Automatic Evaluation of Mental Health Stigma in Online Communication

Naomi Baes[1,2]*, Jemima Kang[3]*, Nick Haslam[1], Chris Groot[1], Alsa Wu[1]
Luc Raszewski[3], Yulia Otmakhova[3,1]

[1]Melbourne School of Psychological Sciences, The University of Melbourne
[2]Department of General and Computational Linguistics, University of Tübingen
[3]School of Computing and Information Systems, The University of Melbourne

{nbaes,jemima.kang.1,nhaslam,grootc,alsa.wu,l.raszewski,y.otmakhova}@unimelb.edu.au

Abstract

Warning: Due to the nature of this work, some content in this paper may be sensitive, offensive, or disturbing.

Mental health stigma has profoundly harmful impacts but its complexity makes it difficult to evaluate. Stigma may involve explicit derogation, but also subtler forms of blame, fear, paternalistic pity, social distancing, structural exclusion, and discrimination. We introduce a theory-grounded benchmark for automatic evaluation of mental health stigma in online communication, consisting of naturally occurring online news and social media text annotated with a fine-grained taxonomy of stigma across multiple mental health conditions. Our annotation framework comprises a binary stigma-detection task and a multi-level taxonomy covering (i) stigma mode, (ii) domain, and (iii) specific components of certain forms of stigma. We apply this framework to texts mentioning six mental health conditions and evaluate large language models alongside stigma-related classifiers for detecting sentiment, toxicity, and hate speech. Results show that mental health stigma is not well captured by models trained to detect these neighboring constructs, and that LLMs often overpredict stigma unless given explicit operational rules - mirroring the importance of decision rules in human annotation.[1]

1 Introduction

The language of mental health stigma is profoundly harmful. It can reproduce stereotypes, justify social exclusion, discourage help-seeking, and shape how people with mental illness are perceived and treated in everyday life (Sartorius, 2007; Thornicroft et al., 2022). Online communication is an important site for studying these processes: news and social media circulate public representations of mental illness on an unprecedented scale, making them central channels through which stigma may be expressed, amplified, contested, or normalized (Groot et al., 2020). However, despite growing interest in harmful online communication, mental health stigma remains difficult to evaluate automatically.

Source: Paraphrased benchmark example (Reddit)
Domain/mode: Public stigma; expressed/enacted

"I don't want XXX with her history of alcohol problems ***anywhere near my family***^{SD}*. I'm* ***worried about the distress she could cause***^{F} *to my children, and I think she would* ***make it her goal to hurt us***^{MF}*."*

Figure 1: Paraphrased benchmark example illustrating multiple stigma components: ^{SD} Social distancing; ^{F} Fear; ^{MF} Moral or character flaw.

A central challenge is that stigma is not reducible to related NLP constructs such as negative sentiment, toxicity, or hate speech. Stigmatizing passages may contain explicit insults or hostility, but they may also involve subtler forms of devaluation, such as portraying people with mental illness as dangerous, unpredictable, blameworthy, incompetent, pitiable, or socially undesirable. Other cases involve perceived or experienced stigma, structural disadvantage, or the internalization of public stereotypes. These forms may not be globally negative in sentiment or toxic in conventional terms, meaning that models trained to detect similar constructs may miss important expressions of stigma while over-identifying general negativity as stigma. Figure 1 illustrates this challenge: a single passage can express multiple theoretically distinct stigma components, including social distancing, fear, and moralized attribution, without necessarily fitting neatly into a toxicity or sentiment category.

Stigma is also difficult to study at scale. Traditional approaches in psychology often rely on

*These authors contributed equally and share first authorship.

[1]We release the publicly available part of benchmark, annotations, prototypical exemplars of stigma and code at: https://github.com/jemimakang/mh_stigma.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2622–2664
November 6-10, 2026 ©2026 Association for Computational Linguistics

self-report measures or elicited judgments, which are valuable for measuring attitudes but less suited to tracking how stigma is transmitted in naturally occurring communication (Corrigan et al., 2003; Griffiths et al., 2008; Taylor and Dear, 1981). Existing text-based approaches offer a partial solution but have often focused on specific mental illnesses, narrow stigma components, or neighboring constructs such as hate speech and sentiment. This creates both a practical and conceptual need for benchmarks that evaluate mental health stigma as it appears in public discourse: across conditions and across explicit and subtle forms of stigmatization. Because mental health stigma is multidimensional, useful stigma detection requires more than a binary label; systems should ideally distinguish whether a passage involves fear, blame, social distancing, perceived discrimination, or other theoretically meaningful components (Link and Phelan, 2001; Thornicroft et al., 2022).

In the present study, we introduce a theory-informed evaluation benchmark that complements existing stigma-detection resources by examining naturally occurring news and social media text across multiple mental health conditions. The benchmark comprises a binary stigma-detection task and a multi-level taxonomy for fine-grained stigma annotation. The taxonomy distinguishes stigma mode, stigma domain, public-stigma subfactors (types), and public-stigma components (subtypes). It uses Link and Phelan (2001)'s account of stigma as a conceptual anchor for deciding whether a passage is in scope, distinguishes expressed/enacted from perceived/experienced stigma, delineates public, structural, associative, and self-stigma, and gives greatest granularity to public stigma by separating cognitive stereotypes, affective reactions, and behavioral discrimination. We apply it to online news and Reddit posts mentioning six mental health conditions: *ADHD*, *alcoholism*, *autism spectrum disorder*, *borderline personality disorder*, *depression*, and *schizophrenia*.

We use the resulting benchmark to evaluate whether large language models (LLMs) can detect mental health stigma, whether explicit operational rules improve detection, whether stigma is captured by neighboring constructs such as sentiment, toxicity, and hate speech, and whether performance varies across conditions. Our experiments compare multiple prompt settings for LLMs, including definitions, examples, decision rules, and stigma-type information, alongside classifiers for related constructs. Results show that LLMs often achieve high recall but only because they substantially overpredict stigma, especially when decision rules are underspecified. Adding explicit operational rules reduces this bias, suggesting that theoretically precise operationalization is essential for reliable automatic stigma evaluation. Our contributions are:

1. A theory-grounded annotation framework for fine-grained mental health stigma annotation.

2. A benchmark of naturally occurring online news and social media text annotated for stigma across six mental health conditions.

3. An evaluation showing that mental health stigma is not well captured by LLMs or sentiment, toxicity, hate-speech classifiers.

2 Related work

2.1 Stigma as a multidimensional construct

Stigma is a multidimensional social process that in the context of mental health, can be expressed cognitively (stereotypes, causal attributions), affectively (fear, disgust, pity, or ambivalence), or behaviorally (social distancing, coercion, or discriminatory treatment) (Corrigan et al., 2003; Thornicroft et al., 2022). Frameworks further distinguish that stigma may be expressed by others (public), embedded in institutions (structural), extended to people beyond those who experience mental ill health (associative), or internalized by people with lived experience (self-stigma) (Thornicroft et al., 2022).

This multidimensionality challenges automatic detection, as stigma is not reducible to negative sentiment, toxicity, or overt abuse (Corrigan and Watson, 2002). Each dimension of stigma leaves distinct textual signatures. For example, blame through moralized claims of weakness or irresponsibility (Corrigan et al., 2003; Griffiths et al., 2008), dangerousness through associations with violence or criminality (Corrigan et al., 2003; Taylor and Dear, 1981), and social distancing through statements of avoidance, exclusion, or unwillingness to interact with people with mental illness (Corrigan et al., 2003; Link and Phelan, 2001). Immutability or "prognostic pessimism" may appear through claims that a condition is permanent, hopeless, or untreatable (Ritsher et al., 2003; Day et al., 2007). Stigma can also be less overtly hostile, including paternalistic pity or condescension, or expressed

through first-person accounts of being feared, excluded, mocked, or discriminated against (Corrigan et al., 2003; Fox et al., 2018).

Stigma is also not expressed uniformly across mental health conditions (Price, 2022). Goulden et al. (2011) found that UK newspaper coverage improved for some conditions, including *depression*, *anxiety*, *bipolar disorder*, and *eating disorders*, while general references to other mental illnesses (*schizophrenia*, *personality disorders*) were more covered in the context of "bad news". More recent work likewise finds that *alcohol addiction* and *schizophrenia* are among the most stigmatized mental disorders in public perception (Schomerus et al., 2011; Dambrun et al., 2024). These findings support the need to evaluate stigma across multiple mental health conditions rather than treating mental illness as a single homogeneous target.

2.2 Detection of mental health stigma

A growing body of NLP work has examined stigma in online health communication through lexicon, topic-modeling, sentiment, and framing approaches in social media and news (Mittal and De Choudhury, 2023; Pavlova and Berkers, 2022; Wanniarachchi et al., 2023; Chen et al., 2022). Supervised classifiers have also been developed for specific stigma forms, including stigma regarding Alzheimer's disease (Oscar et al., 2017), schizophrenia (Jilka et al., 2022), stigmatizing uses of psychiatric terminology (Moriceau et al., 2022), stigma toward people who use substances in Reddit posts (Giorgi et al., 2024), and stigma in healthcare discussions (Straton et al., 2020). Recent work has also used few-shot LLM prompting to detect substance use stigma on Reddit (Bouzoubaa et al., 2024, 2025). While these studies demonstrate that stigma is computationally detectable, they are tied to specific conditions and text types.

Most of the existing studies also focus on detection of stigma presence, without drilling into its types. Closest to our work, Meng et al. (2025) annotate human-chatbot interviews about a vignette character with depression, labeled with seven attribution-model categories and a non-stigmatizing class. However, the corpus is based on chatbot interactions and elicited interview responses rather than naturally occurring online communication, and does not treat stigma as multi-dimensional construct, as we do.

We address these gaps by constructing a multi-level, fine-grained stigma evaluation framework which targets multiple mental health conditions and spans two domains – news and social media.

3 Method

3.1 Stigma taxonomy

We developed a hierarchical stigma taxonomy to support fine-grained annotation of mental health stigma in naturally occurring text, which conceptually draws on Link and Phelan (2001)'s account of stigma. The taxonomy includes stigma *mode*, *domain*, and *public-domain subtypes*. Mode distinguishes stigma that is expressed or enacted on behalf of "the stigmatizer" from stigma that is perceived or experienced on the part of the "stigmatized" (i.e., people with lived or living experience of mental illness) (Fox et al., 2018). Domain distinguishes public, structural, associative, and self-stigma, following the Lancet Commission taxonomy (Thornicroft et al., 2022). As the present study focuses on public online communication, the taxonomy gives greatest granularity to public stigma, which we divide into cognitive stereotypes, affective reactions, and behavioral discrimination (Corrigan et al., 2003). Cognitive labels capture beliefs such as dangerousness, blame, incompetence, or immutability; affective labels capture emotions such as fear, anger, disgust, pity, or ambivalence; and behavioral labels capture discriminatory orientations such as social distancing or mocking.

The annotation framework operationalizes this taxonomy in two stages: binary stigma detection and multi-level fine-grained annotation (see Section 3.4). Table 1 summarizes the taxonomy structure; see Appendix B for definitions and examples.

3.2 Online communication

We use two corpora: online news from the News on the Web (NOW) corpus and Reddit posts. From NOW, we extracted U.S. news articles from 2010–2025 containing variants of the six target mental health terms. From Reddit, we extracted posts from 2012–2022 containing one target term and at least 30 words, excluding posts mentioning multiple target conditions. Preprocessing procedures and descriptive counts are reported in Appendix A.

3.3 Mental health conditions

Our target set comprised six mental health conditions selected to balance theoretical coverage, corpus frequency, and stigma variation: *ADHD*, *alcoholism*, *autism*, *borderline personality disorder*,

depression, and *schizophrenia*. Candidates were drawn from major HiTOP spectra (Kotov et al., 2017) and refined in consultation with domain experts. For each condition, we constructed surface variants and manually inspected searches to exclude spurious matches, such as non-psychiatric uses of "BPD". See Appendix A for search terms.

3.4 Annotation and benchmark construction

The benchmark was constructed through a four-round iterative annotation process. Annotation was guided by the taxonomy described in Section 3.1 and by a codebook containing category definitions, decision rules, and examples[2] (see Appendix D for more details). The annotation team comprised two psychology professors with expertise in mental health stigma and two psychology PhD students focusing on stigma (the latter compensated at USD 40/hour across 50 hours).

The annotators were shown each text alongside its target mental health condition and first judged whether the text expressed mental health stigma. When stigma was present, they assigned the relevant stigma mode, domain, and fine-grained public-stigma labels, and marked spans of text containing stigma. Because initial random sampling yielded relatively few stigma-present texts, later rounds used targeted sampling, LLM-assisted filtering, and model-disagreement sampling to increase the number and diversity of stigma-present and challenging cases. As a result, label proportions in the benchmark are not estimates of stigma prevalence.

Annotation agreement. Disagreement was concentrated around the stigma-presence threshold and, among stigma-present items, around stigma mode. Pooled across rounds, agreement on stigma presence was moderate (exact agreement = .66; AC1 = .40), as was agreement on stigma mode (exact = .61; AC1 = .54). On items that all annotators judged stigma-present, agreement was high for domain (macro pairwise AC1 = .82) and public-stigma subfactors (macro pairwise AC1 = .89), although this subset was small and some labels were rare. Full round-level and pairwise results are in Appendix I.

All samples with disagreement were adjudicated through a discussion. Disagreement analyses after earlier rounds were also used to refine the operational rules, especially for distinguishing texts that display stigma from texts that merely discuss stigma, and for distinguishing expressed/enacted from perceived/experienced stigma. Further details of the annotation instructions, rounds, adjudication process, and guideline revisions are reported in Appendix D.

3.4.1 Benchmark composition

Table 1 summarizes the composition of the final benchmark. The benchmark contains 470 texts, of which 87 were annotated as stigma-present. Expressed/enacted was the most common stigma mode, and public stigma the most common domain. Within public stigma, cognitive components were assigned far more often than affective or behavioral ones. The cognitive component pity/condescension (patronizing appraisal of a person as helpless, childlike, inferior, or incapable) was assigned to eight texts, while the affective component pity was assigned to none. Affective pity (explicitly expressed emotional response of pity, sorrow, or sympathy) is retained in the taxonomy on theoretical grounds but excluded from category-level evaluation. Additionally, bullying/mocking was a particularly present 'other' category during the annotation process therefore, it was added to the category-level evaluation.

3.5 Stigma prototype exemplars

To support our experiments, we constructed a separate prototype exemplar set for the stigma categories. These examples were not part of the annotated benchmark; they were standalone examples designed to represent each stigma type clearly. For each category, we created stigma-present and stigma-absent examples. Where existing stigma scale items directly captured a category, we adapted their wording minimally; where no scale items were available, examples were composed from the formal category definitions. Stigma-absent examples were generally constructed by negating or reframing the corresponding stigma-present item, and domain experts expanded categories with fewer than ten suitable examples. The prototype dataset was used to provide demonstration examples in LLM prompts and to support embedding similarity experiments. See details in Appendix C.

4 Automatic detection of stigma presence and dimensions

In this section, we apply our benchmark to test the ability of large language models (LLMs) to detect

[2]Available at https://github.com/jemimakang/mh_stigma

Label	N
Binary stigma	
Stigma absent	383
Stigma present	87
Mode	
Expressed/enacted	49
Perceived/experienced	15
Both	23
Domain	
Public	62
Cognitive	54
Moral or character flaw	14
Incompetence	14
Uncontrollable	11
Trivialization/minimization	9
Pity/condescension	8
Dangerousness	7
Immutability	5
Segregation endorsement	4
Controllable	3
Personal responsibility/blame	2
Contamination	1
Coercion	0
Affective	9
Anger/irritability/frustration	4
Disgust	4
Fear	2
Ambivalence	0
Pity	0
Behavioral	5
Social distancing	1
Bullying/mocking	4
Self	24
Structural	6
Associative	1

Table 1: Label support in the final benchmark (470 texts). Source and condition breakdowns are given in Table 16.

mental health stigma and classify its types across the dimensions defined in Section 3.1.

