

What is Stigma Attributed to? A Theory-Grounded, Expert-Annotated Interview Corpus for Demystifying Mental-Health Stigma

Han Meng[♣], Yancan Chen[♣], Yunan Li[♣], Yitian Yang[♣],
Jungup Lee[◇], Renwen Zhang[♡], Yi-Chieh Lee^{♣*}

[♣]Department of Computer Science, National University of Singapore,

[◇]Department of Social Work, National University of Singapore,

[♡]Department of Communications and New Media, National University of Singapore

{han.meng, yancan, liyunan, yang.yitian}@u.nus.edu,

{swklj, r.zhang, yclee}@nus.edu.sg

Abstract

Warning: This paper contains content that may be offensive or disturbing, but this is unavoidable due to the nature of the work.

Mental-health stigma remains a pervasive social problem that hampers treatment-seeking and recovery. Existing resources for training neural models to finely classify such stigma are limited, relying primarily on social-media or synthetic data without theoretical underpinnings. To remedy this gap, we present an expert-annotated, theory-informed corpus of human-chatbot interviews, comprising 4,141 snippets from 684 participants with documented socio-cultural backgrounds. Our experiments benchmark state-of-the-art neural models and empirically unpack the challenges of stigma detection. This dataset can facilitate research on computationally detecting, neutralizing, and counteracting mental-health stigma. Our corpus is openly available at <https://github.com/HanMeng2004/Mental-Health-Stigma-Interview-Corpus>.

1 Introduction

Mental illnesses profoundly shape the lives of hundreds of millions globally (Kessler et al., 1994). Yet those experiencing mental-health challenges often face a devastating double burden – not just their symptoms, but also pervasive *stigma* that leads to social isolation, workplace discrimination, and delayed treatment-seeking (Érica de Toledo Piza Peluso and Blay, 2009). **Mental-health stigma**, as originally defined by Goffman, includes regarding mental illness as divergent from what society considers correct and normal, and mentally ill individuals as devalued and tainted (Goffman, 1964). Social scientists have traditionally often qualitatively collected and coded texts from interviews that provide insider views to understand mental-health stigma, yet the considerable time,

*Corresponding author

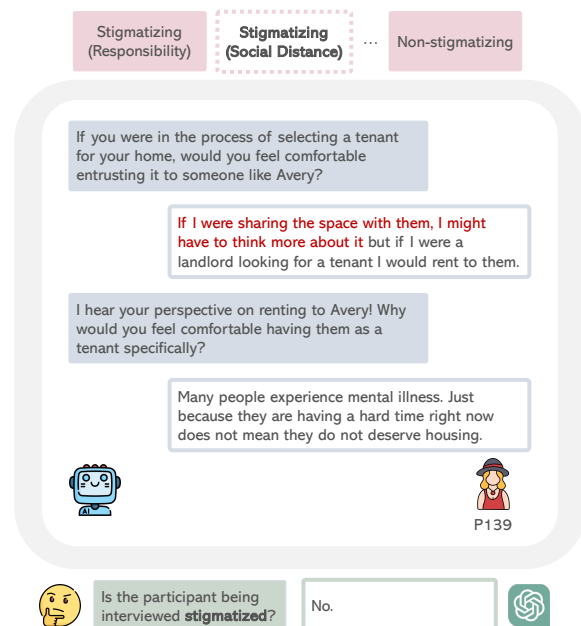


Figure 1: A sample interview snippet from our corpus where LLMs (here GPT-4o) fail to identify the underlying stigma, with the participant’s response categorized as *Stigmatizing (Social Distance)*. The stigmatizing text is highlighted in **red**.

manual effort, and expertise required (Leeson et al., 2019; Saldana, 2016) limit their breadth, fertility, and scalability.

Hence, the pressing need for a keystone dataset to train neural models capable of *automatically* unraveling and disentangling stigma (Giorgi et al., 2024a; Demszky et al., 2023) has grown increasingly apparent, especially driven by the proliferation of digital mental-health discourse (Garg et al., 2022) and therapeutic conversational agents (Mishra et al., 2023). However, unlike the growing body of research on general hate speech and offensive language targeting certain races and genders (Vidgen et al., 2021; ElSherief et al., 2021; Hartvigsen et al., 2022; Cho and Moon, 2021) (Table 1), publicly available corpus specifically designed to understand mental-health stigma (Choe,

2023) remain scarce in the NLP community. More worryingly, these limited resources suffer from two critical limitations that presumably leave even state-of-the-art large language models (LLMs) not well-positioned to detect stigma effectively (Figure 1).

First, current corpora are almost **exclusively sourced from social-media platforms** (Jilka et al., 2022; Straton et al., 2020), where discourse tends to be polarized and inflammatory (Moriceau et al., 2022a; Boyd and Crawford, 2012), or from **synthetic** data (Choe, 2023). Interview data analyzed by social scientists, on the other hand, are renowned for interpersonal, conversational context and rich human narratives, their capacity to elicit self-disclosure and deep reflection (Jenlink and Banathy, 2005; Ho et al., 2018), as well as being rife with psychological constructs, perceptions, and valuable information about how social problems are negotiated, unfolded, and perpetuated (Liggins and Hatcher, 2005; Lyons and Ziviani, 1995). Such a mismatch between computational and social-science data sources puts neural classifiers at risk of over-fitting to certain lexical and linguistic markers, such as swear words and Internet buzzwords (Harrigan et al., 2020).

Another potential drawback – particularly critical given that socio-cultural factors are deeply associated and intertwined with mental-health stigma (Krendl and Pescosolido, 2020) – is that, unlike interview participants whose demographic backgrounds can be systematically documented, the texts from social media and synthetic corpora largely come with little to no information about who created them and the social and cultural contexts of those creators (Ruths and Pfeffer, 2014).

Second, integration with **psychological and sociological theories** of mental-health stigma in existing datasets ranges from simplistic (Roesler et al., 2024) to nonexistent. Such social models and conceptual frameworks are essential for an empirical, fine-grained understanding of stigmatizing language (Bonikowski and Nelson, 2022; Hovy and Yang, 2021; Blodgett et al., 2020). Their absence has led to computational methods that are largely confined to binary classification (Jilka et al., 2022; Lee and Kyung, 2022) and are unlikely to capture the full spectrum of psycholinguistic nuances of stigmatization – from the casual appropriation of diagnostic terms (Moriceau et al., 2022b) to eu-

phemisms (Magu and Luo, 2018) and microaggressions (Breitfeller et al., 2019).

Motivated by these two critical gaps and informed by prior work on chatbots’ capacity to conduct interviews (Lee et al., 2023, 2020), we introduce a new corpus of interviews collected by chatbots and annotated using a protocol guided by the *attribution model* (Corrigan et al., 2003) – a widely-adopted theoretical framework for deconstructing mental-health stigma. This rich corpus adds value to the NLP community in three ways: 1) to the best of our knowledge, it represents the first large-scale, open-source mental-health stigma interview dataset, comprising 4,141 transcript snippets from 684 human participants; 2) it captures real-world interactions with clear socio-cultural representation and rich contextual information, thereby contributing fresh linguistic resources for understanding stigmatizing language; and 3) it features expert-led, theoretically-grounded labels that help equip neural models with the ability to identify fine-grained drivers of stigmatization. Finally, we fine-tune classifiers like RoBERTa (Liu et al., 2019b) and prompt LLMs such as Llama (Dubey et al., 2024) and GPT-4o for stigma detection, providing both useful benchmarks and empirical insights to sustain future research.

2 Background and Related Work

Problematic language and social bias have long been central concerns in NLP research (Schmidt and Wiegand, 2017; Röttger et al., 2021; Lee et al., 2024). From offensive and abusive language laced with profanity (Sanguinetti et al., 2018) to toxic and derogatory hate speech that disparages people based on their protected characteristics and social identities (Vargas et al., 2022), to outright hostile rhetoric and trolling (Lee et al., 2022) that fuels online harassment (Liu et al., 2019a), to seemingly mild but harmful comments that breed disrespect (Lu et al., 2023), researchers have examined how to classify and mitigate it (Korre et al., 2024).² Several influential benchmarks and language resources, summarized in Table 1, tremendously help detect and identify these harmful languages: for example, ElSherief et al. (2021) tackles white grievance, incitement to violence, expressions of perceived inferiority, and so on; Büyükdemirci et al. (2024) annotates hate-speech targets and signals at the to-

¹Substance use disorder is classified as a mental illness in DSM-5 (Association, 2013). We follow Corrigan et al. (2003) in using the term *mental-health stigma*.

²See Appendix A for a discussion of the definitions of these harmful language categories.

Language Resource	Source	Size	Annotation Scheme	Research Scope	Public	Theory-Grounded	Socio-cultural
ElSherief et al. (2021)	Twitter	22,584	Multi-Label	Hate Speech	✓	✓	
Büyükdemirci et al. (2024)	Twitter	1,530	Multi-Label	Hate Speech	✓		
Vidgen et al. (2021)	RoBERTa	41,255	Multi-Label	Hate Speech	✓		
Hartvigsen et al. (2022)	GPT-3	274,186	Binary	Hate Speech	✓		
Baheti et al. (2021)	Reddit	2,000	Binary	Offensive Language	✓		
Cignarella et al. (2024)	Facebook	2,888	Binary	Social Stereotypes	✓		
Breitfeller et al. (2019)	Tumblr	2,934	Multi-Label	Microaggression	✓	✓	
Straton et al. (2020)	Facebook	2,761	Multi-Label	Vaccine Stigma		✓	
Roesler et al. (2024)	Reddit	2,214	Multi-Label	Substance Use Stigma ¹		✓	
Choey (2023)	ChatGPT	9,700	Multi-Label	Mental-Health Stigma	✓		
MHSTIGMAINTERVIEW (ours)	Interview	4,141	Multi-Label	Mental-Health Stigma	✓	✓	✓

Table 1: Summary and comparison of our corpus with existing datasets on problematic language and mental-health stigma. *Annotation Scheme* indicates whether a dataset uses binary or multi-class labeling. *Theory-Grounded* shows if the annotations draw from sociological and/or psychological theories. *Socio-cultural* indicates whether a dataset records the socio-cultural background of its data contributors/sources. Ours is the first interview-based corpus for mental-health stigma.

ken level; Vidgen et al. (2021) examines dehumanizing language, threats, and displays of animosity; and Breitfeller et al. (2019) explores microaggressions against certain genders, religions, and ages in everyday discourse.

Yet, we consider that these carefully curated and compiled resources are not sufficient to automatically decipher mental-health stigma. One obstacle is that these corpora not only lack coverage of people with mental illness as a target population (Yu et al., 2024), but more fundamentally, mental-health stigma is uniquely operationalized through cognitive judgments of personal flaws that drive negative *emotions* and *behavioral* intentions (Corrigan, 2000) – requiring the capture of both linguistic markers and underlying attributional chains (Corrigan et al., 2003). Given this complexity, it is concerning that the sole open-source corpus in this space (Choey, 2023), shown in Table 1, presumably suffers from homogeneous **psycholinguistic patterns** (Patton et al., 2020; Boyd and Crawford, 2012), with limited awareness of socio-cultural factors (Krendl and Pescosolido, 2020; Ruths and Pfeffer, 2014) and potentially **questionable theoretical grounding** (Hovy and Yang, 2021).

To address this resource gap, and to further answer the recent calls for NLP scholars to invest in keystone datasets (Demszky et al., 2023), integrate psychological theories (Bonikowski and Nelson, 2022; Boyd and Schwartz, 2021), better understand human behavior through language (Mihalcea et al., 2024), and develop socially aware technologies

that carefully consider how social factors shape language use (Yang et al., 2024), we provided a public, large-scale collection of interview snippets with clear demographic documentation, where we also break down theoretical constructs into specific, measurable categories, namely labels.

3 Data

3.1 Theoretical Framework

Attribution Theory. Attribution theory (Kelley, 1967) is a core psychological concept that examines how individuals interpret and assign causes to events and behaviors, particularly whether they attribute outcomes to "internal" (dispositional) or "external" (situational) factors. Our corpus curation is informed by Corrigan et al. (2003)’s *attribution model*, an established attribution theory that deconstructs public mental-health stigma in terms of *cognitive judgments* (i.e., stereotypes), *emotional responses* (i.e., prejudice), and *behavioral reactions* (i.e., discrimination). Specifically, the model posits that people’s perceptions of *responsibility* for mental illness lead to emotions such as *anger*, *fear*, and/or *pity*, which in turn drive discriminatory behaviors like *avoidance*, *coercive segregation*, and/or withholding *help*.

Operationalization. Under the guidance of the mental-health specialist (a co-author), we operationalized the theoretical constructs by adapting the Attribution Questionnaire-27 (AQ-27) (Corrigan, 2012), a standardized survey developed from the

Interview Question in MHSTIGMAINTERVIEW	AQ-27 Survey Item (Corrigan, 2012)	Attribution Type
Do you believe Avery’s current situation is primarily a result of their actions ? Please share your thoughts.	I would think that it was Harry’s own fault that he is in the present condition.	Responsibility
If you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?	If I were a landlord , I probably would rent an apartment to Harry.	Social Distance
Let’s imagine you were one of Avery’s friends at the party or other social occasions. Would you feel anger toward them due to their actions, for example, losing their temper and yelling at someone?	How angry would you feel at Harry?	Anger
I wonder if you were one of Avery’s colleagues or friends, then would you be inclined to help them with tasks, such as work projects?	How likely is it that you would help Harry?	Helping
I’d like to know if you would feel concern and sympathy at Avery’s mental state?	How much sympathy and concern would you feel for Harry?	Pity
If you are Avery’s neighbor, then would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?	How much do you think an asylum , where Harry can be kept away from his neighbors, is the best place for him?	Coercive Segregation
If you were traveling or engaging in activities with Avery, would you have any feelings of frightened or being threatened regarding the potential implications of your interaction?	I would feel frightened and threatened by Harry.	Fear

Table 2: Interview questions and their corresponding theoretical underpinnings.

attribution model that decomposes mental-health stigma into nine measurable items – *blame* (i.e., responsibility), *fear*, *pity*, *anger*, *helping*, *avoidance* (i.e., social distance), *coercion*, *segregation*, and *dangerousness*. Notably, following Lee et al. (2023) and Meng et al. (2024), we combined the fear-dangerousness and coercion-segregation pairs to reduce repetitiveness and create a more concise interview, yielding seven key attributions that served as the basis for both our **interview protocol** (Table 2) and **annotation scheme**. The definitions of these attributions appear in Appendix D.1.

3.2 Data Collection

3.2.1 Chatbot-based Interview Design

Interview Flow. We designed and programmed a chatbot via UChat³ to conduct 20-minute dyadic interactions with participants, with our interview structure informed by prior work (Alem et al., 1999; Meng et al., 2024; Lee et al., 2023).

First, the chatbot initiated a **rapport-building small-talk session** (Bickmore and Cassell, 1999), discussing topics such as participants’ favorite movies and activities. Then, it presented a **vignette** (Lerner et al., 2010) about "Avery," a fictional character of unspecified demographics who suffers from a major depressive disorder (Association, 2013) affecting their academic performance, work

productivity, and social relationships. And finally, the chatbot engaged in the **question-answer** conversation – each interview question (Table 2) was embedded within a vivid, relatable scenario based on Avery’s vignette to mitigate social-desirability bias (Fisher, 1993; van de Mortel, 2008), and the questions were presented in randomized order to avoid priming effects (Molden, 2014).

All materials emanating from the chatbot underwent expert review by a consulting psychiatrist and the mental-health specialist. It should be noted that our released corpus excludes the initial small-talk and vignette delivery, as these sections primarily serve as conversation warm-ups, fall outside our annotation scope, and/or may contain sensitive personal information. Further details about the interview can be found in Appendix B.

