The Simulation-to-Real Gap in Mental Health Chatbot Conversations

Samyam Lamichhane*
slamichhane@virginia.edu
University of Virginia
Department of Computer Science
Charlottesville, Virginia, USA

Hongru Du
hongrudu@virginia.edu
University of Virginia
Systems and Information Engineering
Charlottesville, Virginia, USA

Zhiyuan Wang
vmf9pr@virginia.edu
University of Virginia
Systems and Information Engineering
Charlottesville, Virginia, USA

Colin Depp
cdepp@health.ucsd.edu
University of California, San Diego
Department of Psychiatry
San Diego, California, USA

Subigya Nepal
sknepal@virginia.edu
University of Virginia
Department of Computer Science
Charlottesville, Virginia, USA

arXiv:submit/8388386 [cs.AI] 17 Nov 2026

Abstract

Simulated users, typically LLMs prompted with personas, are increasingly used to build and evaluate mental health tools. Yet it is rarely tested whether they resemble the real users they represent. We compare mental health messages from three public corpora of real chatbot conversations with seven simulated datasets, using a consistent LLM annotator to assess distress, help-seeking, and crisis across all sources. We find that simulated users are easily distinguished from real users. Their messages are more often labeled as distressed or in crisis. They also state their distress more explicitly and appear to express help-seeking as disclosure rather than as practical questions. These differences matter in practice: detectors trained on simulated users score between 0.08 and 0.26 AUROC lower on real users than detectors trained on real users, across four simulators, three tasks and every cross-validation fold, and they miss indirect expressions of distress and crisis. The gap persists after normalizing writing style (correcting apostrophes, capitalization and text-speak), rebalancing personas and changing the labeler and it is several times larger than the loss from moving between real corpora. Adding simulated data to real training data does not improve real-user performance, and validation on simulated users overstates real performance by up to 0.22 AUROC. We conclude that simulated users are a poor substitute for real users when training or validating such detectors, and we recommend guardrails for their use.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Natural language processing**; • **Applied computing** → **Health informatics**; • **Information systems** → *Data mining*; • **Human-centered computing** → Human computer interaction (HCI).

Keywords

Simulated users, synthetic data, mental health, large language models, chatbots, distribution shift, sim-to-real gap, crisis detection, LLM labeling, deep learning

*Corresponding author.

1 Introduction

People increasingly turn to AI chatbots for mental health support. In a 2025 national survey, almost a fifth of US adolescents and young adults reported using chatbots for mental health advice, up by nearly half from a year earlier, and most told no one [25].

Understanding how people use chatbots for mental health and whether chatbots respond safely, is therefore a pressing need. Real conversations are sensitive and difficult to share, so researchers increasingly rely on simulated users, typically LLMs prompted with personas, to train detectors and to evaluate chatbots [4, 9, 10, 22, 40, 44, 48]. For this study alone, we identified seventeen public simulated mental health datasets. These simulators are built for different purposes, from representing everyday interactions to testing responses to high-risk situations, and they are usually validated through expert judgments of how realistic their conversations appear (Section 2). Their value depends on whether they behave like the real users they represent and whether they support the tasks they are used for. Comparisons with real use have so far relied on private production data [35, 39, 53], and no study has tested whether detectors trained on simulated users work on real users. We compare several simulators with public real conversations and test their detectors on real users.

We compare mental health messages from three public corpora of real chatbot conversations [46, 49, 51] with seven simulated datasets, labeling every message with the same LLM labelers. We ask how simulated users differ from real users, whether detectors trained on them work on real users, and what explains the gap. We also test whether the gap comes from writing style, which could be fixed by editing the text, or from content, which could not.

Our contributions are as follows:

- **A comparison of real and simulated users.** The messages of simulated users are more often labeled as distressed or in crisis than those of real users and they appear to express help-seeking as disclosure rather than as practical requests. Simulated users are also easily distinguished from real users by their writing style.
- **A consistent simulation-to-real gap.** Detectors trained on simulated users score 0.08 to 0.26 AUROC lower on real users than detectors trained on real users, across four simulators, three tasks and every fold. The gap persists

after normalizing writing style, rebalancing personas and changing the labeler, so it is not a surface artifact and it is several times larger than the loss from moving between real corpora.
- **Implications for training and evaluation.** Adding simulated data does not improve real-user performance, validation on simulated users overstates real performance by up to 0.22 AUROC and reported results depend on the choice of labels. Safety evaluations that rely on simulated users may therefore not reflect real use and ethically collected real conversations are needed to evaluate chatbot safety.

2 Related Work

Detecting mental health signals in text. Most work on detecting distress and suicide risk from text has used social media. The CLPsych 2019 shared task graded the suicide risk of Reddit users [54], and Gaur et al. evaluated models on psychiatrist-annotated Reddit data using a five-category scheme adapted from the Columbia–Suicide Severity Rating Scale (C-SSRS) [11]. These models often transfer poorly beyond the data on which they were built: cross-dataset evaluations of depression detection have found that classifiers trained on one social media dataset perform worse on another than classifiers trained on the target dataset [13]. Similarly, a review of 75 studies on predicting mental health status from social media found no common approach to evaluating whether the labels captured the intended mental health constructs [6]. Conversations with chatbots provide a newer setting for these tasks and public logs of real chatbot conversations [46, 49, 51] make it possible to study such use directly.

Simulated users in mental health. Because real mental health conversations are sensitive and difficult to share, early work on emotional support dialogue used structured crowdsourcing to collect conversations. For example, ESConv paired crowdworkers seeking support for personal experiences with crowdworkers trained to provide emotional support [19]. More recent work uses LLMs to simulate patients for counselor and therapist training, guiding their behavior with cognitive models developed by mental health clinicians [42] or with rules derived from expert counselors' feedback [21]. LLM-simulated users are also used to build guardrails and to audit chatbot safety at scale [4, 9, 10, 22, 44, 48]. In the largest such evaluation, Transluce tested 77 model variants in more than 50,000 conversations with 157 simulated users [39, 40]. These simulators are validated mainly through expert judgments of how realistic they appear, which do not establish that they match the language or behavior of real users [28]. Three studies have drawn on real use. An analysis of 20,000 real conversations with a mental health AI argued that small simulated benchmarks may not reflect the diversity of real use and called for auditing real conversations as a complement [35] and Transluce validated its simulated users against anonymized data on how real users of ChatGPT and Claude discuss mental health, without access to conversation contents [39]. A study of a mental health chatbot's own simulator found that zero-shot simulated users were overly compliant and that fine-tuning on real sessions narrowed the gap [53]. All three relied on private data that others cannot examine, and whether results from simulated users carry over to real users remains an open problem [28]. To our knowledge, no study has compared several simulators with public real conversations or tested whether models trained on simulated users work on real users. We address both gaps.

Realism of LLM-simulated humans. Whether LLMs can stand in for people has been debated beyond mental health. When prompted with sociodemographic backstories, GPT-3 reproduced some aggregate response patterns across subgroups in U.S. political surveys [2]. Other work found that identity-prompted LLMs can misrepresent groups and understate variation within them, warranting caution when substituting LLMs for human participants whose identities matter to the task [41]. LLM-based user simulators have been found to diverge from human behavior in conversational recommendation [8] and to differ from human users in how they solve problems and engage in task-oriented dialogue [43]. Simulated users also tend to introduce less conversational friction than real users, which may lead simulation-based evaluations to overestimate chatbot performance [20]. A comparison of 24 LLM-based simulators with real WildChat users on coding and writing tasks found a large gap in behavior for every simulator, but did not measure how this gap affects models trained or evaluated on simulated users [27]. Such comparisons remain rare in mental health, where simulated users serve both as training data and as evaluation subjects. It is also unclear whether simulated users differ from real users only in surface language, which could be corrected by editing the text, or in what they say. We separate the two by comparing writing style and content.

Synthetic training data and the sim-to-real gap. The term *sim-to-real* originates in robotics, where models trained in simulation must transfer to the physical world [38]. Across the text classification tasks studied, training on LLM-generated data generally led to larger performance drops relative to training on real-world data for more subjective tasks [18]. Distress, help-seeking and crisis are subjective constructs of this kind, which makes their transfer from simulated to real users an open question that we address empirically.

LLMs as annotators. LLMs can outperform crowdworkers on some text annotation tasks [12], but their performance depends strongly on the dataset and task and must be validated against human labels [32]. Disagreement between annotators can also reflect genuine variation rather than noise [33]. We therefore validate our LLM labels against human annotations where possible, use three LLM labelers and test whether our conclusions depend on the choice of labeler.

3 Experimental Setup

3.1 Real Users

The real user interactions are drawn from three public corpora of chatbot interactions: WildChat [49] (conversations with ChatGPT), ShareChat [46] (conversations with ChatGPT, Perplexity, Grok, Gemini and Claude) and LMSYS-Chat-1M [51] (chatbot interactions from Chatbot Arena). Post-acquisition, we extract mental health (MH) conversations using an LLM labeler as outlined in Section 3.3, which produced a total of 12,875 conversations and 21,009 MH user messages (Table 1).

Table 1: Sources used in this study. MH msgs counts user messages that meet the MH rule (Section 3.3); the 21,009 real messages form the real-user set.

Source	Written by	Conv.	MH msgs
WildChat	Real users	5,530	8,019
ShareChat	Real users	1,339	3,513
LMSYS-Chat	Real users	6,006	9,477
SimMH	LLM	45,402	332,169
MHDialog	LLM	978	8,685
SIM-VAIL	LLM	810	5,437
VERA-MH	LLM	2,000	18,546
MindGuard	Psychologists	67	788
MHB	Experts + LLM	865	1,744
HealthBench-Psych	Physicians + LLM	611	748

3.2 Simulated Users

Our simulated users are derived from seven sources as outlined in Table 1. The first four form the core of our analysis:

- SimMH-Chat [40] (hereafter SimMH), released by Transluce to test how chatbots respond to users in crisis, in which an LLM plays 157 personas in conversation with chatbots.
- MHDialog [48], a corpus of LLM-generated user-chatbot dialogues.
- MindGuard [9], in which clinical psychologists role-play patients.
- MentalHealthBench (MHB) [22], an expert-built benchmark from OpenAI that we use only as a test set.

