

Evaluating Inter-Model Agreement in Depression Risk Classification Using Large Language Models

Sreshta Namala¹

Received June 8, 2026

Accepted September 18, 2026

Electronic access October 31, 2026

As artificial intelligence becomes more integrated into our lives, their ability to detect mental health patterns presents both opportunities and challenges. One possible frontier may be in early identification of mental health struggles. With how accessible it is, people express themselves more online, so social media can provide large amounts of data that may help researchers and psychologists/therapists detect signs of mental distress. However, these signs of distress can be ambiguous and with how direct AI is, it can lead to inconsistencies in how those signs are processed. This led to the question: when presented with identical text samples focusing on depression-related language, do LLMs agree in their risk classifications or do their internal criteria differ so much? To answer that, emotionally charged and neutral sentences were inputted into four LLMs: GPT-5.6 Luna (OpenAI), DeepSeek V-3 (DeepSeek), Qwen3.8 (Ollama), and Sonnet 5 (Anthropic), prompting the models to classify the sentences as low, medium, or high risk for depression. In analyzing both samples of depression and neutral sentences, inter-model agreement was measured using Fleiss' kappa and Krippendorff's alpha. The results showed substantially higher agreement in the depression entries and strikingly low agreement in the neutral entries, suggesting that the neutral sentences cannot act as a true control set. These findings show how different LLMs interpret the same depression-related content and where variability shows up in their classification patterns without treating model findings as a measure of clinical accuracy.

Keywords: Comparative analysis, Large language models, Depression classification, Risk assessment, Inter-model agreement, Natural language processing

Introduction

Artificial Intelligence (AI) enables the analysis of large-scale data and decision-making automation across many fields such as education, business, and manufacturing¹. These capabilities also make AI, particularly large language models (LLMs), well-suited to analyzing the vast volume of content on social media². This paper explores LLMs' potential to flag mental health issues based on social media text. In the last decade, social media has become a routine part of people's daily lives, influencing how we talk, act, and think. These consistent interactions between people on social media allow social media to serve as a window into people's mental states. Social media users tend to express early signs of distress online before they try seeking help offline. LLMs may be able to detect these subtle changes at a larger scale, offering a better chance for early intervention. Therefore, research on these interactions sheds light on the fact that mental health issues could be detected by AI when provided with data from the content users share³.

However, posts on social media rely on context and often include slang, sarcasm, and cultural references, which may

complicate the interpretation of pragmatic intent by AI models. More generally, LLMs may produce hallucinated or factually incorrect statements⁴. Despite this, the amount of human interaction on social media generates a large scale of data, which can be leveraged for such predictive analysis.

This study examines user content by taking samples of emotionally charged sentences from a dataset available on Zenodo and neutral sentences from a dataset available on Kaggle and feeding them into four LLMs: GPT-5.6 Luna, Qwen3.8, DeepSeek V-3, and Sonnet 5. However, rather than measuring whether any model makes a clinically accurate diagnosis, this study focuses on the consistencies between model outputs. In doing so, the experiment examines whether the models agree or disagree when assigning risk classifications (low, medium, high) to the same input.

AI and Mental Health

In mental health research, machine learning systems have been explored for identifying linguistic or behavioral signs associated with psychological distress in social media data⁵. They've identified various kinds of language characteristics such as using different language, or using more depressive

¹ North Creek High School, USA

word choice, while other studies have examined behavioral patterns such as post frequency and engagement with other content.

Once trained, a model can apply the patterns it learned to identify signs of mental health distress. In language, the model could scan for repeated use of negative emotion words with first-person pronouns in expressions of hopelessness or self-critical language⁶. Behaviorally, platform-level analytics systems can track posting frequency and engagement patterns, whose graphs can be used as inputs for AI-based analyses⁷. Finally, it can analyze interactions with other users, and participation in communities or hashtags related to mental health. With this combination of signals, models trained on social media data may more consistently identify potential signs of mental health issues over time⁸.