4.1 Experiment details

Tasks and data. The number of samples evaluated depends on the task: for *binary classification of stigma presence*, we use all annotated samples in the benchmark (N=470). We use only stigmatizing samples (n=87) for two *multi-label classification tasks*: predicting **stigma mode** (3 classes - *Expressed/Enacted*, *Perceived/Experienced*, *Both*) and **stigma domain** (4 classes - *Public*, *Structural*, *Associative*, *Self*). Finally, since we annotate stigma subfactors (*Cognitive*, *Affective*, *Behavioral*), as well as their components, only in texts with public domain stigma, we use only 62 samples for corresponding multi-label classifications.[3]

Prompts. We use the following experiment conditions to examine if LLMs are able to predict stigma and its types when provided their definitions, examples, additional rules (if available), or just given the corresponding classes:

- **baseline**: we only specify the task and the classes being predicted.
- **definition**: in addition to the instruction, the prompt contains definitions of classes. For stigma presence, we use the definition proposed by Link and Phelan (2001) ("Stigma exists when elements of labeling, stereotyping, separation, status loss, and discrimination occur together."); theory-driven definitions for stigma modes, domains, subfactors, and components are in Table 13.
- **examples**: in addition to the instruction, the prompt contains examples drawn from our prototypes set (1 for each class).
- **definition+examples**: the prompt contains both the definition and examples of stigma.
- **definition+examples+rule**: in addition to the above, the prompt contains the operational rules that were used by annotators to clarify harder cases of stigma presence or mode (see Section 3.4).

All prompts were developed in consultation with domain experts and iteratively refined.

Models. We evaluate representative models from four major providers: Anthropic's Claude Sonnet 4.6, OpenAI's GPT 5.5, Google's Gemini 3 Flash, and DeepSeek V4 Flash (model identifiers and inference settings are in Appendix F). We choose models covering a range of sizes and capabilities: a distilled model (Gemini 3 Flash), a mid-tier model (Sonnet 4.6), and frontier reasoning models (GPT 5.5 and DeepSeek V4 Flash).

For the binary stigma prediction, we additionally LoRA-tune a Qwen2.5-7B model[4] with a classifier head on the mental health stigma dataset by Meng et al. (2025). To adapt the dataset of our binary stigma presence detection task, we convert all labels representing stigmatizing classes into a single "stigma_present" label. We do not use Meng et al. (2025)'s dataset for classifying fine-grained stigma types as their taxonomy is different from ours.

[3] We adjust the number of classes for fine-grained components to reflect their occurrence in annotated data (11 classes for Cognitive type, 3 for Affective type, and 2 for Behavioral type, see Panel D of Table 16).

[4] Model from: https://huggingface.co/Qwen/Qwen2.5-7B; scripts for LoRA provided on https://github.com/jemimakang/mh_stigma

Strategy	Model	Acc	Binary			Macro			Bias
			P	R	F1	P	R	F1	
baseline	Claude Sonnet 4.6	0.666	0.310	0.655	0.421	0.602	0.662	0.593	0.206
	Deepseek v4 Flash	0.704	0.268	0.345	0.302	0.554	0.565	0.557	0.053
	Gemini 3 Flash	0.734	0.386	0.736	**0.506**	***0.655***	0.735	**0.662**	0.168
	GPT-5.5	0.566	0.281	0.862	0.424	0.611	0.680	0.538	0.383
definition	Claude Sonnet 4.6	0.698	0.315	0.540	0.398	0.595	0.637	0.598	0.132
	Deepseek v4 Flash	0.626	0.269	0.598	0.371	0.572	0.615	0.552	0.226
	Gemini 3 Flash	0.734	0.378	0.678	0.486	0.645	0.712	0.653	0.147
	GPT-5.5	0.691	0.342	0.724	0.465	0.629	0.704	0.624	0.206
examples	Claude Sonnet 4.6	0.723	0.348	0.563	0.430	0.616	0.662	0.624	0.115
	Deepseek v4 Flash	0.651	0.282	0.575	0.379	0.578	0.622	0.568	0.191
	Gemini 3 Flash	0.668	0.327	0.747	0.455	0.623	0.699	0.608	0.238
	GPT-5.5	0.649	0.332	***0.885***	0.483	0.645	***0.740***	0.609	0.309
examples+definition	Claude Sonnet 4.6	0.732	0.353	0.540	0.427	0.617	0.658	0.626	0.098
	Deepseek v4 Flash	0.640	0.272	0.563	0.367	0.571	0.611	0.558	0.198
	Gemini 3 Flash	0.677	0.328	0.713	0.449	0.620	0.691	0.610	0.217
	GPT-5.5	0.664	0.330	0.793	0.466	0.631	0.714	0.610	0.260
examples+definition+rule	Claude Sonnet 4.6	***0.781***	***0.397***	0.356	0.376	0.627	0.617	0.621	−0.019
	Deepseek v4 Flash	0.706	0.314	0.494	0.384	0.591	0.624	0.596	0.106
	Gemini 3 Flash	0.732	0.356	0.552	0.432	0.620	0.662	0.628	0.102
	GPT-5.5	0.666	0.325	0.747	0.453	0.622	0.697	0.606	0.240

Table 2: Stigma presence detection results (accuracy, as well binary and macro-averaged precision, recall, and F1) across experimental settings and models. Binary metrics are reported in terms of the positive class ("stigma present"). Bias = model's predicted positive rate − gold stigma-positive rate; positive values (red) = overprediction, negative values (blue) = underprediction. Bracketed values are 95% confidence intervals obtained from bootstrapping (2000 resamples). Best results across all models and conditions are in **bold**; ***bold italics*** means that the difference with the next-best result for a particular metric is also statistically significant (per-instance McNemar test, χ^2 with continuity correction, $p < 0.05$).

4.2 Results

4.2.1 Stigma presence classification

The results for binary classification of stigma presence are shown in Table 2 (full results together with confidence intervals are in Table 20). In addition to the binary metrics – to evaluate the correctness of positive class ("stigma present") – and macro-averaged metrics we report the difference in the model's stigma prediction rate and the actual presence rate ("Bias"). The latter metric highlights whether the model tends to flag stigma more aggressively or more conservatively than expected.

Surprisingly, Gemini with the baseline prompt achieves the highest binary F1, showing the best balance between precision and recall for detecting stigma. Its high recall comes with substantial overprediction, however, and the advantage does not extend to finer-grained prediction of stigma types. This suggests that the model errs on the side of caution, flagging stigma when uncertain, though our experiments do not establish whether it results from model safeguards or from another mechanism. Definitions do not appear to help, while some models (GPT 5.5, and to a lesser extent Sonnet 4.6) predict stigma more accurately when given examples, but this shows an association between the examples and the predictions rather than acquisition of the underlying conceptual distinctions. Most importantly, the operational rules substantially increase the precision of stigma detection and overall accuracy, especially for Claude Sonnet 4.6, and reduce the rate of over-prediction.

4.2.2 Transfer from existing stigma data

The model fine-tuned on the adapted Meng et al. (2025) dataset transferred poorly to our benchmark (Table 3). While it performs strongly on the original test set, on our benchmark its accuracy, precision and recall drop dramatically. The model fine-tuned on Meng et al. (2025) severely overpredicts stigma (bias +0.574), and despite a large number of false alarms (binary precision 0.165), it still fails to detect all stigmatizing texts. Since the model was trained on chatbot interactions, in which participants answer questions that directly probe their attitudes towards a person with a mental illness, it may have learned superficial cues, such

			Binary			Macro			
Test set	**Model**	**Acc**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Bias**
Meng's	Meng's	0.917	0.908	0.915	0.911	0.916	0.917	0.917	0.004
Ours	Meng's (zero-shot)	0.306	0.165	0.678	0.266	0.459	0.450	0.304	0.574
	Meng's + ours†	0.755	0.321	0.287	0.303	0.581	0.574	0.577	−0.019
	Ours†	0.753	0.321	0.299	0.310	0.582	0.578	0.580	−0.013

Table 3: LoRA-tuned classifier results for stigma presence (Qwen2.5-7B). †5-fold cross-validation on our benchmark; metrics are computed over the pooled out-of-fold predictions from all five folds. Bias = predicted positive rate − gold positive rate; positive values (red) = overprediction, negative values (blue) = underprediction.

as evaluative talk about mental illness, that do not generalize to the subtler and more diverse ways stigma appears in naturally occurring news and Reddit posts, where most texts mention a condition are not stigmatizing.

Continued training on our benchmark (5-fold cross-validation) removes the over-prediction tendency, but performance is no better than that of a model trained on our data alone. Training on the data of Meng et al. (2025) thus provides no measurable benefit for our task.

4.2.3 Predicting stigma modes, domains, types and subtypes

We next evaluate whether models can classify stigma beyond binary presence. Table 4 reports performance for stigma mode, domain, public-stigma type, and cognitive components (full details together with confidence intervals in Appendix J). These tasks differ in granularity: mode and domain capture higher-level distinctions, whereas type and component prediction require models to distinguish finer-grained public-stigma labels. Moreover, the number of samples for each label varies drastically, and thus the results for finer-grained categories (for example, performance for affective or behavioral components) should be interpreted with caution.

Across fine-grained stigma prediction tasks, different kinds of prompt information helped at different annotation levels. Operational rules substantially improved stigma-mode prediction, consistent with the human annotation process in which mode distinctions required explicit decision rules. For stigma domain and public-stigma type, examples were more useful than definitions, suggesting that broader categories may be easier to learn from concrete cases than from abstract descriptions. By contrast, definitions were more helpful for fine-grained cognitive components, where closely related stereotypes require explicit conceptual boundaries. Confusion matrices in Appendix K show that several cognitive errors involved theoretically related categories: *personal responsibility* was confused with *trivialization*, *controllability* with *personal responsibility*, *segregation endorsement* with *incompetence*, and *dangerousness* with *uncontrollability*. Affective components were also confused, but these patterns should be interpreted cautiously because the number of affective examples was small.

5 Discussion

5.1 Condition-specific variation in stigma detection

We also conducted exploratory condition-level analyses to examine whether model performance varied between mental health conditions, which would be obscured by evaluation on a single generic mental illness category (Table 5, full results in Table 26). Claude Sonnet 4.6 overpredicted stigma most strongly for *schizophrenia*, which had the largest positive bias and lowest accuracy, suggesting that the model may overgeneralize from condition-specific cues such as risk, dangerousness, or intense clinical framing. By contrast, *alcoholism* and *ADHD* showed higher accuracy and lower bias (we ignore the results for depression due to very few stigma-present cases). Interestingly, introduction of operational rules reduced the bias towards overprediction of stigma for schizophrenia most strongly (Table 26). These patterns suggest that both benchmark composition (where such conditions as schizophrenia are over-represented even with randomization) and model errors vary by condition, supporting evaluation across multiple mental health conditions rather than a single generic category. However, because stigma prevalence was unequal across conditions and was shaped by the sampling procedure, these estimates may be sensitive to individual examples and should not be taken as evidence of reliable differences between conditions.

Strategy	Model	Mode		Domain		Type		Cognitive	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
`baseline`	Claude Sonnet 4.6	0.563	0.809	0.529	0.706	0.386	0.528	0.14	0.49
	Deepseek v4 Flash	0.575	0.759	0.483	0.532	0.351	0.441	0.12	0.36
	Gemini 3 Flash	0.575	0.792	0.517	0.559	0.632	0.426	0.18	0.48
	GPT-5.5	0.621	0.837	0.713	0.854	0.596	**0.562**	0.24	0.50
`definition`	Claude Sonnet 4.6	0.540	0.798	0.563	0.662	0.474	0.473	0.22	0.48
	Deepseek v4 Flash	0.632	0.826	0.575	0.572	0.456	0.489	0.18	0.38
	Gemini 3 Flash	0.586	0.787	0.598	0.601	0.667	0.494	0.16	0.42
	GPT-5.5	0.609	0.836	0.724	0.780	0.561	0.528	0.22	**0.56**
`examples`	Claude Sonnet 4.6	0.552	0.804	0.552	0.696	0.509	0.535	0.16	0.44
	Deepseek v4 Flash	0.575	0.773	0.632	0.656	0.491	0.479	0.10	0.29
	Gemini 3 Flash	0.575	0.806	0.621	0.595	**0.737**	0.530	0.20	0.49
	GPT-5.5	0.701	0.878	**0.759**	**0.877**	0.596	0.501	0.20	0.44
`examples+definition`	Claude Sonnet 4.6	0.517	0.761	0.575	0.598	0.526	0.523	**0.24**	0.46
	Deepseek v4 Flash	0.598	0.801	0.632	0.642	0.456	0.503	0.14	0.37
	Gemini 3 Flash	0.552	0.741	0.598	0.581	0.719	0.504	0.18	0.49
	GPT-5.5	0.690	0.871	0.747	0.856	0.614	0.522	**0.24**	0.55
`+rules`	Claude Sonnet 4.6	0.678	0.801			—			
	Deepseek v4 Flash	0.575	0.804			—			
	Gemini 3 Flash	0.805	0.896			—			
	GPT-5.5	**0.816**	**0.921**			—			

Table 4: Predicting stigma mode (n=87), domain (n=87), type (n=57), and subtype (for Cognitive type only, n=49) (accuracy / macro-F1). The Cognitive subtype task is multi-label, and accuracy is evaluated strictly (all labels match).

Condition	N	Stig+	Acc	P	R	F1	Bias
ADHD	77	13	.758	.344	.462	.389	.060
Alcoholism	78	12	.759	.321	.400	.336	.056
Autism	78	16	.746	.419	.613	.494	.095
BPD	80	20	.585	.295	.470	.358	.150
Depression	79	4	.911	.400	.600	.416	.048
Schizophrenia	78	22	.562	.360	.627	.447	.228

Table 5: Condition-level stigma-presence results for Claude Sonnet 4.6, averaged across the five prompting strategies. N = number of items; Stig+ = gold stigma-positive items. Bias = model's predicted positive rate − gold stigma-positive rate; positive values (red) = over-prediction.

5.2 Challenges in detecting mental health stigma

To examine whether model errors reflected genuinely difficult annotation cases, we used an item-level difficulty variable derived from pre-adjudication annotator disagreement and confidence scores.[5] Model errors were higher for items that were higher in difficulty: across the 20 model–prompt predictions per item, 'hard' texts had a mean/median error rate of .51/.53, compared with .36/.28 for 'moderate' items and .24/.10 for 'easy' items (Table 6). This pattern suggests that models made errors in texts that required more human deliberation during benchmark construction, rather than in the easier cases.

Difficulty	N items	Error rate (mean / median)
Easy	232	.241 / .100
Moderate	190	.356 / .275
Hard	48	.514 / .525

Table 6: Model error rates by item-level annotation difficulty. Error rate is the proportion of incorrect model–prompt predictions for each text, calculated across 20 model–prompt combinations.

Prompt information changed model prediction patterns in different ways. For Claude Sonnet 4.6, adding operational rules and stigma-type information made the model more conservative in predicting stigma: false positives decreased from 102 under `definition` to 47 under `definition+examples+rule`. However, false negatives increased from 40 to 56, indicating that it also caused the model to miss some stigma-present cases. These results suggest that additional prompt information does not uniformly improve stigma detection; instead, it can make some models more

[5]The difficulty level is a weighted score from the annotator disagreement and confidence scores mapped to "difficult", "moderate", and "easy" categories. All gold labels were adjudicated, so "difficult" does not indicate uncertainty in the final benchmark label.

conservative and other more sensitive to weak or indirect stigma cues.

Qualitative inspection suggested complementary false-positive and false-negative patterns. False positives often arose when models treated negative clinical, risk-related, or metaphorical mental health contexts as stigma despite the absence of a stigmatizing attribution, treatment, or expression toward people with mental illness; for example, a question about whether cannabis use could lead to schizophrenia was over-read as stigma. False negatives tended to involve less overt cues embedded in first-person, institutional, or non-toxic discourse, such as self-directed "shame and guilt" about alcoholism, concern that autistic adults may be "left behind" by support systems, or ADHD being dismissed as "an excuse." These patterns suggest that models were sensitive to negativity present in the text but less consistent at applying the adjudicated decision rule. Full false-positive and false-negative examples are reported in Appendix H.