Moreover, the remaining three sources broaden the range of risks considered in this study:

- SIM-VAIL [44], in which Claude Sonnet 4.5 plays users with mania, paranoia, obsessive-compulsive disorder, insecure attachment or depression.
- VERA-MH [4], a user agent that simulates suicide risk, from passive ideation to acute plans.
- HealthBench-Psych [10], the MH subset of HealthBench [3], written with physicians.

Section 4 also compares the writing style of ten additional simulated datasets. Only the seven sources listed above are labeled and used to train and evaluate detectors.

3.3 Labels

The Gemini 2.5 Flash-Lite model [37] annotates all the real and simulated conversations using one prompt and one output schema to maintain consistency (Appendix A). Each user message is labeled on five fields: *topicality* (0–2), the *subject* of the difficulty described (self, other or none), *distress* (0–2), *help-seeking* (0–2) and one of twelve MH *categories*, where 0 indicates that the attribute is absent and 2 that it is central or severe. For help-seeking, level 1 marks general well-being talk without an explicit request and level 2 an explicit request for advice, treatment or support. A message is help-seeking if its level is at least 1; the task therefore captures engagement with help or well-being rather than explicit requests alone. The annotation fields, MH categories and calibration examples employed in the prompt were adapted from the classification rubric of McBain et al. [26], developed to identify MH conversations in WildChat [49]. We label each message rather than each conversation, drop clinical language and add a field for whose difficulty is described. Throughout, the unit of analysis is the user message and we refer to the authors of these messages as real or simulated users. Specifically, a message is considered MH if *topicality* ≥ 1 and the *subject* $\in \{self, other\}$, and the binary categories are defined as follows:

- **distress:** *distress* ≥ 1
- **help-seeking:** *help-seeking* ≥ 1
- **crisis:** *category* equal to crisis

Together, these categories cover two complementary dimensions of a user's state. Distress and crisis lie on a single spectrum of severity, from noticeable negative emotion to imminent risk, whereas help-seeking is a separate dimension that captures whether the user seeks assistance.

Two additional LLMs further relabel messages following the same prompt to assess the dependence of our results on the labeler: Claude Haiku 5.5 [1] (without extended thinking, default sampling) and GPT-4.1 mini [30] (temperature=0). They relabel 20,000 user messages (7,754 real messages from the held-out conversations of an earlier fixed split and 12,246 from randomly sampled SimMH conversations) as well as all conversations of SIM-VAIL, VERA-MH and HealthBench-Psych. Analyses that involve several labelers use the messages that at least two of the three labelers consider MH, which leaves 3,447 real and 9,954 SimMH messages. A consensus label keeps a message only where all three labelers agree on the task label and discards it otherwise.

3.4 Human Validation of the LLM Labels

We validated the LLM labels on publicly available datasets with human annotations using the same prompt across different tasks.

- **Crisis:** We evaluated crisis on CRADLE Bench [5], which contains clinician annotations of suicidal ideation and self-harm, and on MindGuard, in which psychologists marked each turn as safe or unsafe. The findings indicate moderate to substantial agreement on CRADLE Bench (Cohen's $\kappa = 0.57$ for Gemini and 0.74 for Haiku), although Gemini labeled only 52% of ongoing self-harm posts as crisis versus 91% for Haiku. Gemini also agreed with MindGuard annotations at $\kappa = 0.61$.
- **Distress:** For distress, we used the Depression Severity dataset [29] and the SAD corpus [24]. Both labelers proved lenient: they labeled 75% (Gemini) and 70% (Haiku) of posts annotated as minimal depression as distressed, while recalling 97–98% of posts annotated as mild or worse. Nevertheless, the labels align with humans' severity ratings, as indicated by Spearman correlations of 0.35–0.38 and 0.43–0.44 on Depression Severity and SAD, respectively.
- **Help-seeking:** For help-seeking, we used M-Help [34], in which mental health professionals marked Reddit posts as help-seeking or not (1,790 unique posts after removing duplicates and posts with conflicting annotations). Agreement was poor ($\kappa = 0.13$ for Gemini and 0.06 for Haiku). Both labelers marked far more posts as help-seeking than the annotators did (73% and 95% versus 56%), including 66%

(Gemini) and 92% (Haiku) of the posts annotated as not help-seeking. Counting only explicit requests (*help-seeking* = 2) changed little ($\kappa = 0.14$ and 0.18). M-Help counts indirect requests as help-seeking and gives no explicit rule for the negative class, so this comparison mixes labeler error with a difference in definitions.

We therefore consider crisis a credible label and distress a lenient one and we could not validate help-seeking, so we interpret help-seeking results with caution.

3.5 Detector

We use MentalBERT [15], a BERT variant further pretrained on mental health posts from Reddit, as the detector. For each task, training set and fold, we fine-tune a fresh copy of MentalBERT as a binary classifier for five epochs, with a learning rate of 2×10^{-5}, an effective batch size of 32 and a maximum input length of 512 tokens. Moreover, within each comparison, training sets are matched in size and class mix, and all detectors are evaluated with grouped five-fold cross-validation such that real users are assigned to folds by conversation and simulated users by persona where personas exist and otherwise by conversation, to avoid data leakage.

We assess every difference with two complementary tests. A paired t-test over the five folds reflects variation in the training data, whereas a paired conversation bootstrap (1,000–2,000 resamples) on the pooled out-of-fold predictions yields more optimistic confidence intervals. We describe a result as robust when both tests agree. When several pairwise differences are compared within a task, the p-values of both tests are Holm-corrected [14].

Choice of detector. We use a fine-tuned classifier rather than an LLM as the detector for three reasons. First, our question is whether training data transfers, which requires a model whose decisions are learned from that data; a prompted LLM has no training set to vary. Second, guardrails must screen every message, and a model of this size is inexpensive and can run locally, so sensitive messages need not be sent to an external service. Third, our labels come from LLMs, so an LLM detector would be scored largely against the judgments of a similar model. A classifier also inherits the biases of its training labels, but the architecture is the same for every training source, so differences between detectors reflect the training data.

4 How Simulated Users Differ from Real Users

4.1 Simulated Users Are Readily Detectable

We describe each message by 87 linguistic features covering length, mechanics, punctuation, informal markers, vocabulary, voice, sentiment, syntax and readability (Appendix B). A gradient-boosted classifier (LightGBM [16]) trained on these features to distinguish real from simulated messages flags 96% of SimMH messages as simulated, compared with 65% for MHDialog, 54% for MindGuard and 13% for MHB. These rates partly reflect SimMH's 97% share of the simulated training messages: when each simulator receives equal weight, detection of MHDialog and MindGuard rises to 95% and 79% (see Appendix D).

We then compare twenty sources with real users: the three real corpora, the seven labeled simulators and ten further public simulated datasets that we did not label. We found the ten through a keyword search of the web, Hugging Face and GitHub (Appendix C). We use two measures: writing-style distance, the mean absolute Cliff's δ [7] between the source and the pooled real users over the 87 features, and detectability, the AUROC of a logistic-regression classifier that separates the source from real users (Figure 1; values in Table 8).

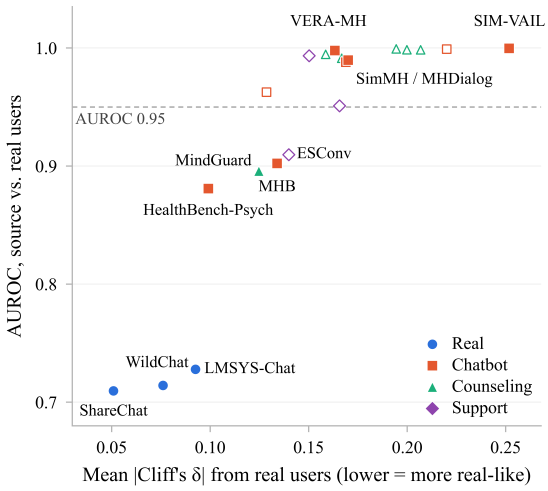


Figure 1: Writing-style distance from real users (mean absolute Cliff's δ over 87 features) against how easily a classifier separates each source from real users (AUROC). Lower left is more like real users. Marker shape gives the setting; hollow markers are sources used only in this comparison.

Three findings emerge from the comparative analysis:

- Every fully LLM-written source reaches an AUROC of at least 0.95, whereas human-written sources (MindGuard and ESConv) and expert-curated sources (MHB and HealthBench-Psych) fall below this level at 0.88–0.91.
- SIM-VAIL, the most recent simulator, is the easiest to separate from real users (AUROC 1.00) and its messages are far longer (median 139 words per message, compared with 17–37 for real sources).
- Simulators fail in opposite directions: 83% of SimMH messages are entirely lowercase, whereas MHDialog has none, yet the two are equally distant from real users.

4.2 Language and Content

To test whether simulated users differ from real users mainly in *how* they write or in *what* they write about, we train three LightGBM classifiers to separate real from SimMH messages: one on the 87 writing features (language), one on the LLM labels (content) and one on both. Table 2 reports their AUROC on a standard test set and on a *content-matched* test set, in which real and SimMH messages have the same mix of labels, so that content alone cannot separate them.

Language alone separates SimMH from real users almost perfectly (AUROC 0.995) and still does so when content is matched (0.993). Content also separates the two (0.878) but adds almost nothing once language is included (0.998), so style is the most visible

Table 2: AUROC of real-versus-SimMH classifiers trained on language features, content labels or both. Matched test sets contain equal numbers of real and simulated messages for every combination of labels. MHB rows apply the SimMH classifiers to MHB without retraining.

Test set	Language	Content	Both
Real vs. SimMH	0.995	0.878	0.998
Real vs. SimMH (*matched*)	0.993	0.500	0.991
Real vs. MHB	0.730	0.666	0.753
Real vs. MHB (*matched*)	0.699	0.500	0.682

Table 3: Label rates among MH user messages (Gemini 2.5 Flash-Lite labels). D: distress. Columns are not mutually exclusive.