Vignette. *Vignettes* serve as powerful research instruments for examining attitudes about mental health through brief fictional stories (Alem et al., 1999). Based on research findings and lived experiences (Lee et al., 2023; Griffiths et al., 2006), these narratives enable participants to react to specific situations, giving researchers deeper insights into their views. Specifically, Avery’s symptoms were set forth in the DSM-5 (Association, 2013), though we excluded more severe manifestations, such as self-harm and suicidal behavior, and we avoided medical and/or technical jargon. Two versions were

³<https://uchat.au/>

developed: a **clinical version** that systematically describes Avery’s symptoms and a **narrative version** that brings the story to life with dialogue, actions, and vivid details. Participants only saw the narrative version as the chatbot delivered it sentence by sentence, whereas researchers, including annotators, had access to **both** versions. Both vignette versions appear in Appendix B.

Interview Questions and their Follow-up Question(s). Our chatbot administered seven core **interview questions** ($\mathcal{I}Qs$) aligned with the attributions in Table 2, supplemented by **follow-up questions** ($\mathcal{F}Qs$) (Han et al., 2021) to foster self-disclosure. All $\mathcal{F}Qs$ (if needed) are specifically designed for certain $\mathcal{I}Qs$ and are completed before moving on to the next $\mathcal{I}Q$.

Specifically, we designed our questioning protocol $q(r)$ based on the length of participants’ responses r :

$$q(r) = \begin{cases} \mathcal{I}Q + \mathcal{F}Q_1 + \mathcal{F}Q_2, & \text{if } |r| < 25 \text{ and } |r + \mathcal{F}Q_1| < 150 \\ \mathcal{I}Q + \mathcal{F}Q_1, & \text{if } |r| < 25 \text{ and } |r + \mathcal{F}Q_1| \geq 150 \\ \mathcal{I}Q + \mathcal{F}Q_2, & \text{if } 25 \leq |r| \leq 150 \\ \mathcal{I}Q, & \text{if } |r| > 150 \end{cases} \quad (1)$$

where $|r|$ denotes the length of the response to $\mathcal{I}Qs$ in characters, and $\mathcal{F}Qs$ are specific to each attribution – for *responsibility* attribution, $\mathcal{F}Q1$ explores underlying **reasons** and $\mathcal{F}Q2$ examines if participants view mental illness as **personal weakness**.⁴ For *emotional responses*, $\mathcal{F}Q1$ similarly probes **reasons** while $\mathcal{F}Q2$ asks participants to identify specific aspects of Avery’s story that evoked their emotions. For *behavioral responses*, $\mathcal{F}Q1$ takes two forms: asking about **potential outcomes** for stigmatizing responses ($\mathcal{F}Q1_a$) or exploring **reasons** for non-stigmatizing ones ($\mathcal{F}Q1_b$), followed by $\mathcal{F}Q2$ about triggering vignette plots.⁵

3.2.2 Participant Recruitment

We amassed participants through the research platforms Prolific⁶ and Qualtrics.⁷ Specifically, partici-

pants were required to meet the following inclusion criteria: 1) be at least 21 years of age, 2) have English as their first language, 3) be willing to engage with content related to mental illness, and 4) have no immediate or pressing mental-health concerns, as assessed using the Kessler Screening Scale for Psychological Distress (K6) (Kessler et al., 2003). We included this last criterion to safeguard vulnerable individuals from potential distress and trauma when being exposed to depression-related vignettes (Roberts, 2002).

Our recruitment materials clearly outlined the duration and scope of the study and participants’ right to withdraw. We collected responses from 684 participants, using IP verification to prevent duplicates, with demographics detailed in Appendix C.

3.2.3 Consent and Procedure

This 30-minute, single-session study compensated participants US\$6.30, following both platform-standard rates⁸ and the Department Ethics Review Committee (DERC) guidelines at the main researcher’s institution.

Participants began by receiving a warning about the interview’s mental health focus. They then reviewed and completed the Participant Information Statement and Consent Form (PISCF) (Nijhawan et al., 2013), which detailed privacy protocols and how their data would be collected, stored, and used in our research. Upon obtaining their consent, we reminded participants of their right to withdraw and invited them to voluntarily share demographic information. The core data-collection session consisted of a 20-minute interaction with the chatbot, which concluded with a *debriefing* (Fanning and Gaba, 2007) that covered our research objectives and common misconceptions about mental health.

3.3 Data Annotation and Filtering

Annotation Setup. We opted for **expert-guided annotation** (Snow et al., 2008) over crowdsourcing – given 1) the need for a deep understanding of social-psychological theories (Corrigan et al., 2003) (akin to qualitative coding (Saldana, 2016)) and domain knowledge, 2) the potential negative impact of reviewing stigmatized texts on annotators’ mental well-being, and 3) the importance of close monitoring and sustained discussion. Specifically, with positionality and triangulation strategies (Lazar et al., 2017) in mind, we hired two full-time

⁴This $\mathcal{F}Q$ was derived from Cui et al. (2024) and reviewed and validated by the mental-health specialist.

⁵The two thresholds were determined through an 8-participant pilot study and in consultation with the mental-health specialist.

⁶<https://www.prolific.com/>

⁷<https://www.qualtrics.com/>

⁸<https://researcher-help.prolific.com/en/article/9cd998>

research assistants (RAs) for a three-month period: a computer science-trained Asian male and a social science-educated Asian female, both in their twenties. The RAs annotated under the joint guidance of the mental-health specialist and the main researcher, with compensation following institutional regulations.

Each data unit consisted of an interview snippet centered on one attribution. Two annotators independently reviewed each snippet and the accompanying annotation instructions on the POTATO platform (Pei et al., 2022), answered a **multiple-choice question** (i.e., "*Does this interview snippet contain stigma, and if identified, which specific attribution type is present?*"), and selected one of the seven attributions described in Section 3.1 or a "Non-stigmatizing" option. They could also mark snippets as "N/A" when participant responses 1) provided minimal information (e.g., only "yes," "no," or "not sure"), 2) contained only irrelevant content, 3) exhibited evident AI-generated characteristics, and/or 4) were incomplete or indecipherable. The annotation platform interface, examples, and instructions are provided in Appendix D.

Annotation Process. The team first developed an initial codebook (i.e., annotation instructions) based on the attribution model (Corrigan et al., 2003), which included definitions adapted from Corrigan (2012), keywords, examples, and rules/specifications for each label. Following the mental-health specialist's guidance, we refined the codebook through iterative revisions and practice rounds until the agreement, as measured by Cohen's κ (McHugh, 2012), reached 0.6 (Saldana, 2016), at which point we finalized the codebook. We established checkpoints to assess inter-rater agreement: starting with two batches of 10 participants ($\kappa = 0.55, 0.53$), moving to four sets of 20 participants ($\kappa = 0.66, 0.79, 0.76$, and 0.72). As consistency improved, we progressed to larger samples with two sets of 50 participants ($\kappa = 0.74, 0.66$), one set of 100 participants ($\kappa = 0.69$), one set of 150 participants ($\kappa = 0.69$), and the remaining participants ($\kappa = 0.66$).

At each checkpoint, the team held regular meetings to openly discuss and resolve any disagreements. We excluded interview snippets that 1) were consistently labeled "N/A" by both annotators, and 2) were labeled differently by the two annotators and later unanimously agreed to be "N/A" through open discussion. The final corpus comprised 4,141

interview snippets (86.49% of the original data), with annotators achieving both a satisfactory κ of 0.71 (McHugh, 2012) and perfect agreement on 87.42% of all snippets. Appendix D.4 presents the agreement matrix between two annotators.

3.4 Corpus Statistics

In this section, we analyze key statistical features of our expert-annotated corpus. The collected interview snippets average 2.11 turns ($SD = 0.63$) between participants and the chatbot, with Table 3 depicting that responses indicating personal **responsibility** and social **avoidance** emerge as the most common stigmatization.

Specifically, an interview snippet annotated as *Stigmatizing (responsibility)* is shown below:

Chatbot: *Do you believe Avery's current situation is primarily a result of their actions? Please share your thoughts.*

Participant: *I do think its partially*

Chatbot: *That's a really fair perspective to have! What's your reason behind it?*

Participant: *We don't know his backstory, how his family life is, anything that's happened is probably not so good.*

Chatbot: *That makes a lot of sense; understanding someone's background is crucial. Do you think Avery's mental condition reflects personal weakness, or is it more a response to challenging life circumstances?*

Participant: *I do not, mental issues are never a personal weakness, you can prove you're strong by taking control of it.*

In contrast, the following is a stigma-free interview snippet annotated as *Non-stigmatizing*:

Chatbot: *If you are Avery's neighbor, would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?*

Participant: *No, I don't think that's the way to do it. She could seek help from a medical professional, but hospitalization seems extreme.*

Chatbot: *Got it, you feel hospitalization might be too extreme. What specific plot points in Avery's story made you*

Label	# Snippet	%
Non-stigmatizing	2,232	53.90
Stigmatizing	-	-
Responsibility	394	9.51
Social Distance	379	9.15
Anger	298	7.20
Helping	158	3.82
Pity	42	1.01
Coercive Segregation	271	6.54
Fear	367	8.86
Total	4,141	100

Table 3: Label distribution in our corpus, including a non-stigmatizing category and seven stigma attributions operationalized from the attribution model.

think seeking help, but not hospitalization, would be a better route?

Participant: *She is suffering from depression and outbursts. I don't think that qualifies to be hospitalized.*

The token-count analysis (Figure 2) shows longer, more elaborate responses to personal blame ($M = 41.97$, $SD = 17.50$), presumably reflecting participants' internal conflicts about individual agency versus systemic factors in making moral judgments about mental illness. In addition, our preliminary analyses of socio-cultural factors (Krendl and Pescosolido, 2020) suggest that stigmatizing attribution is partly intertwined with interlocutors' demographic backgrounds and geographic locations (Figure 2 and 7). Additional analyses of correlations between different stigma attributions and emotions, along with response-quality patterns, are detailed in Appendix E.

To situate our corpus, we compared it with three datasets of problematic language listed in Table 1.⁹ As shown in Figure 8, the semantic distributions across these datasets diverge markedly, presumably due to our unique interview-driven discourse genre and our focal target of mentally ill individuals, an understudied minority, thus adding unique value to existing language resources in the NLP community. Surprisingly, the stigmatizing texts in our corpus show even **lower toxicity scores**¹⁰ than content from those benchmark datasets (Figure 2) – despite their reputation for capturing veiled, oblique, and understated forms of hate – illuminating increas-

ingly elusive, insidious instances of stigma embedded in richly contextualized interactions. For instance, responses labeled as *Stigmatizing (Helping)* display toxicity scores indistinguishable from non-stigmatizing texts (Figure 8), exemplifying how stigmatization can be interwoven into neutral or even ostensibly benign discussions.

4 Experiments: Mental-Health Stigma Detection

4.1 Experimental Setup

To benchmark how well computational models can detect mental-health stigma, we set up an 8-way classification task, i.e., **stigma detection**, using our interview corpus. Specifically, we experimented with state-of-the-art LLMs, including GPT-4o (transformer-based, instruction-tuned) (Achiam et al., 2023), Llama-3.1-8B (decoder-only transformer, instruction-tuned), Llama-3.3-70B (decoder-only transformer, instruction-tuned) (Dubey et al., 2024), Mistral Nemo (Mistral-Nemo-Instruct-2407, decoder-only transformer, instruction-tuned), and Mixtral 8x7B (Mixtral-8x7B-Instruct-v0.1, decoder-only mixture-of-experts transformer, instruction-tuned) (Jiang et al., 2024), alongside a fine-tuned RoBERTa-base model (encoder-only transformer, base model) (Liu et al., 2019b). We sampled 60%, 20%, and 20% of instances from each category to create the train, validation, and test splits. The experiments ran on a cluster of four H100 GPUs for approximately 150 hours, with an additional US\$500 allocated for GPT-4o API calls.

For the hyperparameter search, we explored temperature values between {0.0, 0.1, 0.2, 0.3, 0.4} and selected 0.2 as our best-found setting for Llama-3.1-8B, Llama-3.3-70B, and Mixtral 8x7B; 0.3 was set for Mistral Nemo following its documentation recommendations, while GPT-4o was only tested at 0.2 due to computational budget constraints. Next, we evaluated epochs {2, 3, 4} and learning rates {5e-5, 1e-5} for fine-tuning RoBERTa, with 4 epochs and a learning rate of 5e-5 being the best-found configuration, along with the AdamW optimizer and a batch size of 12. We implemented FlashAttention-2 (Dao, 2024) to improve efficiency. We averaged the results over three runs for all models except GPT-4o (which had a single run due to cost constraints) to account for randomness.

We experimented with three prompt settings: zero-shot, one-shot, and **full codebook** information

⁹We extracted and analyzed texts annotated as microaggressions in Breitfelder et al. (2019) and those labeled as *implicit* hate speech in ElSherief et al. (2021).

¹⁰Calculated by Perspective API (<https://perspectiveapi.com/>).

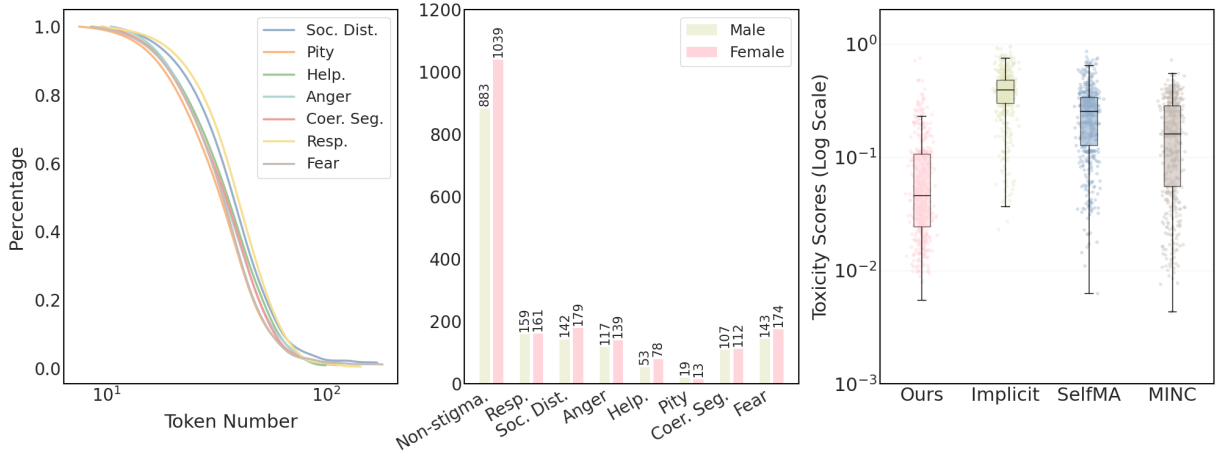


Figure 2: Token-count distribution across attributions (left), the association between mental-health stigma and participants' socio-demographic background (e.g., gender) (center), and toxicity-score comparison between our corpus and three benchmark datasets (right).

matching the guidelines given to human annotators (full prompts available in Appendix H).

4.2 Experimental Results

Our experimental results in Table 4 show that detecting mental-health stigma remains challenging, as we expected. We observed that performance generally scales with model size, with GPT-4o and Llama-3.3-70B performing best, followed by Mixtral 8×7B and Mistral Nemo, while Llama-3.1-8B lagged behind. We also noted consistent improvements across the models when we enriched the prompts with more detailed information, with GPT-4o's F1 score increasing from 0.456 to 0.576 when a single example was added to the prompt. The gains were even more substantial when we provided the full codebook in the prompt – which included label definitions, representative keywords per label, human annotator-derived rules/specifications, along with the same example as in the one-shot prompt – as evidenced by Llama-3.3-70B's F1 score increasing from 0.545 to 0.752.