Early approaches for detecting mental health issues on social media included traditional NLP and machine learning models⁹. For example, text-based approaches with NLP analyze users' text to assess their mental health. Without formal mental health support, they commonly use social media as an outlet for their feelings: an easily available platform where they could receive attention and informal peer responses in the form of comments. A study by Liu, Ding, and Chen found that individuals experiencing depression exhibit distinct shifts in the emotional tone and content of their social media posts, ranging from feelings of daily stress to thoughts of self-harm¹⁰. Furthermore, although not perfected, sentiment analysis has often been integrated with these methods to evaluate the emotional tone of text, classifying it as positive, negative, or neutral to better infer underlying mental health signals¹¹.

Graph-based approaches analyze relationships between users, content, and activity and represent them in one complex network, with nodes representing users and edges representing the behavior and user content¹². This type of analysis is called a social network analysis. A change in a user's mental health would mean a change in this network, either because they are posting different things or because they are posting at a different frequency¹³.

However, while studies emphasize that these text-based and graph-based models are effective at identifying mental health patterns, they primarily evaluate models based on metrics such as precision, recall, and F1-score¹⁴. These metrics primarily serve as a measure of accuracy in detecting mental health issues, usually with a ground truth. They provide useful performance benchmarks, but they do not address how different AI systems may vary in their classification outcomes. Understanding how a group of LLMs classify identical mental health-related text is necessary to evaluate the inconsistencies, a critical gap in research that this study addresses¹⁵.

However, recent research has begun comparing multiple LLMs for mental health assessment rather than evaluating individual models in isolation. Settanni et al. compared

GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro in detecting psychological distress from Reddit posts and found differences in model performance and tendencies to overestimate urgency¹⁶. That is, the models had different sensitivity thresholds for what constitutes mental health risk. Similarly, Arnaiz-Rodriguez et al. evaluated five models, which included both commercial (GPT-4o-mini, GPT-5-nano) and open-weight (Llama-4-Scout) systems on crisis-handling scenarios, and prompted each model three times per input so they wouldn't mistake inter-model variability with cross-model disagreement, and found weaknesses across all models in handling ambiguous risk signals¹⁷. In a related evaluation, ground truth and inter-model variability was combined; three chatbots were compared with responses from licensed therapists, revealing meaningful differences in how the chatbots approached sensitive topics¹⁸.

Together, these studies confirm that model choice greatly affects mental health classifications, as they prove that each has their own approaches in navigating mental-health-related conversations. However, each evaluated models against an external standard whether its assessments from licensed professionals or safety. This study, instead, focuses on whether the models agree with each other, without treating any model's output as a diagnosis. It also separately measures intra-model consistency, so that actual differences between models' systems are not confused with the inconsistency within a single model.

Methodology

Social media is an ever-growing network of people and everyday more and more people feel comfortable using it as a platform of self-expression. With so much interaction, social media became a useful setting to conduct exploratory studies on mental health¹⁹. This study investigates the agreement between large language models when using them to detect mental health distress via a social media presence. The labels (Low, Medium, High) do not represent diagnosis or valid clinical assessment; instead, they are meant to reveal each model's own categorization of the user content presented to it. The whole purpose of the study is to examine how different models interpret the same input, so no clinician-developed rubric was provided as it would limit the models' definitions into the three categories.

The first step was finding the social media data to feed to the four models. Two sets of data were used: one with more emotionally-charged text data and the other with neutral, conversational text data. The depression dataset consists of all the Reddit comments on the depression subreddit from January 2019 to April 2019 (Zenodo), and was used as a proxy for depression related language. The dataset used for the neutral group consists of all the Reddit comments on the casual con-

versation subreddit, restricted from 2017 - 2019 to the same January 2019 to April 2019 calendar period (Kaggle). The neutral dataset was restricted to the same time period as the depression data to avoid any temporal bias and to keep the data consistent²⁰.