Source	D ≥ 1	D = 2	Help-seeking	Crisis
WildChat	75%	15%	36%	3%
ShareChat	78%	33%	32%	6%
LMSYS-Chat	77%	28%	43%	13%
SimMH	97%	59%	13%	21%
MHB	93%	37%	31%	12%
MHDialog	92%	34%	31%	6%
MindGuard	92%	29%	13%	4%
SIM-VAIL	88%	53%	29%	2%
VERA-MH	98%	70%	73%	61%
HealthBench-Psych	84%	22%	70%	7%

difference between simulated and real users. The SimMH classifier transfers poorly to MHB (0.699–0.730 from language), which suggests that the cues that reveal a simulated user are specific to each simulator. Sections 5 and 6 show, however, that style is not the difference that impairs detectors trained on simulated data.

4.3 Label Distributions Across Sources

Table 3 reports the label distribution among MH messages for all ten labeled sources. Among real messages, 75–78% are labeled as distressed, 15–33% as severely distressed, 32–43% as help-seeking and 3–13% as crisis. Every simulator has a higher rate of LLM-labeled distress (84–98%). In SimMH, which was built to test responses to crisis, 59% of messages are labeled as severely distressed, about twice the rate of the most severe real corpus. SIM-VAIL messages are severely distressed almost as often (53%) but are rarely labeled as crisis (2.3%). HealthBench-Psych is the only simulator with severity close to that of real users (22% severely distressed, 7.4% crisis) but 70% of its messages are help-seeking, compared with 32–43% for real users. No simulator reproduces the label distribution of real users, and these patterns persist under other labelers: on the messages labeled by all three, Haiku and GPT label 79% and 38% of VERA-MH messages as crisis, compared with 6% and 7% of real messages (see Table 9 in Appendix E).

4.4 Help-Seeking in Real and Simulated Users

Although MHB is among the simulators that write most like real users (Figure 1), it differs from real users in how help-seeking is expressed. To test this, we ask whether the writing features that rise or fall with a label in real messages do the same in MHB: for each feature, we compute Cliff's δ between messages with and without the label, separately for real users and MHB, and measure agreement as the Spearman correlation between the two sets of values. Agreement is high for distress and crisis ($\rho = 0.70$ and 0.68) but absent for help-seeking ($\rho = -0.10$, 95% CI = -0.28 to $+0.15$), and no subgroup of MHB restores it (see Table 10 in Appendix F). Because we could not validate the help-seeking label (Section 3.4), these results are suggestive.

We further classify help-seeking messages, which include general well-being talk as well as explicit requests (Section 3.3), by request type with regular expressions (see Table 11 in Appendix G). Real users make explicit how-to requests twice as often as MHB users (25.8% vs. 11.9%) and request practical resources four times as often (7.6% vs. 1.7%). By contrast, 70% of MHB help-seeking matches no request type, compared with 52% for real users: it is mainly a disclosure of distress with an implicit request, and 42% of it falls in the crisis or other acute categories, compared with 19% for real users.

Overall, these results suggest that simulated users differ from real users in more than style: their messages are more often labeled as distressed, some simulators are far more often in crisis and they appear to express need through disclosure.

5 The Sim-to-Real Gap

5.1 Experimental Design

The central question is whether a detector trained on simulated users performs well on real users. We therefore train two detectors on training sets of equal size and class balance, one on a simulator and one on real users, and evaluate both on the same held-out real users. We quantify the difference as the *gap*: the AUROC of the real-trained detector minus that of the simulator-trained detector, both measured on held-out real users. We use two designs:

- **SimMH design (Section 5.2).** We train detectors on real users and on SimMH as released. Within each fold and task, both use the same training size: 3,400–3,540, 5,500–5,600 and 1,130–1,210 messages per class for distress, help-seeking and crisis, respectively.
- **Pairs design (Section 5.3).** We train detectors on SimMH, MHDialog, SIM-VAIL and VERA-MH and pair each with a detector trained on the same number of real messages, capped at 4,000 messages per class. We omit a simulator for a task when it has fewer than 150 messages in either class, which excludes VERA-MH for distress (too few non-distressed messages) and SIM-VAIL for crisis (too few crisis messages). In this design, MHB, MindGuard and HealthBench-Psych serve only as additional test sets.

The SimMH design uses all the data that SimMH offers per class without a cap and is the basis for the modified versions of SimMH and the re-scoring under other labelers in Section 6. The pairs design extends the comparison to other simulators and gives each its own size-matched real partner, because simulators differ greatly in how many messages of each class they provide.

Both designs use grouped five-fold cross-validation, dividing real messages by conversation and simulated messages by persona

Table 4: AUROC on held-out real users in the SimMH design (mean ± standard deviation over five folds). The gap is the fold-mean difference between the real-trained and SimMH-trained detectors, with its 95% t-interval.

Task	Real	SimMH	Gap (95% CI)
Distress	0.857 ± 0.008	0.700 ± 0.012	0.157 (0.138, 0.176)
Help-seeking	0.864 ± 0.007	0.709 ± 0.037	0.154 (0.114, 0.195)
Crisis	0.903 ± 0.015	0.725 ± 0.018	0.178 (0.145, 0.212)

where personas exist and otherwise by conversation. Each of the 21,009 real MH messages is tested exactly once; 76.5% are labeled as distressed, 38.4% as help-seeking and 8.1% as crisis. Each detector is tuned on validation data from its own training source: real data for real-trained detectors and simulated data for simulator-trained ones. This mirrors a practitioner who has no real data.

5.2 SimMH-Trained Detectors

Detectors trained on SimMH lose 0.15 to 0.18 AUROC on real users relative to detectors trained on real users (Table 4). The real-trained detector is better in every fold for every task ($p < 0.001$ for each task) and the pooled bootstrap agrees. The loss also appears in AUPRC: for crisis, it falls from 0.60 with real training to 0.22 with SimMH training, against a base rate of 0.08. Results are less stable with SimMH training, with a fold-to-fold standard deviation of up to 0.037 for help-seeking, compared with 0.007 for real training.

The loss matters in practice once the decision threshold is set. With the threshold tuned on simulated validation data, the SimMH-trained detector recalls only 25% of real messages labeled as distressed, compared with 76% for the real-trained detector, because a threshold set on simulated users is too strict for real users. Tuning the threshold on real validation data raises recall only to 59% (Table 12 in Appendix H).

5.3 The Gap Across Simulators

Every simulator trails its real-trained partner on every task and in every fold (Figure 2; Table 13 in Appendix I): all fifty fold-level gaps are positive and every pooled confidence interval excludes zero. Gaps range from 0.079 (MHDialog, help-seeking) to 0.261 (VERA-MH, help-seeking). Because training size also differs across simulators (348 to 4,000 messages per class on average), differences between gaps may partly reflect size. Newer simulators do not narrow the gap. On help-seeking, the frontier-LLM simulator SIM-VAIL is no better than SimMH and the specialist simulator VERA-MH is significantly worse, possibly because both express help-seeking through disclosure (Section 4.4). On crisis, VERA-MH likely transfers worst, although this ordering is significant only in the pooled bootstrap (fold-level Holm $p = 0.15$). MHDialog has the smallest gap on every task, and its advantage over SimMH is robust for distress and help-seeking (Table 14 in Appendix I).

The pattern largely holds on the three test-only datasets. Each real-trained detector outperforms its simulator-trained partner on HealthBench-Psych and MHB for every task, and on MindGuard for eight of ten pairs (Appendix J). The exceptions are the MHDialog-trained detectors for distress and crisis.

5.4 Error Analysis

We compare the errors of the SimMH-trained and real-trained detectors on 2,111 held-out real messages from a single fixed split. Both detectors use thresholds tuned on real validation data, so an error reflects how the detector ranks a message rather than a poorly chosen cutoff (see Table 16 in Appendix K).

- **Distress.** The SimMH-trained detector misses 35.2% of distressed messages, compared with 21.6% for the real-trained detector. The missed messages are not milder: 31% of missed and 34% of caught messages carry the highest distress level. The SimMH-trained detector also raises false alarms on calm health questions, such as a user listing their blood-pressure medication; physical-health and treatment topics account for 57% of the false alarms that only this detector makes.
- **Help-seeking.** The SimMH-trained detector misses 42.5% of help-seeking messages, compared with 25.1% for the real-trained detector. It misses long, detailed requests (median 39 words, compared with 16 for caught requests) and flags short emotional statements.
- **Crisis.** The main failure is over-flagging: the SimMH-trained detector flags twice as many non-crisis messages as the real-trained detector. It misses a similar number of crisis messages, but different ones. It flags short emotional questions, such as a user asking how to become fearless, and misses indirect expressions of crisis, such as questions about news of a suicide, slang, quoted lyrics and threats toward others.

These errors mirror the differences described in Section 4: real users often express need through practical requests and express crisis indirectly, whereas simulated users rarely do. Part of the real-trained detector's advantage may, however, reflect artifacts of labeling real messages in context, which we discuss in Section 9.

6 What Does Not Explain the Gap

6.1 Style

Simulated users are most easily recognized by their style (Section 4.2), so style is the most obvious candidate. Three results indicate that it does not account for the gap.

Few style features act as shortcuts. A style feature could mislead a detector if it changed with the label more strongly in simulated users than in real users. We call such a feature a *shortcut* (Appendix L). For SimMH and MHB, at most five of the 87 writing features per task qualify as shortcuts, including at most two of the 36 punctuation, mechanics, informal and formatting features. Those in SimMH disappear when features are compared within each persona, which suggests that they reflect which personas tend to receive the label rather than how simulated users write.

Normalizing the style does not help. We retrain detectors on a style-normalized version of SimMH (Fix A) that corrects apostrophes, text-speak, the pronoun "I" and sentence capitalization, keeping the training size of the SimMH design. The effect relative to SimMH is indistinguishable from zero for every task (Table 17 in Appendix M), and real-trained detectors remain ahead by 0.15 to 0.18 AUROC.

