.92	59.08
emo_pos	1.79	0.00	4.79	0.0	100.00	71.60	28.40
emo_neg	0.58	0.00	2.20	0.0	100.00	84.70	15.30
emo_anx	0.10	0.00	0.91	0.0	75.00	96.74	3.26
emo_anger	0.13	0.00	0.98	0.0	100.00	95.74	4.26
emo_sad	0.10	0.00	0.94	0.0	66.67	97.21	2.79
swear	0.53	0.00	2.57	0.0	100.00	90.24	9.76
prosocial	0.98	0.00	3.55	0.0	100.00	81.83	18.17
polite	0.69	0.00	3.34	0.0	100.00	90.24	9.76
conflict	0.29	0.00	1.49	0.0	100.00	91.53	8.47
moral	0.29	0.00	1.54	0.0	100.00	91.45	8.55
socrefs	7.24	5.88	7.55	0.0	100.00	29.30	70.70
illness	0.10	0.00	0.92	0.0	66.67	97.17	2.83
wellness	0.05	0.00	0.62	0.0	100.00	98.33	1.67
mental	0.04	0.00	0.55	0.0	66.67	98.92	1.08
death	0.16	0.00	1.15	0.0	66.67	95.70	4.30
need	0.42	0.00	1.80	0.0	75.00	88.80	11.20
want	0.38	0.00	1.64	0.0	66.67	88.49	11.51
acquire	0.96	0.00	2.43	0.0	66.67	74.02	25.98
lack	0.14	0.00	1.16	0.0	80.00	96.27	3.73
fulfill	0.16	0.00	1.01	0.0	66.67	93.81	6.19
fatigue	0.04	0.00	0.61	0.0	100.00	98.70	1.30
reward	0.13	0.00	0.96	0.0	66.67	95.53	4.47
risk	0.24	0.00	1.30	0.0	66.67	91.88	8.12
curiosity	0.37	0.00	1.85	0.0	100.00	90.56	9.44
allure	7.71	6.67	7.28	0.0	100.00	23.38	76.62
attention	0.54	0.00	2.13	0.0	100.00	86.28	13.72
feeling	0.46	0.00	1.90	0.0	100.00	86.82	13.18
focuspast	3.73	0.76	5.47	0.0	100.00	49.52	50.48
focuspresent	6.31	5.45	6.28	0.0	80.00	27.84	72.16
focusfuture	1.36	0.00	3.40	0.0	100.00	72.22	27.78

Table 7: Descriptive statistics of LIWC features on the final RSDD corpus (n $\approx$ 4.73M posts). Values are computed on raw LIWC-22 output, where each score represents the percentage of words in a post matching the corresponding dictionary category, except for summary variables (WC, Analytic, Clout, Authentic, Tone) which follow LIWC's own scaling. % Zero indicates the percentage of posts with a value of zero for the corresponding feature; % Non-zero is its complement. SD = Standard Deviation, Min= Minimum, Max = Maximum

CD	Core markers	Global markers
AoN	emo_anger(+), socrefs(-), focuspast(-), Authentic(+), fatigue(+), differ(-), you(+), death(+), wellness(+), allure(+), swear(+), Analytic(-), emo_anx(-), achieve(+), want(+), tentat(-), emo_pos(-,+), need(+), focuspresent(+)	i(+), mental(+), insight(-), Tone(-), moral(+,-), allnone(+), emo_neg(+), they(-), polite(-), cause(-), conflict(-), risk(-), WC(+), negate(+), focusfuture(-)
ER	emo_anx(+), socrefs(-), i(+), discrep(+), Authentic(+), feeling(+), you(+), death(+), emo_pos(-), focusfuture(+), want(+), swear(+,-), mental(+), differ(-), polite(-), focuspresent(+), Tone(-), illness(+), allure(+), focuspast(-), need(+), auxverb(-), risk(+), emo_neg(+,-), negate(+,-), power(+), adverb(-)	moral(-), affiliation(+), WC(+), cause(+), allnone(+,-), Analytic(-), conflict(-), achieve(+), wellness(-)
FT	emo_anx(+), focusfuture(+), swear(+,-), death(+), need(-), focuspast(-), emo_pos(-), risk(+), tentat(+), Tone(-), differ(-), adverb(-), allnone(+,-), you(-), Analytic(-), focuspresent(-), Authentic(+), discrep(+), Clout(-), auxverb(+), want(-), conflict(-), insight(+,-)	moral(-), emo_neg(-), negate(-), polite(-), socrefs(-), affiliation(+), WC(+), feeling(+), achieve(+,-), cause(+), power(+), i(-), prosocial(-), emo_anger(-), fatigue(-), wellness(-)