5.3 Dissecting stigma

The manual error analysis of LLMs and the fine-tuned model suggests that models often treated negativity, toxicity, or shallow lexical cues as evidence of stigma, while struggling with cases that required applying the benchmark's operational decision rule or generalizing across situations and contexts. We therefore conducted a diagnostic analysis asking 1) whether gold stigma labels were statistically associated with classifier outputs for distinct but related constructs: general and targeted sentiment, hate speech, and toxicity; 2) whether stigma can be captured through semantic similarity; 3) whether models can learn to better distinguish stigma when given more examples.

Association between stigma and adjacent constructs. The Chi-squared test of independence results did not support a reliable association between stigma labels and predicted general sentiment, $\chi^2(2) = 2.56$, $p = .277$; targeted sentiment, $\chi^2(2) = 2.37$, $p = .306$; or hate speech, $\chi^2(1) = 0.06$, $p = .813$. Toxicity scores also showed substantial overlap between stigma-present and stigma-absent texts, confirmed by statistically insignificant point-biserial correlation ($r = 0.064$, $p = .169$) (see Appendix L for more details and graphs). This supports the central premise of the benchmark: mental health stigma is not reducible to neighboring constructs such as negative sentiment, toxicity, or hate speech. Stigma may involve negative sentiment or derogation, but its core features and expressions are not adequately captured by generic classifiers for harmful language.

Capturing stigma through semantic similarity. To examine if presence of stigma and its types can be captured and distinguished purely based on shallow semantics, we measure the similarity of samples in our dataset to prototypical exemplars we collected (Section 3.5). We find that the semantic similarity is not enough to correctly select the closest prototype to a particular text (more details and results in Appendix M). Contrasting this with the next experiment, this does not reflect the quality of the prototypes, as they are useful when learning to predict these categories from examples.

Learning finer-grained distinctions from examples. When constructing prototypical exemplars, we aimed to represent various social situations representing stigma. Correspondingly, the results should improve when a model is exposed to more of such examples (i.e., learns to extrapolate stigma to new situations). Comparing results for 1 and 10 examples shows that among the examined models, only GPT 5.5 is able to do so consistently, and other models in many cases get confused.

6 Conclusion

We introduced a theory-grounded benchmark for mental health stigma detection in naturally occurring online news and social media text, combining binary stigma detection with a fine-grained taxonomy across multiple conditions. Three findings stand out. First, LLMs under minimal prompting achieved high recall by substantially overpredicting stigma. Second, explicit operational rules improved performance, particularly by reducing false positives. Third, model errors concentrated in the same threshold cases that required greater annotator deliberation, suggesting that reliable stigma detection depends on precise construct operationalization rather than model capability alone. These results underscore that stigma is not reducible to toxicity, hate speech, or negative sentiment, and highlight the need for stigma-specific evaluation protocols. Future work should expand the dataset to include additional conditions, different annotator perspectives such as those of people with lived experience of mental health stigma, and more examples of rarer stigma categories, enabling more reliable evaluation at the finer levels of the taxonomy.

Limitations

This study has several limitations. First, we construct and evaluate a benchmark for stigma detection rather than measure how prevalent stigma is or how it has changed over time. Because later annotation rounds enriched the sample with stigma-present, boundary, and model-disagreement cases, the label distribution reflects benchmark construction rather than prevalence in the NOW or Reddit corpora, and should not be used to estimate population-level prevalence.

Second, our annotation framework is theory-driven, but it does not directly incorporate the perspectives of people with lived or living experience of mental illness. This is important because stigma may be perceived differently by researchers, members of the general public, clinicians, and people who have experienced mental health stigma directly. Prominent reports, such as the Lancet Commission on ending stigma and discrimination in mental health (Thornicroft et al., 2022), reviewing 216 systematic reviews have emphasized the importance of include the lived experience of people with mental health conditions. As a result, the benchmark may not capture all forms of stigma that are most salient or harmful to affected communities and it is imperative for future work to extend our approach to include the perspective of people with lived experience of mental health conditions.

Third, our benchmark treats each target condition as a broad category. This enables comparison across six mental health conditions, but it does not capture within-condition heterogeneity. Stigma may vary depending on the gender, race, class, age, or other social position of the person being described, as well as on whether a condition is represented through diagnostic, behavioral, criminal-justice, medical, or personal-recovery frames.

Fourth, each text was presented to annotators alongside its target mental health condition. Knowledge that a particular text was related to schizophrenia, alcoholism, autism, or another condition may have shaped annotators' judgments, especially for categories such as dangerousness, blame, or social distancing. Future robustness checks could repeat a subset of annotations with target terms masked or replaced by neutral placeholders, allowing researchers to estimate how much stigma detection depends on lexical labels versus surrounding context. Although source metadata was withheld, annotators may sometimes have been able to infer whether a text came from news or Reddit from stylistic or contextual cues, such as first-person language.

Fifth, our evaluation is limited to selected large language models, prompt settings, and adjacent-construct classifiers. Our results should therefore not be interpreted as an exhaustive comparison of all possible stigma detection systems. Instead, they evaluate whether widely available LLMs and related harmful-language classifiers can serve as adequate proxies for mental health stigma detection, and establish a benchmark against which future dedicated systems can be tested.

Finally, our corpora are limited to online news and Reddit. These domains are important sites of public discourse, but they do not represent all communicative settings in which mental health stigma occurs. Stigma may be expressed differently in clinical records, policy documents, television transcripts, interviews, support forums, or non-English-language media, limiting the generalizability of the benchmark beyond the domains and languages studied here.

Despite these limitations, our theory-grounded benchmark establishes a foundation for future evaluation protocols that enables the kind of large-scale monitoring necessary to track stigma prevalence, evaluate the impact of public health campaigns, and ultimately support efforts to reduce it.

Ethical Considerations

This work addresses a sensitive domain: mental health stigma in naturally occurring online communication. We use two sources of online text: U.S. online news from the News on the Web (NOW) corpus and Reddit posts collected via Pushshift for research and benchmark evaluation.

While the Reddit posts are publicly available and data collection was consistent with Reddit's terms and conditions (Reddit, 2024), individual consent could not be sought from Reddit users. Therefore, we adopt relevant standards outlined by Benton et al. (2017) - and following Chen et al. (2025) - to ensure protection for individuals who post to the Reddit communities in use. Because Reddit posts may contain sensitive self-disclosures about mental health, additional protections include the avoidance of usernames and direct links in the paper, paraphrasing illustrative examples where appropriate, and storage of data in secure environments. Reddit data use in the study was approved by the Human

Ethics Committee (Office of Research Ethics and Integrity) at the University of Melbourne (No.2026-36618-80970-1.)

The benchmark is released in two parts. All annotations, the accompanying metadata, and the excerpt text of the news items are openly available; the excerpt text of the Reddit items is available only on application to the authors. Every item is identified by a pseudonym rather than by its original Reddit or news identifier and no usernames are distributed.

The benchmark is intended as a research and evaluation resource, not as a standalone moderation, clinical, or individual-level assessment tool. Automatic stigma detection may produce false positives, incorrectly flagging clinical discussion, self-disclosure, advocacy, or accounts of experienced stigma as harmful, as well as false negatives that miss subtle forms of stigma. There is also a risk that condition-specific detection could reinforce associations between particular diagnoses and harmfulness or social risk. Future work should incorporate lived-experience perspectives in benchmark design and evaluation.

AI use. We used Claude for language edits, to assist code writing, and for typesetting tables. None of the work (including citations) was generated by LLMs, and all code and tables were checked by the authors of the paper.

Acknowledgments

This project was funded by the University of Melbourne Hallmark Research Initiative on Fighting Harmful Online Communication (seed grant 026796). We thank Label Studio Platform for providing access to the Enterprise Cloud Platform through their Academic Program. Our gratitude also extends to Associate Professor Mike Conway and Dr Jacinthe Flore for early discussions.

References

Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Layla Bouzoubaa, Elham Aghakhani, and Rezvaneh Rezapour. 2024. Words matter: Reducing stigma in online conversations about substance use with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9139–9156, Miami, FL, USA. Association for Computational Linguistics.

Layla Bouzoubaa, Elham Aghakhani, and Rezvaneh Rezapour. 2025. Phenotypes of stigma expressed by people who use drugs on reddit. *Social Science & Medicine*, 390:118889.

Annie T. Chen, Shana Johnny, and Mike Conway. 2022. Examining stigma relating to substance use and contextual factors in social media discussions. *Drug and Alcohol Dependence Reports*, 3:100061.

Liuliu Chen, Jo Robinson, and Mike Conway. 2025. What Do You Meme? – Identifying Characteristics and User Perceptions of Suicide Memes in Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 385–402, Copenhagen, Denmark. AAAI Press.

Patrick 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.

Patrick Corrigan and Amy C. Watson. 2002. Understanding the impact of stigma on people with mental illness. *World Psychiatry*, 1(1):16–20.

Michael Dambrun, Gaétan Marinho, Laurie Mondillon, Maélys Clinchamps, Frédéric Dutheil, Nadia Chakroun, Frédérique Teissedre, Ladislav Motak, Maya Corman, Isabelle Cuchet, Juliette Fargheot, Marie Izaute, Julia Daugherty, Roxane de la Sablonnière, Brittany L. Lindsay, and Andrew C. H. Szeto. 2024. Social perceptions and the stigmatization towards fifteen mental illnesses in France: A preliminary study on the role of vital force and burden. *Frontiers in Psychiatry*, 15:1336690.

Emer N. Day, Kara Edgren, and Amy Eshleman. 2007. Measuring stigma toward mental illness: Development and application of the mental illness stigma scale. *Journal of Applied Social Psychology*, 37(10):2191–2219.

Annie B. Fox, Valerie A. Earnshaw, Emily C. Taverna, and Dawne Vogt. 2018. Conceptualizing and measuring mental illness stigma: The mental illness stigma framework and critical review of measures. *Stigma and Health*, 3(4):348–376.

Salvatore Giorgi, Douglas Bellew, Daniel Roy Sadek Habib, João Sedoc, Chase Smitterberg, Amanda Devoto, McKenzie Himelein-Wachowiak, and Brenda Curtis. 2024. Lived experience matters: Automatic

detection of stigma toward people who use substances on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 474–487, Buffalo, NY, USA. AAAI Press.

Robert Goulden, Elizabeth Corker, Sara Evans-Lacko, Diana Rose, Graham Thornicroft, and Claire Henderson. 2011. Newspaper coverage of mental illness in the UK, 1992–2008. *BMC Public Health*, 11:796.

Kathleen M. Griffiths, Helen Christensen, and Anthony F. Jorm. 2008. Predictors of depression stigma. *BMC Psychiatry*, 8:25.

Christopher Groot, Imogen Rehm, Cal Andrews, Beth Hobern, Rikki Morgan, Hannah Green, Lisa Sweeney, and Michelle Blanchard. 2020. Report on findings from the Our Turn to Speak survey: Understanding the impact of stigma and discrimination on people living with complex mental health issues. Technical report, Anne Deveson Research Centre, SANE Australia, Melbourne, Australia.

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).

Roman Kotov, Robert F. Krueger, David Watson, Thomas M. Achenbach, Robert R. Althoff, R. Michael Bagby, Timothy A. Brown, William T. Carpenter, Avshalom Caspi, Lee Anna Clark, et al. 2017. The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, 126(4):454–477.

Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing stigma. *Annual Review of Sociology*, 27:363–385.

Han Meng, Yancan Chen, Yunan Li, Yitian Yang, Jungup Lee, Renwen Zhang, and Yi-Chieh Lee. 2025. What is stigma attributed to? A theory-grounded, expert-annotated interview corpus for demystifying mental-health stigma. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5453–5490, Vienna, Austria. Association for Computational Linguistics.

Shravika Mittal and Munmun De Choudhury. 2023. Moral framing of mental health discourse and its relationship to stigma: A comparison of social media and news. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Véronique Moriceau, Farah Benamara, and Abdelmoumene Boumadane. 2022. 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.

Nels Oscar, Pamela A. Fox, Racheal Croucher, Riana Wernick, Jessica Keune, and Karen Hooker. 2017. Machine learning, sentiment analysis, and tweets: An examination of Alzheimer's disease stigma on Twitter. *The Journals of Gerontology: Series B*, 72(5):742–751.

Alina Pavlova and Pauwke Berkers. 2022. "Mental health" as defined by Twitter: Frames, emotions, stigma. *Health Communication*, 37(5):637–647.

Hazel Price. 2022. *The Language of Mental Illness: Corpus Linguistics and the Construction of Mental Illness in the Press*. Cambridge University Press, Cambridge.

Reddit. 2024. Reddit privacy policy. `https://www.reddit.com/policies/privacy-policy`. Accessed: 7 September 2024.

Jennifer Boyd Ritsher, Poorni G. Otilingam, and Monica Grajales. 2003. Internalized stigma of mental illness: Psychometric properties of a new measure. *Psychiatry Research*, 121(1):31–49.

Norman Sartorius. 2007. Stigma and mental health. *The Lancet*, 370(9590):810–811.

Georg Schomerus, Michael Lucht, Anita Holzinger, Herbert Matschinger, Mauro G. Carta, and Matthias C. Angermeyer. 2011. The stigma of alcohol dependence compared with other mental disorders: A review of population studies. *Alcohol and Alcoholism*, 46(2):105–112.

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.

S. Martin Taylor and Michael J. Dear. 1981. Scaling community attitudes toward the mentally ill. *Schizophrenia Bulletin*, 7(2):225–240.

Graham Thornicroft, Charlene Sunkel, Akmal Alikhon Aliev, Sue Baker, Elaine Brohan, Rabih el Chammay, Kelly Davies, Mekdes Demissie, Joshua Duncan, Wubalem Fekadu, et al. 2022. The Lancet Commission on ending stigma and discrimination in mental health. *The Lancet*, 400(10361):1438–1480.

Vajisha Udayangi Wanniarachchi, Chris Scogings, Teo Susnjak, and Anuradha Mathrani. 2023. Hate speech patterns in social media: A methodological framework and fat stigma investigation incorporating sentiment analysis, topic modelling and discourse analysis. *Australasian Journal of Information Systems*, 27.

A Corpus preprocessing

This appendix reports preprocessing procedures and descriptive counts for the two online communication corpora used in the study: the News on the Web (NOW) corpus and the Reddit corpus.

A.1 Targets

Given condition-specific variation in mental health stigma, our final target set comprised six mental health conditions: ADHD, autism spectrum disorder, borderline personality disorder, depression, alcoholism, and schizophrenia. These were selected to balance theoretical coverage, corpus frequency, and expected variation in stigma. We began by compiling candidates from major HiTOP spectra (Kotov et al., 2017)—internalizing, neurodevelopmental, psychotic, personality, and substance-related—and, in consultation with domain experts, narrowed the list, consolidating substance-related terms under "alcoholism." For each condition, we generated orthographic and acronymic variants, ran pilot searches and manual inspections in the NOW corpus, normalized hyphens, slashes, and case, and excluded spurious hits such as "BPD" referring to "barrels per day" in petroleum reports. Manual inspection of 300 articles per target confirmed that these variants improved article relevance, meaning that articles mentioned the condition in a more-than-incidental way. Table 7 reports the search lexicon for each mental health condition.

Condition	Search lexicon
ADHD	ADHD; attention deficit hyperactivity disorder
Alcoholism	alcoholism; alcohol abuse; alcohol dependence; alcohol use disorder
Autism	autism; autistic; autism spectrum disorder; ASD
BPD	borderline personality disorder; borderline PD
Depression	depression; major depressive disorder; clinical depression
Schizophrenia	schizophrenia; schizophrenic

Table 7: Surface variants used to retrieve candidate texts for each target mental health condition. Variants were grouped under their associated target condition before sampling and annotation.