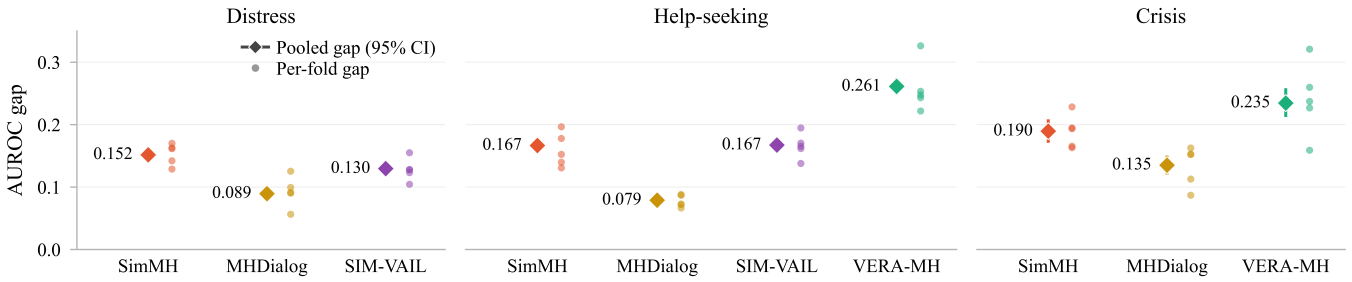


Figure 2: AUROC gap between real-trained and simulator-trained detectors on held-out real users, per simulator and task. A larger gap means worse transfer, and zero means no loss. Diamonds show pooled gaps with 95% bootstrap confidence intervals; small dots show the five per-fold gaps. VERA-MH for distress and SIM-VAIL for crisis are omitted because one class has too few messages.

Clean formatting does not help either. MHDialog is uniformly well formatted, the opposite of SimMH, yet it still leaves a gap on every task (Section 5.3). Because MHDialog also differs from SimMH in content, this result supports the conclusion but does not test it in isolation.

6.2 Persona Balance

A second explanation is that SimMH's labels are confounded with its personas: some personas are almost always distressed and others almost never, so a detector could learn to recognize personas rather than distress. We test this with a persona-balanced version of SimMH (Fix B) that keeps personas with at least 20 messages per class and balances classes within each persona. It has no reliable benefit: no fold-level effect is significant, the pooled bootstrap shows only a small gain for distress and a loss for crisis (Table 17), and real-trained detectors remain ahead by 0.15 to 0.20 AUROC.

Two features of this version plausibly explain why rebalancing does not help. Only personas with enough messages of both classes qualify (57–62 of 157 personas for distress, 74–78 for crisis), so Fix B trains on less varied data. Its negative examples also come mostly from personas that are distressed most of the time and these are atypical of calm real users.

6.3 The Labeler

All detectors so far were trained and tested on Gemini labels, so the gap could reflect the behavior of one labeler. We test this in two steps.

Changing only the test labels. We re-score the five-fold detectors of the SimMH design against five label sources: Gemini, Haiku, GPT, a majority vote and consensus. The gap holds under every source, ranging from 0.097 to 0.173 (Table 18 in Appendix N). A paired conversation bootstrap shared across sources finds no significant difference in gap size between any two sources (0 of 30 comparisons after Holm correction; smallest adjusted $p = 0.37$).

Changing the training labels too. We then train detectors on each labeler's labels (the *labeler design*), using the 3,447 real MH messages labeled by all three labelers, split into five folds by conversation. For each task and label source, we train one detector on real users (real-only) and one on simulated users (simulated-only), pooling SimMH, SIM-VAIL, VERA-MH and HealthBench-Psych. The two training sets are matched in size and class balance and we score both detectors on real users against the labels they were trained on. The gap holds for every label source and task (Table 19 in Appendix N): every confidence interval excludes zero, and the gap is positive in all five folds except for Haiku crisis (four of five). Consensus labels give the smallest distress gap (0.057) but the largest crisis gap (0.257). When detectors are scored against another labeler's labels, the gap stays significant in almost every combination; the exceptions all involve detectors trained on Haiku's labels (Appendix N).

None of the four explanations accounts for the gap. This points to differences in content described in Section 4, which cannot be removed by editing the text or rebalancing the data, and which are specific to simulated users: moving between real corpora costs far less.

6.4 Real-to-Real Transfer

A fourth explanation is that the gap is not specific to simulation: a detector might lose as much when moved from one group of real users to another. We test this with the three real corpora, using the folds of Section 5.1. In each fold and task, we train one detector on the training folds of all three corpora (*in-distribution*) and, for each corpus, one detector on the other two corpora only (*cross-corpus*). All four detectors use the same class-balanced training size, about 1,960, 2,800 and 320 messages per class for distress, help-seeking and crisis, limited by the pool without LMSYS-Chat. Each cross-corpus detector is tuned only on its two training corpora, and both detectors of a comparison are scored on the same test messages of the held-out corpus.

Moving between real corpora costs little (Table 5). The gap is at most 0.027 in eight of nine comparisons, and its confidence interval includes zero for every task on ShareChat, the smallest corpus. The exception is crisis with LMSYS-Chat held out (0.075): LMSYS-Chat contains most of the real crisis messages (13% of its MH messages, compared with 3–6% for the other corpora; Table 3), so a detector trained without it sees few of them. Even this gap is about half the smallest crisis gap of any simulator (Section 5.3). On the same test messages of each corpus, simulator-trained detectors lose more than real-to-real transfer in all but one of 30 comparisons. Real-to-real

Table 5: Real-to-real transfer: AUROC on each real corpus's held-out test messages (mean ± standard deviation over five folds). In-distribution: one detector trained on the training folds of all three corpora. Cross-corpus: a detector trained on the other two corpora only. Both use the same class-balanced training size within each fold and task. The gap is the fold-mean difference, with its 95% t-interval.

Task	Held-out corpus	In-distribution	Cross-corpus	Gap (95% CI)
Distress	WildChat	0.850 ± 0.008	0.823 ± 0.007	0.027 (0.023, 0.032)
	ShareChat	0.819 ± 0.026	0.796 ± 0.029	0.023 (−0.041, 0.086)
	LMSYS-Chat	0.852 ± 0.016	0.833 ± 0.013	0.020 (0.005, 0.034)
Help-seeking	WildChat	0.872 ± 0.010	0.849 ± 0.018	0.023 (0.012, 0.034)
	ShareChat	0.795 ± 0.056	0.796 ± 0.050	−0.001 (−0.015, 0.013)
	LMSYS-Chat	0.853 ± 0.014	0.840 ± 0.008	0.013 (0.004, 0.022)
Crisis	WildChat	0.811 ± 0.026	0.785 ± 0.027	0.026 (−0.010, 0.062)
	ShareChat	0.840 ± 0.079	0.816 ± 0.096	0.025 (−0.003, 0.053)
	LMSYS-Chat	0.884 ± 0.008	0.809 ± 0.052	0.075 (0.019, 0.132)

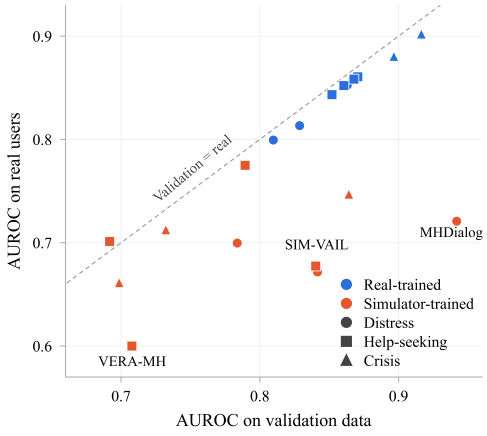


Figure 3: AUROC on each detector's validation data against AUROC on held-out real users (fold means). Color gives the training source and shape the task; real-trained detectors are validated on real users and simulator-trained detectors on their own simulator. On the dashed line, validation predicts real performance; below it, validation overstates performance.

transfer therefore accounts for only a small part of the sim-to-real gap.

7 Implications for Training, Validation and Measurement

7.1 Simulated Data as a Supplement

Simulated users may still help as additional training data. Using the labeler design of Section 6.3, we train a pooled detector on real and simulated users, with VERA-MH capped at 4,000 training messages, so that each fold has about 2,400 real and seven times as many simulated training messages. Pooling never improves real-user AUROC: relative to the real-only detector, the difference ranges from −0.041 to +0.007 and is never significantly positive (Table 20 in Appendix O). It is significantly negative for Gemini distress, consensus help-seeking and consensus crisis. Pooled detectors exceed simulated-only ones by 0.047 to 0.215, so the real messages carry the performance.

Transfer also fails in the opposite direction. For distress and help-seeking, detectors trained only on real users perform poorly on held-out simulated personas and pooled detectors score highest (Table 21 in Appendix O). Detectors are therefore specific to the population they were trained on.

7.2 Validation on Simulated Users

A detector's score on simulated users is an unreliable guide to its score on real users (Figure 3). Validation on real users predicts real performance to within 0.02 AUROC. In contrast, validation on simulated users overstates it by an amount that varies with the simulator and task, from almost none for SimMH on help-seeking to 0.22 for MHDialog on distress. Scores on simulated validation also rank detectors only weakly (Spearman $\rho = 0.44$ across the ten simulator-trained detectors).

7.3 Choice of Labels

Agreement between labelers is limited on real users ($\kappa = 0.24$–0.71): it is lower for distress and help-seeking than for crisis. It is generally lower still on simulated text (Table 22 in Appendix P). For detectors trained on real users, the evaluation labels change AUROC more than the training labels do: changing the evaluation labels moves AUROC by up to 0.20 for distress, 0.23 for help-seeking and 0.10 for crisis, whereas changing the training labels moves it by at most 0.07, 0.12 and 0.05, respectively (Table 23 in Appendix P). Much of the first effect comes from consensus labels, which keep only unanimous and therefore easier messages. A reported AUROC thus reflects the evaluation labels as well as the model, most of all for help-seeking, so studies should state which labels they evaluate against and ideally report more than one.

8 Discussion

8.1 What Simulators Get Wrong

The simulators we study encode a picture of MH users organized around crises. Their users express more severe distress and do so more explicitly, and when they ask for help they appear to disclose a crisis rather than ask a question. Real users of general-purpose chatbots differ. Most of their MH messages are labeled as distressed but few as crisis. Their help-seeking more often takes the form of a short practical question and their distress is often indirect or embedded in an everyday request. Detectors trained on simulated users learn the simulated picture and miss real users who do not match it. Even HealthBench-Psych, which is closest to real users in writing and severity, differs in help-seeking: 70% of its messages seek help, compared with 32–43% for real users, as expected of a benchmark built from health questions.