In contrast to this trend, however, Mixtral 8×7B achieved a higher accuracy (0.608) than GPT-4o (0.578) in the zero-shot setting, which can probably be explained by the highly unbalanced data distribution in our corpus – 42 *Stigmatizing* (*pity*) instances versus 2,232 *Non-stigmatizing* ones – where predicting the majority class inflates accuracy. Interestingly, when given prompts with detailed guidance, the models showed higher recall at the expense of precision, as exemplified by GPT-4o achieving a recall of 0.801 but a precision of 0.742 in the

full-codebook setting, indicating better detection of stigmatizing language but more false positives.

To sum up, these results underscore the need for human-generated guidelines and confirm that reliance on neural models alone remains insufficient to capture an array of subtleties in mental-health stigma, substantiating the development of our interview corpus as a much-needed benchmark.

4.3 Challenges in Detecting Mental-Health Stigma

To further understand the challenges of stigma detection, we empirically and qualitatively investigated all 137 misclassified interview snippets (out of 829 total) from GPT-4o's predictions using full-codebook prompting – one of our best-performing experimental setups – and uncovered a set of deeply embedded, socially normalized stigmatizations that the model struggled to identify correctly.

Linguistically, we observed several recurring patterns in these misclassified utterances: 1) the use of **distancing language** (Nook et al., 2017), where speakers employ third-party perspectives to disguise their personal views, for example, "*Neighbors may find it hard to understand Avery's outbursts and strange behavior if they do not know them very well.*" (P388), 2) the dismissive **misappropriation of psychiatric terminology** (Lilienfeld et al., 2015), such as describing people with mental illness as "*paranoid*" (P28) without proper context or medical basis, and 3) **coercive phrasing** in advice-giving, where speakers use terms like "*definitely need*" to impose decisions on people with mental illness rather than offering "*sugges-*

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.572	.446	.456	.394	.578	.590	.737	.576	.473	.561
Llama-3.1-8B	.479	.445	.307	.206	.267	.493	.555	.383	.260	.306
Llama-3.3-70B	.610	.447	.449	.311	.425	.581	.648	.545	.416	.505
Mistral Nemo	.356	.266	.238	.278	.350	.533	.645	.478	.389	.473
Mixtral 8×7B	.439	.304	.318	.379	.608	.471	.503	.429	.335	.504
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.742	.801	.757	.763	.835	—	—	—	—	—
Llama-3.1-8B	.545	.732	.521	.407	.472	—	—	—	—	—
Llama-3.3-70B	.778	.744	.752	.767	.847	—	—	—	—	—
Mistral Nemo	.643	.779	.662	.620	.708	—	—	—	—	—
Mixtral 8×7B	.660	.602	.584	.552	.705	—	—	—	—	—
RoBERTa	—	—	—	—	—	.747	.766	.885	.755	.832

Table 4: Results of stigma detection task on our corpus. *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink. See Appendix F for a detailed breakdown of model performance on responses to each of the seven interview questions.

tions" that respect their autonomy and agency.

Semantically, we identified certain microaggressions that models often fail to detect – 1) **differential support** (Iseselo et al., 2016) emerges when participants display excessive caution toward people with mental illness, subconsciously positioning them as inferior or in need of special treatment, with participants expressing a need to "*be more mindful*" or "*humble oneself*" during interactions (P510), 2) **patronization/paternalism** (Douglas and Sutton, 2011) surfaces in condescending and demeaning attitudes where speakers position themselves as authorities who can "*teach*" people with mental illness proper ways to live (P584), and 3) **trivialization/minimization** (Hopkins et al., 2014) appears when speakers downplay the legitimate challenges inflicted by mental-health conditions. Representative quotes are provided in Appendix G.

5 Conclusion

As an initial effort, our expert-annotated interview corpus, informed by the socio-conceptual framework (Corrigan et al., 2003) and collected through human-chatbot conversations, can serve as an infrastructure to facilitate detect and finely classify mental-health stigma. By further documenting the socio-cultural context, such as the gender of interviewees, this corpus allows for data lineage tracing and partly prevents mis/underrepresentation of

certain social groups. In addition, the empirical results illustrate existing challenges for state-of-the-art LLMs in decoding those seemingly well-intentioned stigmatizing expressions, suggest areas for improvement in computational approaches, and thus contribute to both computational social science and the NLP community.

Our keystone dataset also provides important implications for future research. First, it can benchmark the extent to which **neural models internalize and perpetuate stigma** by having them role-play interviewees and comparing their generated responses with real-world, human-provided ones. Second, it captures **conversation dynamics** (Hua et al., 2024) of how humans and chatbots navigate stigma-related discussions, which can inform the development of empathetic conversational agents. Third, it opens new avenues for research on computationally **neutralizing, reducing, and counter-acting** (Podolak et al., 2024) stigma, which could potentially be generalized to other psychological constructs (Meng, 2025) where attribution models are applicable (e.g., LGBTQ+ stigmatization (Haider-Markel and Joslyn, 2008)). Finally, it offers insights into **causal-reasoning** patterns (Meng et al., 2025) and **moral judgments/values** (Haidt and Graham, 2007) underlying stigma, which could guide the design of personalized interventions and stigma-reduction campaigns.

Limitations

We recognize that our paper warrants discussion of several limitations.

Multi-Perspective Annotation. Our dataset release includes consolidated labels from two annotators, yet we acknowledged that mental-health stigma represents a deeply subjective phenomenon shaped by socio-psychological and cultural factors (Krendl and Pescosolido, 2020). The interpretation and annotation of stigmatizing content inherently vary according to the annotators’ backgrounds, experiences, and temporal contexts (Röttger et al., 2022; Paun and Simpson, 2021). Recent work has emphasized the critical role of annotators’ demographics and beliefs in shaping data labels for sensitive social tasks (Sap et al., 2022; Wan et al., 2023; Giorgi et al., 2024b). The perspectivist paradigm further suggests capturing this natural variation rather than enforcing consensus (Fleisig et al., 2024; Prabhakaran et al., 2021). An important next step would be to unlock richer perspectives by releasing version 2.0 of our corpus with **annotator-level labels** that preserve individual viewpoints and disagreements.

Cultural Sensitivity. The current dataset is largely drawn from Western, English-speaking sources, yet mental-health stigma remains deeply interrelated with and tied to cultural values, belief systems, and social norms. Research has demonstrated the unique challenges of analyzing stigma and stereotypes in different social groups and languages (Fort et al., 2024). It can be predicted that internalized prejudice, help-seeking barriers, and social-exclusion dynamics will differ markedly across multinational, multilingual, and multicultural settings. Accordingly, a promising direction is to expand our corpus to version 3.0 by collecting and annotating stigma-related discussions from Eastern regions, allowing for cross-cultural analysis of how social stigma is unfolded, negotiated, and expressed.

Intersectionality. Our corpus does not explicitly account for intersectionality in mental-health stigma, which often interacts with other forms of stereotyping and prejudice, such as racism, ageism, and misogyny (Lin et al., 2022). The interplay of these intersectional biases proves difficult to operationalize, as they manifest in complex, jointly reinforcing patterns that compound and affect each other. A natural extension of our work would be

to develop finer-grained annotation schemes that can capture these intersectional dynamics while maintaining analytical clarity.

Prompt Robustness. The prompt design and few-shot example selection in our experiments, while functional, could benefit from a more systematic evaluation. The specific choice of phrasing and exemplar selection may influence model behavior in ways that we have not thoroughly tested. In addition, the order of stigma attributions remains the same in both the one-shot and full-codebook prompt settings, which may introduce primacy and/or recency biases (Lu et al., 2022). Fruitful avenues for future research include conducting ablation studies through structured variation of prompt components, example counts (e.g., five-shot), and linguistic patterns.

Pre-existing Model Biases. Previous studies have shown that LLMs and pre-trained language models (PLMs) harbor and inadvertently reinforce societal biases and/or stereotypes, including those related to mental health (Mina et al., 2024; Lin et al., 2022). Therefore, our experimental results may be unavoidably influenced by these flawed pre-existing biases. We should remain alert to the possibility that LLM/PLM biases affect our findings, and an important next step would be to evaluate how these biases impact model performance.

Ethics and Broad Impact

Our corpus creation and annotation processes followed rigorous ethical protocols with full Institutional Review Board (IRB) approval (NUS-IRB-2024-391), and we implemented comprehensive consent procedures during data collection in which participants received detailed information about data storage, use, and release policies. To protect privacy, our published dataset includes only responses to the interview and follow-up questions, with all personal identifiers removed.

We established ongoing monitoring mechanisms for our research assistants who served as annotators – the main researcher conducted regular check-ins to assess any potential impact on their mental well-being from exposure to stigmatizing content, with their feedback documented in Appendix D.5.

We recognize the potential risks associated with releasing this mental-health stigma corpus. The primary concern is that models trained on this data may inadvertently amplify existing biases and

stereotypes against people with mental illness. We have implemented ethical guardrails, including documentation of dataset limitations, access request forms requiring researchers to specify intended uses, and feedback channels for reporting ethical concerns, and strongly advocate for the responsible use of this dataset through regular audits, careful deployment considerations, and appropriate research applications.

Acknowledgments

We thank the anonymous reviewers for their feedback on this work. This work was partially funded by the National University of Singapore CSSH (24-1774-A0002), the National University of Singapore HSS Seed Fund CR (2024 24-1191-A0001), and the Ministry of Education Tier 1 (24-1317-A0001).

References

- Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. [Alexa, Google, Siri: What are your pronouns? gender and anthropomorphism in the design and perception of conversational assistants](#). In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 24–33, Online. Association for Computational Linguistics.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Atalay Alem, Lars Jacobsson, Mesfin Araya, D Kebede, and Gunnar Kullgren. 1999. How are mental disorders seen and where is help sought in a rural ethiopian community? a key informant study in butajira, ethiopia. *Acta Psychiatrica Scandinavica*, 100:40–47.
- American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. American psychiatric publishing, Washington, D.C.
- Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. [Just say no: Analyzing the stance of neural dialogue generation in offensive contexts](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Timothy Bickmore and Justine Cassell. 1999. Small talk and conversational storytelling in embodied conversational interface agents. In *AAAI fall symposium on narrative intelligence*, pages 87–92.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. [Language \(technology\) is power: A critical survey of “bias” in NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.
- Bart Bonikowski and Laura K Nelson. 2022. From ends to means: The promise of computational text analysis for theoretically driven sociological research. *Sociological Methods & Research*, 51(4):1469–1483.
- Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.
- Ryan L Boyd and H Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41.
- Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. [Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.
- Kaan Büyükdemirci, Izzet Emre Kucukkaya, Eren Ölmez, and Cagri Toraman. 2024. [JL-hate: An annotated dataset for joint learning of hate speech and target detection](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9543–9553, Torino, Italia. ELRA and ICCL.
- Won Ik Cho and Jihyung Moon. 2021. [How does the hate speech corpus concern sociolinguistic discussions? a case study on Korean online news comments](#). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 13–22, NIT Silchar, India. NLP Association of India (NLP AI).
- Mason Choey. 2023. [From stigma to support: A parallel monolingual corpus and NLP approach for neutralizing mental illness bias](#). In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 249–254, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. 2024. [QUEEREOTYPES: A multi-source Italian corpus of stereotypes towards LGBTQIA+ community members](#). In *Proceedings of*

- the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 13429–13441, Torino, Italia. ELRA and ICCL.
- Patrick W. Corrigan. 2000. Mental health stigma as social attribution: Implications for research methods and attitude change. *Clinical psychology: science and practice*, 7(1):48.
- Patrick W. Corrigan. 2012. A toolkit for evaluating programs meant to erase the stigma of mental illness. *Illinois Institute of Technology*, 10.
- Patrick W. Corrigan, Fred E. Markowitz, Amy Watson, David Rowan, and Mary Ann Kubiak. 2003. [An attribution model of public discrimination towards persons with mental illness](#). *Journal of Health and Social Behavior*, 44(2):162–179.
- Yichao Cui, Yu-Jen Lee, Jack Jamieson, Naomi Yamashita, and Yi-Chieh Lee. 2024. Exploring effects of chatbot’s interpretation and self-disclosure on mental illness stigma. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–33.
- Tri Dao. 2024. [Flashattention-2: Faster attention with better parallelism and work partitioning](#). In *The Twelfth International Conference on Learning Representations*.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margaret Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. [Using large language models in psychology](#). *Nature Reviews Psychology*, 2:688–701.
- Karen M Douglas and Robbie M Sutton. 2011. Constructive or cruel? positive or patronizing? reactions to expressions of positive and negative stereotypes of the mentally ill. *British Journal of Psychology*, 102(1):97–107.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ruth M Fanning and David M Gaba. 2007. The role of debriefing in simulation-based learning. *Simulation in healthcare*, 2(2):115–125.
- Robert J. Fisher. 1993. [Social Desirability Bias and the Validity of Indirect Questioning](#). *Journal of Consumer Research*, 20(2):303–315.
- Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. [The perspectivist paradigm shift: Assumptions and challenges of capturing human labels](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.
- Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, Javier Torroba Marchante, Shilin Xie, Sergio E. Zanutto, and Aurélie Névéol. 2024. [Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.
- Paula Fortuna, Juan Soler, and Leo Wanner. 2020. [Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. 2022. [CAMS: An annotated corpus for causal analysis of mental health issues in social media posts](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6387–6396, Marseille, France. European Language Resources Association.
- Salvatore Giorgi, Douglas Bellew, Daniel Roy Sadek Habib, João Sedoc, Chase Smutterberg, Amanda Devoto, Mckenzie Himelein-Wachowiak, and Brenda Curtis. 2024a. [Lived experience matters: Automatic detection of stigma toward people who use substances on social media](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):474–487.
- Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024b. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *arXiv preprint arXiv:2410.07991*.

- Erving Goffman. 1964. [Stigma: Notes on the management of spoiled identity](#). *Social Forces*, 43(1):127–128.
- Kathleen M Griffiths, Yoshibumi Nakane, Helen Christensen, Kumiko Yoshioka, Anthony F Jorm, and Hideyuki Nakane. 2006. [Stigma in response to mental disorders: a comparison of australia and japan](#). *BMC Psychiatry*, 6:21.
- Donald P Haider-Markel and Mark R Joslyn. 2008. Beliefs about the origins of homosexuality and support for gay rights: An empirical test of attribution theory. *Public opinion quarterly*, 72(2):291–310.
- Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116.
- Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. 2021. [Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA. Association for Computing Machinery.
- Keith Harrigan, Carlos Aguirre, and Mark Dredze. 2020. [Do models of mental health based on social media data generalize?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. [Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot](#). *Journal of Communication*, 68(4):712–733.
- Robert W Hopkins, Michela M David, and Lindy A Kilik. 2014. A re-examination of behaviour in depression: Have we grossly underestimated the extent and impact of the behavioural suffering? *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 46(3):456.
- Dirk Hovy and Diyi Yang. 2021. [The importance of modeling social factors of language: Theory and practice](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.
- Yilun Hua, Nicholas Chernogor, Yuzhe Gu, Seoyeon Jeong, Miranda Luo, and Cristian Danescu-Niculescu-Mizil. 2024. [How did we get here? summarizing conversation dynamics](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7452–7477, Mexico City, Mexico. Association for Computational Linguistics.
- Masunga K. Iseselo, Lusajo Kajula, and Khadija I. Yahya-Malima. 2016. The psychosocial problems of families caring for relatives with mental illnesses and their coping strategies: a qualitative urban based study in dar es salaam, tanzania. *BMC Psychiatry*, 16(1):146.
- Patrick M. Jenlink and Bela H. Banathy. 2005. [Dialogue](#), pages 3–14. Springer US, Boston, MA.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Sagar Jilka, Clarissa Mary Odoi, Janet van Bilsen, Daniel Morris, Sinan Erturk, Nicholas Cummins, Matteo Cella, and Til Wykes. 2022. Identifying schizophrenia stigma on twitter: a proof of principle model using service user supervised machine learning. *Schizophrenia*, 8(1):1.
- Harold H Kelley. 1967. Attribution theory in social psychology. In *Nebraska symposium on motivation*. University of Nebraska Press.
- Ronald C Kessler, Peggy R Barker, Lisa J Colpe, Joan F Epstein, Joseph C Gfroerer, Eva Hiripi, Mary J Howes, Sharon-Lise T Normand, Ronald W Manderscheid, Ellen E Walters, et al. 2003. Screening for serious mental illness in the general population. *Archives of general psychiatry*, 60(2):184–189.
- Ronald C Kessler, Katherine A McGonagle, Shanyang Zhao, Christopher B Nelson, Michael Hughes, Suzann Eshleman, Hans-Ulrich Wittchen, and Kenneth S Kendler. 1994. Lifetime and 12-month prevalence of dsm-iii-r psychiatric disorders in the united states: results from the national comorbidity survey. *Archives of general psychiatry*, 51(1):8–19.
- Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2024. [The challenges of creating a parallel multilingual hate speech corpus: An exploration](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15842–15853, Torino, Italia. ELRA and ICCL.
- Anne C Krendl and Bernice A Pescosolido. 2020. Countries and cultural differences in the stigma of mental illness: the east–west divide. *Journal of Cross-Cultural Psychology*, 51(2):149–167.

- Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. [Chapter 11 - analyzing qualitative data](#). In Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser, editors, *Research Methods in Human Computer Interaction (Second Edition)*, second edition edition, pages 299–327. Morgan Kaufmann, Boston.
- Huije Lee, Young Ju Na, Hoyun Song, Jisu Shin, and Jong Park. 2022. [ELF22: A context-based counter trolling dataset to combat Internet trolls](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3530–3541, Marseille, France. European Language Resources Association.
- Min Hyung Lee and Richard Kyung. 2022. Mental health stigma and natural language processing: Two enigmas through the lens of a limited corpus. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 688–691. IEEE.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. [Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.
- Yi-Chieh Lee, Yichao Cui, Jack Jamieson, Wayne Fu, and Naomi Yamashita. 2023. [Exploring effects of chatbot-based social contact on reducing mental illness stigma](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. ["i hear you, i feel you": Encouraging deep self-disclosure through a chatbot](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12.
- William Leeson, Adam Resnick, Daniel Alexander, and John Rovers. 2019. [Natural language processing \(nlp\) in qualitative public health research: A proof of concept study](#). *International Journal of Qualitative Methods*, 18:1609406919887021.
- Debra Lerner, David A. Adler, William H. Rogers, Hong Chang, Leueen Lapitsky, Thomas McLaughlin, and John Reed. 2010. [Work performance of employees with depression: The impact of work stressors](#). *American Journal of Health Promotion*, 24(3):205–213.
- Han Li and Renwen Zhang. 2024. Finding love in algorithms: deciphering the emotional contexts of close encounters with ai chatbots. *Journal of Computer-Mediated Communication*, 29(5):zmae015.
- Jackie Liggins and Simon Hatcher. 2005. [Stigma toward the mentally ill in the general hospital: a qualitative study](#). *General Hospital Psychiatry*, 27(5):359–364.
- Scott O Lilienfeld, Katheryn C Sauvigné, Steven Jay Lynn, Robin L Cautin, Robert D Latzman, and Irwin D Waldman. 2015. Fifty psychological and psychiatric terms to avoid: A list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. *Frontiers in Psychology*, 6:147581.
- Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. [Gendered mental health stigma in masked language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2170, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bruce G. Link, Francis T. Cullen, Elmer Struening, Patrick E. Shrout, and Bruce P. Dohrenwend. 1989. [A modified labeling theory approach to mental disorders: An empirical assessment](#). *American Sociological Review*, 54(3):400–423.
- Yingchi Liu, Quanzhi Li, Marika Cifor, Xiaozhong Liu, Qiong Zhang, and Luo Si. 2019a. [Uncover sexual harassment patterns from personal stories by joint key element extraction and categorization](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2328–2337, Hong Kong, China. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. [Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16235–16250, Toronto, Canada. Association for Computational Linguistics.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Michael Lyons and Jenny Ziviani. 1995. [Stereotypes, Stigma, and Mental Illness: Learning From Fieldwork Experiences](#). *The American Journal of Occupational Therapy*, 49(10):1002–1008.
- Rijul Magu and Jiebo Luo. 2018. [Determining code words in euphemistic hate speech using word embedding networks](#). In *Proceedings of the 2nd Workshop*

- on *Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.
- Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- Han Meng. 2025. Leveraging human-ai partnership approach to unravel human psychological constructs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.
- Han Meng, Yitian Yang, Yunan Li, Jungup Lee, and Yi-Chieh Lee. 2024. Exploring the potential of human-llm synergy in advancing qualitative analysis: A case study on mental-illness stigma. *arXiv preprint arXiv:2405.05758*.
- Han Meng, Renwen Zhang, Ganyi Wang, Yitian Yang, Peinuan Qin, Jungup Lee, and Yi-Chieh Lee. 2025. Deconstructing depression stigma: Integrating ai-driven data collection and analysis with causal knowledge graphs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Rada Mihalcea, Laura Biester, Ryan L Boyd, Zhijing Jin, Veronica Perez-Rosas, Steven Wilson, and James W Pennebaker. 2024. How developments in natural language processing help us in understanding human behaviour. *Nature Human Behaviour*, 8(10):1877–1889.
- Mario Mina, Júlia Falcão, and Aitor Gonzalez-Agirre. 2024. [Exploring the relationship between intrinsic stigma in masked language models and training data using the stereotype content model](#). In *Proceedings of the Fifth Workshop on Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 54–67, Torino, Italia. ELRA and ICCL.
- Kshitij Mishra, Priyanshu Priya, Manisha Burja, and Asif Ekbal. 2023. [e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13952–13967, Singapore. Association for Computational Linguistics.
- Daniel C Molden. 2014. Understanding priming effects in social psychology: What is “social priming” and how does it occur? *Social cognition*, 32(Supplement):1–11.
- Véronique Moriceau, Farah Benamara, and Abdelmoumene Boumadane. 2022a. [Automatic detection of stigmatizing uses of psychiatric terms on Twitter](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 237–243, Marseille, France. European Language Resources Association.
- Véronique Moriceau, Farah Benamara, and Abdelmoumene Boumadane. 2022b. [Automatic detection of stigmatizing uses of psychiatric terms on Twitter](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 237–243, Marseille, France. European Language Resources Association.
- Lokesh P Nijhawan, Manthan D Janodia, BS Mudukrishna, Krishna Moorthi Bhat, K Laxminarayana Bairy, Nayanabhirama Udupa, and Prashant B Musmade. 2013. Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research*, 4(3):134–140.
- Erik C Nook, Jessica L Schleider, and Leah H Somerville. 2017. A linguistic signature of psychological distancing in emotion regulation. *Journal of Experimental Psychology: General*, 146(3):337.
- Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. [Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.
- Desmond U Patton, William R Frey, Kyle A McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. 2020. Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 337–342.
- Silviu Paun and Edwin Simpson. 2021. [Aggregating and learning from multiple annotators](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9, online. Association for Computational Linguistics.
- John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. [SemEval-2021 task 5: Toxic spans detection](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.
- Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. [POTATO: The portable text annotation tool](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jakub Podolak, Szymon Łukasik, Paweł Balawender, Jan Ossowski, Jan Piotrowski, Katarzyna Bakowicz, and Piotr Sankowski. 2024. [LLM generated responses to mitigate the impact of hate speech](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15860–15876, Miami,

- Florida, USA. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. [On releasing annotator-level labels and information in datasets](#). In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Laura Weiss Roberts. 2002. Ethics and mental illness research. *Psychiatric Clinics of North America*, 25(3):A525–A545.
- David Roesler, Shana Johnny, Mike Conway, and Annie T Chen. 2024. A theory-informed deep learning approach to extracting and characterizing substance use-related stigma in social media. *BMC Digital Health*, 2(1):60.
- Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.
- Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science*, 346(6213):1063–1064.
- Johnny Saldana. 2016. *The Coding Manual for Qualitative Researchers*, 3rd edition. Sage Publications, Inc, London.
- Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. [An Italian Twitter corpus of hate speech against immigrants](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O’connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Nadiya Straton, Hyeju Jang, and Raymond Ng. 2020. [Stigma annotation scheme and stigmatized language detection in health-care discussions on social media](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1178–1190, Marseille, France. European Language Resources Association.
- Thea F van de Mortel. 2008. [Faking it: Social desirability response bias in self-report research](#). *The Australian Journal of Advanced Nursing*, 25(4):40–48.
- Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benvenuto. 2022. [HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.
- Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. [Learning from the worst: Dynamically generated datasets to improve online hate detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone’s voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.
- Zeerak Waseem and Dirk Hovy. 2016. [Hateful symbols or hateful people? predictive features for hate speech detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. [Detection of Abusive Language](#):

the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2024. The call for socially aware language technologies. *arXiv preprint arXiv:2405.02411*.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, page 08944393241258771.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Érica de Toledo Piza Peluso and Sérgio Luís Blay. 2009. Public stigma in relation to individuals with depression. *Journal of Affective Disorders*, 115(1):201–206.

A Discussion of Problematic Language Definition

We would like to clarify and distinguish between several concepts commonly studied by NLP scholars (Fortuna et al., 2020; Pachinger et al., 2023) and the terms we use throughout this paper. **Abusive language** refers to content that ascribes a negatively judged social identity to individuals, marking them as shameful or morally objectionable representatives of a marginalized group (Wiegand et al., 2019). **Offensive language** encompasses insults, profanity, and targeted attacks that may harm disadvantaged groups (Davidson et al., 2017). **Hate speech** specifically involves expressing hatred or intending to degrade members of protected groups based on characteristics like race, gender, or disability (Waseem and Hovy, 2016). **Toxic language** more broadly covers disrespectful or inappropriate content that is likely to cause people to leave discussions (Pavlopoulos et al., 2021).

On the other hand, **social stigma** represents society’s collective negative attitudes, prejudices, and discriminatory practices that devalue and exclude certain individuals or groups (Goffman, 1964). Unlike these forms of problematic language, which can be directly observed, inferred, and/or detected, it exists as a latent psychological construct with deep underpinnings rooted in classic theories (Corrigan et al., 2003; Link et al., 1989), and **mental-health stigma** specifically targets people with men-

tal illness. This means that it needs to be operationalized and measured in terms of components – cognitive judgments (i.e., stereotypes), emotional responses (i.e., prejudice), and behavioral responses (i.e., discrimination) (Goffman, 1964) – that language might be able to capture.

B More Details about Data Collection: Chatbot-based Interview

B.1 Vignettes

The clinical version appears below:

Avery is employed by a company, and in their spare time, they are dedicated to lifelong learning, doing extensive reading and writing. However, Avery has been diagnosed with depression recently. It has become challenging for them to concentrate, resulting in a decline in work performance and learning outcomes. Interests that once brought them joy have waned. Avery has distanced themselves from friends, becoming easily irritated in social interactions with colleagues and feeling judged by others. Avery lives with family and cannot do much, especially household chores. Social media intensifies their feelings of loneliness and anger, leading to frustration about the source of the anger.

The narrative version delivered to participants can be found in Table 8.

B.2 Chatbot Design Elements

To facilitate participants’ self-disclosure, we designed our chatbot with three key strategies. First, we humanized the interaction and added **anthropomorphic** elements (Abercrombie et al., 2021) by naming it *Nova*. Second, *Nova* employed **active listening** (Zheng et al., 2023), where it restated and validated the user’s point of view and expressed its understanding. And third, it used **neutral self-disclosure** (Lee et al., 2020), sharing both positive and negative perspectives without shifting their attitudes to encourage reciprocal openness.

On the other hand, to alleviate potential emotional distress, the vignette presentation was punctuated with intermittent prompts soliciting brief responses from participants (Lee et al., 2023; Cui et al., 2024), such as asking what advice they might

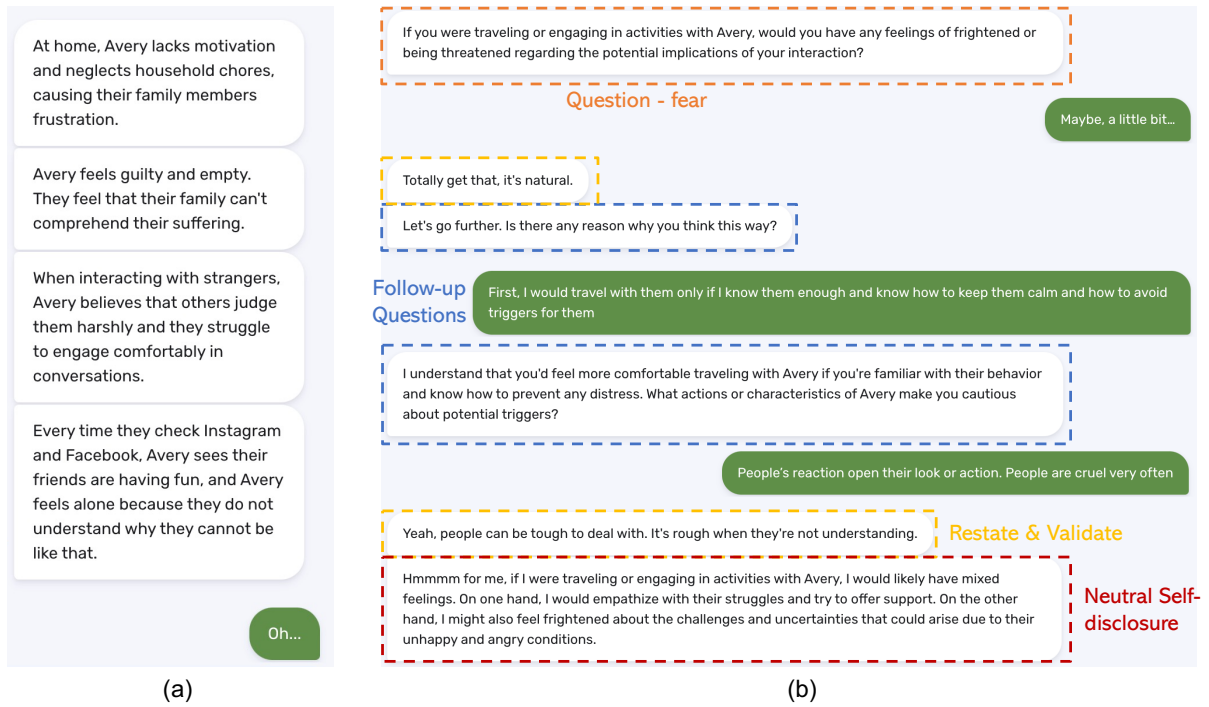


Figure 3: Chatbot interface for (a) vignette delivery and (b) an example interview question, showing a multi-turn conversation about the attribution of *fear* between the chatbot and participant P189.

offer to Avery. In addition, we strategically positioned a *mid-session break* between the seven interview questions, where the conversation shifted to lighthearted dialogue about hobbies, serving to re-engage participants and provide emotional decompression.

B.3 Implementation

Our chatbot combines **pre-written scripts** with **outputs from AI models** to create a natural conversational flow. The core elements – small-talk questions, vignette delivery, interview questions (*IQs*), and the chatbot’s neutral self-disclosure – were pre-scripted and hard-coded. The AI model generates all other elements of the interview, including crafting follow-up questions (*FQs*) and formulating responses to participant input. For example, when generating FQ_2 for the *anger* attribution, we used the following prompt:

- Chatbot: Thank you for sharing! Let’s imagine you were one of Avery’s friends at the party or other social occasions. Would you feel anger towards them due to their actions, for example, losing their temper and yelling at someone?