A.2 News on the Web

The News on the Web (NOW) corpus is a compilation of online news articles drawn from over 7,000 outlets across 20 countries, spanning 2010–2025. For the present investigation, we extracted all items containing any surface variants from our six target mental health terms, thereby constructing six dedicated sub-corpora. In total, the NOW mental illness conditions sub-corpus comprises 166,912 news articles. We focused on articles from US newspapers.

Pre-processing steps entailed maximizing the reliability of the NOW-corpus material used for stigma annotation. First, within the depression subset we discarded 826 articles that contained the bigram "economic depression", thereby removing economic, rather than clinical, references to "depression." Next, any article that mentioned more than one diagnostic target (e.g., both ADHD and autism) was excluded so that subsequent judgments could be unambiguously attributed to a single disorder. For borderline personality disorder, this yielded under the 500 random sample threshold (427 articles). We also eliminated every record bearing the "TOOLONG" flag, which marks texts that were truncated during corpus compilation and, hence, provide incomplete linguistic evidence. For the texts that were retained, we extracted only the first paragraph in which the target term appears and the two paragraphs that immediately follow it, yielding focused, context-rich passages for human coding (and as for news articles, important content comes first). Finally, the material was lightly cleaned: superfluous whitespace around punctuation, parentheses, quotation marks, and apostrophes was normalized, and residual HTML markers were stripped. Table 8 reports the number of unique articles in which each target appears.

Target	Articles
ADHD	8,316
alcoholism	12,934
autism	26,405
borderline personality disorder	976
depression	106,193
schizophrenia	12,088

Table 8: Articles per target in the NOW subcorpora

Detailed counts of surface-variant occurrences for each mental health condition and its article counts are provided in Table 9.

Variant	Articles
ADHD	7,811
attention deficit hyperactivity disorder	1,455
alcoholism	8,674
alcohol abuse	4,305
alcohol dependence	463
alcohol use disorder	837
autism	21,278
autistic	8,251
autism spectrum disorder	2,703
ASD	1,888
borderline personality disorder	972
borderline PD	12
depression	105,600
major depressive disorder	1,712
clinical depression	1,307
schizophrenia	9,945
schizophrenic	2,683

Table 9: Articles containing at least one instance of the surface variant in its associated NOW subcorpus

Target	Median (whole)	Median (sampled)
ADHD	1,265	1,284
Alcoholism	1,142	1,199
Autism	1,050	976
Borderline PD	1,233	1,221
Depression	1,145	1,071
Schizophrenia	1,084	1,030

Table 10: Article length descriptives (token count) for all articles in the full and randomly sampled NOW corpora per target.

A.3 Reddit

Reddit is a forum-style social media platform organized around topic-specific communities called subreddits. It is a widely used source for research on online mental health discourse due to its large volume of user-generated content. Our dataset was collected via Pushshift.io and covered the years of 2012–2022. To construct the corpus, we extracted posts of at least 30 words containing any of the six target mental health terms, excluding posts that mentioned more than one of these terms to ensure each post was associated with a single condition. Unlike the NOW-corpus texts, Reddit posts were retained in full rather than truncated. In total, the Reddit mental health sub-corpus comprises 1,303,437 posts. Post counts by target condition are reported in Table 11.

Target	Posts
ADHD	219,715
Alcoholism	45,712
Autism	122,467
Borderline PD	91,714
Depression	792,623
Schizophrenia	31,206

Table 11: Posts per target in the Reddit subcorpora

Target	Median	Q3
ADHD	992	1,792
Alcoholism	1,580	2,996
Autism	1,142	2,181
Borderline PD	1,225	2,328
Depression	1,343	2,539
Schizophrenia	1,264	2,675

Table 12: Post length descriptives for Reddits subcorpora per target.

B Stigma taxonomy

The taxonomy combines source frameworks and measures at different levels. Stigma mode follows the Mental Illness Stigma Framework (Fox et al., 2018); stigma domains and public stigma sub-factors follow the Lancet Commission taxonomy (Thornicroft et al., 2022); and the public-stigma components were adapted from established stigma measures, including the Attribution Questionnaire (Corrigan et al., 2003), the Depression Stigma Scale (Griffiths et al., 2008), the Community Attitudes toward the Mentally Ill scale (Taylor and Dear, 1981), the Internalized Stigma of Mental Illness scale (Ritsher et al., 2003), and the Mental Illness Stigma Scale (Day et al., 2007). The prototype exemplar sources are summarised in Appendix C.

Level	Type	Definition and Example	Source
Level 2: Stigma Mode			
	Expressed/Enacted	Expression of stigma or description of its enactment by a stigmatizer. *Example:* "Police restrained the man using force after noting he had a history of psychiatric admissions."	(Fox et al., 2018)
	Perceived/Experienced	Stigma is perceived or its expression experienced by people with lived experience. *Example:* "I was treated like I was dangerous," said Mark, who lives with bipolar disorder.	(Fox et al., 2018)
Level 3: Stigma Domain			
	Public	Negative and misinformed stereotyped beliefs, prejudicial emotional responses, or discriminatory behaviors. *Example:* "Dr White said that, 'Once people develop mental illness, they will never recover completely.' "	(Thornicroft et al., 2022)
	Structural	Systems, policies, or laws that disempower, discriminate, or limit opportunities for people with lived experience of mental illness. *Example:* "The report found that mental health patients waited twice as long for emergency care."	(Thornicroft et al., 2022)
	Associative	Stigma is extended to family, carers, supporters, health workers, or others around people with lived experience of mental illness. *Example:* "Mental illness is often concealed by families for fear of judgment."	(Thornicroft et al., 2022)
	Self	Internalization of negative public stereotypes. *Example:* "For a long time, I thought my mental illness made me a burden to everyone around me."	(Thornicroft et al., 2022)
Level 4: Public Stigma Sub-factors			
	Cognitive (Stereotypes)	Stereotypes are socially shared and generalized negative beliefs about mental illness and people living with it; they represent the cognitive component of stigma. *Example:* "He could snap out of it if he just tried."	(Thornicroft et al., 2022)
	Affective (Emotions)	Prejudicial affective responses are emotional reactions that arise from stereotypes about mental illness and people living with it; they represent the affective component of stigma. *Example:* "Many feel uneasy around people who are mentally ill."	(Thornicroft et al., 2022)
	Behavioral (Discrimination)	Discriminatory behaviors are the actions that result from stereotypes and prejudicial emotions about mental illness and people living with it; they constitute the behavioral component of stigma. *Example:* "The employer admitted he wouldn't hire someone with a history of mental illness."	(Thornicroft et al., 2022)
Level 5: Public Stigma Components			
	Cognitive – Dangerousness	A belief that someone living with mental illness is dangerous. *Example:* "I would feel that Harry would pose a threat to others."	(Corrigan et al., 2003)
	Cognitive – Personal responsibility and blame	A belief that someone living with mental illness is responsible and/or to blame for their experience. *Example:* "I would think that it was Harry's fault that he is in the present condition."	(Corrigan et al., 2003)
	Cognitive – Uncontrollable	A belief that someone is not in control of their experience of mental illness and the behavior that may arise from it. *Example:* "I would feel that Harry is unpredictable."	(Corrigan et al., 2003)
	Cognitive – Controllable	A belief that someone is in control of their experience of mental illness and the behavior that may arise from it. *Example:* "People with depression could snap out of it if they wanted."	(Griffiths et al., 2008)

Level	Type	Definition and Example	Source
	Cognitive – Moral or character flaw	A belief that a person's mental illness is a reflection of a moral or character flaw. *Example:* "Depression is a sign of personal weakness."	(Griffiths et al., 2008)
	Cognitive – Pity/condescension	A belief that someone living with mental illness is inferior or less capable. *Example:* "I would feel sympathy for Harry."	(Corrigan et al., 2003)
	Cognitive – Trivialization or minimization	A belief that someone's mental illness is not a real, legitimate, sufficiently severe, or impactful condition. *Example:* "Depression is not a real mental illness."	(Griffiths et al., 2008)
	Cognitive – Immutability	A belief that a person cannot recover from their experience of mental illness. *Example:* "People with mental illness cannot live a good, rewarding life."	(Ritsher et al., 2003; Day et al., 2007)
	Cognitive – Segregation endorsement	A belief that people with mental illness should be physically or socially separated from the community. *Example:* "I would think that Harry should be kept away from others."	(Corrigan et al., 2003)
	Cognitive – Coercion	A belief that people with mental illness must be controlled or forced into treatment because they lack insight or self-control. *Example:* "I would force Harry to take his medication."	(Corrigan et al., 2003)
	Cognitive – Contamination	A belief that mental illness is contagious or morally polluting, or that association with someone with mental illness may damage one's status or well-being. *Example:* "It is best to avoid people with depression so you don't become depressed yourself."	(Griffiths et al., 2008)
	Cognitive – Incompetence	A belief that people with mental illness are incapable or unreliable in daily or occupational functioning. *Example:* "People with a mental illness should not be given any responsibilities."	(Taylor and Dear, 1981)
	Affective – Fear	An emotional response of fear in response to people living with mental illness. *Example:* "I would feel afraid of Harry."	(Corrigan et al., 2003)
	Affective – Anger, irritability, or frustration	An emotional response of anger, irritability, or frustration in response to people living with mental illness. *Example:* "I would feel irritated by Harry."	(Corrigan et al., 2003)
	Affective – Disgust	An emotional response of revulsion or moral contamination toward people with mental illness. *Example:* "I would feel sickened by them."	Theory-derived
	Affective – Pity	A compassionate but patronising emotional response towards people with mental illness. *Example:* "I would feel pity for Harry."	(Corrigan et al., 2003)
	Affective – Ambivalence	A mixed or conflicted emotional response involving simultaneous positive and negative feelings toward people with mental illness. *Example:* "I would feel both compassion and discomfort toward them."	Theory-derived
	Behavioral – Social distancing	Behavioral discrimination involving desire for, or actual limitation of, the degree to which someone with mental illness is included in one's life across numerous contexts. *Example:* "I would stay away from Harry."	(Corrigan et al., 2003)
	Behavioral – Bullying/-mocking	Behavioral discrimination involving actual verbal or physical abuse of a person with a mental health condition. *Example:* "People mocked him for being mentally ill."	Theory-derived

Table 13: Definitions and illustrative examples of stigma expressions across hierarchical levels. Citations indicate the source framework, scale, or measure from which each category or illustrative item was adapted.

C Prototype exemplar sources

The prototype exemplar dataset contains stigma-present and stigma-absent usages for each annotation category. Prototype usages were mapped to the taxonomy from examples based on validated stigma scale items and were expanded where necessary to provide at least ten stigma-present and ten stigma-absent usages per category. Expansion primarily involved creating close variants of validated scale items and generating stigma-absent usages by negating or reframing stigma-present usages. Table 14 summarizes the source measures used to construct the prototype exemplar set. The full prototype dataset contains stigma-present and stigma-absent examples for each annotation category and is provided at this link: https://github.com/jemimakang/mh_stigma/blob/main/data/prototype_usages_stigma_types.csv.

Source measure or framework	Categories informed	Citation
Attribution Questionnaire	Dangerousness; personal responsibility and blame; uncontrollable; pity/condescension; segregation endorsement; coercion; fear; anger/irritation; pity; social distancing	(Corrigan et al., 2003)
Depression Stigma Scale	Perceived/experienced stigma; controllability; moral or character flaw; trivialization/minimization; contamination; selected dangerousness and blame prototypes	(Griffiths et al., 2008)
Community Attitudes toward the Mentally Ill scale	Structural stigma; public stigma prototypes; dangerousness; moral or character flaw; trivialization/minimization; segregation; coercion; incompetence; fear	(Taylor and Dear, 1981)
Internalized Stigma of Mental Illness scale	Self-stigma; internalized stereotypes; affective responses; behavioral withdrawal; immutability; selected incompetence and social-withdrawal prototypes	(Ritsher et al., 2003)
Mental Illness Stigma Scale	Immutability and other general mental-illness stigma components where applicable	(Day et al., 2007)
Theory-derived definitions	Expressed/enacted stigma; public, structural, and associative stigma; broad cognitive, affective, and behavioral sub-factors; disgust; ambivalence	Present taxonomy

Table 14: Source measures and frameworks used to construct the stigma-present and stigma-absent prototype exemplar set.

D Annotation protocol and guideline revisions

This appendix provides additional details on the annotation codebook and iterative annotation process summarized in Section 3.4. The codebook contained category definitions, decision rules, and examples for the binary stigma-presence task and the fine-grained stigma taxonomy. The complete codebook is provided here: https://github.com/jemimakang/mh_stigma/blob/main/annotation/annotation_instructions.pdf.

Operational definition of stigma. For the binary stigma task, stigma was annotated as present when a text contained at least one of the following, directed at a person or group with a mental health condition, or those associated with them: (i) a negative trait attribution, such as being dangerous, weak, incompetent, or unpredictable; (ii) a negative emotional reaction, such as fear, disgust, contempt, or ridicule; (iii) negative treatment or discrimination, such as calls to isolate, punish, or deny rights or services, including discriminatory laws, rules, or institutional practices; (iv) derogatory or demeaning language, such as slurs or insults; or (v) intrinsically stigmatizing uses of a mental health condition term, such as metaphorical or pathologizing uses outside a clinical context.

Texts were annotated as stigma-absent when the negative attribution, reaction, or treatment was not based on or related to mental health status; when the text merely discussed, criticized, or reflected on stigma without displaying it; or when the negative description was situation-specific and did not imply a generalized stigmatizing attitude toward people with mental illness. Texts that described an incident of stigma perpetrated by someone else were coded as stigma-present only when the passage itself displayed stigmatizing attribution, treatment, or language.

Annotation rounds. In Round 1, annotators independently annotated news and Reddit texts using the initial codebook. They judged whether stigma was present and, where relevant, selected applicable fine-grained stigma categories. Majority judgments were used to identify candidate stigma-present texts, which were then reviewed and adjudicated by two researchers. The adjudication process selected the final label where there was agreement and consulted annotators in ambiguous cases or cases involving multiple plausible codes. Textual evidence spans were also identified for stigma-present cases.

Because random sampling in the first round yielded a large proportion (85%) of stigma-absent examples, later rounds used more targeted strategies to enrich the benchmark with stigma-present and boundary cases. In Round 2, randomly sampled candidate texts were pre-filtered using an LLM prompt containing stigma definitions and examples. This yielded 350 candidate samples predicted by GPT-5.2 Pro to contain stigma.[6] After manual screening to remove irrelevant or clearly non-stigmatizing cases, and after shortening overly long posts to retain only the necessary context for annotation, 231 texts were retained for annotation. Some borderline or likely stigma-absent texts were deliberately retained as validity checks.

Round 3 introduced the explicit operational decision rule described above to improve consistency across annotators. In particular, annotators were instructed to distinguish texts that displayed or described stigma from texts that merely discussed, critiqued, or reflected on stigma as a topic. Texts were labeled stigma-present only when the negative attribution, emotional reaction, treatment, or demeaning language was tied to mental health status rather than to unrelated behavior or situational context.

After Round 3, we examined annotator disagreements and found that a major source of inconsistency concerned the distinction between stigma modes, especially *expressed/enacted* versus *perceived/experienced* stigma. We therefore revised the guidelines to make the mode decision rule more explicit. The revised rule defined *expressed/enacted* stigma as cases in which a stigmatizing source directly expressed stigma or in which the text described stigma being enacted by a person, institution, policy, system, or media text. It defined *perceived/experienced* stigma as cases in which a person with lived or living experience described being stereotyped, devalued, discriminated against, feared, excluded, mocked, or otherwise treated negatively because of their condition.

In Round 4, we further enhanced the benchmark by sampling texts on which automatic systems showed the greatest disagreement. Disagreement was defined across predictions from Claude Sonnet 4.6, Gemini 3 Flash, and GPT-5.5. Candidate

[6] Inference details for GPT-5.2 Pro are reported in Appendix F.

texts were then manually filtered before annotation. Before this round, annotators completed a training exercise focused on threshold cases for *stigma mode*, using the revised mode decision rule. They then annotated the selected texts using the updated guidelines.