These differences lie in what simulated users say, not only in how they write. Normalizing style did not narrow the gap (Section 6.1), so surface fixes are unlikely to help. Closing the gap would require simulators whose severity and help-seeking match those of real users.

8.2 Implications for Safety Evaluation

These results do not make simulated users useless. Several simulators are severe by design: SimMH, VERA-MH and SIM-VAIL test how chatbots respond to suicide risk and other high-risk states. Such stress tests are valuable because acute cases are rare in real data and are difficult if not impossible to collect safely at scale, so simulation may be the only practical way to test how chatbots respond to severe states before deployment. Simulators also allow researchers to study specific scenarios without involving real people. The problem arises when simulated users are treated as representative of real users, whether as training data, as validation data or as a measure of how a system performs in everyday use.

The same concern applies to safety benchmarks. We did not test any chatbot, so the following is a hypothesis. If benchmark users are as severe and explicit as the simulators studied here, a chatbot could score well on them while handling poorly the indirect, mild cases that are common in real conversations. Simulators are usually validated through expert judgments of how realistic they appear (Section 2); our results suggest also checking whether the severity and help-seeking of simulated users match those of real users.

8.3 Recommendations

Our results lead to the following recommendations:

- Audit chatbots with both real and simulated use. Simulated users suit the states they were built to test, such as crisis, whereas real conversations show how a system performs in everyday use.
- Validate and calibrate detectors on real users rather than relying on scores or thresholds from simulated data.
- Do not expect simulated data to improve detectors as a supplement; in our experiments, even seven times as much simulated as real data did not help.
- Report results under more than one labeler or under consensus labels, particularly for help-seeking.
- Build simulators that reproduce the indirect and practical ways in which real users express need, for example by seeding disclosures with more general conversations. A natural next step is to test whether prompting simulators with the severity and help-seeking of real users narrows the gap.
- Expand access to real conversations through open, consented datasets and data donation, since real data remains scarce and most public logs are now dated.

9 Limitations

All labels are LLM outputs, validated against human annotations only on non-chat text, with poor agreement for help-seeking. Because labelers see the whole conversation, harmless messages in MH conversations can inherit MH labels, which may inflate the gap. Our real users come from older, LLM-filtered and mostly English public logs that do not represent all users or current chatbots. Transfer between real corpora was measured only among three older, general-purpose and mostly English chatbot logs; a larger shift, such as to a dedicated mental health app or to newer chatbots, could cost more. We also tested one detector architecture with one seed per fold and no hyperparameter search. The pooled bootstrap is optimistic, the fold tests are conservative and the labeler design has small real training sets. Finally, 54 SIM-VAIL conversations from two unparsed personas were split by conversation rather than by persona, which affects the simulated folds but not the real test sets.

10 Conclusion

Simulated users are a poor substitute for real users when building mental health detectors. Detectors trained on them score 0.08 to 0.26 AUROC lower on real users across every simulator and task, and writing style, persona balance and the choice of labeler do not explain this gap. Simulated users remain useful for testing how chatbots respond to severe states that are hard to study with real data, but detectors should be trained, validated and calibrated on real users.

Ethical Considerations

This work analyzes existing public datasets of chatbot conversations and simulated users. We did not recruit participants, interact with users or attempt to identify anyone. Labeling required sending these public messages to commercial LLM services. All labels are model outputs rather than clinical assessments, and our detectors have not been validated for clinical use or deployment. We did not test any chatbot, and our results do not certify the safety of any system.

Use of Generative AI

LLMs served two roles in this work. As research instruments, Gemini 2.5 Flash-Lite, Claude Haiku 5.5 and GPT-4.1 mini produced the message labels (Section 3.3). As assistants, Claude (Anthropic) helped draft and edit the text, including descriptions of results, summarize related work, write code for figures and check reported numbers against result files. The authors reviewed and verified all such material.

References

[1] Anthropic. 2026. Introducing Claude Haiku 5.5. https://www.anthropic.com/claude-haiku-5-5

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351. doi:10.1017/pan.2023.2

[3] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv:2505.08775 [cs.CL] https://arxiv.org/abs/2505.08775

[4] Luca Belli, Kate H. Bentley, Will Alexander, Emily Ward, Matt Hawrilenko, Kelly Johnston, Mill Brown, and Adam M. Chekroud. 2026. VERA-MH Concept Paper. arXiv:2510.15297 [cs.CY] https://arxiv.org/abs/2510.15297

[5] Grace Byun, Rebecca Lipschutz, Sean T. Minton, Abigail Lott, and Jinho D. Choi. 2026. CRADLE Bench: A Clinician-Annotated Benchmark for Multi-Faceted Mental Health Crisis and Safety Risk Detection. arXiv:2510.23845 [cs.CL] https://arxiv.org/abs/2510.23845

[6] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *npj Digital Medicine* 3 (2020). Article 43. doi:10.1038/s41746-020-0233-7

[7] Norman Cliff. 1993. Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions. *Psychological Bulletin* 114, 3 (1993), 494–509. doi:10.1037/0033-2909.114.3.494

[8] Se eun Yoon, Zhankui He, Jessica Maria Echterhoff, and Julian McAuley. 2024. Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation. arXiv:2403.09738 [cs.CL] https://arxiv.org/abs/2403.09738

[9] António Farinhas, Nuno M. Guerreiro, José Pombal, Pedro Henrique Martins, Laura Melton, Alex Conway, Cara Dochat, Maya D'Eon, and Ricardo Rei. 2026. MindGuard: Guardrail Classifiers for Multi-Turn Mental Health Support. arXiv:2602.00950. https://arxiv.org/abs/2602.00950

[10] Matthew Flathers, Phuong Anh Nguyen, Jill Noorily, Julian Herpertz, Meiting Chen, Jasreen Multani, Samuel Powell, Mason Granof, Mark Kalinch, and John Torous. 2026. HealthBench-Psych: A Mental Health Subset of OpenAI's HealthBench. arXiv:2608.25071 [cs.CL] https://arxiv.org/abs/2608.25071

[11] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 514–525. doi:10.1145/3308558.3313698

[12] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2305016120 doi:10.1073/pnas.2305016120

[13] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do Models of Mental Health Based on Social Media Data Generalize?. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3774–3788. doi:10.18653/v1/2020.findings-emnlp.337

[14] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6 (1979), 65–70. https://api.semanticscholar.org/CorpusID:122415379

[15] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv:2110.15621 [cs.CL] https://arxiv.org/abs/2110.15621

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[17] Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14245–14274. doi:10.18653/v1/2024.findings-emnlp.832

[18] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10443–10461. doi:10.18653/v1/2023.emnlp-main.647

[19] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 3469–3483. doi:10.18653/v1/2021.acl-long.269

[20] Yu Lu Liu, Hyokun Yun, Tanya Roosta, and Ziang Xiao. 2026. Synthetic Users, Real Differences: an Evaluation Framework for User Simulation in Multi-Turn Conversations. arXiv:2605.02624 [cs.CL] https://arxiv.org/abs/2605.02624

[21] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10570–10603. doi:10.18653/v1/2024.emnlp-main.591

[22] Ali Malik, Declan Grabb, Rebecca Soskin Hicks, Rahul K. Arora, Preston Bowman, Michael Sharman, Sahra Ghalebikesabi, Mikhail Trofimov, Vinnie Monaco, and Karan Singhal. 2026. *MentalHealthBench: An Expert-Informed Benchmark of AI Capabilities in Realistic Mental Health Conversations*. Technical Report. OpenAI. https://cdn.openai.com/ctf-cdn/MentalHealthBench_A_Comprehensive_Benchmark_of_AI_Capabilities_in_Realistic_Mental_Health_Conversations.pdf

[23] Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2026. MAGneT: Coordinated Multi-Agent Generation of Synthetic Multi-Turn Mental Health Counseling Sessions. arXiv:2509.04183 [cs.CL] https://arxiv.org/abs/2509.04183

[24] Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 399, 7 pages. doi:10.1145/3411763.3451799

[25] Ryan K. McBain, Jonathan H. Cantor, Joshua Breslau, Melissa Diliberti, Li Ang Zhang, Fang Zhang, Alyssa Burnett, Aaron Kofner, Benjamin Rader, Pat Pataranutaporn, Bradley D. Stein, Ateev Mehrotra, and Hao Yu. 2026. AI Chatbot Use and Disclosure for Mental Health Among US Adolescents and Young Adults. *JAMA Pediatrics* 180, 8 (2026), 884–890. doi:10.1001/jamapediatrics.2026.2015

[26] Ryan K. McBain, Li Ang Zhang, Alyssa Burnett, Jonathan H. Cantor, and Hao Yu. 2026. Prevalence of Mental Health Discussions in Publicly Available Generative AI Conversations. *JAMA Network Open* 9, 8 (2026), e2630635. doi:10.1001/jamanetworkopen.2026.30635

[27] Shuhaib Mehri, Philippe Laban, Sumuk Shashidhar, Marwa Abdulhai, Sergey Levine, Michel Galley, and Dilek Hakkani-Tür. 2026. Measuring and Mitigating the Distributional Gap Between Real and Simulated User Behaviors. arXiv:2605.07847 [cs.CL] https://arxiv.org/abs/2605.07847

[28] Jared Moore, Jacy Reese Anthis, William Agnew, Andrea Mock, Kevin Klyman, Sanmi Koyejo, Nick Haber, Stevie Chancellor, Eric Lin, Ryan Louie, et al. 2026. Open Problems in Chatbots and Mental Health. *arXiv preprint arXiv:2611.10997* (2026).