La	swear(+), moral(+), i(+), you(+), Tone(-), allnone(+,-), mental(+), Authentic(-), emo_neg(+,-), focuspast(-), negate(+,-), discrep(-), they(-), focusfuture(-), risk(-), death(-), adverb(-), feeling(-), affiliation(-), emo_sad(-), socrefs(+)	Analytic(+), emo_anger(+), emo_anx(-), achieve(+,-), focuspresent(+), emo_pos(-,+), conflict(-), insight(+), WC(+), prosocial(-), allure(-)
Mag	death(+), socrefs(-), emo_anx(+), Tone(-), illness(+), negate(-), Authentic(+), mental(+), focusfuture(+), discrep(+), differ(-), focuspast(-), insight(-), you(+), swear(+,-), adverb(-), conflict(-), allnone(+,-), auxverb(-), feeling(+), risk(+), Analytic(-), WC(+), emo_neg(+,-)	moral(-), i(+), achieve(-), emo_pos(-), focuspresent(+), cause(+), power(+), polite(-), emo_anger(-), prosocial(-), want(-), fatigue(-), Clout(+), allure(-)
Over	swear(+,-), emo_anx(-), moral(+,-), i(-), allnone(+), you(-), Tone(-,+), negate(+), focuspast(-,+), feeling(-), affiliation(-,+), conflict(-,+)	focuspresent(-), differ(-), wellness(+), achieve(+), discrep(-), socrefs(+), emo_neg(+), insight(+), WC(+,-), power(+), Analytic(-), emo_pos(-), polite(-), focusfuture(-), Authentic(+), death(-), want(-), mental(-), adverb(+), they(+)
Per	moral(+), i(+), emo_neg(+), Tone(-), WC(+), emo_anx(-), allnone(+,-), Authentic(-), cause(+), prosocial(+), emo_sad(-), death(-), feeling(+), you(+), illness(-), achieve(+,-), socrefs(+), polite(-,+)	focuspast(+), focuspresent(+), negate(+,-), risk(-), affiliation(-,+), Analytic(-), discrep(-), emo_pos(-), insight(+), power(+), swear(+), focusfuture(-), emo_anger(-), differ(+), need(+), fatigue(-), auxverb(-), they(+), wellness(-)
SS	i(+), auxverb(+), WC(+), Tone(-,+), need(+), focuspast(-), focuspresent(-), discrep(+), prosocial(+), socrefs(+), allure(+), death(-), focusfuture(-), moral(+), affiliation(-), swear(+,-), allnone(+,-), emo_neg(+,-), risk(-), tentat(-), feeling(+), want(+), they(-), negate(+,-), differ(+), illness(-)	achieve(+), conflict(+), power(+), polite(-), Analytic(-), insight(+), emo_pos(-), cause(+), mental(-), adverb(-)
MR	socrefs(+), insight(+), swear(+,-), you(-), Authentic(-), focusfuture(+), i(+), tentat(+), affiliation(-), Analytic(-), emo_anx(+), moral(+,-), emo_anger(+), fatigue(+), focuspresent(+), death(-), emo_pos(-), Clout(-), want(+), need(-), discrep(-), auxverb(-), they(+), negate(+,-)	allnone(-), emo_neg(-), polite(-), prosocial(+), achieve(-), conflict(+), power(-), WC(+,-), illness(-), focuspast(-,+), Tone(-), allure(-)

Table 8: Linguistic profiles of the nine CD types. Core markers are the most CD-specific features, while Global markers capture features that are relevant for the CD but align more closely with broader distorted-language patterns or weaker CD-specific effects. Features within each cell are ordered by descending profile strength. The sign indicates the direction of the effect; when two signs are shown, the first corresponds to the TASK(XCD/ND) and the second to the TASK(XCD/¬XCDS).