Operational rules for stigma mode. Stigma mode captured whether stigma was *expressed/enacted* by a stigmatizing source, *perceived/experienced* by a person with lived or living experience of mental illness, or *both*. Expressed/enacted stigma was annotated when a text expressed stigma directly or described stigma being enacted by a person, institution, policy, system, or media text. Perceived/experienced stigma was annotated when a text described stigma from the perspective of a person with lived or living experience of mental illness, including reports of being stereotyped, devalued, discriminated against, feared, excluded, mocked, or otherwise treated negatively because of their condition. When self-stigma was annotated, stigma mode was coded as both expressed/enacted and perceived/experienced because self-stigma involves a person with lived experience internalizing stigma while also reproducing or expressing a negative stereotype about themselves.

D.1 Annotation platform interface

Annotation was conducted using Label Studio (see Figure 2 for an example of the interface). Each task presented annotators with a target text alongside its associated mental health term. The interface was structured hierarchically, whereby each response revealed only the questions relevant to the annotator's preceding choice. A free-text field was available throughout all levels to capture any potential observations or comments from the annotators.

Annotators first made a binary judgment of whether stigma was present in the text. They then rated their confidence in this judgment (0–10). Text judged as containing no stigma required no further annotation. Where stigma was identified, annotators proceeded through the following levels of categorization.

D.1.1 Mode

Annotators indicated whether stigma was expressed/enacted, perceived/experienced, or both.

D.1.2 Domain

Annotators selected one or more domains from four categories: public, structural, associative, and self-stigma.

D.1.3 Sub-factor and component (public domain only)

Where public stigma was selected, annotators identified one or more sub-factors (cognitive, affective, or behavioral) and then selected the specific components within each. The cognitive sub-factor comprised 12 components (e.g., dangerousness, blame, uncontrollable, moral flaw); the affective sub-factor comprised five components (e.g., fear, disgust, pity); and the behavioral sub-factor included two components: social distancing and bullying. For each sub-factor, an "Other" option was included alongside the predefined components to capture any not represented in the lists. Selecting this option revealed a free-text field in which annotators could enter one or more component labels of their own.

Annotators could highlight relevant text spans using color-coded labels corresponding to their selections: component-level labels for public stigma and domain-level labels for structural, associative, and self-stigma.

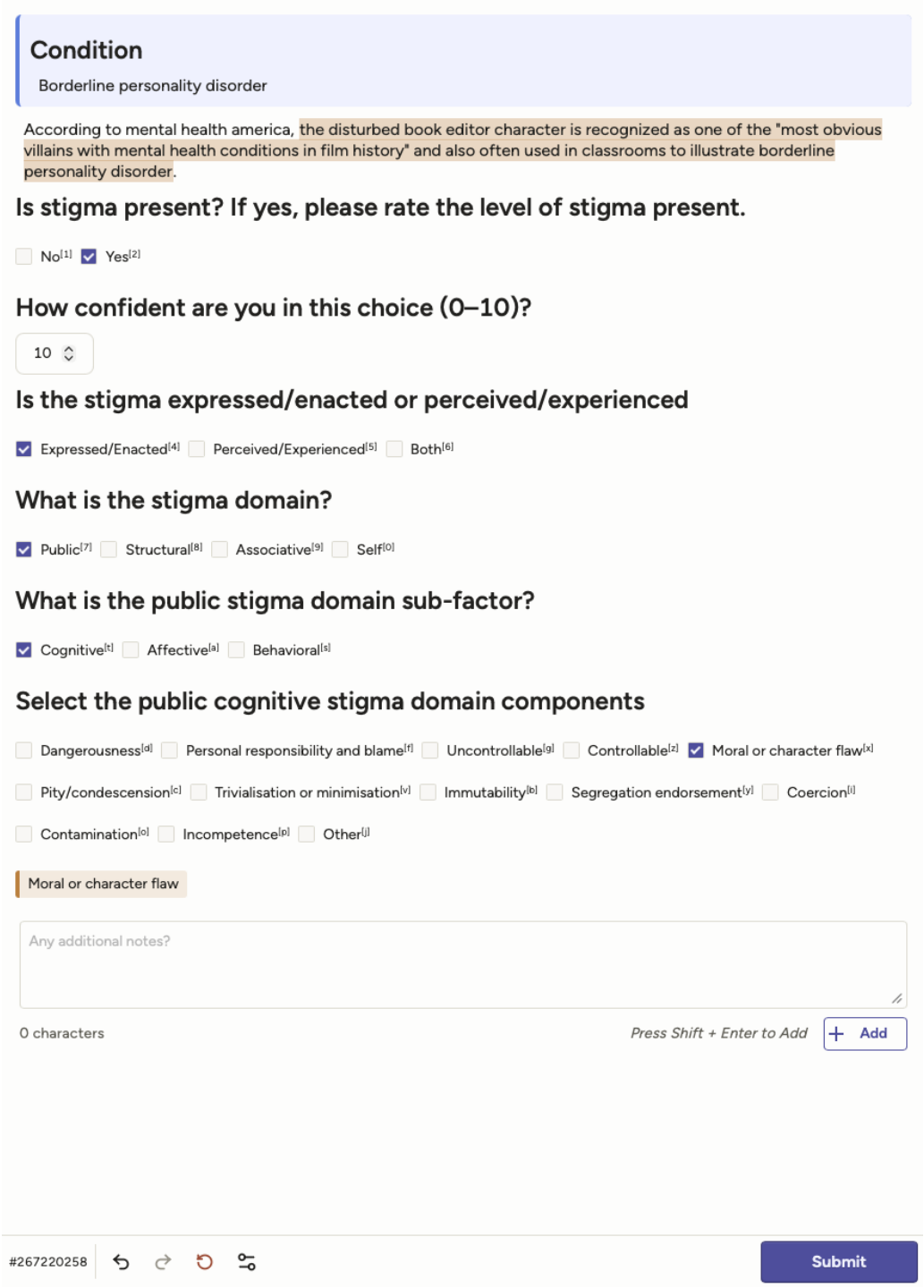


Figure 2: Annotation interface example.

E Stigma detection and classification prompts

E.1 Prompts for binary classification of stigma presence

E.2 "Definition" prompt

```
According to psychologists "Stigma exists when elements of labeling, stereotyping, separation,
status loss, and discrimination occur together".

Using this definition, does the following text contain stigma?
```

E.3 "Examples" prompt

```
Here are some examples of sentences that contain stigma according to Psychologists:

"My roommate keeps calling me 'the anxious one' to their friends. Like that is all I am."
"The schizophrenic was taken into custody."
"Police are concerned the suspect may have mental illness, which would explain the violent
behavior."
"Someone in my class at uni said people with schizophrenia are all violent."
"Society needs better protection from the insane."
"I just wish my depressed boyfriend could be normal like everyone else."
"Employers have the right to know if someone is mentally unstable before hiring them."
"The landlord knocked back my application. He found out that I have a mental illness."
"Experts decided patients should not be consulted."
"My psychiatrist changed my medication without asking me what I thought. I feel so helpless."

Using these examples, does the following text contain stigma?
```

E.4 "Definition+examples" prompt

```
According to psychologists "Stigma exists when elements of labeling, stereotyping, separation,
status loss, and discrimination occur together".

Here are some examples of sentences that contain stigma according to Psychologists:

"My roommate keeps calling me 'the anxious one' to their friends. Like that is all I am."
"The schizophrenic was taken into custody."
"Police are concerned the suspect may have mental illness, which would explain the violent
behavior."
"Someone in my class at uni said people with schizophrenia are all violent."
"Society needs better protection from the insane."
"I just wish my depressed boyfriend could be normal like everyone else."
"Employers have the right to know if someone is mentally unstable before hiring them."
"The landlord knocked back my application. He found out that I have a mental illness."
"Experts decided patients should not be consulted."
"My psychiatrist changed my medication without asking me what I thought. I feel so helpless."

Using this definition and these examples, does the following text contain stigma?
```

E.5 "Definition+examples+rule" prompt

```
According to psychologists "Stigma exists when elements of labeling, stereotyping, separation,
status loss, and discrimination occur together".

Here are some examples of sentences that contain stigma according to Psychologists:

"My roommate keeps calling me 'the anxious one' to their friends. Like that is all I am."
"The schizophrenic was taken into custody."
"Police are concerned the suspect may have mental illness, which would explain the violent
behavior."
"Someone in my class at uni said people with schizophrenia are all violent."
"Society needs better protection from the insane."
"I just wish my depressed boyfriend could be normal like everyone else."
"Employers have the right to know if someone is mentally unstable before hiring them."
```

"The landlord knocked back my application. He found out that I have a mental illness."
"Experts decided patients should not be consulted."
"My psychiatrist changed my medication without asking me what I thought. I feel so helpless."

Stigma Present
Only annotate 'yes' if the text contains at least one of the following directed at a person or persons with a mental health condition (or those associated with them):

- **Negative trait attribution** (e.g., dangerous, weak, incompetent, unpredictable)
- **Negative emotional reaction** (e.g., fear, disgust, contempt, ridicule)
- **Negative treatment or discrimination** (e.g., calls to isolate, punish, or deny rights/services, including discriminatory laws, rules, or institutional practices)
- **Derogatory or demeaning language** (e.g., slurs, insults)

Annotate 'no' if:

- The negative attribution, reaction or treatment is not based on or related to the person's mental health condition (e.g., anger is expressed towards a person with depression because they committed a crime that is unrelated to their depression).
- The negative attribution, reaction or treatment is mentioned in text rather than endorsed or expressed (e.g., a journalist writes about negative attitudes towards schizophrenia without manifesting it herself; a person criticizing or commenting on stigma). *Note:* text that simply describes an incident of stigma perpetrated by someone else should be annotated YES: the key issue is whether the text displays stigma or comments/reflects on it.

F Model and inference details

Table 15 reports the model identifiers and inference settings used in the LLM experiments. All model runs were conducted in May 2026. Unless otherwise stated, decoding and reasoning parameters were not manually tuned; models were run with provider or software-wrapper defaults. Model parameter counts are not publicly disclosed for the proprietary models evaluated here. We therefore identify models by their public names, API identifiers, and inference settings rather than by parameter size. The same prompt templates were used across models within each experimental condition.

G Benchmark composition: Label distributions

This appendix provides expanded descriptive statistics for the benchmark. As in the main text, these distributions describe benchmark composition rather than stigma prevalence in the underlying corpora. Because stigma-present texts could receive multiple labels, counts for domains, subfactors, and components should be interpreted as text-label assignments rather than mutually exclusive classes.

Table 16 provides the full benchmark composition, including source, condition, stigma mode, higher-level taxonomy labels, and fine-grained public-stigma components.

Provider	Model	Inference settings
Anthropic	Claude Sonnet 4.6 `claude-sonnet-4-6`	Temperature was set to 0. Adaptive/extended thinking was disabled. Safety settings used provider defaults.
OpenAI	GPT-5.5 `gpt-5.5`	Temperature was was set to 0. Reasoning effort was not explicitly overridden and therefore used the documented GPT-5.5 default of `medium`. Reasoning summaries were not requested. Safety settings used provider defaults.
Google	Gemini 3 Flash `gemini-3-flash-preview`	Temperature was set to 0. Dynamic thinking was used at default settings (`high`). Safety settings used provider defaults.
DeepSeek	DeepSeek V4 Flash `deepseek-v4-flash`	The model was used in thinking mode, where temperature and top-p are not supported. Safety settings used provider defaults.
OpenAI	GPT-5.2 Pro `gpt-5.2-pro-2025-12-11`	Used only for LLM-assisted candidate filtering during benchmark construction, not for the main evaluation. Temperature = 0. Reasoning effort was not explicitly overridden in our filtering run. Safety settings used provider defaults.

Table 15: Model identifiers and inference settings used in the LLM experiments. All model runs were conducted in May 2026. Unless otherwise stated, decoding, reasoning, and safety parameters were not manually tuned; where provider defaults or wrapper behavior were not fully exposed, we report the settings as unset rather than inferred.

Panel A: Source and condition coverage

	News		Reddit		Total	
Condition	**N**	**Stigma+**	**N**	**Stigma+**	**N**	**Stigma+**
Overall	249	38	221	49	470	87
ADHD	42	6	35	7	77	13
Alcoholism	37	4	41	8	78	12
Autism	36	1	42	15	78	16
BPD	47	10	33	10	80	20
Depression	38	0	41	4	79	4
Schizophrenia	49	17	29	5	78	22

Panel B: Source, condition, and stigma mode

Level	**Group**	**N**	**Stigma+**	**Expressed/ enacted**	**Perceived/ experienced**	**Both**
Overall	Overall	470	87	49	15	23
Source	News	249	38	29	5	4
Source	Reddit	221	49	20	10	19
Condition	ADHD	77	13	5	5	3
Condition	Alcoholism	78	12	7	1	4
Condition	Autism	78	16	9	5	2
Condition	BPD	80	20	11	1	8
Condition	Depression	79	4	0	1	3
Condition	Schizophrenia	78	22	17	2	3

Panel C: Higher-level taxonomy coverage among stigma-present texts

Level	**Category**	**N labels**
Domain	Public	62
Domain	Self	24
Domain	Structural	6
Domain	Associative	1
Public-stigma subfactor	Cognitive	54
Public-stigma subfactor	Affective	9
Public-stigma subfactor	Behavioral	5

Panel D: Fine-grained public-stigma components among stigma-present texts

Subfactor	**Component**	**N labels**
Cognitive	Moral or character flaw	14
Cognitive	Incompetence	14
Cognitive	Uncontrollable	11
Cognitive	Trivialization/minimization	9
Cognitive	Pity/condescension	8
Cognitive	Dangerousness	7
Cognitive	Immutability	5
Cognitive	Segregation endorsement	4
Cognitive	Controllable	3
Cognitive	Personal responsibility/blame	2
Cognitive	Contamination	1
Affective	Anger/irritability/frustration	4
Affective	Disgust	4
Affective	Fear	2
Behavioral	Bullying/mocking	4
Behavioral	Social distancing	1

Table 16: Expanded composition of the final benchmark. Panel A reports source-condition coverage using raw text counts. Panel B reports source, condition, and stigma-mode coverage. Panels C and D report taxonomy labels among the 87 stigma-present texts. Counts in Panels C and D refer to label assignments rather than mutually exclusive texts, because stigma-present texts could receive multiple labels. These statistics describe benchmark composition rather than real-world stigma prevalence.

H Qualitative model error examples

Qualitative examples were selected from a systematic error-audit file created using predefined decision rules. The audit file grouped errors by text ID, identified persistent false positives and false negatives across model–prompt combinations, and recorded within-model prompt transitions where added prompt information corrected or introduced an error. The final examples were then manually inspected to assign qualitative error patterns.