[29] Usman Naseem, Adam G. Dunn, Jinman Kim, and Matloob Khushi. 2022. Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2563–2572. doi:10.1145/3485447.3512128

[30] OpenAI. 2025. Introducing GPT-4.1 in the API. OpenAI. https://openai.com/index/gpt-4-1/

[31] Sam Paech. 2025. Spiral-Bench. GitHub repository. https://github.com/sam-paech/spiral-bench

[32] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated Annotation with Generative AI Requires Validation. arXiv:2306.00176 [cs.CL] https://arxiv.org/abs/2306.00176

[33] Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 10671–10682. doi:10.18653/v1/2022.emnlp-main.731

[34] MSVPJ Sathvik, Zuhair Hasan Shaik, and Vivek Gupta. 2025. M-HELP: Using Social Media Data to Detect Mental Health Help-Seeking Signals. arXiv:2508.15440 [cs.CL] https://arxiv.org/abs/2508.15440

[35] Caitlin A. Stamatis, Jonah Meyerhoff, Richard Zhang, Olivier Tieleman, Matteo Malgaroli, and Thomas D. Hull. 2026. Beyond Simulations: What 20,000 Real Conversations Reveal About Mental Health AI Safety. arXiv:2601.17003 [cs.CY] https://arxiv.org/abs/2601.17003

[36] Ian Steenstra, Paola Pedrelli, Weiyan Shi, Stacy Marsella, and Timothy W. Bickmore. 2026. Assessing Risks of Large Language Models in Mental Health Support:

A Framework for Automated Clinical AI Red Teaming. arXiv:2602.19948 [cs.CL] https://arxiv.org/abs/2602.19948
[37] Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261 [cs.CL] https://arxiv.org/abs/2507.06261
[38] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 23–30. doi:10.1109/IROS.2017.8202133
[39] Transluce. 2026. Announcing Transluce's Mental Health Evaluation. https://transluce.org/announcing-mental-health-evaluation. Published August 31, 2026. Accessed November 16, 2026.
[40] Transluce. 2026. SimMH-Chat. Hugging Face dataset. doi:10.57967/hf/10224
[41] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence* 7, 3 (2025), 400–411. doi:10.1038/s42256-025-00986-z
[42] Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 12772–12797. doi:10.18653/v1/2024.emnlp-main.711
[43] Zhefan Wang, Ning Geng, Zhiqiang Guo, Weizhi Ma, and Min Zhang. 2025. Human vs. Agent in Task-Oriented Conversations. arXiv:2509.17619 [cs.IR] https://arxiv.org/abs/2509.17619
[44] Veith Weilnhammer, Kevin YC Hou, Lennart Luettgau, Christopher Summerfield, Raymond Dolan, and Matthew M Nour. 2026. A clinically validated framework for auditing AI chatbot behavior in mental health interactions. arXiv:2602.01347 [q-bio.NC] https://arxiv.org/abs/2602.01347
[45] Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance. arXiv:2503.13509 [cs.LG] https://arxiv.org/abs/2503.13509
[46] Yueru Yan, Tuc Nguyen, Bo Su, Melissa Lieffers, and Thai Le. 2025. ShareChat: A Dataset of Chatbot Conversations in the Wild. arXiv:2512.17843. https://arxiv.org/abs/2512.17843
[47] Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. 2025. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. arXiv:2509.10970 [cs.LG] https://arxiv.org/abs/2509.10970
[48] Yihe Zhang, Cheyenne N. Mohawk, Kaiying Han, Vijay Srinivas Tida, Manyu Li, and Xiali Hei. 2026. MHDash: An Online Platform for Benchmarking Mental Health-Aware AI Assistants. arXiv:2602.00353. https://arxiv.org/abs/2602.00353
[49] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations (ICLR)*.
[50] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1552–1568. doi:10.18653/v1/2023.findings-acl.99
[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. arXiv:2309.11998 [cs.CL] https://arxiv.org/abs/2309.11998
[52] Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building Emotional Support Chatbots in the Era of LLMs. arXiv:2308.11584 [cs.CL] https://arxiv.org/abs/2308.11584
[53] Ziyi Zhu, Olivier Tieleman, Caitlin A Stamatis, Luka Smyth, Thomas D Hull, Daniel R Cahn, and Matteo Malgaroli. 2025. Adversarial Training for Failure-Sensitive User Simulation in Mental Health Dialogue Optimization. *arXiv e-prints* (2025), arXiv–2512.
[54] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, Kate Niederhoffer, Kristy Hollingshead, Philip Resnik, Rebecca Resnik, and Kate Loveys (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 24–33. doi:10.18653/v1/W19-3003

A Labeling Prompt and Task Definitions

All three labelers receive the same system prompt and a strict JSON output schema, with one entry per message. Each message is judged on its own content, with earlier messages used only as context, and the labeler is told to choose the higher score when uncertain between two adjacent scores. Table 6 lists the user-message fields as defined in the prompt. The prompt also scores valence and arousal for user messages and referrals for chatbot messages, which we do not use. It closes with calibration examples that separate fiction and professional tasks (*subject* `none`), disclosures with no request (help-seeking 0) and passing mentions or idioms (topicality at most 1). The full prompt is released with our code.

A user message is MH if topicality ≥ 1 and subject $\in$ {`self`, `other`}. The three tasks are distress ≥ 1, help-seeking ≥ 1 and category = `crisis`.

Table 6: User-message fields in the labeling prompt (wording abridged).

Field	Definition
Topicality	0: not about MH; 1: mentioned in passing; 2: MH is the central focus
Help-seeking	0: none; 1: general well-being talk, no request for coping, treatment or support; 2: explicitly seeking or giving MH advice, treatment or crisis support
Distress	0: everyday emotions only; 1: noticeable distress or impairment; 2: severe distress, crisis, suicide, self-harm or explicit risk
Subject	`self`; `other` (a real person in the user's life); `none` (fiction, role-play, exam, academic or design tasks, hypotheticals)
Category	One of crisis, distress, mood, trauma, psychosis, substance, treatment, neurodiversity, psychosocial, well-being, physical or logistics; none if topicality is 0. Crisis: imminent or serious self-harm or suicide risk, or safety planning

B Writing Features

Table 7 lists the 87 writing features used in Sections 4 and 6. Rates are per word or per sentence, and binary features are 0 or 1 per message.

C Writing-Style Distance and Detectability by Source

Table 8 reports the values plotted in Figure 1 for the ten labeled sources. The ten further datasets in the figure are English, publicly downloadable sets of simulated MH users: Spiral-Bench [31] and psychosis-bench [47] (chatbot users); CACTUS [17], MAGneT [23], AI-Psychotherapy-Eval [36], Mental-health-CBT-dialogues[1] and the synthetic part of MentalChat16K [45] (counseling clients); and ESConv [19], AugESC [50] and ExTES [52] (emotional support).

D Effect of Simulator Weighting on Detection

In the classifier of Section 4.1, SimMH supplies 97% of the simulated training messages. Giving each simulator equal weight raises detection of MHDialog and MindGuard to 95% and 79%. Figure 4 also shows detection when each simulator is held out of training.

[1] https://huggingface.co/datasets/yuana1234567/Mental-health-CBT-dialogues

Table 7: The 87 writing features, by group.

Group	n	Features
Length	8	characters, words, sentences, lines, paragraphs, mean and SD of sentence length, short message ($\leq$ 3 words)
Mechanics	12	all-lowercase message, lowercase "i", lowercase sentence starts, all-caps words, missing apostrophes, no terminal punctuation, missing space after or space before punctuation, double spaces, repeated letters, typos, out-of-vocabulary words
Punctuation	14	overall density; periods, commas, question and exclamation marks, colons, semicolons, ellipses, parentheses, em and en dashes, spaced hyphens, repeated punctuation, curly quotes
Formatting	5	bold, code, lists, Markdown line share, URLs
Informal	5	emoji, emoticons, fillers, profanity, text-speak
Vocabulary	7	contractions, mean word length, long words, rare words, mean Zipf frequency, MTLD, words typical of LLM output
Voice	12	first-person singular and plural, second person, feeling statements, emotion words, hedges, intensifiers, negation, politeness, greeting, acknowledgement opening, question sentences
Sentiment	3	VADER compound, positive and negative scores
Syntax	18	shares of 14 parts of speech, fragments, dependency depth, passives and subordinate clauses per sentence
Readability	3	Flesch reading ease, Flesch–Kincaid grade, Dale–Chall score

Table 8: Writing-style distance from real users (mean absolute Cliff's δ over 87 features) and detectability (AUROC of a logistic-regression classifier separating each source from real users). Each real source is compared with the other two.

Source	Mean $\lvert\delta\rvert$	AUROC
ShareChat	0.051	0.71
WildChat	0.076	0.71
LMSYS-Chat	0.093	0.73
HealthBench-Psych	0.099	0.881
MindGuard	0.125	0.896
MHB	0.134	0.902
VERA-MH	0.163	0.998
SimMH	0.169	0.988
MHDialog	0.170	0.990
SIM-VAIL	0.252	1.000

E Label Rates Across Labelers

Table 9 repeats the label rates of Section 4.3 under all three labelers. Every labeler rates simulated users as more distressed than real users, and the labelers disagree most on help-seeking in simulated text.

F Agreement of Help-Seeking Cues Between Real Users and MHB

Table 10 gives the correlations of Section 4.4. A high ρ means that the writing features that change with a label in real users change the same way in MHB. For help-seeking, no MHB profile or acuity level reaches the agreement seen for distress and crisis.

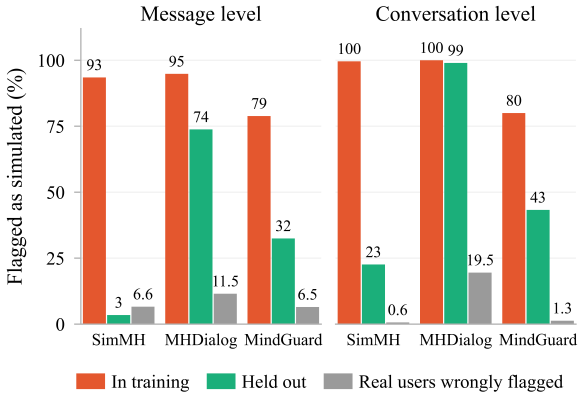


Figure 4: Share of each simulator's messages (left) or conversations (right) that a real-versus-simulated classifier flags as simulated. Orange: the simulator is in training, with each simulator weighted equally. Green: the classifier is retrained without it. Gray: the highest share of any real corpus wrongly flagged by the held-out model.