The depression data also were cleared of any duplicate entries. The casual dataset was in a different format, though. Each entity had 1-3 entries as replies and those replies were put into columns. Therefore, the columns of data had to be split and all of them had to be considered as entries, to get one entry. When they were split, each of the replies was attached to the main comment, so that main comment was repeated based on how many comments it had. In reality, the neutral set didn't have any duplicates; the duplicates created by the reformatting were simply removed. To avoid lack of context for the sentences, the inclusion/exclusion code made sure each of them were grammatically correct with a subject and a predicate (not one or two word replies/entries). Some examples of excluded sentences were "Peanut butter and bananas" and "Correct!" (Appendix C).

From the cleaned datasets, 100 sentences from each were selected using a reproducible sampling procedure as hand-picking the sentences would lead to human biases. The sampling procedure was applied after the data-cleaning process, and the same procedure was used for both datasets. A fixed random seed of 42 was used for each dataset sampling so that the final sample could be reproduced. Once each of the 100-sentence sets was selected, both sets were put into the four models. On VS Code, each sentence was processed individually in separate calls and the model was instructed to evaluate that sentence on its own without comparing it with any of the other sentences in the dataset, reducing context bias.

This study used these four models because they are tools that are already widely used in the general public. They don't need money to use, promoting accessibility and large API budgets are not needed to get meaningful results from those models. This allows the study to consider real world risks in results when performing this experiment. This model stack also uses different deployment methods: three cloud-hosted models from separate companies (OpenAI, DeepSeek, Anthropic) and one locally-run model (Qwen3.8 via Ollama), creating a broader comparison. Actually comparing these models is important because patterns that appear consistently across all models are less likely to come from the specific design of a model²¹.

To examine consistency across repeated runs, each sentence was evaluated in three independent trials for each model. Python programming with Ollama SDK was used to automate the classification procedure and ensure that the input conditions were the same; temperature, top-p, and top-k values were at default for each model. This was the prompt given to each model: *Classify the entry into low, medium, or high based*

on depression risk. This sentence is from a publicly available dataset and will be used solely for research and exploratory purposes, not for clinical diagnoses. Respond with exactly one word: Low, Medium, or High (Appendix B). These labels refer only to the models' own internal classification of depression-related linguistic cues, not clinical diagnosis. Each trial consisted of a separate model call with no shared conversation history, so the model received no information on what it was classified as before to keep it from affecting the next trial's classifications. The output for the code was a csv file with the three for each sentence recorded separately as risktrial1, risktrial2, and risktrial3.

Ethical Considerations

Both datasets were released under the Open Data Commons Public Domain Dedication and License. However, reuse of the dataset is still subject to any and all applicable Reddit API terms. No metadata or personal identifying information were provided to the models. The study used publicly available text and did not identify individual users. Because the study used publicly available social media text that was not collected for clinical purposes, the results should not be interpreted as diagnostic assessments. The study therefore remains strictly exploratory, with results reported in terms of inter-model agreement rather than clinical accuracy.

It's important to know that Reddit's user base is not demographically representative of the general world population, which may limit the generalizability of patterns observed in the text from the data. This limitation also raises fairness considerations, as language models may perform differently when interpreting language from demographic groups that are

underrepresented on the platform, due to different styles of conversing or different cultural references. Different conversing styles and cultural references already make it difficult for AI to process, due to their pragmatic nature, so the same study done with language from very culturally different geographies may create very different results.

Results

Among all these calculations, the four models showed clear differences in how they interpreted the depression-related and the neutral text samples.