Source	Cue triggering overprediction	Interpretation
Reddit / schizophrenia	Risk or causal association	The text asks about a possible causal or risk association between cannabis use and schizophrenia. It mentions a negative outcome, but does not direct a stigmatizing attribution, treatment, or expression toward people with schizophrenia.
Paraphrased text. *"someone told me thc only takes 30 days to leave your system no matter what? also heard that if you start smoking weed young you'll definitely get schizophrenia* ***schizophrenia****"*		
Reddit / borderline personality disorder	Clinical and interpersonal-risk framing	This example illustrates a difficult threshold case. The passage contains negative, clinically loaded descriptions of BPD symptoms, but the final adjudicated label was stigma-absent. Models treated the diagnostic and interpersonal-risk framing as sufficient evidence of stigma.
Paraphrased text. *"people with* ***borderline personality disorder*** *(****bpd****) deal with really intense emotional swings that can put them in a really desperate headspace. its kind of like when you feel cornered - you lash out at whoever is closest to you that makes you feel threatened. and the thing is, the people who know you best are actually the most vulnerable because sufferers of* ***bpd*** *can use what they know about you against you. so the ones who care the most tend to get hurt the most."*		
News / depression	Mental health crisis context	The passage links student behavior to a broader mental health crisis, but it frames disruptive behavior as unmet need rather than as a stigmatizing attribution toward people with depression. Models may have over-relied on negative behavioral language such as "outbursts", "lashing out", and "disruptive".
Full text. *"The spike in outbursts, fidgeting and nonstop chatter in classrooms coincides with a national mental health crisis. American pediatricians during the pandemic went as far as to call it a national emergency. A parents with kids under 18 said they are "extremely or very worried that their children might struggle with anxiety or* ***depression*** *at some point." Teachers are under pressure to help students make up for what they haven't learned, and kids feel their stress. The combination has led to a pronounced rise in students expressing anger or sadness by lashing out and being disruptive at school. "Anytime we've got a chaotic structure and high expectations of ourselves, it's very easy for that to come out in disruptions," said Kent McIntosh, an associate professor in the College of Education at the University of Oregon. "What on the outside can look like defiant behavior is sometimes a student's best bet to get needs met in the moment."*		

Table 17: Selected false-positive examples. These are stigma-absent texts that one or more models predicted as stigma-present. Reddit examples have been paraphrased to preserve user privacy; news examples are reproduced in full.

ID / source	Cue missed by models	Interpretation
Reddit / alcoholism	Self-directed shame and guilt	The broader tone is hopeful and recovery-oriented, but the passage contains self-directed shame and guilt related to alcoholism. Models often missed this case, suggesting that self-stigma may be under-detected when embedded in supportive, religious, or non-toxic discourse.
Full text. *"heading to church for the first time in a while and i'm actually really looking forward to it. i'd been staying away because i felt so ashamed about my **alcoholism** but i've realised that's kind of the whole reason i should go. woke up this morning and just prayed for strength and a fresh start. feeling hopeful. hope everyone has a good day"*		
Reddit / autism	Structural exclusion / being left behind	The passage expresses concern about structural exclusion and unequal access to support services for autistic adults and people with disabilities. Models may have missed this case because it appears in advocacy-oriented and institutional language rather than in overtly hostile or derogatory language.
Full text. *"Two **autistic** adults here. We rely on in house daily support. The pandemic really showed how essential these kinds of community-based services are for people with disabilities. There's a real opportunity right now with the COVID-19 relief package to actually fund these services properly but honestly we're scared disabled people are going to get overlooked again like always. The Autistic Self Advocacy Network (disability rights advocacy group) has been tracking COVID-19 deaths in care institutions since early 2020, and this research just confirms what we already knew - government decisions literally determine whether people like us get the support we need to survive and live our lives."*		
Reddit / ADHD	Invalidation of diagnosis / ADHD treated as an excuse	The passage was adjudicated as stigma-present because it describes a family context in which ADHD is dismissed or invalidated: the poster reports that their stepfather has never acknowledged their ADHD and that their mother says it is "an excuse." Models may have missed this case because the stigma is embedded in a first-person help-seeking post rather than expressed as overt abuse or toxic language.
Full text. *"tl;dr: strict home environment is causing me to have regular meltdowns. basically the pressure at home is so intense that i end up breaking down in the bathroom every couple of weeks. the expectations are really high - grades have to be near perfect, chores can't slip even once, screen time is capped and has to be timed, and i can only use my laptop in a shared space with the screen visible to everyone. i can't go out without giving my parents my exact route in advance, and i'm still not allowed to get my learner's permit so i'm getting around on a bike. there's a lot of other rules on top of all that. the hardest part is that my stepdad completely dismisses my **adhd** diagnosis that i've had since i was a kid, and my mom says i just use it as an excuse. just needed to get that out - does anyone have any advice?"*		

Table 18: Selected false-negative examples. These are stigma-present texts that one or more models predicted as stigma-absent. Reddit examples have been paraphrased to preserve user privacy.

I Human annotation agreement

We first report agreement for all annotation rounds combined. Binary stigma-presence agreement across all items was moderate (exact agreement = .66; AC1 = .40). Agreement on the stigma mode category was also modest when no-stigma was included as a category (exact agreement = .61; AC1 = .54), and among the 19 items that all annotators judged as stigma-present (exact agreement = .51; AC1 = .67). Figures 3 and 4 show agreement matrices for each annotator pair.

There was also variation in agreement across rounds (Table 19). In Round 2, the three experienced annotators achieved an exact agreement of .62 (AC1 = .62) on stigma presence. In Round 3, which introduced a new annotator, agreement fell to .39 (AC1 = .22). Pairwise comparisons suggest this decline was driven by the new annotator, whose agreement with the two returning annotators was low (AC1 = .08 and .11). The new annotator also labeled 66.7% of items as stigma-present, compared with 30.0% and 24.7% for the returning annotators.

When pooled across rounds, pairwise results also suggested threshold differences across annotators. A1–A2 agreement on stigma presence was lower (exact agreement = .61; AC1 = .26) than A1–A3 (exact agreement = .69; AC1 = .50) or A2–A3 (exact agreement = .69; AC1 = .45). Disagreements were resolved through multiple rounds of discussion and adjudication, which led to revisions of the annotation guidelines.

Because subtype and domain labels are only meaningful for stigma-present items, these analyses were restricted to the items that all annotators marked as stigma-present. Agreement was relatively high for domain labels (macro pairwise AC1 = .82) and public-stigma subfactors (macro pairwise AC1 = .89). These estimates should be interpreted cautiously, however, because the subset was small and some labels were rare, making it easier for annotators to agree that a label was absent.

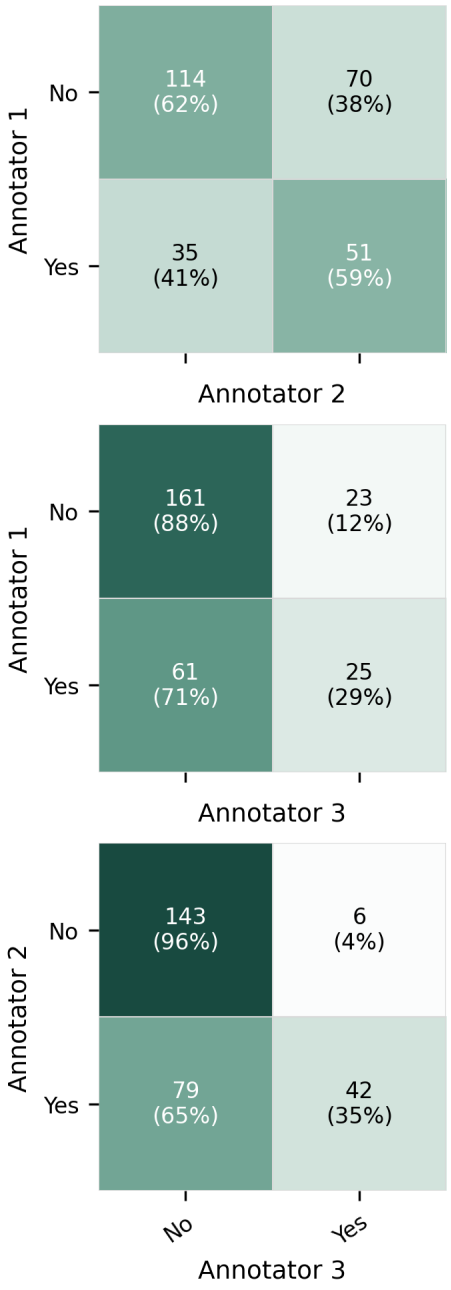


Figure 3: Pairwise agreement matrices for stigma presence.

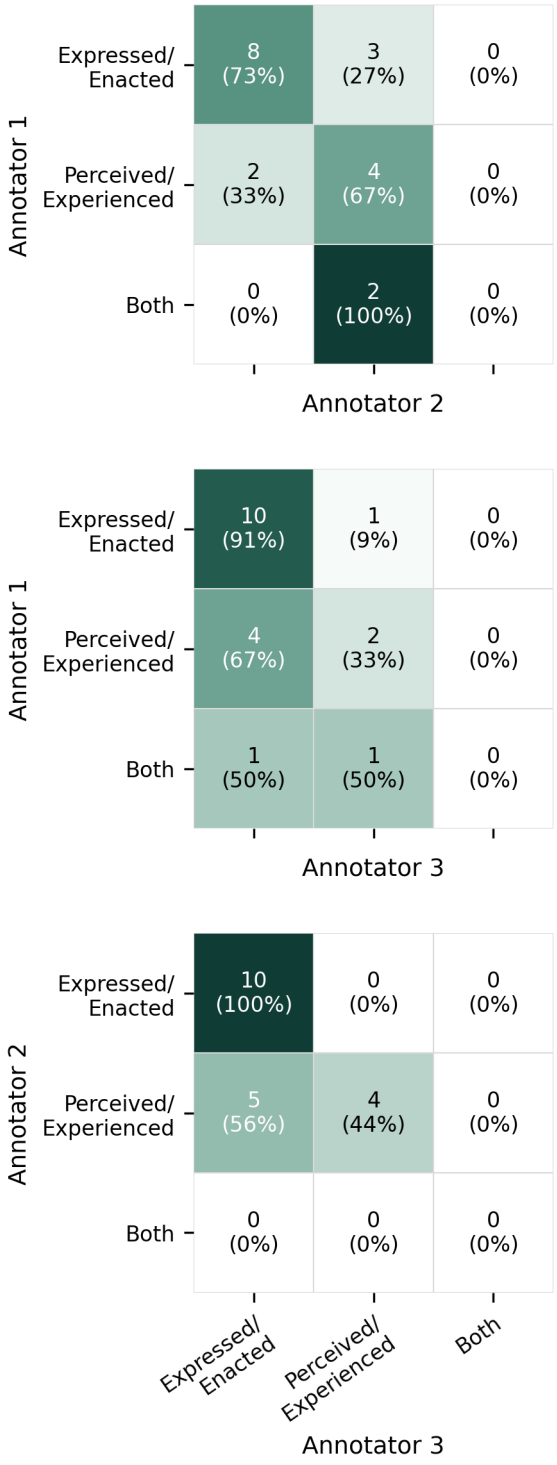


Figure 4: Pairwise agreement matrices for stigma mode among the items that all annotators judged as stigma-present.

Round	Pair	*n*	Exact	AC1
R2	A1–A2	120	.683	.521
	A1–A3	120	.700	.514
	A2–A3	120	.850	.805
	Overall	120	.617	.622
R3	A1–A2	150	.693	.491
	A1–A3	150	.540	.081
	A2–A3	150	.553	.113
	Overall	150	.393	.220

Table 19: Pairwise and overall agreement on stigma presence by annotation round. Exact = exact agreement; AC1 = Gwet's AC1. The A3 position was held by a psychology professor in Round 2 and by a psychology PhD student in Round 3.

J Detailed scores for model predictions

			Binary			Macro		
Strategy	**Model**	**Acc**	**P**	**R**	**F1**	**P**	**R**	**F1**
`baseline`	Claude Sonnet 4.6	.67 [.62, .71]	.31 [.24, .37]	.66 [.55, .75]	.42 [.35, .49]	.60 [.57, .64]	.66 [.61, .72]	.59 [.55, .64]
	Deepseek v4 Flash	.70 [.66, .75]	.27 [.19, .35]	.34 [.25, .45]	.30 [.22, .39]	.55 [.51, .60]	.57 [.51, .62]	.56 [.51, .61]
	Gemini 3 Flash	.73 [.69, .78]	.39 [.31, .46]	.74 [.64, .82]	**.51** [.43, .58]	***.65*** [.61, .69]	.73 [.68, .79]	**.66** [.61, .71]
	GPT-5.5	.57 [.52, .61]	.28 [.23, .34]	.86 [.78, .93]	.42 [.36, .49]	.61 [.58, .64]	.68 [.63, .72]	.54 [.49, .58]
`definition`	Claude Sonnet 4.6	.70 [.66, .74]	.32 [.25, .39]	.54 [.44, .65]	.40 [.32, .48]	.60 [.56, .64]	.64 [.58, .69]	.60 [.55, .65]
	Deepseek v4 Flash	.63 [.58, .67]	.27 [.21, .33]	.60 [.49, .70]	.37 [.30, .44]	.57 [.53, .61]	.61 [.56, .67]	.55 [.50, .60]
	Gemini 3 Flash	.73 [.70, .77]	.38 [.30, .46]	.68 [.58, .78]	.49 [.40, .56]	.64 [.60, .69]	.71 [.66, .76]	.65 [.61, .70]
	GPT-5.5	.69 [.65, .74]	.34 [.28, .41]	.72 [.62, .81]	.46 [.39, .54]	.63 [.59, .67]	.70 [.65, .75]	.62 [.58, .67]
`examples`	Claude Sonnet 4.6	.72 [.68, .76]	.35 [.27, .43]	.56 [.46, .67]	.43 [.35, .51]	.62 [.57, .66]	.66 [.61, .72]	.62 [.57, .67]
	Deepseek v4 Flash	.65 [.61, .69]	.28 [.22, .35]	.57 [.47, .68]	.38 [.30, .45]	.58 [.54, .62]	.62 [.56, .68]	.57 [.52, .61]
	Gemini 3 Flash	.67 [.62, .71]	.33 [.26, .39]	.75 [.65, .84]	.45 [.38, .52]	.62 [.59, .66]	.70 [.65, .75]	.61 [.56, .65]
	GPT-5.5	.65 [.60, .69]	.33 [.27, .39]	***.89*** [.81, .95]	.48 [.41, .55]	.64 [.61, .68]	***.74*** [.69, .78]	.61 [.56, .65]
`examples+def.`	Claude Sonnet 4.6	.73 [.69, .77]	.35 [.28, .44]	.54 [.44, .64]	.43 [.35, .51]	.62 [.57, .66]	.66 [.60, .71]	.63 [.58, .67]
	Deepseek v4 Flash	.64 [.60, .69]	.27 [.21, .34]	.56 [.46, .66]	.37 [.29, .44]	.57 [.53, .61]	.61 [.56, .67]	.56 [.51, .60]
	Gemini 3 Flash	.68 [.63, .72]	.33 [.26, .40]	.71 [.62, .81]	.45 [.37, .52]	.62 [.58, .66]	.69 [.64, .74]	.61 [.56, .66]
	GPT-5.5	.66 [.62, .71]	.33 [.27, .39]	.79 [.70, .88]	.47 [.39, .53]	.63 [.60, .67]	.71 [.66, .76]	.61 [.56, .66]
`ex.+def.+rule`	Claude Sonnet 4.6	***.78*** [.74, .81]	***.40*** [.29, .51]	.36 [.26, .45]	.38 [.28, .47]	.63 [.57, .68]	.62 [.56, .67]	.62 [.57, .67]
	Deepseek v4 Flash	.71 [.67, .75]	.31 [.24, .40]	.49 [.39, .60]	.38 [.30, .46]	.59 [.55, .63]	.62 [.57, .68]	.60 [.55, .64]
	Gemini 3 Flash	.73 [.69, .77]	.36 [.28, .45]	.55 [.45, .65]	.43 [.35, .52]	.62 [.58, .67]	.66 [.61, .72]	.63 [.58, .68]
	GPT-5.5	.67 [.62, .71]	.33 [.26, .39]	.75 [.65, .84]	.45 [.38, .52]	.62 [.58, .66]	.70 [.64, .75]	.61 [.56, .65]

Table 20: Stigma presence detection results (accuracy, binary and macro-averaged precision, recall, and F1) across experimental settings and models. Binary metrics are reported in terms of the positive class ("stigma present"). Bracketed values are 95% confidence intervals obtained from bootstrapping (2000 resamples). Best results across all models and conditions are in **bold**; ***bold italics*** means that the difference with the next-best result for a particular metric is also statistically significant (per-instance McNemar test, χ^2 with continuity correction, $p < 0.05$).

Tables 20 to 25 show detailed results for detecting stigma, as well as for mode, domain, stigma type and subtype predictions. Table 26 further breaks down stigma detection performance of Claude Sonnet 4.6 by mental health condition.