Table 9: Positive rates under each labeler (Gemini / Haiku / GPT) on the messages labeled by all three and called MH by at least two.

Source	Messages	Distress (%)	Help-seeking (%)	Crisis (%)
Real users	3,447	76 / 52 / 75	39 / 46 / 48	9 / 6 / 7
SimMH	9,954	78 / 93 / 95	14 / 66 / 41	17 / 26 / 10
SIM-VAIL	6,258	80 / 92 / 94	25 / 80 / 72	2 / 6 / 2
VERA-MH	19,316	95 / 98 / 100	70 / 94 / 93	58 / 79 / 38
HealthBench-Psych	780	80 / 80 / 84	68 / 92 / 83	7 / 6 / 5

Table 10: Spearman correlation (ρ) between real users and MHB in how writing features change with each label (top) and within MHB subgroups for help-seeking (bottom). HS: share of messages labeled as help-seeking.

Subset	Messages	HS (%)	ρ
All MHB messages, by label			
Distress	1,744	—	0.70
Crisis	1,744	—	0.68
Help-seeking	1,744	30.9	−0.10
Help-seeking, by MHB subgroup			
Profile: adult	1,205	32.2	−0.05
Profile: teen	394	28.9	+0.10
Profile: clinician	54	37.0	−0.19
Acuity: non-acute	707	28.9	+0.08
Acuity: high	489	26.6	−0.24
Acuity: emergent	548	37.4	−0.10

Table 11: Request types among help-seeking messages, real users vs. MHB. A message can match several types. Confidence intervals are from a conversation bootstrap.

Request type	Real	MHB	MHB − real (95% CI)
How-to coping	25.8%	11.9%	−13.9 (−16.7, −10.6)
Information	10.3%	7.6%	−2.7 (−5.0, −0.0)
Practical resources	7.6%	1.7%	−5.9 (−7.3, −4.4)
Third party	7.6%	6.5%	−1.1 (−3.2, +1.3)
Medical / medication	5.6%	4.6%	−1.0 (−3.1, +1.3)
Therapy / professional	3.4%	1.9%	−1.6 (−3.0, +0.1)
Legal / police	2.8%	0.2%	−2.7 (−3.2, −2.1)
Fiction framing	1.7%	0.2%	−1.6 (−2.0, −1.0)
Crisis services	1.4%	1.3%	−0.1 (−1.3, +1.3)
None of the above	52%	70%	—

G Request Types in Help-Seeking Messages

Table 11 covers 8,066 real and 539 MHB help-seeking messages. Each request type is a case-insensitive keyword pattern; for example, how-to coping matches phrases such as "how do I", "what should I do", "tips" and "advice". The patterns are released with our code. Confidence intervals use 1,000 resamples of conversations.

H Decision Thresholds

Each detector's threshold maximizes balanced accuracy on either its own (simulated) validation data or real validation data (Table 12). Fix A and Fix B are defined in Sections 6.1 and 6.2. At the real threshold, no simulator-trained detector reaches the recall of the real-trained detector on either class.

Table 12: Recall (%) on held-out real users in the SimMH design, averaged over five folds. Pos. is recall on positive messages and Neg. is recall on negative messages. The real-trained detector uses real validation data for both checkpoint and threshold. Fix A: style-normalized SimMH; Fix B: persona-balanced SimMH.

		Simulated threshold		Real threshold	
Task	Training data	Pos.	Neg.	Pos.	Neg.
Distress	Real	—	—	76.4	79.0
	SimMH	25.0	93.1	59.1	71.5
	Fix A	21.9	95.0	62.0	66.9
	Fix B	41.3	83.7	60.5	71.7
Help-seeking	Real	—	—	77.0	80.0
	SimMH	71.4	57.9	61.3	68.7
	Fix A	68.3	61.5	63.6	67.7
	Fix B	61.3	67.2	69.5	61.6
Crisis	Real	—	—	80.7	83.2
	SimMH	57.6	73.4	64.7	67.4
	Fix A	60.4	71.0	56.0	75.6
	Fix B	57.8	69.7	58.5	71.2

I The Gap Across Simulators

Table 13 gives the values behind Figure 2. Because each simulator has its own size-matched real partner, the real-trained AUROC differs between pairs. Pooled and fold-mean gaps agree to within 0.008. Table 14 compares gaps between simulators. A difference is robust when both the fold t-test and the pooled bootstrap are significant after Holm correction within a task.

An earlier single split agreed in direction but not always in size: the SimMH help-seeking gap moved from 0.114 to 0.167, crisis gaps of about 0.13–0.14 separated into the ordering in Table 14, and an apparent SIM-VAIL disadvantage on help-seeking and an apparent +0.02 gain for Fix A on crisis disappeared. We therefore report five-fold results throughout.

J External Test Sets

The pairs-design detectors are also scored on three sources never used for training: HealthBench-Psych (737 messages), MHB (1,744) and MindGuard (726). These counts are below those in Table 1 because 11 HealthBench-Psych and 62 MindGuard MH messages lack a label for at least one task and are excluded. Each fold's detector scores every message, and Table 15 reports the mean over the five folds. The real-trained detector scores higher in 28 of 30 comparisons. The exceptions are the MHDialog-trained detectors for distress and crisis on MindGuard. These comparisons are descriptive and were not tested for significance. Crisis estimates rest on few positives (55, 210 and 27 messages).

K Error Rates of the SimMH-Trained Detector

Table 16 gives the error rates behind Section 5.4. Each detector's threshold maximizes balanced accuracy on real validation data. For distress, help-seeking and crisis, the SimMH-trained detector alone misses 321, 193 and 24 positive messages that the real-trained detector catches, and alone flags 113, 233 and 519 negative messages; Section 5.4 describes these messages.

L Style Shortcuts

For each writing feature, we compute Cliff's δ between label-positive and label-negative messages, separately for real users and each simulator. A feature is a *shortcut* when its association with the label is stronger in the simulator by at least the small-effect threshold: $|\delta_{\text{sim}} - \delta_{\text{real}}| \geq 0.147$ and $|\delta_{\text{sim}}| > |\delta_{\text{real}}|$. The few shortcuts in SimMH come from its personas. Pooled over personas, SimMH shows no agreement with real users in how features change with distress (Spearman $\rho = 0.00$); within each persona, agreement rises to $\rho = 0.52$ and shortcuts disappear for distress and crisis. Logistic-regression detectors trained on the writing features place only 3–6% of their weight on shortcut features (single split).

M Effect of the Style and Persona Fixes

Table 17 compares detectors trained on the two modified versions of SimMH with detectors trained on SimMH as released. Fix A normalizes style (Section 6.1) and Fix B balances classes within personas (Section 6.2).

Table 13: Gap per simulator in the pairs design (five-fold). "Per class" is the mean training size per class over folds. AUROC values are pooled out-of-fold estimates on real users.

Task	Simulator	Per class	Real-trained	Simulator-trained	Gap, pooled (95% CI)	Gap, fold mean (95% CI)	Folds > 0
Distress	SimMH	3,479	0.839	0.688	0.152 (0.141, 0.163)	0.153 (0.132, 0.174)	5/5
	MHDialog	500	0.811	0.721	0.089 (0.078, 0.102)	0.093 (0.062, 0.123)	5/5
	SIM-VAIL	348	0.797	0.668	0.130 (0.118, 0.141)	0.128 (0.105, 0.150)	5/5
Help-seeking	SimMH	4,000	0.859	0.693	0.167 (0.156, 0.177)	0.159 (0.126, 0.193)	5/5
	MHDialog	2,041	0.845	0.766	0.079 (0.067, 0.089)	0.077 (0.065, 0.089)	5/5
	SIM-VAIL	1,197	0.841	0.673	0.167 (0.156, 0.178)	0.166 (0.141, 0.191)	5/5
	VERA-MH	3,573	0.855	0.594	0.261 (0.248, 0.275)	0.258 (0.209, 0.308)	5/5
Crisis	SimMH	1,204	0.896	0.706	0.190 (0.171, 0.209)	0.189 (0.156, 0.222)	5/5
	MHDialog	417	0.878	0.742	0.135 (0.120, 0.150)	0.133 (0.093, 0.173)	5/5
	VERA-MH	1,204	0.895	0.661	0.235 (0.212, 0.258)	0.241 (0.168, 0.313)	5/5

Table 14: Differences in gap between simulators (five-fold pairs design). Bold: significant under both tests after Holm correction. Bootstrap Holm $p < 0.001$ for all rows except SimMH − SIM-VAIL on help-seeking ($p = 0.914$).

Task	A − B	Δ gap (95% CI)	p (fold)	Folds A > B
Distress	**SimMH − MHDialog**	+0.062 (0.049, 0.076)	0.012	5/5
	SimMH − SIM-VAIL	+0.022 (0.009, 0.036)	0.221	4/5
	MHDialog − SIM-VAIL	−0.040 (−0.048, −0.031)	0.221	1/5
Help-seeking	**SimMH − MHDialog**	+0.088 (0.075, 0.101)	0.011	5/5
	SimMH − SIM-VAIL	−0.001 (−0.011, 0.010)	0.559	1/5
	SimMH − VERA-MH	−0.095 (−0.107, −0.081)	0.025	0/5
	MHDialog − SIM-VAIL	−0.088 (−0.101, −0.077)	0.009	0/5
	MHDialog − VERA-MH	−0.182 (−0.196, −0.169)	0.005	0/5
	SIM-VAIL − VERA-MH	−0.094 (−0.107, −0.082)	0.018	0/5
Crisis	SimMH − MHDialog	+0.054 (0.035, 0.075)	0.152	5/5
	SimMH − VERA-MH	−0.045 (−0.072, −0.019)	0.152	1/5
	MHDialog − VERA-MH	−0.099 (−0.124, −0.073)	0.152	0/5

Table 15: AUROC on external test sets (real-trained / simulator-trained; mean over five folds; Gemini labels).