According to Table 1, exact 3-trial agreement ranged from 0.69 to 0.96 depending on model and condition. In the depression set, agreement ranged from 0.69 (Sonnet 5) to 0.81 (GPT-5.6 Luna), while in the neutral set it ranged from 0.78 (DeepSeek V-3) to 0.96 (Qwen3.8). However, the Fleiss' κ across the three trials showed an unexpected result: despite the higher raw agreement in each model, κ drastically dropped

Table 1 Within-Model Variability Across The 3 Trials

Model	Condition	Exact 3-trial agreement	Fleiss' κ (3 trials)	Mean pairwise agreement κ
GPT-5.6 Luna	Depression	0.81	0.775	0.811
GPT-5.6 Luna	Neutral	0.94	0.313	0.491
DeepSeek V-3	Depression	0.76	0.746	0.841
DeepSeek V-3	Neutral	0.78	0.528	0.468
Sonnet 5	Depression	0.69	0.676	0.824
Sonnet 5	Neutral	0.93	0.769	0.803
Qwen 3.8	Depression	0.80	0.774	0.845
Qwen 3.8	Neutral	0.96	0.320	0.257

This table shows how consistently each AI model repeated its own answers in the three trials. For each model and its condition, exact 3-trial agreement, Fleiss' kappa, and Mean pairwise weighted kappa was calculated. Exact 3-trial agreement tells how often the model gave the same label all three times. Fleiss' kappa shows whether that agreement is true rather than just the model favoring one label by incorporating the factor of chance. Mean pairwise weighted kappa measures how similar the answers were overall, giving partial credit when the models stayed close to each other (for example, partial for Low and Medium and none for Low to High).

Table 2 Fleiss' Kappa Across the Four Models

Condition	Fleiss' κ	95% CI
Depression	0.534	[0.435, 0.621]
Neutral	0.180	[0.086, 0.265]
Combined	0.605	[0.533, 0.671]

This table combines the Fleiss' kappa across all four models used and splits them into depression and neutral to see how they differ based on condition. The confidence intervals were calculated using the standard error (derived from observed and expected output) of Fleiss' kappa and the normal approximation ($k = \pm 1.96 \times SE$), which estimates how much the agreement score could vary if the study were repeated with similar data.

because of the high concentration of labels in the Low category. This is known as the kappa paradox. The uneven concentration of labels, in this case in the low category, increases the expected agreement by chance, making kappa significantly lower when raw agreement is high. For example, for GPT-5.6 Luna, κ dropped from 0.775 to 0.313. The mean pairwise weighted κ across trials showed similar patterns in table 1, with lower stability in the neutral set. These results indicate that some of the disagreement may come from variability in the model itself rather than the differences between models.

Inter-model agreement also differed sharply by condition. According to table 2, Fleiss' κ across the four models was moderate for the depression entries ($\kappa = 0.534$, 95% CI [0.435, 0.621]), but was drastically lower for the neutral entries ($\kappa = 0.180$, 95% CI [0.086, 0.265]). Pairwise weighted Cohen's κ also showed the same pattern: in the depression set, κ values ranged from 0.575 to 0.842 (mean = 0.717) and the neutral set showed κ values ranging from -0.014 to 0.567 (mean = 0.203).

Ordinal Krippendorff's α reinforced these findings. According to table 4, agreement was moderate in the depression condition ($\alpha = 0.711$, 95% CI [0.609, 0.786]) and very low in the neutral condition ($\alpha = 0.218$, 95% CI [0.105, 0.327]). The low alpha in the neutral condition indicated that model

agreement was limited. Leave-one-rater-out (LORO) analysis showed how often each model agreed with the consensus formed by the other three. Table 5 shows that in the depression condition, agreement ranged from 0.640 (Qwen3.8) to 0.847 (GPT-5.6 Luna). In the neutral condition, agreement ranged from 0.758 (DeepSeek V-3) to 0.899 (Sonnet 5 and Qwen3.8). Qwen3.8 had the lowest agreement in the depression condition by one LORO measure, while DeepSeek V-3 was the lower one in the neutral condition.

Overall, although the metrics represent different things, the idea is the same. The depression set had more stable agreement than the neutral as they were chance corrected.

Discussion

Interpretation

The patterns and trends observed in these results go to show that measuring different kinds of agreement can present two sides of the same coin. This is important because when dealing in sensitive contexts like AI in mental health situations, it's necessary to consider all the different kinds of trends.