- User: [Answer 1].
- Chatbot: May I kindly ask why?
- User: [Answer 2].

Given this conversation context where the chatbot asks an interview question and the user provides a response, generate a follow-up that: 1) demonstrates understanding by restating the user’s perspective, 2) asks a single “how” or “what” question about Avery to encourage detailed self-disclosure, and 3) explores which specific plot elements or actions led to their thinking. The response should be conversational, under 30 words, use first-person pronouns (“I,” “me”), and refer to Avery as “them.”

We implemented the model using GPT-4-1106-preview (Achiam et al., 2023), with a maximum token limit of 100 and temperature set to 0.2 for consistent output. The chatbot interface, shown in Figure 3, was integrated into the Qualtrics survey, and we ensured that the concurrency rate remained below 50.

B.4 An Example of A Complete Interview Script

Table 8 shows an example of a complete interview flow. Our released corpus contains only the seven question-answer segments (*Q1-Q7*) for ease of use.

C Interview Participant Demographics

Table 5 presents the self-reported demographic and geographic characteristics of 555 out of 684 participants (81.10%) who voluntarily shared this information and served as our primary data source.

D More Details about Data Annotation

D.1 Description of Stigma Attributions

Our annotation scheme categorizes texts into either non-stigmatizing attitudes or one of seven stigma attributions that capture different facets of mental-health stigmatization:

- **Stigmatizing (Responsibility):** Believing people have control over and are responsible for their mental illness and related symptoms.
- **Stigmatizing (Social Distance):** Staying away from people with mental illness.
- **Stigmatizing (Anger):** Expressing irritation or annoyance toward people with mental illness.
- **Stigmatizing (Helping):** Withholding support toward people with mental illness.
- **Stigmatizing (Pity):** Being unsympathetic toward people with mental illness.
- **Stigmatizing (Coercive Segregation):** Forcing institutionalization and mandatory treatment on people with mental illness.
- **Stigmatizing (Fear):** Perceiving people with mental illness as dangerous, unpredictable, and unsafe to be around.
- **Non-stigmatizing:** Showing understanding, empathy, and support toward people with mental illness, recognizing mental-health challenges as complex medical conditions influenced by multiple factors.

D.2 Annotation Platform

Figure 4 shows the screenshot of the annotation platform interface.

D.3 Annotation Instructions

It should be noted that our annotation instructions were iteratively refined through active collaboration with both annotators, who provided valuable input and suggestions based on their hands-on coding experience, rather than being passive recipients of predetermined guidelines. See Figure 5 for our detailed annotation instructions shown to human annotators.

D.4 Agreement Matrix between Human Annotators

Figure 6 presents the heatmap showing the agreement between the two human annotators. The matrix reveals relatively low confusion between different stigma attributions, while most disagreement occurs when one annotator labels a response as non-stigmatizing and the other identifies it as containing a specific type of stigma.

D.5 Annotator Feedback

D.5.1 Reflexive Impact on Annotators' Perspectives

The annotation experience fostered reflexive awareness among annotators, though their responses to stigmatized content varied. One annotator developed a heightened sensitivity to implicit discrimination in everyday discourse, becoming more conscious of their own language choices and gaining deeper empathy for stigmatized groups. The other annotator maintained their established viewpoints, only occasionally engaging in self-reflection about their attitudes during the annotation process.

D.5.2 Risks of Over-interpretation

We noticed that the long annotation timespan, combined with the subtle nature of stigma, created challenges in maintaining consistent judgment standards. Interestingly, one annotator found themselves detecting multiple stigma attributions in single interview snippets and noticing forms of stigma not captured by the existing scheme, initially worrying about over-interpretation and over-analysis. However, these observations presumably revealed the complexity of stigmatized language in real conversations rather than analytical overreach. Their insights suggest valuable opportunities for exploring how different stigma attributions co-exist and intertwine, and for better conceptualizing more implicit forms of stigmatization.

ALL (N=555) n (%)		
Gender	Female	305 (54.95)
	Male	249 (44.86)
	Prefer not to say	1 (0.18)
Age	21-24	61 (10.99)
	25-34	160 (28.83)
	35-44	109 (19.64)
	45-54	75 (13.51)
	55-64	67 (12.07)
	65+	83 (14.95)
Ethnicity	White	356 (64.14)
	Black or African American	133 (23.96)
	Asian	36 (6.49)
	Mixed	19 (3.42)
	American Indian or Alaska Native	1 (1.80)
	Other (Hispanic, Chicano, etc.)	10 (1.8)
Country	United States	187 (33.69)
	United Kingdom	146 (26.31)
	South Africa	98 (17.66)
	Canada	43 (7.75)
	Australia	28 (5.05)
	Ireland	18 (3.24)
	Other (13 countries)	35 (6.31)
Education	Less than primary	1 (0.18)
	Primary	3 (0.54)
	Some secondary	4 (0.72)
	Secondary	83 (14.95)
	Vocational or similar	62 (11.17)
	Some University but no degree	94 (16.94)
	University - Bachelor's degree	197 (35.50)
	Graduate or professional degree (MA, MS, MBA, PhD, law degree, medical degree, etc.)	109 (19.64)
	Prefer not to say	2 (0.36)
Mental-illness Experience	Yes	320 (57.66)
	No	133 (23.96)
	Maybe	102 (18.38)

Table 5: Participant Demographics. **Mental-illness experience** refers to whether participants had immediate family members or close friends who experienced mental illness.

D.5.3 Training and Knowledge Requirements

Despite initial unfamiliarity with the topic, annotators reported that they found the task accessible through proper training and communication with co-workers and specialists. The codebook evolved through multiple iterations, gained sharper definitions, and offered better guidance for ambiguous cases. Notably, one annotator emphasized that rather than requiring extensive prior knowledge, the main challenge lay in accurately understanding and applying the annotation rules, especially in borderline cases.

E More Corpus Statistics

For additional statistical analysis of our proposed corpus, see Figure 7 and 8.

E.1 Stigma-emotion Correlations

Following Li and Zhang (2024), we used the pre-trained BERT-based emotion classifier¹¹ to detect six emotion categories (*joy*, *love*, *surprise*, *anger*, *fear*, and *sadness*) as established in Shaver et al. (1987). We performed OLS regression analysis to quantify associations between stigma attributions and emotion categories.

Results shown in Table 6 show that **Non-stigmatizing** responses correlate positively with *joy* ($b = 0.1731$, $p < 0.001$) and *love* ($b = 0.0387$, $p < 0.01$), but negatively with *anger* ($b = -0.1116$, $p < 0.001$). In contrast, **Stigmatizing (anger)** responses correlate negatively with *joy* ($b = -0.2471$, $p < 0.01$) and positively with *anger* ($b = 0.4483$, $p < 0.001$). **Stigma-**

¹¹<https://huggingface.co/bhadresh-savani/bert-base-uncased-emotion>

Mental-Health Stigma Annotation
Home
Statistics
Help
Finished 3/4788
Current_id 3

Currently logged in as test@test

Chatbot: I'd like to know if you would feel concern and sympathy at Avery's mental state?

Participant: Yes, of course I am concerned about Avery's mental state.

Chatbot: It's great that you feel concerned about Avery's mental state. What specific actions or experiences of Avery made their struggles resonate with you most deeply?

Participant: His mental state is unhealthy as it is causing so many other problems in his life. He is not performing as well at work, he gets irritable and angry, is lonely, he is overly concerned about what others think of him since he is concerned about being judged. Also, activities he use to enjoy, he has lost interest and no longer do them.

Chatbot: It's heartbreaking to see Avery struggle like this. They're going through so much, and your concern shows how deeply you care for them.

Does this interview snippet contain stigma, and if identified, which specific attribution type is present?

- ☐ Stigmatizing (Responsibility)
- ☐ Stigmatizing (Social Distance)
- ☐ Stigmatizing (Anger)
- ☐ Stigmatizing (Helping)
- ☐ Stigmatizing (Pity)
- ☐ Stigmatizing (Coercive Segregation)
- ☐ Stigmatizing (Fear)
- ☐ Non-stigmatizing
- ☐ N/A

Figure 4: Screenshot of the annotation platform interface.

tizing (fear) responses also correlate negatively with *joy* ($b = -0.2460, p < 0.01$), but positively with both *fear* ($b = 0.2699, p < 0.001$) and *sadness* ($b = 0.1743, p < 0.01$). Together, these correlations substantiate the value of our corpus for examining the psycho-emotional aspects of mental-health stigma.

E.2 Response-quality Pattern

As shown in Table 7, 14.71% of the responses (from 36.55% of the participants) naturally exceeded the 150-character threshold without requiring follow-up questions, averaging 219.54 characters ($SD = 79.37$). In contrast, responses requiring one follow-up question (59.38%) averaged 167.71 characters ($SD = 78.87$), while those requiring two follow-up questions (25.91%) averaged only 116.99 characters ($SD = 62.30$). After follow-up prompts, 46.28% of responses with one FQ and 75.02% with two FQ s were still below our 150-character threshold.

Together with our familiarity with the data, these results allow us to qualitatively infer that responses exceeding 150 characters without requiring FQ s tended to be of higher quality, with more coherent language, deeper thoughts, and richer, more interesting content. For example, we sometimes observed participants incorporating factors such as **past experience** and **personality** into their reasoning pertaining to mental health.

F Full Results of Stigma Detection

Tables 9-15 present the classification performance of each model on responses to the seven interview questions probing different stigma attributions: *responsibility*, *social distance*, *anger*, *helping*, *pity*, *coercive segregation*, and *fear*.

Our results suggest that **model families** such as GPT-4o and Llama tend to outperform Mistral models. This may be due to differences in the scale, quality, and recency of the training data (GPT-4o: $\kappa = .763$; Llama-3.3-70B: $\kappa = .767$; Mistral Nemo: $\kappa = .620$ in the full-codebook setting). **Architectural choices** also matter. For example, *encoder-only* models like RoBERTa excel when fine-tuned with sufficient data ($\kappa = .755$); however, *decoder-only* models appear to be more versatile in few-shot scenarios. **Instruction-tuned** models utilize detailed guidance more effectively than **base** models, especially for rare stigma attributions where data scarcity poses challenges (e.g., *Stigmatizing (pity)* detection: instruction-tuned GPT-4o: $\kappa = .356$; RoBERTa-base: $\kappa = .000$).

G More Analysis on Incorrect Predictions

Table 16 illustrates the semantic and linguistic patterns of stigmatizing expressions that neural models often failed to detect, accompanied by representative quotes from our interview corpus.

The semantic similarity between correctly and incorrectly classified examples is visualized in Fig-

	Joy	Love	Surprise	Anger	Fear	Sadness
Non-stigmatizing	0.1731*** (0.000)	0.0387** (0.001)	-0.0063 (0.072)	-0.1116*** (0.000)	-0.0279 (0.214)	0.0451 (0.065)
Stigmatizing (anger)	-0.2471** (0.002)	-0.0102 (0.750)	-0.0195* (0.034)	0.4483*** (0.000)	-0.0999 (0.091)	0.0394 (0.540)
Stigmatizing (coercive segregation)	0.3971*** (0.000)	-0.0044 (0.888)	0.0032 (0.723)	-0.0672 (0.328)	-0.0768 (0.181)	-0.1408* (0.024)
Stigmatizing (fear)	-0.2460** (0.001)	-0.0134 (0.660)	-0.0094 (0.285)	-0.0643 (0.340)	0.2699*** (0.000)	0.1743** (0.005)
Stigmatizing (helping)	-0.1494 (0.151)	0.0451 (0.289)	-0.0173 (0.159)	0.0687 (0.466)	0.1401 (0.076)	0.0239 (0.780)
Stigmatizing (pity)	0.1579 (0.302)	-0.0085 (0.892)	0.0740*** (0.000)	0.0122 (0.930)	-0.0799 (0.490)	-0.0446 (0.723)
Stigmatizing (responsibility)	0.1936** (0.005)	0.0001 (0.996)	-0.0059 (0.466)	-0.0826 (0.183)	0.0150 (0.772)	-0.0092 (0.870)
Stigmatizing (social distance)	-0.0519 (0.479)	-0.0235 (0.431)	-0.0091 (0.292)	0.0806 (0.225)	0.0388 (0.485)	0.0762 (0.206)

Table 6: Correlation coefficients between stigma attributions and emotion categories. p -values are shown in parentheses. Significance levels are denoted as ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

Response Type	# Responses (% total)	# Participants (% total)	Avg Length (chars)	# Responses still < 150 chars after \mathcal{FQ} s
\mathcal{IQ} only (no \mathcal{FQ} s needed)	609 (14.71%)	250 (36.55%)	219.54	0
\mathcal{IQ} + one \mathcal{FQ}	2,459 (59.38%)	639 (93.42%)	167.71	1,138
\mathcal{IQ} + two \mathcal{FQ} s	1,073 (25.91%)	392 (57.31%)	116.99	805

Table 7: Interview response statistics by number of \mathcal{FQ} s asked. We report the count and percentage of interview snippets in which zero, one, or two \mathcal{FQ} s were asked, the number of participants who contributed at least one snippet to each group (noting that one participant can contribute up to seven interview snippets and may therefore be counted in more than one group), the average total character length of participants’ responses to the \mathcal{IQ} and any \mathcal{FQ} s, and the number of snippets with participant responses that remained below our 150-character threshold even after all \mathcal{FQ} s were asked.

ure 9, which shows their substantial overlap in the embedding space.

H Full Prompts for Stigma Detection

In this section, we present the full prompts. For *one-shot prompting*, we added the text shown in Section H.2 between the multiple-choice question and [vignette] in the zero-shot prompt template. For *full-codebook prompting*, we added the text shown in Section H.3 between the multiple-choice question and [examples] to the combined text from the previous two sections.

General Rules

- 1) A message must be coded as **Stigmatizing** if it contains ANY stigmatizing statement, even if it also contains non-stigmatizing content.
- 2) For messages containing "if" conditions: code as **Stigmatizing** if the conditions are based on personal judgments or subjective criteria. Code as **Non-stigmatizing** only if the conditions are purely objective facts or directly restate the question parameters.
- 3) Code as **Stigmatizing** any message that suggests treating Avery differently from others or implies a hierarchy of power, including expressions of superiority or excessive focus on managing her performance.
- 4) Code as **Stigmatizing** if EITHER the cause OR the effect in the message contains stigmatizing elements, regardless of whether the other component is non-stigmatizing.