Feature	AoN $\tilde{\beta}$	AoN RSE	ER $\tilde{\beta}$	ER RSE	FT $\tilde{\beta}$	FT RSE	La $\tilde{\beta}$	La RSE	Mag $\tilde{\beta}$	Mag RSE	MR $\tilde{\beta}$	MR RSE	Over $\tilde{\beta}$	Over RSE	Per $\tilde{\beta}$	Per RSE	SS $\tilde{\beta}$	SS RSE
Tone	−0.73	● ● ●	−0.81	● ● ●	−0.86	● ● ●	−0.83	● ● ●	−1.17	● ● ●	−0.53	● ● ●	−0.50	● ● ●	−0.96	● ● ●	−0.26	●●
socrefs	−0.42	●●	−0.50	●●	—	—	+0.32	●●	−0.64	●●	+1.32	● ● ●	+0.22	●	+0.32	●●	+0.46	●●
emo_anx	−0.15	●	+1.10	●●	+1.76	● ● ●	−0.10	●	+0.68	●●	+0.26	●	−0.19	●●	−0.26	●●	—	—
i	+0.16	●	+0.73	● ● ●	—	—	+0.39	●●	+0.28	●●	+0.39	●●	−0.15	●	+0.46	●●	+0.52	● ● ●
Authentic	+0.24	●●	+0.42	● ● ●	+0.20	●●	−0.19	●●	+0.40	●●	−0.39	●●	—	—	−0.22	●●	—	—
moral	+0.32	●●	—	—	—	—	+0.96	●●	—	—	+0.18	●●	+0.28	●●	+1.09	● ● ●	+0.63	● ● ●
focuspast	−0.47	●●	−0.34	●●	−0.50	● ● ●	−0.36	●●	−0.43	● ● ●	−0.09	●	−0.09	●	—	—	−0.51	● ● ●
swear	+0.82	● ● ●	+0.35	●●	+0.11	●	+1.77	● ● ●	+0.43	●●	+0.13	●	+0.43	●●	+0.65	● ● ●	+0.47	● ● ●
you	+0.22	●	+0.38	●●	−0.22	●●	+0.26	●●	+0.23	●●	−0.42	●●	−0.13	●	+0.14	●	—	—
emo_neg	+0.44	●●	+0.21	●	—	—	+0.15	●	+0.30	●	—	—	+0.32	●●	+0.72	●●	+0.20	●
WC	+0.42	●●	+0.28	●●	+0.55	●●	+0.43	●●	+0.47	●●	+0.33	●●	+0.43	●●	+0.70	●●	+0.80	● ● ●
death	+0.21	●●	+0.36	●●	+0.39	●●	−0.11	●	+1.12	● ● ●	−0.18	●●	—	—	−0.16	●	−0.19	●●
allnone	+0.52	● ● ●	+0.29	●●	+0.18	●●	+0.20	●●	+0.23	●●	—	—	+0.55	● ● ●	+0.17	●●	+0.25	●●
focusfuture	—	—	+0.33	●●	+1.52	● ● ●	−0.14	●	+0.34	●●	+0.39	● ● ●	—	—	—	—	−0.19	●●
focuspresent	+0.19	●	+0.33	●●	−0.10	●	+0.17	●●	+0.17	●●	+0.29	●●	—	—	+0.25	●●	−0.18	●●
Analytic	−0.31	●●	−0.25	●●	−0.38	●●	—	—	−0.24	●●	−0.43	●●	−0.16	●	−0.19	●	−0.10	●
polite	−0.27	●	−0.42	●●	−0.47	●	—	—	−0.18	●	−0.45	●●	−0.18	●	−0.11	●	−0.25	●
negate	+0.29	●●	+0.17	●	—	—	+0.11	●	−0.16	●	+0.23	●●	+0.39	●●	+0.16	●	+0.19	●●
discrep	—	—	+0.36	●●	+0.11	●	−0.24	●●	+0.24	●●	−0.22	●●	−0.11	●	−0.11	●	+0.18	●●
affiliation	—	—	—	—	—	—	−0.25	●●	—	—	−0.46	●●	−0.11	●	−0.12	●	−0.34	●●