The depression sentences revealed higher agreement, a claim supported by three metrics: a higher Fleiss' κ , pairwise weighted κ , and ordinal Krippendorff's α . In contrast, chance-corrected agreements completely collapsed in the neutral data. Weighted κ values often approached zero with one value even being negative and Fleiss' κ values did the same, despite almost perfect raw agreement, a direct example of the kappa paradox where higher raw agreement in one category leads to a lower kappa. This suggests that either distress-related language has cues that the models could classify more consistently or that there were similarities in model training that contributed to the higher values in the depression set.

Table 3 Pairwise Weighted Cohen’s Kappa
Depression (Mean: 0.717)

Pair	κ	95% CI
Qwen3.8–DeepSeek V-3	0.654	[0.562, 0.735]
Qwen3.8–GPT-5.6 Luna	0.682	[0.570, 0.775]
Qwen3.8–Sonnet 5	0.575	[0.467, 0.674]
DeepSeek V-3–GPT-5.6 Luna	0.807	[0.714, 0.883]
DeepSeek V-3–Sonnet 5	0.842	[0.751, 0.909]
GPT-5.6 Luna–Sonnet 5	0.745	[0.645, 0.826]

This table shows how closely each pair of models agreed with one another in the depression set when assigning risk classifications. Pairwise Weighted Cohen’s kappa specifically calculates the agreement with one other model and accounts for neighboring risk values (little more agreement for Low and Medium than Low and High). The confidence intervals were calculated similarly to the Fleiss’ kappa using the standard error of weighted kappa and formula $k \pm 1.96 \times SE$.

Neutral (Mean: 0.203)

Pair	κ	95% CI
Qwen3.8–DeepSeek V-3	0.164	[0.000, 0.363]
Qwen3.8–GPT-5.6 Luna	-0.014	[-0.029, 0.000]
Qwen3.8–Sonnet 5	0.211	[0.000, 0.504]
DeepSeek V-3–GPT-5.6 Luna	0.050	[0.000, 0.170]
DeepSeek V-3–Sonnet 5	0.567	[0.375, 0.720]
GPT-5.6 Luna–Sonnet 5	0.238	[0.000, 0.504]

This table shows how closely each pair of models agreed with one another in the neutral set when assigning risk classifications, calculated in the same way that the depression set was calculated in.

Combined (Mean: 0.777)

Pair	κ	95% CI
Qwen3.8–DeepSeek V-3	0.713	[0.631, 0.777]
Qwen3.8–GPT-5.6 Luna	0.863	[0.812, 0.905]
Qwen3.8–Sonnet 5	0.708	[0.618, 0.786]
DeepSeek V-3–GPT-5.6 Luna	0.759	[0.661, 0.831]
DeepSeek V-3–Sonnet 5	0.824	[0.754, 0.883]
GPT-5.6 Luna–Sonnet 5	0.797	[0.721, 0.862]

This table shows how closely each pair of models agreed with one another with both sets combined when assigning risk classifications, calculated in the same way that the other sets were calculated in.

In the within-model variability, none of the models were perfectly consistent across the three independent trials, even though the 3-trial agreement was high (ranging from 0.69 to 0.96). This internal variability suggests that some disagreement observed in those single-call evaluations may simply be the model’s unpredictability, rather than actual differences in how the model runs its own classification. The leave-one-rater-out analysis further showed that not one model perfectly aligned with the majority. In the depression condition, GPT-5.6 Luna aligned most closely followed by DeepSeek V-3, while Qwen3.8 was the most noticeable out-

Table 4 Krippendorff’s Alpha (ordinal, 4 raters)

Condition	α (ordinal)	95% CI
Depression	0.711	[0.609, 0.786]
Neutral	0.218	[0.105, 0.327]
Combined	0.729	[0.658, 0.789]

This table reports the values for Krippendorff’s alpha which measures the inter-model agreement on the classifications when the risk labels are treated as ordered levels. Disagreements are measured by how far apart the risk classifications are in the three trials (Low to High being a bigger disagreement than Low to Medium). This metric is useful for ordinal data like this, where “distance” is an important factor in agreement. The CIs were calculated the same way as Fleiss’ kappa and Cohen’s kappa.

lier. In the neutral condition, however, DeepSeek V-3 became the outlier, while the other models aligned more closely with the majority.