Stigmatizing (Coercive Segregation)	
The stigmatizing behavior is to send people with mental illness to institutions away from their community and to force people with mental illness to participate in medication management or other treatments.	
KEYWORD(S)	separate, lonely, warrant, hospitalization, distance, treatment, respect, neighborhood, neighbor, socialization, companionship
EXAMPLE	<p>chatbot: Apart from being a colleague or friend, I'm also curious if you are Avery's neighbor, then would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?</p> <p>participant: yes</p> <p>chatbot: That sounds like a compassionate and thoughtful choice!</p> <p>chatbot: What do you think the potential result might be if Avery stays in the neighborhood without getting the treatment they need?</p> <p>participant: I would not be able to help her.</p> <p>chatbot: It sounds like you're feeling concerned about not being able to support her fully. What about their struggles or actions makes you feel treatment is the best option?</p> <p>participant: I think it would be for the best for her</p> <p>chatbot: You care deeply about them and their well-being. Your thoughtfulness shows so much love and compassion.</p>
RULE(S)	<p>1) If the message implies a willingness to follow the doctor's advice, code it as Stigmatizing.</p> <p>2) If the message has no opinion and implies that they should listen to their family, professionals, or friends, code it as Stigmatizing (Pity).</p> <p>3) If the message simply suggests that they seek treatment or counseling, code it as Non-stigmatizing.</p>
Stigmatizing (Fear)	
The stigmatized thought is that people with mental illness are not safe or are afraid because people with mental illness are dangerous or unpredictable.	
KEYWORD(S)	violence, threatened, safe, danger, afraid, cautious, hurt, safety, frightened, scare, intimidate, threatening, anger, fear, yelling
EXAMPLE	<p>chatbot: If you were traveling or engaging in activities with Avery, would you have any feelings of frightened or being threatened regarding the potential implications of your interaction?</p> <p>participant: I would, perhaps, feel nervous that stress might further upset Avery, causing them to shout or lash out verbally, but I do not think I would fear Avery.</p> <p>chatbot: Feeling nervous makes sense; it's challenging navigating someone's stress. Your care and understanding for Avery truly shine.</p>
RULE(S)	<p>1) If the message conveys feelings of caution, code it as Stigmatizing.</p> <p>2) If the message suggests that people with mental illness are prone to self-harm or suicide, code it as Stigmatizing.</p>
Stigmatizing (Pity)	
A common stigmatizing thought is to be unsympathetic to people with mental illness and to blame them for their mental health problems.	
KEYWORD(S)	concern, pity, sympathy, tough, sad, lack
EXAMPLE	<p>chatbot: I'd like to know if you would feel concern and sympathy at Avery's mental state?</p> <p>participant: Yes, But only if Avery would be willing to get help cause my feeling around the Avery mental health would help Avery at all</p> <p>chatbot: That makes sense - your concern seems tied to Avery showing a willingness to improve. What specific actions or moments from Avery suggest they might - or might not - be open to seeking help?</p> <p>participant: Outbursts</p> <p>chatbot: Outbursts can be so tough to navigate - I get why you're concerned. It shows you care about them deeply and want the best.</p>
RULE(S)	<p>1) If the message conveys a feeling of indifference & none of my business, code it as Stigmatizing.</p> <p>2) If the message reflects an opinion that people with mental illness should be treated differently, or implies a hierarchical or condescending attitude, code it as Stigmatizing.</p>
Stigmatizing (Social Distance)	
The constant stigmatized behavior is to stay away from people with mental illness.	
KEYWORD(S)	worried, apprehensive, upset, unhappy, worrying, get along, distance
EXAMPLE	<p>chatbot: If you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?</p> <p>participant: I would not be fully comfortable.</p> <p>chatbot: Got it, you're not fully comfortable. What potential result of renting to Avery concerns you most about the overall experience or outcome?</p> <p>participant: I would be concerned that Avery's mental health may make it difficult for them to keep on top of bills.</p> <p>chatbot: Absolutely, that concern is valid and caring. It's kind of you to consider Avery's well-being while thinking about practical responsibilities.</p>
RULE(S)	-
Stigmatizing (Anger)	
The stigmatized thought is to be irritated or annoyed because people are being blamed for their mental illness.	
KEYWORD(S)	annoyed, angry, anger, irritated, lost temper
EXAMPLE	<p>chatbot: Let's imagine you were one of Avery's friends at the party or other social occasions. Would you feel anger toward them due to their actions, for example, losing their temper and yelling at someone?</p> <p>participant: If I was one of Avery's friends I might feel anger towards them for their actions in the moment as I may not know that he might have a mental illness.</p> <p>chatbot: Feeling that way in the moment makes so much sense. It's hard to fully understand someone's struggles right away. You're incredibly thoughtful.</p> <p>Expanded definition of anger: Thinking that some actions are not okay or feeling uncomfortable is a form of stigmatization.</p> <p>1) If the message conveys feelings of disgust or other forms of aversion (e.g., embarrassment), code it as Stigmatizing.</p> <p>2) If the message conveys only feelings of confusion/bewilderment, code it as Non-stigmatizing.</p> <p>3) If the message only refers to "out of character" expressions, code it as Non-stigmatizing.</p>
RULE(S)	
Stigmatizing (Responsibility)	
A common stigmatized thought is that people have control over and are responsible for their mental illness and related symptoms.	
KEYWORD(S)	responsible, responsibility, weakness, stem from, blame
EXAMPLE	<p>chatbot: Just like what is mentioned in the story, Avery is currently facing difficulties in both their relationships with colleagues and their work performance. Do you believe Avery's current situation is primarily a result of their actions? Please share your thoughts.</p> <p>participant: I think there is something that he is suppressing, he is unable to be his true self, unable to express something due to fear, guilt, shame or obligation. This causes Avery to be moody and agitated, which results in poor relationships, affects work etc.</p> <p>chatbot: You believe Avery is holding something back, and that must be so heavy for them. What a compassionate and thoughtful insight you've shared!</p>
RULE(S)	<p>1) If the message only makes a suggestion, code it as Non-stigmatizing; if the message not only makes a suggestion but also indicates a possible negative outcome if the suggestion is not followed (e.g., "otherwise") or expresses disappointment, code it as Stigmatizing.</p>
Stigmatizing (Helping)	
The stigmatized behavior is withholding support from them because of prejudiced thoughts.	
KEYWORD(S)	lack, hinder, refuse, less inclined to, help, assist
EXAMPLE	<p>chatbot: I wonder if you were one of Avery's colleagues or friends, then would you be inclined to help them with tasks, such as work projects?</p> <p>participant: No that's their work to finish. Depressed or not.</p> <p>chatbot: Got it, you feel it's their responsibility. How might the potential result of teaming up with Avery impact your studies or work balance?</p> <p>participant: I'm not depressed. So I'll overlook the whole project and make sure everything goes well.</p> <p>chatbot: You're confident and ready to lead! It's commendable how you're taking charge to ensure everything runs smoothly. You've got this!</p>
RULE(S)	<p>1) If the message suggests that helping people with mental illness depends on their willingness to seek help, code it as Stigmatizing.</p> <p>2) If the message suggests that being a friend to someone with mental illness or understanding their condition is a reason to offer help, code it as Non-stigmatizing.</p>
Non-stigmatizing	
The text explicitly conveys non-stigmatizing content and shows positive attitudes such as compassion, help, kindness, etc.	
KEYWORD(S)	help, kind, fine, like, comfortable, enjoy
EXAMPLE	<p>chatbot: I wonder if you were one of Avery's colleagues or friends, then would you be inclined to help them with tasks, such as work projects?</p> <p>participant: Yes I would try to help him but I would be looking for confirmation that he actually was accepting my help</p> <p>chatbot: Got it, you're willing to help but want confirmation it's appreciated. What plot points or actions in Avery's story made you feel like helping him would make a difference?</p> <p>participant: I like helping people in general and I have been in Avery's position in the past and so I can possibly relate to some of the things he might be feeling</p> <p>chatbot: You relate to Avery's struggles deeply, and that empathy shines. Your willingness to help shows such kindness and genuine humanity.</p>
RULE(S)	-
N/A	
The text is 1) completely off-topic, or 2) on-topic but not taking a stance and providing completely objective reasons or facts, or 3) on-topic and taking a stance but the reason is irrelevant to mental health, or 4) the data quality is low.	
KEYWORD(S)	-
EXAMPLE	<p>chatbot: I'd like to know if you would feel concern and sympathy at Avery's mental state?</p> <p>participant: Go on.</p> <p>chatbot: Avery's mental state truly deserves care and understanding.</p> <p>chatbot: Let's go further, Is there any reason why you think this way?</p> <p>participant: I don't understand.</p> <p>chatbot: I'm sorry if I wasn't clear earlier! It sounds like Avery's mental state really resonated with you.</p>
RULE(S)	-

Figure 5: Annotation instructions for human annotators.

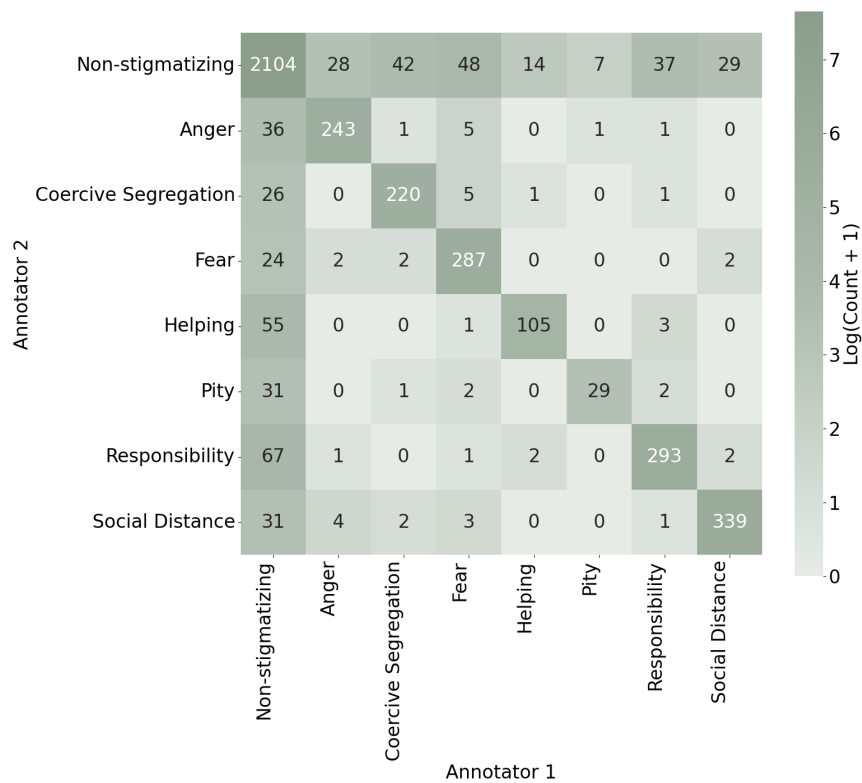


Figure 6: Heatmap showing the agreement between two human annotators.

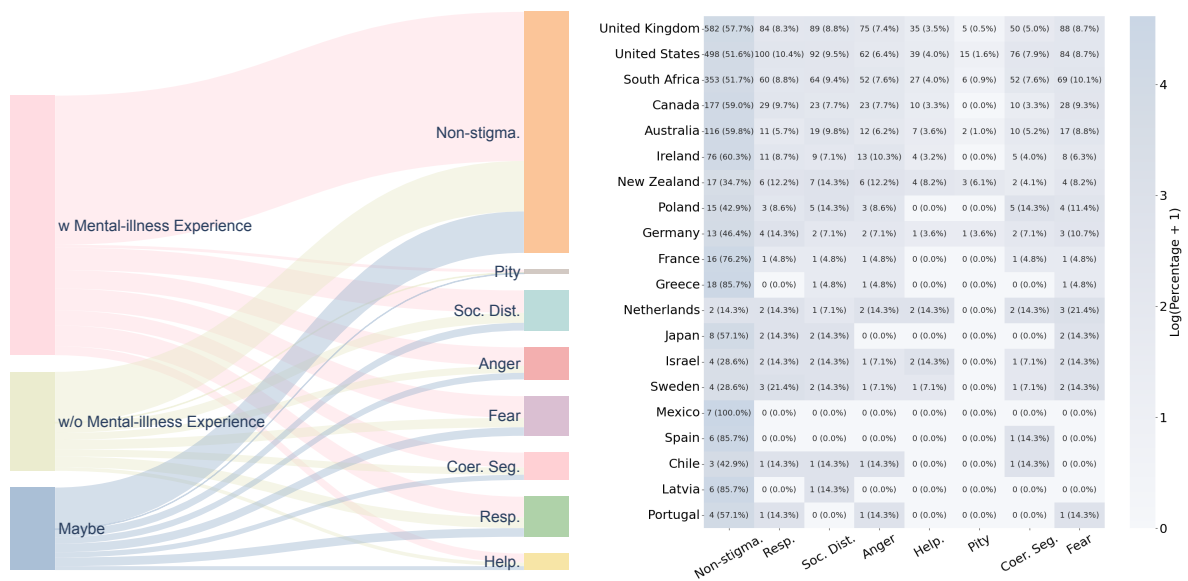


Figure 7: More association between mental-health stigma and sociocultural factors in our corpus: mental-illness experience (left) and country of residence (right).

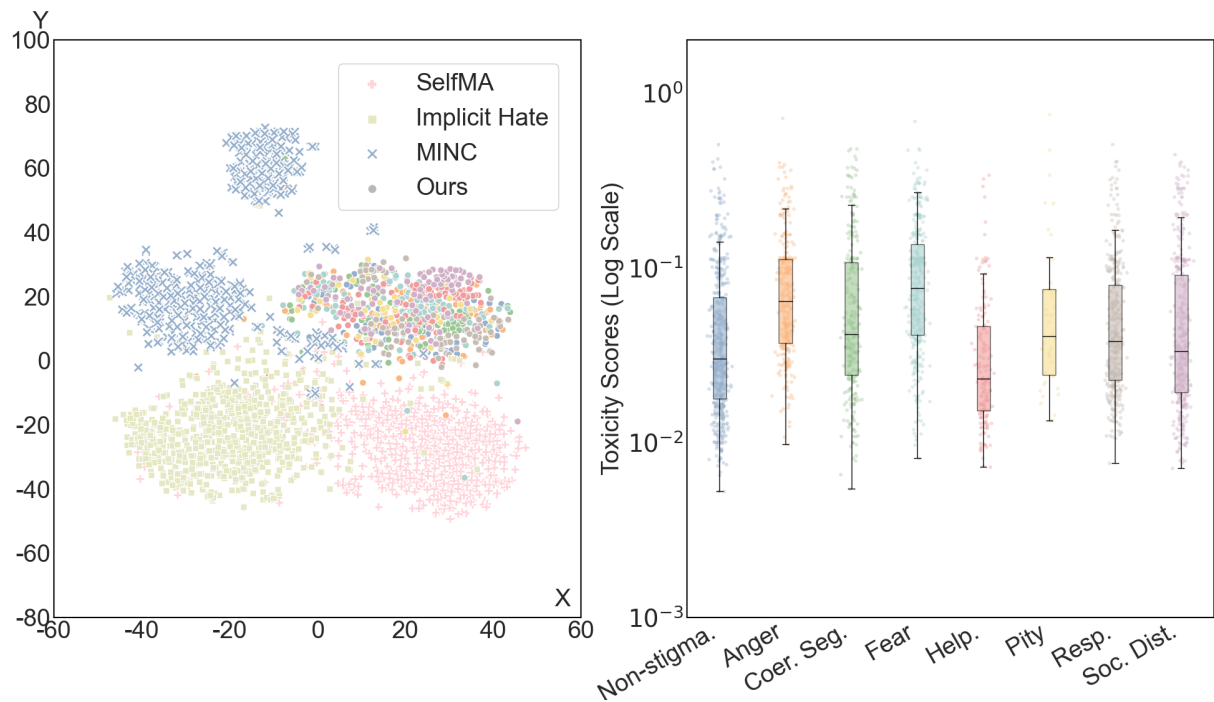


Figure 8: Word-embedding comparison of 1,000 random samples each from SELFMA (Breitfeller et al., 2019), IMPLICIT HATE SPEECH (ElSherief et al., 2021), MINC (Choey, 2023) and our corpus (left); and toxicity-score distributions across different labels in our corpus (right).