emo_pos	−0.08	●	−0.53	●●	−0.47	●●	−0.12	●	−0.26	●●	−0.35	●●	−0.18	●●	−0.15	●	−0.21	●●
insight	—	—	—	—	+0.08	●	+0.16	●●	−0.09	●	+0.96	● ● ●	+0.14	●●	+0.13	●	+0.19	●●
differ	−0.22	●●	−0.26	●●	−0.25	●●	—	—	−0.26	●●	—	—	−0.08	●	—	—	+0.09	●
feeling	—	—	+0.42	●●	+0.09	●	−0.09	●	+0.10	●	—	—	−0.07	●	+0.15	●	+0.12	●●
achieve	+0.31	●●	+0.18	●●	+0.10	●	+0.08	●	—	—	—	—	+0.21	●●	+0.08	●	+0.31	●●
need	+0.09	●	+0.16	●●	−0.38	●●	—	—	—	—	−0.14	●●	—	—	—	—	+0.33	●●
tentat	−0.10	●	—	—	+0.25	●●	—	—	—	—	+0.37	●●	—	—	—	—	−0.12	●
prosocial	—	—	—	—	—	—	—	—	—	—	+0.20	●	—	—	+0.20	●●	+0.24	●●
mental	+0.16	●	+0.27	●	—	—	+0.20	●	+0.37	●●	—	—	—	—	—	—	—	—
emo_anger	+0.70	●●	—	—	—	—	+0.13	●	—	—	+0.20	●	—	—	—	—	—	—
illness	—	—	+0.18	●●	—	—	—	—	+0.55	●●	−0.09	●	—	—	−0.12	●	−0.08	●
auxverb	—	—	−0.14	●	+0.14	●	—	—	−0.15	●	−0.11	●	—	—	—	—	+0.50	● ● ●
risk	−0.07	●	+0.12	●	+0.27	●●	−0.12	●	+0.09	●	—	—	—	—	−0.09	●	−0.13	●●
fatigue	+0.28	●	—	—	—	—	—	—	—	—	+0.19	●	—	—	—	—	—	—
conflict	−0.20	●	−0.15	●	−0.22	●●	−0.17	●	−0.31	●●	—	—	−0.11	●	—	—	—	—
want	+0.11	●	+0.31	●●	−0.10	●	—	—	—	—	+0.15	●	—	—	—	—	+0.12	●
cause	—	—	+0.20	●●	+0.11	●	—	—	+0.11	●●	—	—	—	—	+0.29	●●	+0.11	●●
adverb	—	—	−0.09	●	−0.25	●●	−0.10	●	−0.19	●●	—	—	—	—	—	—	—	—
Clout	—	—	—	—	−0.15	●	—	—	—	—	−0.17	●	—	—	—	—	—	—
power	—	—	+0.23	●●	+0.12	●	—	—	+0.09	●	—	—	+0.12	●	+0.10	●	+0.23	●●
they	−0.09	●	—	—	—	—	−0.15	●●	—	—	+0.09	●	—	—	—	—	−0.10	●
allure	+0.18	●●	+0.17	●●	—	—	—	—	—	—	—	—	—	—	—	—	+0.20	●●
emo_sad	—	—	—	—	—	—	−0.09	●	—	—	—	—	—	—	−0.16	●	—	—
wellness	+0.20	●	—	—	—	—	—	—	—	—	—	—	+0.07	●	—	—	—	—

Table 9: $\tilde{\beta}$ coefficients and Relative Standard Error (RSE) stability per feature and CD for TASK(XCD/ND). '—' indicates non-significant result. $\widetilde{\text{OR}}$ can be calculated as $\exp(\tilde{\beta})$. %Δ Odds can be computed as $(\widetilde{\text{OR}} - 1) \times 100$. RSE encode stability categories: ● ● ● = extremely stable (RSE ≤ 0.10), ●● = stable (0.10–0.25], ● = moderately stable (0.25–0.50].