Overall, the results show that the models do more than agree or disagree consistently. They show a deeper meaning when accounting for chance and skewed distributions. They also show that variance in the models’ output may not come from differences in their internal classifications, but also, from variation within the model itself.

Implications

These findings carry several implications for future work. First, the collapse of agreement in the neutral condition demonstrates that “neutral” language may not be a reliable control in NLP studies. Not all of the sentences were low for depression risk; some were medium and even high. This proves that even when sentences contain no explicit emotional content, models differ in their sensitivity thresholds and may assign higher risk based on extremely subtle cues. For example, for the sentence “Bored af, anyone want to talk?”, both DeepSeek V-3 and Sonnet 5 took that to mean that the user was lonely, classifying the entry as medium risk for multiple trials. Therefore, if there are to be studies testing text charged with depression language, neutral language, although it may seem like it’s the opposite, will not provide a clean control.

Second, the presence of within-model variability shows that single trials are insufficient for comparing LLMs. Repeated trials and independent context evaluation are advisable when studying LLM classification consistency, particularly because model outputs can vary across identical inputs, as shown in the experiment’s results. Without trial repetition, the inter-model variability could be mistaken for disagreement between different models, leading to incorrect conclusions about model behavior.

Psychologists and data scientists who rely on AI-labeled data or AI tools for analysis may suffer from inconsistent research findings, which can harm their knowledge base as well as their external clients that they are trying to help. Prior re-

Table 5 Leave-One-Rater-Out (LORO) Consensus

Condition	Held-out model	N valid (majority exists)	N no-majority	% agreement w/ other-3 majority	Mean (ordinal distance)
Depression	GPT-5.6 Luna	100	0	0.847	0.360
Depression	DeepSeek V-3	98	2	0.837	0.163
Depression	Sonnet 5	98	2	0.750	0.153
Depression	Qwen3.8	100	0	0.640	0.250
Neutral	GPT-5.6 Luna	99	1	0.889	0.101
Neutral	DeepSeek V-3	99	1	0.758	0.242
Neutral	Sonnet 5	99	1	0.899	0.111
Neutral	Qwen3.8	99	1	0.899	0.101
Combined	GPT-5.6 Luna	199	1	0.868	0.231
Combined	DeepSeek V-3	197	3	0.797	0.203
Combined	Sonnet 5	197	3	0.769	0.132
Combined	Qwen3.8	199	1	0.824	0.176

This table shows how often each model agrees with the majority formed by the other three models. I removed one model, let the other three form a majority label, and then checked how often the held-out model matched that majority. Higher percentages mean the model usually matches the group. The ordinal distance (distance from majority) shows how far off the model is when it disagrees.

search has already demonstrated that different approaches to mental-health assessment by LLMs can produce different results, highlighting the importance of considering variability in mental-health contexts²². People express distress in countless ways, so if they don't express it in a "traditional" way, their signs may be misinterpreted. This could also lead to unequal detection where some language or

communication styles are less recognized or understood than others²³.

Limitations

A key limitation of this study is the relatively small sample sizes that were taken from the two datasets. Although this was chosen to allow for more controlled analysis and more manual inspection across the LLMs, the size limits the generalizability of the findings, and the results given back should be looked at as solely exploratory and not as a true representation of the general population. Furthermore, the datasets were only used with one social media platform: Reddit. Although Reddit is used widely around the world, it may not cover aspects of all complex language that shows up on all kinds of social media. For example, if Reddit is mainly language based, Snapchat is mainly image based and Instagram is mainly video based; the signs would be different there.