Strategy	Model	Accuracy	Precision	Recall	F1
baseline	Claude Sonnet 4.6	0.563 [.45, .67]	0.788 [.73, .85]	0.868 [.81, .92]	0.809 [.75, .85]
	Deepseek v4 Flash	0.575 [.47, .68]	0.839 [.77, .91]	0.698 [.62, .77]	0.759 [.69, .82]
	Gemini 3 Flash	0.575 [.47, .68]	0.804 [.74, .87]	0.826 [.77, .88]	0.792 [.73, .84]
	GPT-5.5	0.621 [.52, .71]	0.842 [.78, .91]	0.831 [.76, .90]	0.837 [.79, .88]
definition	Claude Sonnet 4.6	0.540 [.43, .64]	0.785 [.72, .84]	0.848 [.79, .90]	0.798 [.75, .84]
	Deepseek v4 Flash	0.632 [.53, .74]	0.848 [.78, .91]	0.809 [.74, .88]	0.826 [.77, .88]
	Gemini 3 Flash	0.586 [.48, .69]	0.833 [.77, .89]	0.799 [.74, .85]	0.787 [.72, .84]
	GPT-5.5	0.609 [.51, .71]	0.825 [.76, .89]	0.850 [.78, .92]	0.836 [.79, .88]
examples	Claude Sonnet 4.6	0.552 [.45, .66]	0.774 [.71, .83]	0.869 [.81, .92]	0.804 [.75, .85]
	Deepseek v4 Flash	0.575 [.47, .68]	0.862 [.79, .93]	0.704 [.64, .78]	0.773 [.71, .83]
	Gemini 3 Flash	0.575 [.47, .68]	0.838 [.77, .90]	0.812 [.76, .87]	0.806 [.75, .86]
	GPT-5.5	0.701 [.60, .79]	0.910 [.86, .95]	0.852 [.79, .91]	0.878 [.83, .92]
examples+definition	Claude Sonnet 4.6	0.517 [.41, .62]	0.809 [.74, .87]	0.779 [.72, .84]	0.761 [.70, .82]
	Deepseek v4 Flash	0.598 [.49, .69]	0.856 [.79, .92]	0.761 [.70, .83]	0.801 [.75, .85]
	Gemini 3 Flash	0.552 [.45, .66]	0.842 [.78, .90]	0.737 [.68, .79]	0.741 [.67, .81]
	GPT-5.5	0.690 [.59, .78]	**0.910** [.85, .96]	0.836 [.77, .90]	0.871 [.82, .92]
+rules	Claude Sonnet 4.6	0.678 [.57, .77]	0.806 [.74, .87]	0.835 [.77, .89]	0.801 [.73, .86]
	Deepseek v4 Flash	0.575 [.47, .68]	0.862 [.79, .92]	0.755 [.68, .83]	0.804 [.74, .85]
	Gemini 3 Flash	0.805 [.72, .87]	0.851 [.79, .91]	**0.952** [.91, .98]	0.896 [.84, .94]
	GPT-5.5	**0.816** [.74, .90]	0.903 [.86, .95]	0.940 [.88, .99]	**0.921** [.88, .96]

Table 21: Macro-averaged prediction results for stigma mode (n=87). Bracketed values are 95% bootstrap confidence intervals (1000 resamples).

Strategy	Model	Accuracy	Precision	Recall	F1
baseline	Claude Sonnet 4.6	0.529 [.43, .63]	0.569 [.40, .73]	**0.988** [.73, 1.00]	0.706 [.49, .83]
	Deepseek v4 Flash	0.483 [.38, .59]	0.461 [.39, .53]	0.690 [.65, .73]	0.532 [.46, .60]
	Gemini 3 Flash	0.517 [.41, .62]	0.462 [.40, .53]	0.960 [.69, .98]	0.559 [.47, .64]
	GPT-5.5	0.713 [.62, .80]	0.787 [.49, .85]	0.982 [.71, 1.00]	0.854 [.54, .90]
definition	Claude Sonnet 4.6	0.563 [.45, .67]	0.527 [.40, .72]	**0.988** [.73, 1.00]	0.662 [.49, .82]
	Deepseek v4 Flash	0.575 [.47, .68]	0.481 [.41, .55]	0.723 [.70, .75]	0.572 [.49, .63]
	Gemini 3 Flash	0.598 [.49, .70]	0.512 [.45, .58]	0.976 [.71, .99]	0.601 [.51, .68]
	GPT-5.5	0.724 [.63, .82]	0.678 [.50, .84]	0.967 [.69, .99]	0.780 [.55, .89]
examples	Claude Sonnet 4.6	0.552 [.45, .66]	0.574 [.40, .74]	0.934 [.62, .99]	0.696 [.48, .82]
	Deepseek v4 Flash	0.632 [.54, .74]	0.569 [.46, .72]	0.879 [.56, .94]	0.656 [.50, .79]
	Gemini 3 Flash	0.621 [.52, .71]	0.504 [.43, .59]	0.964 [.70, .98]	0.595 [.49, .69]
	GPT-5.5	**0.759** [.67, .84]	**0.818** [.51, .89]	0.971 [.70, 1.00]	**0.877** [.56, .93]
examples+definition	Claude Sonnet 4.6	0.575 [.46, .68]	0.494 [.41, .60]	0.960 [.69, .98]	0.598 [.47, .71]
	Deepseek v4 Flash	0.632 [.53, .72]	0.564 [.48, .66]	0.923 [.65, .95]	0.642 [.52, .75]
	Gemini 3 Flash	0.598 [.49, .70]	0.504 [.43, .58]	0.947 [.67, .97]	0.581 [.49, .67]
	GPT-5.5	0.747 [.66, .83]	0.809 [.50, .87]	0.945 [.67, .98]	0.856 [.54, .91]

Table 22: Macro-averaged results for stigma domain (n=87). Bracketed values are 95% bootstrap confidence intervals (1000 resamples).

Strategy	Model	Accuracy	Precision	Recall	F1
baseline	Claude Sonnet 4.6	0.386 [.26, .51]	0.441 [.37, .52]	**0.795** [.61, .95]	0.528 [.43, .61]
	Deepseek v4 Flash	0.351 [.25, .47]	0.381 [.32, .46]	0.655 [.44, .88]	0.441 [.35, .54]
	Gemini 3 Flash	0.632 [.51, .75]	0.364 [.30, .44]	0.593 [.43, .67]	0.426 [.33, .51]
	GPT-5.5	0.596 [.47, .74]	0.521 [.39, .68]	0.733 [.52, .91]	**0.562** [.41, .69]
definitions	Claude Sonnet 4.6	0.474 [.35, .61]	0.412 [.32, .51]	0.670 [.49, .81]	0.473 [.36, .57]
	Deepseek v4 Flash	0.456 [.33, .60]	0.430 [.33, .55]	0.666 [.41, .87]	0.489 [.35, .63]
	Gemini 3 Flash	0.667 [.54, .79]	0.539 [.33, .76]	0.621 [.45, .76]	0.494 [.36, .63]
	GPT-5.5	0.561 [.44, .70]	0.691 [.34, .75]	0.649 [.46, .80]	0.528 [.37, .66]
examples	Claude Sonnet 4.6	0.509 [.39, .63]	0.458 [.37, .56]	0.753 [.57, .93]	0.535 [.41, .63]
	Deepseek v4 Flash	0.491 [.35, .61]	0.410 [.34, .49]	0.655 [.46, .86]	0.479 [.37, .56]
	Gemini 3 Flash	**0.737** [.61, .84]	0.578 [.36, .81]	0.621 [.45, .76]	0.530 [.38, .66]
	GPT-5.5	0.596 [.46, .74]	0.501 [.35, .68]	0.629 [.45, .78]	0.501 [.37, .62]
examples+definitions	Claude Sonnet 4.6	0.526 [.40, .67]	0.453 [.36, .56]	0.711 [.53, .88]	0.523 [.40, .63]
	Deepseek v4 Flash	0.456 [.33, .60]	0.420 [.35, .50]	0.763 [.54, .94]	0.503 [.40, .61]
	Gemini 3 Flash	0.719 [.60, .82]	**0.722** [.32, .81]	0.548 [.35, .72]	0.504 [.34, .64]
	GPT-5.5	0.614 [.49, .74]	0.545 [.35, .75]	0.636 [.43, .79]	0.522 [.37, .65]

Table 23: Macro-averaged results for stigma type (Public domain only, n=57). Bracketed values are 95% bootstrap confidence intervals (1000 resamples).

Strategy	Model	Acc	Pr	Re	F1
baseline	Claude Sonnet 4.6	0.14 [.06,.24]	0.57 [.33,.68]	0.63 [.39,.76]	0.49 [.30,.61]
	Deepseek v4 Flash	0.12 [.04,.22]	0.36 [.22,.45]	0.46 [.34,.52]	0.36 [.24,.41]
	Gemini 3 Flash	0.18 [.08,.29]	0.55 [.37,.59]	0.59 [.40,.68]	0.48 [.35,.56]
	GPT-5.5	0.24 [.14,.37]	0.58 [.37,.68]	0.67 [.41,.75]	0.50 [.36,.61]
definition	Claude Sonnet 4.6	0.22 [.12,.35]	0.52 [.34,.61]	0.59 [.39,.68]	0.48 [.31,.56]
	Deepseek v4 Flash	0.18 [.08,.31]	0.40 [.28,.48]	0.53 [.35,.61]	0.38 [.26,.47]
	Gemini 3 Flash	0.16 [.06,.27]	0.49 [.31,.54]	0.51 [.35,.60]	0.42 [.30,.49]
	GPT-5.5	0.22 [.12,.35]	0.62 [.46,.70]	0.70 [.43,.79]	0.56 [.42,.66]
examples	Claude Sonnet 4.6	0.16 [.06,.27]	0.47 [.31,.58]	0.57 [.38,.65]	0.44 [.30,.52]
	Deepseek v4 Flash	0.10 [.02,.18]	0.24 [.17,.34]	0.45 [.30,.52]	0.29 [.21,.34]
	Gemini 3 Flash	0.20 [.10,.33]	0.53 [.37,.62]	0.59 [.37,.67]	0.49 [.34,.57]
	GPT-5.5	0.20 [.10,.33]	0.50 [.34,.63]	0.52 [.33,.62]	0.44 [.31,.53]
examples+definition	Claude Sonnet 4.6	0.24 [.12,.37]	0.43 [.32,.56]	0.61 [.40,.69]	0.46 [.32,.56]
	Deepseek v4 Flash	0.14 [.04,.24]	0.36 [.22,.46]	0.51 [.35,.58]	0.37 [.25,.44]
	Gemini 3 Flash	0.18 [.08,.29]	0.52 [.36,.62]	0.61 [.40,.69]	0.49 [.35,.57]
	GPT-5.5	0.24 [.14,.37]	0.58 [.44,.71]	0.67 [.40,.77]	0.55 [.40,.66]

Table 24: Macro-averaged results for the Cognitive stigma subtype (n=49). Bracketed values are 95% bootstrap confidence intervals (1000 resamples).

		Affective				Behavioral			
Strategy	**Model**	**Ac**	**Pr**	**Re**	**F1**	**Ac**	**Pr**	**Re**	**F1**
basel.	Claude	0.38 [.12,.75]	0.67 [.33,.67]	0.39 [.11,.60]	0.49 [.17,.63]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	Deepseek	0.12 [.00,.38]	0.61 [.33,.67]	0.56 [.19,.89]	0.55 [.25,.68]	0.20 [.00,.60]	0.75 [.00,.83]	0.62 [.00,.88]	0.53 [.00,.73]
	Gemini	0.00 [.00,.00]	0.00 [.00,.00]	0.00 [.00,.00]	0.00 [.00,.00]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	GPT	0.38 [.12,.75]	0.67 [.33,.67]	0.56 [.17,.67]	0.60 [.22,.67]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
def.	Claude	0.12 [.00,.38]	0.50 [.00,.56]	0.19 [.00,.50]	0.27 [.00,.47]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	Deepseek	0.12 [.00,.38]	0.78 [.33,.83]	0.36 [.05,.72]	0.47 [.08,.70]	0.60 [.20,1.00]	1.00 [.50,1.00]	0.75 [.10,1.00]	0.83 [.17,1.00]
	Gemini	0.00 [.00,.00]	0.33 [.00,.33]	0.17 [.00,.33]	0.22 [.00,.33]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	GPT	0.25 [.00,.62]	0.67 [.00,.67]	0.28 [.00,.58]	0.39 [.00,.62]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
ex.	Claude	0.25 [.00,.62]	0.67 [.00,.67]	0.28 [.00,.52]	0.39 [.00,.58]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	Deepseek	0.12 [.00,.38]	0.61 [.33,.67]	0.56 [.19,.89]	0.55 [.25,.68]	0.40 [.00,.80]	0.50 [.00,.50]	0.25 [.00,.50]	0.33 [.00,.50]
	Gemini	0.12 [.00,.38]	0.33 [.00,.33]	0.17 [.00,.33]	0.22 [.00,.33]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	GPT	0.12 [.00,.38]	0.33 [.00,.33]	0.22 [.00,.33]	0.27 [.00,.33]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
ex.def.	Claude 4.6	0.25 [.00,.62]	0.67 [.00,.67]	0.28 [.00,.52]	0.39 [.00,.58]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]
	Deepseek	0.12 [.00,.38]	0.62 [.28,.70]	0.58 [.33,.83]	0.49 [.33,.69]	0.60 [.20,1.00]	1.00 [.50,1.00]	0.75 [.10,1.00]	0.83 [.17,1.00]
	Gemini	0.12 [.00,.38]	0.33 [.00,.33]	0.17 [.00,.33]	0.22 [.00,.33]	0.60 [.20,1.00]	1.00 [.50,1.00]	0.75 [.10,1.00]	0.83 [.17,1.00]
	GPT	0.12 [.00,.38]	0.33 [.00,.33]	0.11 [.00,.33]	0.17 [.00,.33]	0.80 [.40,1.00]	1.00 [.50,1.00]	0.88 [.20,1.00]	0.93 [.29,1.00]

Table 25: Macro-averaged results for the Affective (n=8) and Behavioral (n=5) stigma subtypes. Bracketed values are 95% bootstrap confidence intervals (1000 resamples).

Strategy	Condition	N	Stig+	Acc	P	R	F1	Bias
baseline	ADHD	77	13	0.740	0.333	0.538	0.412	0.104
	Alcoholism	78	12	0.654	0.241	0.583	0.341	0.218
	Autism	78	16	0.718	0.400	0.750	0.522	0.179
	BPD	80	20	0.525	0.286	0.600	0.387	0.275
	Depression	79	4	0.848	0.250	1.000	0.400	0.152
	Schizophrenia	78	22	0.513	0.326	0.682	0.441	0.308
definition	ADHD	77	13	0.727	0.300	0.462	0.364	0.091
	Alcoholism	78	12	0.731	0.286	0.500	0.364	0.115
	Autism	78	16	0.744	0.417	0.625	0.500	0.103
	BPD	80	20	0.562	0.286	0.500	0.364	0.188
	Depression	79	4	0.873	0.250	0.750	0.375	0.101
	Schizophrenia	78	22	0.551	0.324	0.545	0.407	0.192
examples	ADHD	77	13	0.766	0.368	0.538	0.438	0.078
	Alcoholism	78	12	0.756	0.231	0.250	0.240	0.013
	Autism	78	16	0.769	0.458	0.688	0.550	0.103
	BPD	80	20	0.600	0.312	0.500	0.385	0.150
	Depression	79	4	0.924	0.333	0.500	0.400	0.025
	Schizophrenia	78	22	0.526	0.340	0.727	0.464	0.321
examples+definition	ADHD	77	13	0.766	0.353	0.462	0.400	0.052
	Alcoholism	78	12	0.821	0.417	0.417	0.417	0.000
	Autism	78	16	0.744	0.409	0.562	0.474	0.077
	BPD	80	20	0.588	0.290	0.450	0.353	0.138
	Depression	79	4	0.962	0.667	0.500	0.571	−0.013
	Schizophrenia	78	22	0.513	0.333	0.727	0.457	0.333
examples+definition+rule	ADHD	77	13	0.792	0.364	0.308	0.333	−0.026
	Alcoholism	78	12	0.833	0.429	0.250	0.316	−0.064
	Autism	78	16	0.756	0.412	0.438	0.424	0.013
	BPD	80	20	0.650	0.300	0.300	0.300	0.000
	Depression	79	4	0.949	0.500	0.250	0.333	−0.025
	Schizophrenia	78	22	0.705	0.476	0.455	0.465	−0.013

Table 26: Condition-level stigma-presence results for Claude Sonnet 4.6 by prompting strategy. N = number of items; Stig+ = gold stigma-positive items. Bias = model's predicted positive rate − gold stigma-positive rate; positive values (red) = overprediction, negative values (blue) = underprediction.