Feature	AoN $\tilde{\beta}$	AoN RSE	ER $\tilde{\beta}$	ER RSE	FT $\tilde{\beta}$	FT RSE	La $\tilde{\beta}$	La RSE	Mag $\tilde{\beta}$	Mag RSE	MR $\tilde{\beta}$	MR RSE	Over $\tilde{\beta}$	Over RSE	Per $\tilde{\beta}$	Per RSE	SS $\tilde{\beta}$	SS RSE
Tone	−0.09	●	−0.14	●●	−0.31	●●	−0.26	●●	−0.40	●●●	—	—	+0.24	●●	−0.38	●●●	+0.31	●●
socrefs	−0.70	●●●	−0.85	●●●	−0.17	●	+0.19	●	−0.95	●●●	+1.21	●●●	—	—	+0.11	●	+0.22	●●
emo_anx	−0.17	●●	+1.21	●●●	+1.56	●●●	—	—	+0.80	●●	+0.38	●●	−0.46	●●	−0.31	●●	—	—
i	—	—	+0.46	●●	−0.24	●●	+0.29	●●	—	—	+0.29	●●	−0.59	●●●	+0.31	●●	+0.36	●●
Authentic	+0.26	●●	+0.37	●●	+0.21	●●	−0.21	●●	+0.42	●●●	−0.44	●●	+0.15	●	−0.23	●●	—	—
moral	−0.07	●	−0.60	●●	−0.52	●●	+0.32	●●	−0.49	●●	−0.32	●●	−0.22	●●	+0.35	●●	+0.07	●
focuspast	−0.28	●●	−0.16	●●	−0.31	●●	−0.20	●●	−0.24	●●	+0.10	●	+0.22	●●	+0.14	●●	−0.38	●●●
swear	+0.10	●	−0.14	●●	−0.56	●●	+0.74	●●●	−0.13	●●	−0.54	●●●	−0.31	●●	—	—	−0.07	●
you	+0.24	●●	+0.49	●●	−0.15	●	+0.30	●●	+0.27	●●	−0.36	●●	−0.28	●●	+0.15	●	—	—
emo_neg	—	—	−0.23	●	−0.54	●●	−0.18	●●	−0.13	●	−0.42	●●	—	—	+0.33	●●	−0.15	●
WC	—	—	—	—	+0.09	●●	—	—	+0.15	●●	−0.10	●	−0.09	●	+0.20	●●	+0.19	●●
death	+0.15	●●	+0.21	●●	+0.26	●●	−0.08	●	+1.17	●●●	−0.21	●●	−0.14	●	−0.14	●●	−0.21	●●
allnone	+0.07	●	−0.10	●	−0.25	●●	−0.13	●●	−0.15	●●	−0.37	●●	+0.24	●●	−0.17	●●	−0.13	●●
focusfuture	−0.06	●	+0.33	●●	+1.72	●●●	−0.19	●●	+0.30	●●	+0.39	●●●	−0.08	●	−0.09	●	−0.21	●●
focuspresent	+0.11	●	+0.21	●●	−0.29	●●	+0.08	●	—	—	+0.15	●	−0.12	●	+0.10	●	−0.34	●●●
Analytic	−0.19	●●	−0.09	●	−0.21	●●	+0.13	●	−0.12	●	−0.28	●●	—	—	—	—	—	—
polite	—	—	−0.08	●	—	—	—	—	—	—	—	—	—	—	+0.12	●	—	—
negate	—	—	−0.11	●	−0.40	●●	−0.15	●	−0.44	●●	−0.18	●●	+0.23	●●	−0.08	●	−0.10	●
discrep	—	—	+0.46	●●●	+0.23	●●	−0.19	●●	+0.33	●●	−0.15	●●	—	—	—	—	+0.33	●●
affiliation	—	—	+0.09	●	+0.23	●●	−0.14	●	—	—	−0.21	●●	+0.11	●	+0.07	●	−0.19	●●
emo_pos	+0.10	●	−0.19	●●	−0.11	●	+0.08	●	—	—	−0.08	●	−0.07	●	—	—	—	—