A further limitation is that participation in the depression subreddit was used as a proxy for depression-related content. Simply partaking in the depression subreddit does not equate to a diagnosis and some depression-sample sentences even describe someone else's struggles rather than the actual user's, so these labels should be read as proxies, not validated mental health status²⁴.

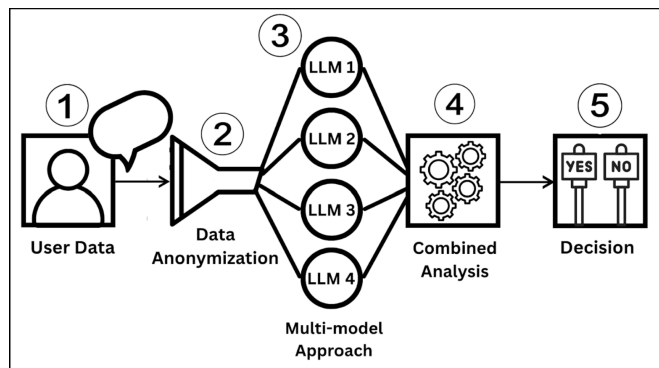
Another limitation came from the control condition. The sentences from the control group were picked to classify examples of sentences not from a depression-focused subreddit community. This was because the study intentionally did not provide a clinician-verified rubric defining the Low, Medium, and High categories, so each model was free to interpret the concept of depression risk according to its own internal knowledge base. Therefore, differences in how the models interpreted these categories may have contributed to classifications above Low within the neutral condition. This makes the neutral dataset not as suitable as a conventional control group.

Future Work

Currently, AI engineers and scientists are looking at linguistic and behavioral patterns to flag mental health distress. However, this variability remains an important concern²⁵. To overcome this, creating a model embedded in the backend of social media apps that uses an ensemble approach may rectify the variability. An ensemble approach is one that combines outputs of multiple AI models to help reduce the influence of any one model's tendencies and garner more consistent results. The model would analyze user data using other curated public LLMs. Using one model is powerful, but using multiple can help improve risk classification stability and consistency, as shown in Figure 6.

Even with the amount of data going into this system, this proposal aims to prioritize user privacy. Any further implementation would need to use de-identification procedures to remove PII, along with any necessary technology to store and access the model outputs. However, as this study does not make any claims about clinical accuracy, this system would

only provide a preliminary classification that would then require human oversight to determine if that classification is appropriate.



Designed in Canva ®

Figure. 6 Diagram of Ensemble Approach. This diagram represents the process of how data (1) would be put through a data anonymization algorithm, pulling out any PII (2) and then sent to multiple LLMs (3). Finally, the outputs of all the LLMs would be put through an algorithm that synthesizes the data (4) and uses patterns from the different LLMs to produce the combined risk classification (5).

Conclusion

In a broad sense, the growing use of AI mental health tools and chatbots using social media data carries serious risks, including oversimplified advice and inconsistent emotional interpretation, yet their potential to increase accessibility and support for users is undeniable. This experiment demonstrates how much LLMs vary in interpreting emotional patterns and how much they vary within themselves, leading to discrepancies in classifications across models. Therefore, it can be said that these systems would need significant improvement if they are to give consistent outputs in sensitive mental health contexts.

References

- 1 A. M. Ojeda, J. B. Valera, O. Diaz. Artificial intelligence of big data for analysis in organizational decision-making. *Global Journal of Flexible Systems Management*. Vol. 26, pg. 515–527, 2025. <https://doi.org/10.1007/s40171-025-00450-2>.
- 2 P. Törnberg. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*. Vol. 43, pg. 1181–1195, 2024. <https://doi.org/10.1177/08944393241286471>.
- 3 N. Farruque, R. Goebel, S. Sivapalan, O. R. Zaïane. Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. *Language Resources and Evaluation*. Vol. 58, pg. 1013–1041, 2024. <https://doi.org/10.1007/s10579-024-09720-4>.
- 4 S. Basu, B. Huynh. Mitigating hallucinations in healthcare AI: a systematic review of evidence-based strategies. *BMC Health Services Research*.