Task	Simulator	HealthBench-Psych	MHB	MindGuard
Distress	SimMH	0.826 / 0.749	0.890 / 0.794	0.805 / 0.768
	MHDialog	0.813 / 0.790	0.876 / 0.826	0.773 / 0.799
	SIM-VAIL	0.811 / 0.751	0.867 / 0.762	0.753 / 0.733
Help-seeking	SimMH	0.848 / 0.643	0.766 / 0.655	0.741 / 0.704
	MHDialog	0.844 / 0.743	0.754 / 0.730	0.736 / 0.702
	SIM-VAIL	0.827 / 0.609	0.753 / 0.633	0.726 / 0.654
	VERA-MH	0.846 / 0.629	0.763 / 0.661	0.741 / 0.670
Crisis	SimMH	0.897 / 0.634	0.891 / 0.833	0.838 / 0.814
	MHDialog	0.866 / 0.736	0.858 / 0.829	0.838 / 0.875
	VERA-MH	0.901 / 0.733	0.891 / 0.641	0.839 / 0.758

N The Gap Under Other Labelers

Table 18 re-scores the SimMH-design detectors, trained on Gemini labels, against other labels. Majority labels take the vote of the three labelers, and consensus labels keep only unanimous messages.

Table 19 reports the gap when detectors are trained on each labeler's labels and scored against the same labels. When the labeler-design detectors are scored against another labeler's labels, the gap remains significant in almost every combination. The exceptions all involve detectors trained on Haiku's labels: for help-seeking scored against Gemini, GPT or consensus labels (gaps of −0.014, +0.014 and +0.041) and for distress scored against Gemini (+0.022).

Table 16: Error rates (%) on real users for the real-trained and SimMH-trained detectors. Both use thresholds tuned on real validation data (single split). Arrows show the change from the real-trained to the SimMH-trained detector.

Task	Positives missed	Negatives flagged
Distress	21.6 → 35.2	25.9 → 37.3
Help-seeking	25.1 → 42.5	20.6 → 29.1
Crisis	25.0 → 23.9	17.0 → 35.2

Table 17: Effect of the fixes relative to SimMH on real users (five-fold). Fix A: style-normalized SimMH; Fix B: persona-balanced SimMH. Fold: mean difference over folds, with the paired t-test p-value. Pooled: difference in pooled out-of-fold AUROC, with its 95% bootstrap confidence interval.

	Fix A − SimMH		Fix B − SimMH	
Task	Fold (p)	Pooled (95% CI)	Fold (p)	Pooled (95% CI)
Distress	−0.005 (0.53)	−0.004 (−0.008, +0.000)	+0.012 (0.23)	+0.022 (+0.011, +0.035)
Help-seeking	+0.002 (0.29)	+0.002 (−0.003, +0.006)	+0.006 (0.83)	−0.003 (−0.013, +0.008)
Crisis	−0.002 (0.68)	−0.003 (−0.008, +0.001)	−0.020 (0.24)	−0.038 (−0.062, −0.015)

Table 18: Real − SimMH gap when only the test labels change (five-fold detectors trained on Gemini labels; 95% bootstrap confidence intervals).

Label source	Distress	Help-seeking	Crisis
Gemini	0.160 (0.130, 0.189)	0.147 (0.119, 0.174)	0.132 (0.105, 0.161)
Haiku	0.130 (0.091, 0.172)	0.148 (0.101, 0.195)	0.132 (0.092, 0.176)
GPT	0.120 (0.082, 0.156)	0.141 (0.108, 0.174)	0.111 (0.079, 0.143)
Majority	0.141 (0.110, 0.172)	0.161 (0.128, 0.190)	0.112 (0.082, 0.143)
Consensus	0.146 (0.092, 0.213)	0.173 (0.114, 0.228)	0.097 (0.063, 0.136)

Table 19: Real-only − simulated-only gap when the training labels change too (five-fold over 3,447 real messages; training labels equal test labels). AUROC values are pooled out-of-fold estimates on real users.

Task	Labels	Real-only	Simulated	Gap (95% CI)	Folds > 0
Distress	Gemini	0.796	0.672	0.124 (0.099, 0.148)	5/5
	Haiku	0.725	0.648	0.077 (0.051, 0.102)	5/5
	GPT	0.749	0.654	0.095 (0.073, 0.120)	5/5
	Consensus	0.887	0.830	0.057 (0.038, 0.078)	5/5
Help-seeking	Gemini	0.782	0.672	0.110 (0.086, 0.136)	5/5
	Haiku	0.749	0.644	0.105 (0.078, 0.133)	5/5
	GPT	0.771	0.694	0.077 (0.052, 0.100)	5/5
	Consensus	0.921	0.821	0.100 (0.075, 0.122)	5/5
Crisis	Gemini	0.868	0.682	0.186 (0.122, 0.247)	5/5
	Haiku	0.865	0.637	0.228 (0.142, 0.313)	4/5
	GPT	0.887	0.667	0.220 (0.145, 0.297)	5/5
	Consensus	0.953	0.696	0.257 (0.171, 0.352)	5/5

O Pooling Real and Simulated Users

Table 20 compares the pooled detector of Section 7.1 with real-only and simulated-only detectors. Table 21 scores the same detectors on held-out simulated personas, where real-only detectors perform worst.

In an earlier fixed-split experiment, multi-label detectors trained on real and SimMH data ranked both populations about equally (distress AUROC 0.81 on each), yet one threshold flagged 28–54% of distressed real users against 75–90% of SimMH users. Mixed training can align ranking but not the threshold.

Table 20: Effect of pooling on real users (five-fold; training labels equal test labels; 95% confidence intervals). Bold: interval excludes zero.

Task	Labels	Pooled − real-only	Pooled − simulated-only
Distress	Gemini	**−0.019** (−0.038, −0.000)	**+0.105** (0.084, 0.128)
	Haiku	−0.012 (−0.028, 0.003)	**+0.065** (0.041, 0.089)
	GPT	−0.011 (−0.027, 0.006)	**+0.084** (0.059, 0.110)
	Consensus	−0.010 (−0.027, 0.004)	**+0.047** (0.028, 0.067)
Help-seeking	Gemini	−0.001 (−0.015, 0.014)	**+0.109** (0.090, 0.131)
	Haiku	−0.016 (−0.033, 0.001)	**+0.089** (0.066, 0.112)
	GPT	−0.010 (−0.028, 0.008)	**+0.067** (0.038, 0.094)
	Consensus	**−0.018** (−0.033, −0.000)	**+0.082** (0.062, 0.103)
Crisis	Gemini	+0.007 (−0.013, 0.029)	**+0.193** (0.138, 0.250)
	Haiku	−0.023 (−0.063, 0.015)	**+0.204** (0.133, 0.276)
	GPT	−0.011 (−0.033, 0.014)	**+0.209** (0.135, 0.289)
	Consensus	**−0.041** (−0.081, −0.009)	**+0.215** (0.144, 0.296)

Table 21: AUROC on held-out simulated personas (fold mean; Gemini labels).

Task	Test source	Real-only	Simulated	Pooled
Distress	SimMH	0.650	0.801	0.813
	SIM-VAIL	0.611	0.754	0.791
Help-seeking	SimMH	0.610	0.722	0.735
	SIM-VAIL	0.738	0.874	0.880

P Labeler Agreement and Evaluation Labels

Table 22 reports Cohen's κ on the messages labeled by all three labelers. On real users, agreement is fair to moderate for distress and help-seeking and highest for crisis. Agreement is lower on simulated text, partly because of extreme base rates in SIM-VAIL and VERA-MH, except on HealthBench-Psych. Consensus retains only 34–36% of the help-seeking labels of SimMH and SIM-VAIL, compared with 49% for real users, so each labeler effectively applies its own definition of help-seeking, most of all to simulated disclosures.

Table 23 crosses training labels (rows) with evaluation labels (columns) for real-only detectors. Help-seeking depends on the labeler even without consensus labels: a model trained on one labeler's labels varies by 0.09–0.11 across the three single-labeler evaluation labels. Crisis depends little on the labeler (AUROC 0.84–0.95). The consensus-trained model is best on consensus labels for every task but second on Gemini and Haiku labels for distress and help-seeking, so we read consensus labels as the safest single choice when the evaluation labels are unknown, not as uniformly best. An earlier fixed-split experiment (504 real test messages) showed the same pattern.

Table 22: Cohen's κ between labeler pairs (Gemini–Haiku / Gemini–GPT / Haiku–GPT) on the messages labeled by all three.

Source	Messages	Distress	Help-seeking	Crisis
Real users	3,447	0.27 / 0.42 / 0.29	0.24 / 0.40 / 0.29	0.54 / 0.71 / 0.64
SimMH	9,954	0.13 / 0.27 / 0.21	0.12 / 0.14 / 0.24	0.45 / 0.35 / 0.36
SIM-VAIL	6,258	0.26 / 0.26 / 0.30	0.14 / 0.20 / 0.35	0.11 / 0.16 / 0.18
VERA-MH	19,316	0.11 / 0.05 / 0.09	0.14 / 0.14 / 0.31	0.41 / 0.41 / 0.26
HealthBench-Psych	780	0.55 / 0.46 / 0.59	0.25 / 0.43 / 0.35	0.76 / 0.76 / 0.78

Table 23: Real-user AUROC of real-only detectors (five-fold, 3,447 messages). Rows: training labels. Columns: evaluation labels (G: Gemini, H: Haiku, P: GPT, C: consensus). Bold: model on its own labels. †: differs significantly from the model trained on that column's labels. Consensus columns use unanimous messages only.

Trained on	G	H	P	C
Distress				
Gemini	**0.796**	0.681†	0.723†	0.879
Haiku	0.727†	**0.725**	0.721†	0.862†
GPT	0.742†	0.692†	**0.749**	0.866†
Consensus	0.776†	0.696†	0.755	**0.887**
Help-seeking				
Gemini	**0.782**	0.669†	0.735†	0.895†
Haiku	0.658†	**0.749**	0.667†	0.840†
GPT	0.759†	0.679†	**0.771**	0.904†
Consensus	0.764†	0.714†	0.763	**0.921**
Crisis				
Gemini	**0.868**	0.859	0.888	0.929
Haiku	0.838†	**0.865**	0.858	0.928
GPT	0.886	0.854	**0.887**	0.946
Consensus	0.861	0.857	0.891	**0.953**