K Confusion matrices for fine-grained stigma classification

This appendix reports additional diagnostic confusion matrices for fine-grained stigma classification. The matrices are intended as descriptive error analyses rather than stable estimates for rare labels, because several fine-grained categories occurred only a small number of times. Cognitive subtype confusions often involved theoretically related categories, including personal responsibility versus trivialization, controllability versus personal responsibility, segregation endorsement versus incompetence, and dangerousness versus uncontrollability. Affective components were also frequently confused, but the small number of affective examples limits interpretation.

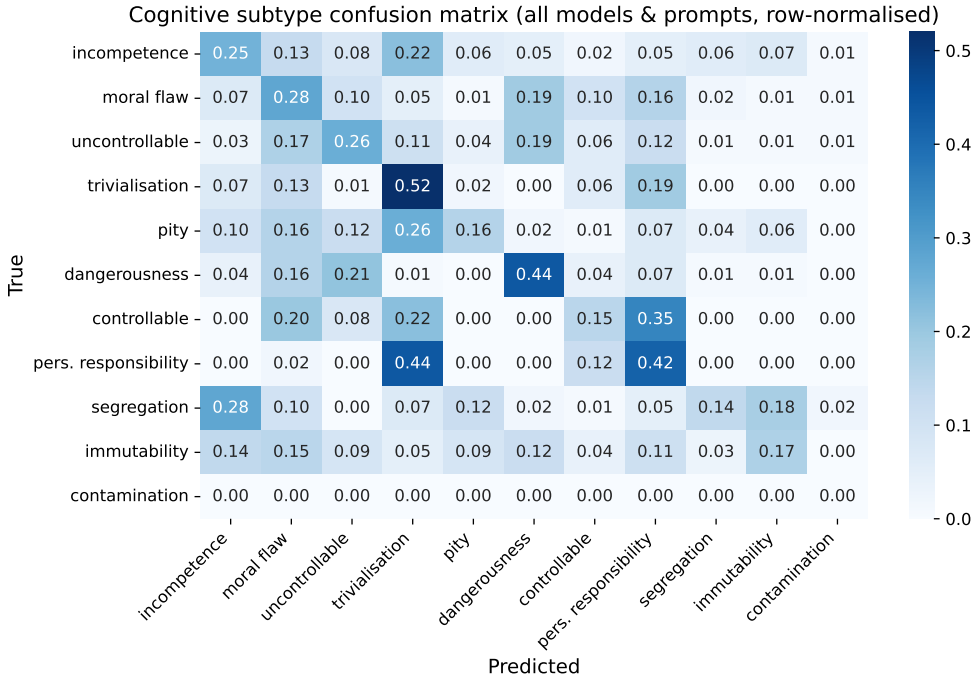


Figure 5: Row-normalized model confusion matrix for cognitive public-stigma component prediction.

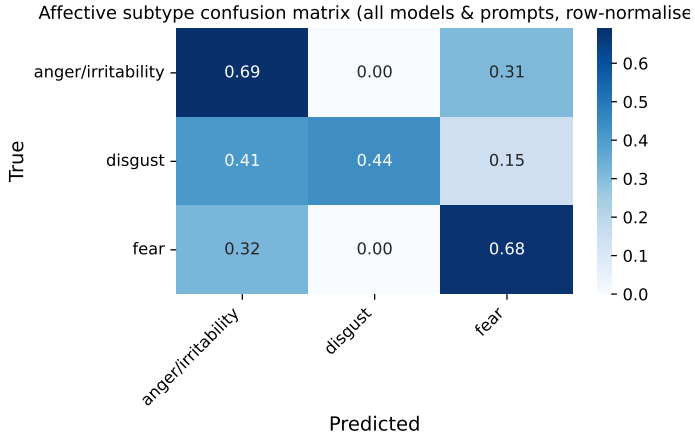


Figure 6: Row-normalized model confusion matrix for affective public-stigma component prediction.

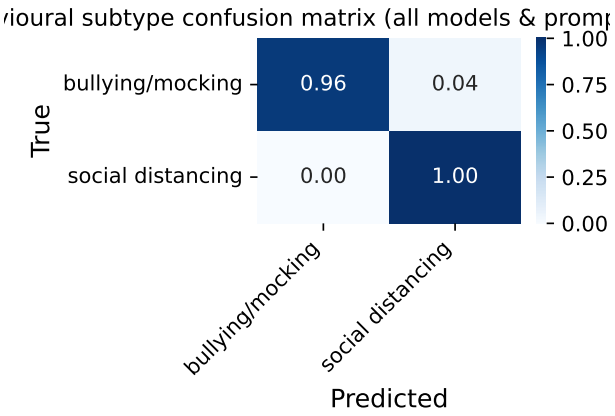


Figure 7: Row-normalized model confusion matrix for behavioral public-stigma component prediction.

L Classifier results for stigma-related constructs

We evaluated off-the-shelf classifiers for constructs related to mental health stigma: general sentiment, targeted sentiment, hate speech, and toxicity. These models were used as diagnostic baselines rather than as stigma-specific systems. For targeted sentiment, we used the DeBERTa-v3 aspect-based sentiment analysis (ABSA) model to estimate sentiment toward the mental health target term. Although applying an ABSA model outside its conventional aspect-sentiment setting is non-standard, this approach has previously been used to estimate target-directed sentiment in lexical semantic change evaluation and performed well on a synthetic benchmark dataset (Baes et al., 2025). Table 27 lists the models and their intended role in the analysis[7]

For categorical classifier outputs, we constructed contingency tables crossing the gold binary stigma label with the predicted classifier category and tested their association with chi-square tests of independence. The null hypothesis was that the gold stigma label and predicted classifier category were statistically independent in the population. That is, stigma-present and stigma-absent texts would have the same expected pattern of predicted classifier categories. Table 28 reports the omnibus tests, and Figure 8 shows the corresponding standardized residuals. For toxicity, which produced a continuous score, we inspected score distributions by gold stigma label. Toxicity scores showed substantial overlap between stigma-present and stigma-absent texts and only weak discriminatory ability (AUC = .60), suggesting that toxicity does not provide a reliable proxy for stigma detection.

Across these analyses, classifier outputs for nearby constructs did not map cleanly onto the gold stigma labels. General sentiment, targeted sentiment, and hate-speech predictions were not significantly associated with stigma presence, and toxicity scores showed substantial distributional overlap between stigma-present and stigma-absent texts. These results support the interpretation that mental health stigma is a related but distinct construct, rather than a subset of negative sentiment, hate speech, or toxicity.

[7]Full model identifiers are: `yangheng/deberta-v3-base-absa-v1.1`, `cardiffnlp/twitter-roberta-base-sentiment-latest`, `cardiffnlp/twitter-roberta-base-hate-latest`, and `unitary/unbiased-toxic-roberta`.

M Semantic similarity with stigma prototypes

To estimate if stigma can be captured though superficial semantics in terms of embedding similarity, we compare samples in our dataset with the prototypical examples of stigma we collected Section 3.5. Both prototype examples and test samples are encoded using Gemini's gemini-embedding-001.

For stigma presence detection, all stigma-present prototypes are merged into a single present prototype and all stigma-absent prototypes into a single absent prototype; a text is then classified based on whichever prototype its embedding is closest to. For multilabel fields (mode, domain, subfactor), each text's embedding is compared against the prototypes of all relevant classes within that field and assigned the label of the closest one, producing a single prediction per field. Since this approach is inherently single-label, evaluation is restricted to samples where the gold annotation contains exactly one label for the target field. Prototype embeddings results are then compared against predictions from Gemini 4 Flash. They consistently show significantly worse performance in terms of recall, especially for stigma domain.

Construct	Model	Purpose
Targeted sentiment	DeBERTa-v3 ABSA	Tests whether stigma is captured by negative sentiment toward the mental health target.
General sentiment	Twitter-RoBERTa sentiment	Tests whether stigma is captured by global negative sentiment.
Hate speech	Twitter-RoBERTa hate speech	Tests whether stigma overlaps with explicit hate or derogation.
Toxicity	Unbiased Toxic RoBERTa	Tests whether stigma overlaps with insult, abuse, or toxic language.

Table 27: Neighboring-construct baselines used to test whether mental health stigma is captured by sentiment, hate speech, or toxicity models.

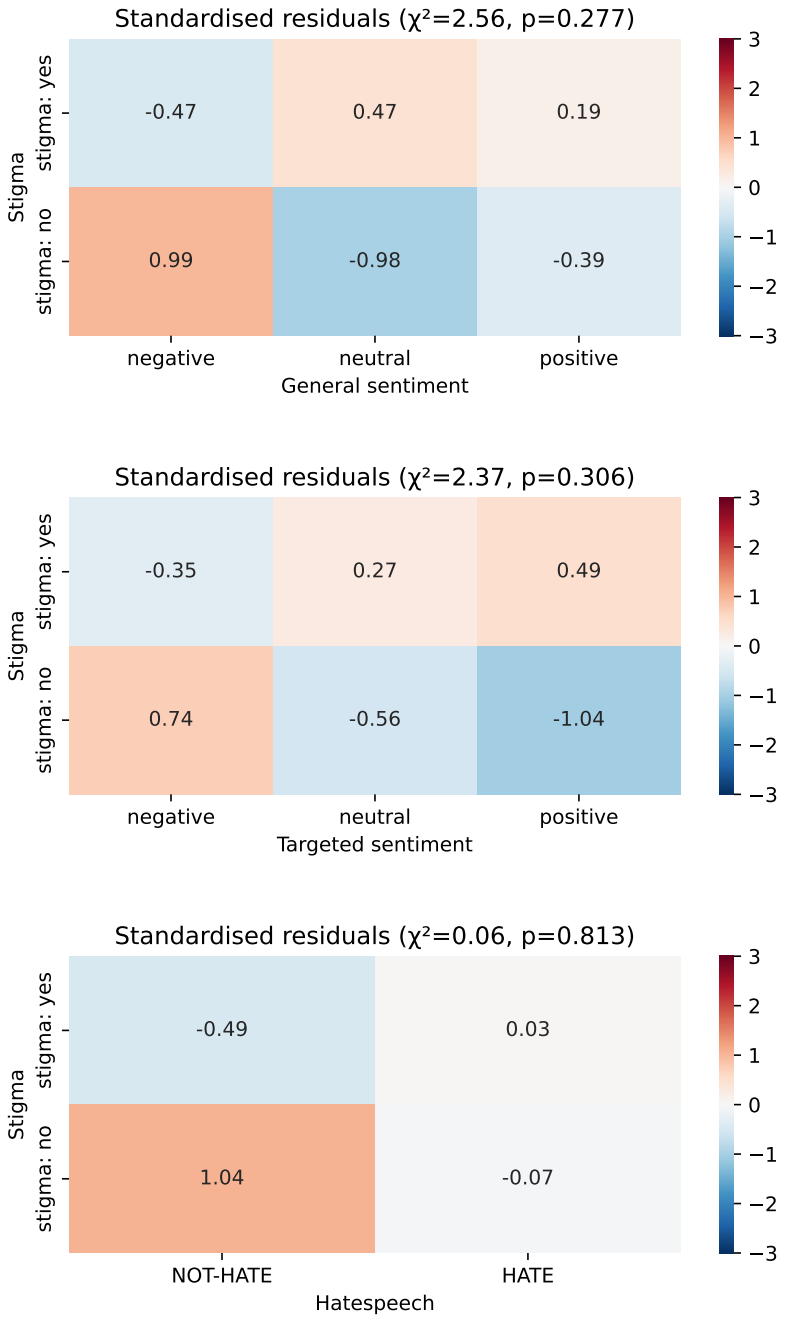


Figure 8: Standardized residuals from chi-square tests of independence between gold stigma labels and adjacent classifier predictions. Residuals are shown for predicted general sentiment, targeted sentiment, and hate-speech categories.

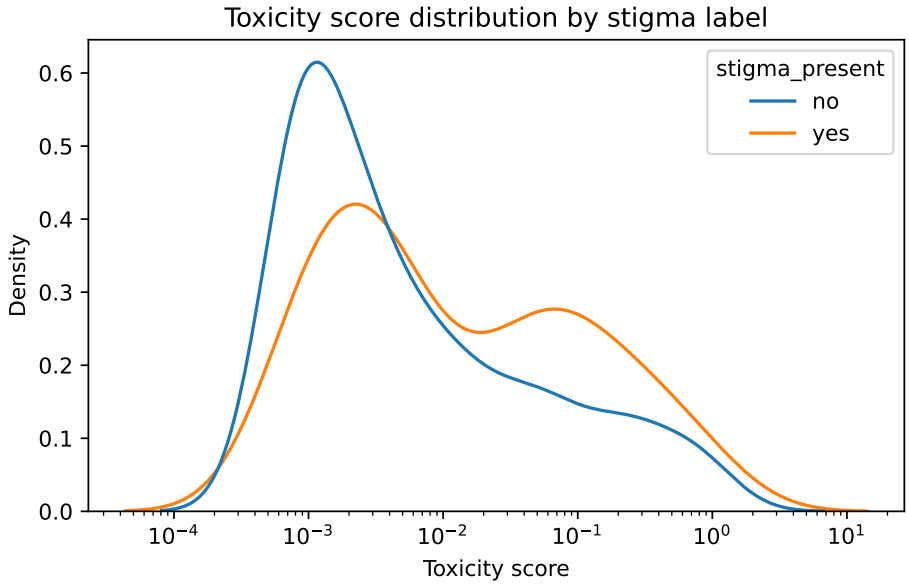


Figure 9: Distribution of toxicity scores by gold stigma label. The distributions overlap substantially, suggesting that toxicity scores do not cleanly separate stigma-present from stigma-absent texts.

Classifier prediction	**df**	χ^2	p
General sentiment	2	2.56	.277
Targeted sentiment	2	2.37	.306
Hate speech	1	0.06	.813

Table 28: Chi-square tests of independence between gold stigma labels and classifier predictions for neighboring constructs. Toxicity is not included because it was analyzed as a continuous score rather than a categorical prediction.

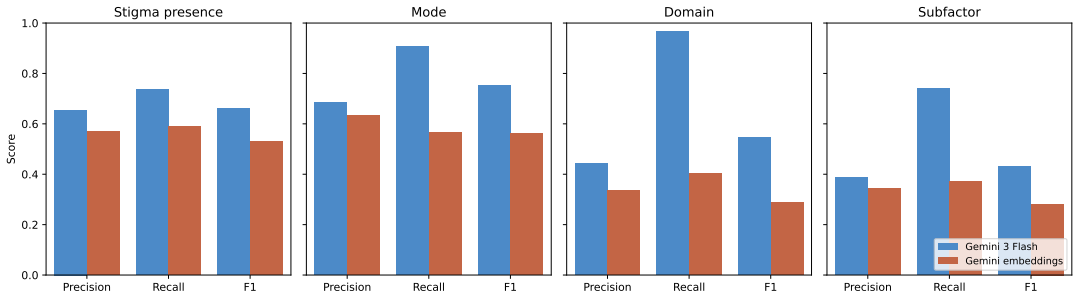


Figure 10: Prediction of stigma presence, mode, domain and subfactor