insight	−0.20	●●	—	—	−0.12	●	—	—	−0.25	●●	+0.84	●●●	—	—	—	—	—	—
differ	−0.18	●●	−0.21	●●	−0.23	●●	—	—	−0.25	●●	—	—	—	—	+0.11	●	+0.12	●●
feeling	—	—	+0.43	●●	—	—	−0.10	●	+0.09	●	—	—	−0.18	●●	+0.21	●●	+0.11	●
achieve	+0.13	●●	—	—	−0.08	●	−0.08	●	−0.21	●●	−0.15	●●	+0.09	●	−0.12	●●	+0.08	●
need	+0.05	●	+0.12	●	−0.41	●●	—	—	—	—	−0.18	●●	—	—	+0.05	●	+0.29	●●
tentat	−0.10	●●	—	—	+0.26	●●	—	—	—	—	+0.41	●●●	—	—	—	—	−0.10	●
prosocial	—	—	—	—	−0.08	●	−0.09	●	−0.12	●	—	—	—	—	+0.16	●●	+0.14	●
mental	—	—	+0.16	●	—	—	+0.11	●	+0.21	●	—	—	−0.10	●	—	—	−0.14	●
emo_anger	+0.50	●●	—	—	−0.17	●	—	—	−0.09	●	+0.16	●	—	—	−0.20	●●	—	—
illness	—	—	+0.15	●	—	—	—	—	+0.57	●●	—	—	—	—	−0.15	●●	−0.09	●
auxverb	—	—	−0.17	●●	+0.16	●	—	—	−0.26	●●	−0.14	●	—	—	−0.13	●	+0.46	●●●
risk	—	—	+0.19	●●	+0.31	●●	−0.07	●	+0.16	●●	—	—	—	—	—	—	−0.09	●
fatigue	+0.19	●	—	—	−0.12	●	—	—	−0.13	●	+0.12	●	—	—	−0.18	●	—	—
conflict	−0.08	●	—	—	−0.11	●	—	—	−0.13	●	+0.10	●	+0.10	●	—	—	+0.06	●
want	+0.08	●	+0.23	●●	−0.16	●	—	—	−0.13	●	+0.15	●	−0.11	●	—	—	+0.08	●
cause	−0.07	●	+0.08	●	—	—	—	—	—	—	—	—	—	—	+0.23	●●	—	—
adverb	—	—	−0.07	●	−0.25	●●	−0.10	●	−0.17	●●	—	—	+0.09	●	—	—	−0.06	●
Clout	—	—	—	—	−0.15	●	—	—	+0.13	●	−0.29	●●	—	—	—	—	—	—
power	—	—	+0.11	●	—	—	—	—	—	—	−0.21	●●	—	—	—	—	+0.09	●
they	—	—	—	—	—	—	−0.18	●●	—	—	+0.07	●	+0.10	●	+0.07	●	−0.10	●
allure	+0.16	●●	+0.06	●	—	—	−0.11	●	−0.09	●	−0.09	●	—	—	—	—	+0.13	●●
emo_sad	—	—	—	—	—	—	−0.10	●	—	—	—	—	—	—	−0.11	●	—	—
wellness	+0.14	●	−0.09	●	−0.07	●	—	—	—	—	—	—	—	—	−0.08	●	—	—

Table 10: $\tilde{\beta}$ coefficients and Relative Standard Error (RSE) stability per feature and CD forTASK(XCD/¬XCDS). '—' indicates non-significant result. $\widetilde{\mathrm{OR}}$ can be calculated as $\exp(\tilde{\beta})$. %Δ Odds can be computed as $(\widetilde{\mathrm{OR}}-1)\times 100$. RSE encode stability categories: ●●● = extremely stable (RSE ≤ 0.10), ●● = stable (0.10–0.25], ● = moderately stable (0.25–0.50].