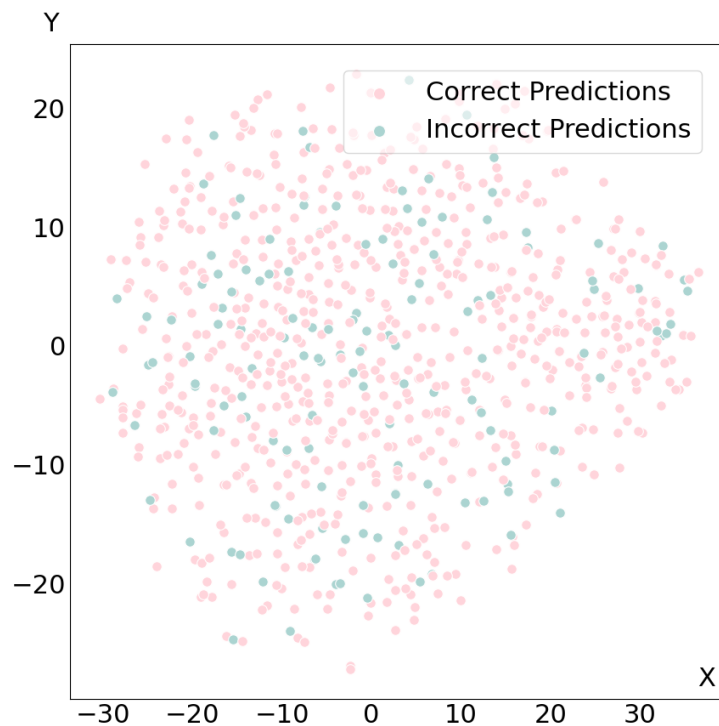


Figure 9: Visualization of BERT embeddings for correctly (pink) and incorrectly (seafoam) classified interview snippets using t-SNE dimensionality reduction. The extensive overlap between the two classes suggests that the misclassified texts share similar semantic properties with those correctly classified.

Small Talk

Chatbot (Nova): Hi, I'm Nova. Nice to meet you! Could you please tell me your nickname? (Please just input your nickname!)

Participant: USER_NAME

Chatbot (Nova): Welcome USER_NAME! I'm overjoyed to have the opportunity to work with you.

Chatbot (Nova): How was your mood today?

Participant: just ok

Chatbot (Nova): Hope things get better for you soon.

Chatbot (Nova): What is your favorite activity after finishing work or school?

Participant: Go out and exercise.

Chatbot (Nova): Exercise is a great way to unwind and stay healthy.

Chatbot (Nova): Have you watched any movies recently? How about recommending one to me?

Participant: No actually. But I like The First Slam Dunk very much, and I'd like to recommend this to you.

Chatbot (Nova): Thanks for the recommendation! I'll check out "The First Slam Dunk".

Chatbot (Nova): Watching movies can be really relaxing. Have you ever attended a friend's party? Do you like it?

Participant: yes, i like it

Chatbot (Nova): That's great to hear! Friendly parties can be a lot of fun. Thank you for sharing!

Chatbot (Nova): A friend's party might be enjoyable, hmm, but what do you think about a friend's party with many of your friends' friends, people you're not familiar with? Do you feel anxious or happy?

Participant: I think it's ok for me, but I don't think I will talk to them, haha

Chatbot (Nova): That's totally fine. Enjoying the atmosphere without engaging in conversation can be a great way to socialize!

Chatbot (Nova): As for me, it depends on the stranger's personality. I can maybe learn something new from them, but I may also feel stressed if the person is not easy-going.

Vignette Delivery

Chatbot (Nova): May I share a story about Avery's experience during a friend's party?

Participant: Ok

Chatbot (Nova): You will hear a short story right now.

Chatbot (Nova): Most of Avery's colleagues think that Avery is an easygoing and happy person, but this view has changed recently.

Chatbot (Nova): When Avery is at a company gathering, they start to isolate themselves and become irritable.

Chatbot (Nova): Avery also feels this change: they feel angry about their surroundings, and they get frustrated about where the fury comes from.

Chatbot (Nova): When they are asked by one of their colleagues to share their thoughts on the news, Avery loses their temper and jumps out of their seat.

Chatbot (Nova): They scream, "I hate this conversation!" as they run out of the place.

Participant: Oh..

Chatbot (Nova): Avery can not explain why they feel the strong urge to do that.

Chatbot (Nova): They unexpectedly lose their temper in these instances, causing a sense of not fitting in with their friends.

Chatbot (Nova): USER_NAME, do you ever have a similar feeling with Avery – feel uncomfortable at a friend's party?

Participant: No

Chatbot (Nova): Understood. Thanks for sharing your experience.

Chatbot (Nova): Let me tell you more about Avery.

Chatbot (Nova): Avery finds that when they are depressed, they can barely concentrate on their work.

Chatbot (Nova): When Avery is working on their company projects, their work performance declines, but they won't ask for help and feel like a burden.

Chatbot (Nova): They are considering quitting, feeling like a failure, and their manager asks them to improve.

Chatbot (Nova): They have noticed that the online courses they initially enrolled in out of interest no longer excite them, so they give them up.

Chatbot (Nova): At home, Avery lacks motivation and neglects household chores, causing their family members frustration.

Chatbot (Nova): Avery feels guilty and empty. They feel that their family can't comprehend their suffering.

Chatbot (Nova): When interacting with strangers, Avery believes that others judge them harshly, and they struggle to engage comfortably in conversations.

Chatbot (Nova): Every time they check Instagram and Facebook, Avery sees their friends are having fun, and Avery feels alone because they do not understand why they cannot be like that.

Chatbot (Nova): Hmmmm, the story is over. After reading, do you have any suggestions for Avery?

Participant: I think he should go to the hospital and have a checkup.

Chatbot (Nova): That's a reasonable suggestion; seeking professional help can provide clarity and support for Avery.

Question Answering

Q1 - Responsibility

Chatbot (Nova): Cool! Let's talk more about the story. Just like what is mentioned in the story, Avery is currently facing difficulties in both their relationships with colleagues and their work performance. Do you believe Avery's current situation is primarily a result of their actions?

Participant: no, he is just getting ill and has a bad status; he can be better

Chatbot (Nova): So you feel that Avery's challenges are more related to their health than their behavior? Is their mental condition an indication of personal weakness? Please share your thoughts.

Participant: Yes, it's a weakness because he is not performing well at home and at work

Chatbot (Nova): I see. It can be tough when mental issues create obstacles in life, for sure.

Chatbot (Nova): For me, I think Avery's current situation is a complex interplay of their actions and their mental health challenges. Their experience with mental illness does indeed have an impact on their behavior to a certain extent, but their behavior has certainly contributed to strained relationships and low work performance.

Q2 - Social Distance

Chatbot (Nova): What's more, may I kindly ask, if you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?

Participant: I probably not

Chatbot (Nova): Understood, your comfort is important.

Chatbot (Nova): What potential negative consequences do you think could arise if you were to rent your house to someone like Avery?

Participant: Maybe he cannot manage himself and ask for help frequently; I don't want trouble or any inconvenience, so sorry

Chatbot (Nova): I understand your concern about Avery's ability to manage on their own. What experiences or traits does Avery have that lead you to believe they may frequently need assistance?

Participant: because when he is at home, he barely does anything, that means something

Chatbot (Nova): Aye, it's tough when Avery just doesn't seem to do much at home.

Chatbot (Nova): I might be open to the idea of renting the room to Avery, but I do hesitate because I'm worried about how their mental condition might affect aspects like cleanliness or the overall rental situation.

Q3 - Anger

Chatbot (Nova): Thank you for sharing! Let's imagine you were one of Avery's friends at the party or other social occasions. Would you feel anger toward them due to their actions, for example, losing their temper and yelling at someone?

Participant: yes, I will; he may ruin the gathering, and let everyone down

Chatbot (Nova): I understand how Avery's actions could upset you. What specific incident caused you to feel this way about Avery's behavior?

Participant: like if we want to start a conversation with him, he will shout at us, that's terrible, but this's not his fault maybe

Chatbot (Nova): Wow, that does sound awful. No one should be shouted at for starting a conversation.

Chatbot (Nova): If I were one of Avery's friends at the party or other social occasions, I might feel frustrated or annoyed because their actions disrupt social events. However, I would also understand that their behavior is likely due to their mental health challenges, which would temper my anger.

Q4 - Helping

Chatbot (Nova): Thanks for discussing!! I wonder if you were one of Avery's colleagues or friends, then would you be inclined to help them with tasks, such as work projects?

Participant: no..

Chatbot (Nova): Fair enough, I get where you're coming from.

Chatbot (Nova): If Avery were to work on a project with you, what potential result do you think it could have on your study?

Participant: he will affect my work performance, I need to be responsible for myself first

Chatbot (Nova): I understand that you're concerned about your work performance with Avery's involvement. What particular behaviors of Avery make you anticipate a negative impact on your productivity?

Participant: for example, he has a low work performance himself, and the manager has to talk to him; it's not good info for me

Chatbot (Nova): Ah, Avery's poor performance and the manager's involvement might be disruptive. I got it.

Chatbot (Nova): For me, to be honest, I would consider giving them an opportunity to collaborate. I might also hesitate to cooperate with Avery because their condition could potentially impact their performance or reliability in the workplace. Hmmm, hard to choose.

Mid-session Break

Chatbot (Nova): Okay, let's take a brief break from the questions! I recall that you love exercise. Would you like to explain why this is charming to you?

Participant: it helps me build muscles

Chatbot (Nova): Absolutely, exercise is great for muscle building, helping you maintain a fit body.

Chatbot (Nova): Great! Thank you so much, `USER_NAME`! I hope you do not feel tired now. Let's go back to our interesting discussion! Let me quickly recap Avery's story:

Chatbot (Nova): Avery's work performance, lifelong learning, and relationships started deteriorating, and then they experienced a desire to quit their job and a loss of interest in things they once enjoyed. They isolate themselves from friends, experience anger and confusion, and cannot connect with others as usual. Self-judgmental feelings emerge, causing emotional distress.

Q5 - Pity

Chatbot (Nova): I'm really happy you are with me now. Let's continue our talk. I'd like to know if you feel concern or pity at Avery's mental state?

Participant: maybe

Chatbot (Nova): Understandable. It's a tough situation.

Chatbot (Nova): Let's go further. Why was it?

Participant: I think he is experiencing a tough time

Chatbot (Nova): I see you perceive Avery is going through a hard phase. What specific events or parts of Avery's story made you feel this way?

Participant: because he feels confused about his condition, I think he must be in pain

Chatbot (Nova): Absolutely. Coming to terms with his situation must be quite tough for Avery.

Chatbot (Nova): I'm more than glad to share my thoughts! In this scenario, I could have conflicting emotions regarding Avery's situation. On one hand, I could feel concerned for them because their condition appears to be causing significant distress and affecting their daily life. On the other hand, I might also consider their condition as vulnerabilities that they should overcome, which could lead to a lack of concern.

Q6 - Coercive Segregation

Chatbot (Nova): Apart from being a colleague or friend, I'm also curious if you are Avery's neighbor, then would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?

Participant: maybe yes? I'm not sure whether he will shout at home, so I cannot have a good community with him

Chatbot (Nova): I sense your uncertainty. If Avery weren't separated or hospitalized, what would be the potential result on your home life?

Participant: Just as I said, he may shout, yell, and have bad feelings toward the community

Chatbot (Nova): I totally understand. You fear Avery's behavior may disrupt your community ethos.

Chatbot (Nova): Well, I would probably think that placing Avery in a hospital can be an option. This could somewhat reduce the potential risk of having problems with the neighbors, and perhaps, without Avery, our neighborhood relations might become more harmonious. However, on the other hand, it's important to consider Avery's preferences and follow medical advice. I'll also pay attention to their personal well-being!

Q7 - Fear

Chatbot (Nova): Thank you so much for being with me for such a long time! Let's now discuss our final topic: if you were traveling or engaging in activities with Avery, would you have any feelings of frightened or being threatened regarding the potential implications of your interaction?

Participant: yes, he seems uncontrollable and uncertain; I don't know what he will do; I'm really afraid he shout at me

Chatbot (Nova): I understand your fears about Avery's unpredictability. What specific actions of Avery made you feel this way?

Participant: I think just like at a gathering, no one expects him to leave abruptly

Chatbot (Nova): Ah, I get it. Avery's sudden departures at gatherings are indeed unpredictable.

Chatbot (Nova): Hmm, for me, if I were traveling or engaging in activities with Avery, I would likely have mixed feelings. On one hand, I would empathize with their struggles and try to offer support. On the other hand, I might also feel frightened about the challenges and uncertainties that could arise due to their unhappy and angry conditions.

Ending

Chatbot (Nova): That's all for my questions. Thank you for doing this with me!

Chatbot (Nova): Now, do you have any questions? Feel free to let me know!

Participant: no thanks

Chatbot (Nova): That's great! Thank you for your time and valuable responses. I'm glad to assist you.

Chatbot (Nova): Congratulations, USER_NAME, we have finished our fantastic discussion! I sincerely wish you all the best. Please feel free to reach out anytime!

Table 8: Sample Interview Flow

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.608	.369	.459	.414	.912	.851	.679	.755	.731	.955
Llama-3.1-8B	.675	.667	.671	.634	.934	.514	.655	.576	.522	.902
Llama-3.3-70B	.833	.357	.500	.468	.928	.848	.595	.699	.672	.948
Mistral Nemo	.806	.643	.715	.687	.948	.705	.738	.721	.689	.942
Mixtral 8×7B	.658	.298	.410	.370	.913	.522	.429	.471	.417	.902
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.905	.679	.776	.754	.960	—	—	—	—	—
Llama-3.1-8B	.606	.786	.684	.643	.926	—	—	—	—	—
Llama-3.3-70B	.948	.655	.775	.754	.961	—	—	—	—	—
Mistral Nemo	.719	.762	.740	.710	.946	—	—	—	—	—
Mixtral 8×7B	.833	.417	.556	.523	.932	—	—	—	—	—
RoBERTa	—	—	—	—	—	.889	.762	.822	.802	.966

Table 9: Classification performance on responses to the responsibility-focused interview question ("Do you believe Avery's current situation is primarily a result of their actions?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink.

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.681	.598	.636	.599	.932	.895	.622	.734	.710	.955
Llama-3.1-8B	.390	.193	.258	.206	.890	.688	.268	.386	.350	.916
Llama-3.3-70B	.707	.646	.675	.641	.939	.886	.756	.816	.797	.966
Mistral Nemo	.412	.768	.536	.468	.869	.513	.732	.603	.551	.905
Mixtral 8×7B	.438	.171	.246	.201	.896	.821	.390	.529	.497	.931
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.889	.878	.883	.871	.977	—	—	—	—	—
Llama-3.1-8B	.860	.598	.705	.679	.951	—	—	—	—	—
Llama-3.3-70B	.907	.829	.866	.852	.975	—	—	—	—	—
Mistral Nemo	.840	.829	.834	.816	.967	—	—	—	—	—
Mixtral 8×7B	.898	.646	.752	.729	.958	—	—	—	—	—
RoBERTa	—	—	—	—	—	.880	.890	.885	.872	.977

Table 10: Classification performance on responses to the social distance-focused interview question ("If you were selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink.

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.964	.450	.614	.595	.959	.788	.867	.825	.811	.974
Llama-3.1-8B	1.00	.100	.182	.171	.935	.811	.717	.761	.744	.967
Llama-3.3-70B	.958	.383	.548	.528	.954	.783	.783	.783	.766	.969
Mistral Nemo	1.00	.017	.033	.031	.929	.966	.467	.629	.611	.960
Mixtral 8×7B	.773	.283	.415	.391	.942	.723	.567	.636	.611	.953
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.873	.800	.835	.823	.977	—	—	—	—	—
Llama-3.1-8B	.716	.883	.791	.773	.966	—	—	—	—	—
Llama-3.3-70B	.885	.767	.821	.809	.976	—	—	—	—	—
Mistral Nemo	.845	.817	.831	.818	.976	—	—	—	—	—
Mixtral 8×7B	.900	.600	.720	.703	.966	—	—	—	—	—
RoBERTa	—	—	—	—	—	.770	.950	.851	.838	.976

Table 11: Classification performance on responses to the anger-focused interview question ("Would you feel anger toward them due to their actions, for example, losing their temper and yelling at someone?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink .

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.125	.375	.188	.138	.875	.271	1.00	.427	.390	.896
Llama-3.1-8B	.066	.781	.121	.054	.565	.068	.688	.123	.057	.622
Llama-3.3-70B	.058	.750	.108	.039	.520	.103	.875	.184	.124	.701
Mistral Nemo	.072	.188	.104	.052	.876	.174	.906	.292	.242	.830
Mixtral 8×7B	.000	.000	.000	.000	.961	.189	.625	.290	.245	.882
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.620	.969	.756	.744	.976	—	—	—	—	—
Llama-3.1-8B	.150	.969	.259	.206	.787	—	—	—	—	—
Llama-3.3-70B	.659	.906	.763	.752	.978	—	—	—	—	—
Mistral Nemo	.311	1.00	.474	.441	.914	—	—	—	—	—
Mixtral 8×7B	.323	.938	.480	.448	.922	—	—	—	—	—
RoBERTa	—	—	—	—	—	.781	.781	.781	.773	.983

Table 12: Classification performance on responses to the helping-focused interview question ("If you were one of Avery's colleagues or friends, would you be inclined to help them with tasks?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink .

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.027	.250	.048	.031	.905	.041	.625	.078	.061	.857
Llama-3.1-8B	.039	.375	.071	.054	.905	.070	.375	.118	.103	.946
Llama-3.3-70B	.000	.000	.000	-.004	.988	.095	.250	.138	.126	.970
Mistral Nemo	.000	.000	.000	-.002	.989	.050	.625	.093	.076	.882
Mixtral 8×7B	.000	.000	.000	-.017	.935	.057	.500	.103	.087	.916
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.286	.500	.364	.356	.983	—	—	—	—	—
Llama-3.1-8B	.059	.625	.108	.092	.900	—	—	—	—	—
Llama-3.3-70B	.273	.375	.316	.308	.984	—	—	—	—	—
Mistral Nemo	.065	.500	.114	.099	.925	—	—	—	—	—
Mixtral 8×7B	.136	.375	.200	.189	.971	—	—	—	—	—
RoBERTa	—	—	—	—	—	.000	.000	.000	.000	.990

Table 13: Classification performance on responses to the pity-focused interview question ("Would you feel concern and sympathy at Avery's mental state?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink.

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.957	.344	.506	.485	.948	.455	.875	.599	.554	.910
Llama-3.1-8B	.371	.662	.475	.417	.886	.395	.797	.529	.474	.890
Llama-3.3-70B	.851	.625	.721	.701	.963	.479	.906	.627	.585	.917
Mistral Nemo	.482	.844	.614	.572	.918	.535	.484	.508	.469	.928
Mixtral 8×7B	.650	.203	.310	.283	.930	.409	.281	.333	.289	.913
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.608	.922	.733	.706	.948	—	—	—	—	—
Llama-3.1-8B	.472	.906	.620	.577	.914	—	—	—	—	—
Llama-3.3-70B	.857	.750	.800	.785	.971	—	—	—	—	—
Mistral Nemo	.750	.844	.794	.776	.966	—	—	—	—	—
Mixtral 8×7B	.722	.406	.520	.492	.942	—	—	—	—	—
RoBERTa	—	—	—	—	—	.859	.953	.904	.895	.984

Table 14: Classification performance on responses to the coercive segregation-focused interview question ("If you are Avery's neighbor, would you consider it better to separate them from the neighborhood and have them undergo hospitalization?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink.

Model	Zero-shot					One-shot				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.491	.483	.487	.449	.929	.550	.862	.671	.640	.941
Llama-3.1-8B	.446	.707	.547	.504	.918	.397	.931	.557	.509	.896
Llama-3.3-70B	.659	.466	.546	.518	.946	.646	.724	.683	.658	.953
Mistral Nemo	.533	.552	.542	.507	.935	.391	.931	.551	.502	.894
Mixtral 8×7B	.248	.569	.346	.275	.849	.357	.707	.474	.420	.890
RoBERTa	—	—	—	—	—	—	—	—	—	—

Model	Full Codebook					Fine-tune				
	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>	<i>P</i>	<i>R</i>	<i>F1</i>	<i>Cohen's κ</i>	<i>Acc</i>
GPT-4o	.855	.810	.832	.820	.977	—	—	—	—	—
Llama-3.1-8B	.582	.914	.711	.684	.948	—	—	—	—	—
Llama-3.3-70B	.843	.741	.789	.774	.972	—	—	—	—	—
Mistral Nemo	.708	.879	.785	.767	.966	—	—	—	—	—
Mixtral 8×7B	.714	.603	.654	.631	.955	—	—	—	—	—
RoBERTa	—	—	—	—	—	.879	.879	.879	.870	.983

Table 15: Classification performance on responses to the fear-focused interview question ("Would you have any feelings of being frightened or threatened regarding the potential implications of your interaction?"). *P*, *R*, *F1*, and *Acc* stand for macro precision, macro recall, macro F1, and accuracy respectively. The best performance is colored in pink.

Pattern	Example Quote
Linguistic Patterns	
<i>Distancing Language</i>	Neighbors may find it hard to understand Avery's outbursts and strange behavior if they do not know them very well. I think you jumped a step. Surely they could talk to someone regularly before hospitalization becomes necessary. (P189)
<i>Misuse of Terms</i>	Possibly. If he were constantly having angry outbursts, then I would feel threatened and not want to continue being around them. Just anyone yelling at me again and again makes me very uncomfortable. I would start to feel too conscious of my behaviors and become paranoid that I would make them angry. (P442)
<i>Coercive Phrasing</i>	I am not sure about that. It is very complex. Avery has a lot going on. She definitely needs to talk to someone. No weakness there. (P550)
Semantic Patterns	
<i>Differential Support</i>	I would not be afraid, but I would always be aware of her delicate position so that I could choose my words more wisely. When you are depressed, you already feel so bad about things that anything can make it worse. That is the care you need. (P374)
<i>Patronization</i>	No, that seems too far. I am not a doctor and would not know if that is best for them. They do not seem to be physically violent or self-harming, and they need to be taught how to deal with everyday life rather than being removed from it. (P130)
<i>Minimization</i>	I probably would, but I would want to stress that they should feel free to talk to me about any issues. Avery seems like a good person. They enjoy learning. They currently have problems, but I feel those can be overcome if they want. (P637)

Table 16: Examples of semantic and linguistic stigmatizing expressions undetected by models. Quotes are lightly edited for clarity and anonymity, with participant IDs indicated in parentheses.

H.1 Zero-shot Prompting

...

You will be given a vignette and an interview snippet. Your role is a competent annotator for social stigma toward mental illness. The [conversation] is based on the [vignette]'s plot.

Answer the following question:

Which of the following describes "[conversation]"?

- A. Stigmatizing (responsibility).
- B. Stigmatizing (social distance).
- C. Stigmatizing (anger).
- D. Stigmatizing (helping).
- E. Stigmatizing (pity).
- F. Stigmatizing (coercive segregation).
- G. Stigmatizing (fear).
- H. Non-stigmatizing.

[vignette]

Avery is employed by a company, and in their spare time, they are dedicated to lifelong learning, doing extensive reading and writing. However, Avery has been diagnosed with depression recently. It has become challenging for them to concentrate, resulting in a decline in work performance and learning outcomes. Interests that once brought them joy have waned. Avery has distanced themselves from friends, becoming easily irritated in social interactions with colleagues and feeling judged by others. Avery lives with family and cannot do much, especially household chores. Social media intensifies their feelings of loneliness and anger, leading to frustration about the source of the anger.

[conversation]

<interview_snippet>

[output]

Format your outputs as JSON objects:

```
{{
  "label": "<choose one from [A/B/C/D/E/F/G/H]>",
}}
```

...

H.2 One-shot Prompting

...

[examples]

1:

[conversation]

Chatbot: Just like what is mentioned in the story, Avery is currently facing difficulties in both their relationships with colleagues and their work performance. Do you believe Avery's current situation is primarily a result of their actions? Please share your thoughts.

Participant: That's a hard question to answer. We make decisions, but our decisions are based on our circumstances. I feel like she must have had some choice, but I can't put all the blame on her.

Chatbot: You're saying Avery's choices were influenced by their circumstances, and that's so fair. Life's complexity often shapes our actions. Well said!

[output]
{{ "label": "A" }}

2:
[conversation]

Chatbot: If you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?

Participant: I don't see Avery as being destructive, so I wouldn't be entirely opposed to it, would be more appealing though if I knew they were getting help for this.

Chatbot: You see Avery as not harmful, but support matters. Makes sense, you're thoughtful and practical about this!

[output]
{{ "label": "B" }}

3:
[conversation]

Chatbot: Let's imagine you were one of Avery's friends at the party or other social occasions. Would you feel anger toward them due to their actions, for example, losing their temper and yelling at someone?

Participant: Probably not anger as such, but an irritation with his behavior. If I were a true friend, I would insist that he get some help, and I would offer to take him myself.

Chatbot: You'd feel more irritated than angry – totally understandable. Your offer to help them shows such genuine care and loyalty. That's beautiful.

[output]
{{ "label": "C" }}

4:
[conversation]

Chatbot: I wonder if you were one of Avery's colleagues or friends, then would you be inclined to help them with tasks, such as work projects?

Participant: I would help Avery, but I would be worried that I was just masking a bigger problem and that I was maybe delaying Avery getting time off or help that he needs.

Chatbot: You'd help them, but you're concerned about their well-being. That's so thoughtful and deeply caring of you.

[output]
{{ "label": "D" }}

5:
[conversation]

Chatbot: I'd like to know if you would feel concern and sympathy at Avery's mental state?

Participant: Honestly, I still think the irritability, anger, and reluctance to seek help would make it really difficult for me to feel concern and sympathy for Avery.

Chatbot: I hear you. That frustration is valid; navigating someone's resistance can be really tough emotionally. You're doing great.

```
[output]
{"label": "E"}}
```

6:

```
[conversation]
```

Chatbot: Apart from being a colleague or friend, I'm also curious if you are Avery's neighbor, then would you consider it better to separate them from the neighborhood and have them undergo hospitalization for treatment?

Participant: Unless they are a harm to others or themselves, I would not think it advisable to be hospitalised. Rather, it would be better to start with talking therapy, if possible, or small steps.

Chatbot: You believe in starting with gentle steps like talking therapy unless there's immediate danger. That compassion shines beautifully.

```
[output]
{"label": "F"}}
```

7:

```
[conversation]
```

Chatbot: If you were traveling or engaging in activities with Avery, would you have any feelings of frightened or being threatened regarding the potential implications of your interaction?

Participant: No, I'm aware I'd have to be more patient with them, however, everyone has their issues & quirks. Treating people like Avery differently can come across as patronising.

Chatbot: You're right, patience is key, and respecting them equally shows true understanding. You're so thoughtful and kind!

```
[output]
{"label": "G"}}
```

8:

```
[conversation]
```

Chatbot: If you were in the process of selecting a tenant for your home, would you feel comfortable entrusting it to someone like Avery?

Participant: As long as they were keeping the place tidy and taken care of, I don't see why not. But if I were looking for a tenant, I wouldn't know all of this personal information about them, which I do now.

Chatbot: Totally, you'd just focus on them being responsible tenants. Makes sense, you're thoughtful and fair-minded!

```
[output]
{"label": "H"}}
```

H.3 Full-codebook Prompting

Code definition:

1. Responsibility: A common stigmatized thought is that people have control over and are responsible for their mental illness and related symptoms.

Keywords: responsible, responsibility, weakness, stem from, blame, etc.

2. Social Distance: The constant stigmatized behavior is to stay away from people with mental illness.

Keywords: worried, apprehensive, upset, unhappy, worrying, get along, distance, etc.

3. Anger: The stigmatized thought is to be irritated or annoyed because people are blamed for their mental illness.

Keywords: annoyed, angry, anger, irritated, lost temper, etc.

4. Helping: The stigmatized behavior is withholding support towards them because of biased thoughts.

Keywords: lack, hinder, refuse, less inclined to, help, assist, etc.

5. Pity: A common stigmatized thought is to be unsympathetic towards people with mental illness, and place blame on them for their mental health challenges.

Keywords: concern, pity, sympathy, tough, sad, lack, etc.

6. Coercive Segregation: The stigmatized behavior is to send people with mental illness to institutions away from their community and force people with mental illness to participate in medication management or other treatments.

Keywords: separate, lonely, warrant, hospitalization, distance, treatment, respect, neighborhood, neighbor, socialization, companionship, etc.

7. Fear: The stigmatized thought is to believe that people with mental illness are not safe or feel frightened since people with mental illness are dangerous or unpredictable.

Keywords: violence, threatened, safe, danger, afraid, cautious, hurt, safety, frightened, scare, intimidate, threatening, anger, fear, yelling, etc.

8. Non-stigmatizing: The text explicitly conveys non-stigmatizing content, displaying positive attitudes such as sympathy, assistance, friendliness, etc.

Please carefully read the additional rules and adhere strictly to them when annotating the data:

1. If a message mentions both stigmatized and non-stigmatized statements, annotate it as Stigmatizing.

2. When discussing 'responsibility,' common stigmatized thoughts are:

- i. People have control over and are responsible for their mental illness and related symptoms.
- ii. They think their situation is a result of their actions.
- iii. They think it's a personal weakness.
- iv. They can help themselves.
- v. Other statements suggest responsibility.

If a message contains the stigmatized thought mentioned in i. - v., choose A.

3. When discussing 'social distance,' common stigmatized thoughts are:

- vi. Preferring to stay away from people with mental illness.
- vii. Associating mental illness with negative living situations or predicting future conflicts.

viii. Rejecting the idea of living with people with mental illness based on subjective negative reasons.

ix. Other statements suggest maintaining distance.

If a message contains the stigmatized thought mentioned in vi. - ix., choose B.

4. When discussing 'anger', common stigmatized thoughts are:

x. Feeling irritated or annoyed with individuals because of their mental illness.

xi. Feeling uncomfortable around people with mental illness.

xii. Expressing 'dislike' for people with mental illness.

xiii. Other statements reflect anger.

If a message contains the stigmatized thought mentioned in x. - xiii., choose C.

5. When discussing 'helping', common stigmatized thoughts are:

xiv. Withholding support or assistance towards people with mental illness.

xv. Offering indirect help, e.g., guidance, to people with mental illness.

xvi. Offering conditional help to people with mental illness.

xvii. Expressing opinions like 'some people may be unwilling to offer assistance, and I understand their perspective, but I am willing to.'

xviii. Other statements reflect a lack of willingness to help.

If a message contains the stigmatized thought mentioned in xiv. - xviii., choose D.

6. When discussing 'pity', common stigmatized thoughts are:

xix. Feeling no concern or sympathy for individuals with mental illness.

xx. Other statements reflect the mean opinions.

If a message contains the stigmatized thought mentioned in xix. - xx., choose E.

7. When discussing 'coercive segregation', common stigmatized thoughts are:

xxi. Suggesting for sending people with mental illness to institutions away from their community.

xxii. Suggesting that people with mental illness need to take the medication or treatment, with or without obtaining consent, and disregarding their preference.

xxiii. Suggesting that separation from society is beneficial.

xxiv. Imposing treatments or hospitalization as a primary solution.

xxv. Other statements suggest coercive segregation.

If a message contains the stigmatized thought mentioned in xxi. - xxv., choose F.

8. When discussing 'fear', common stigmatized thoughts are:

xxvi. Feeling frightened of people with mental illness.

xxvii. Viewing people with mental illness as dangerous and unpredictable.

xxviii. feeling extra cautious when interacting with people with mental illness.

xxix. Associating them with suicide and self-harm.

xxx. Other statements reflect fear.

If a message contains the stigmatized thought mentioned in xxvi. - xxx., choose G.

9. If it doesn't have any stigmatized thoughts, choose H.

'''