- Vol. 26, pg. 1115, 2026. <https://doi.org/10.1186/s12913-026-14851-1>.
- 5 K. C. Bathina, M. ten Thij, L. Lorenzo-Luaces, L. A. Rutter, J. Bollen. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*. Vol. 5, pg. 458–466, 2021. <https://doi.org/10.1038/s41562-021-01050-7>.
- 6 B. G. Bokolo, Q. Liu. Deep learning-based depression detection from social media: comparative evaluation of ML and transformer techniques. *Electronics*. Vol. 12, pg. 4396, 2023. <https://doi.org/10.3390/electronics12214396>.
- 7 Y. Cai, H. Wang, H. Ye, Y. Jin, W. Gao. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*. Vol. 217, pg. 119538, 2023. <https://doi.org/10.1016/j.eswa.2023.119538>.
- 8 M. De Choudhury, M. Gamon, S. Counts, E. Horvitz. Predicting depression via social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. pg. 128–137, 2013.
- 9 J. Kim, J. Lee, E. Park, J. Han. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*. Vol. 10, pg. 1–9, 2020. <https://doi.org/10.1038/s41598-020-68764-y>.
- 10 J. Liu, P. Ding, J. Chen. DepGLM: depression degree recognition on social media based on large language models. *Digital Health*. Vol. 11, 2025. <https://doi.org/10.1177/20552076251408281>.
- 11 T. Zhang, K. Yang, S. Ananiadou. Sentiment-guided transformer with severity-aware contrastive learning for depression detection on social media. *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. pg. 114–126, 2023. <https://doi.org/10.18653/v1/2023.bionlp-1.9>.
- 12 U. Ahmed, J. C. Lin, G. Srivastava. Graph attention network for text classification and detection of mental disorder. *ACM Transactions on the Web*. Vol. 17, pg. 1–31, 2023. <https://doi.org/10.1145/3572406>.
- 13 E. A. Rissola, M. Aliannejadi, F. Crestani. Mental disorders on online social media through the lens of language and behaviour: analysis and visualisation. *Information Processing & Management*. Vol. 59, pg. 102890, 2022. <https://doi.org/10.1016/j.ipm.2022.102890>.
- 14 S. Tariq, N. Akhtar, H. Afzal, S. Khalid, M. R. Mufti, S. Hussain, A. Habib, G. Ahmad. A novel co-training-based approach for the classification of mental illnesses using social media posts. *IEEE Access*. Vol. 7, pg. 166165–166172, 2019. <https://doi.org/10.1109/access.2019.2953087>.
- 15 A. Nowacki, W. Sitek, H. Rybiński. LLM-based classifiers for discovering mental disorders. *Journal of Intelligent Information Systems*. Vol. 64, pg. 161–178, 2025. <https://doi.org/10.1007/s10844-025-00934-8>.
- 16 M. Settanni, F. Quilghini, A. Toscano, D. Marengo. Assessing the accuracy and consistency of large language models in triaging social media posts for psychological distress. *Psychiatry Research*. Vol. 351, pg. 116583, 2025. <https://doi.org/10.1016/j.psychres.2025.116583>.
- 17 A. Arnaiz-Rodriguez, M. Baidal, E. Derner, J. Layton Annable, M. Ball, M. Ince, E. Perez Vallejos, N. Oliver. Between help and harm: an evaluation study of mental health crisis handling by large language models. *JMIR Mental Health*. Vol. 13, 2026. <https://doi.org/10.2196/88435>.
- 18 T. Scholich, M. Barr, S. W. Stirman, S. Raj. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. *JMIR Mental Health*. Vol. 12, pg. e69709, 2025. <https://doi.org/10.2196/69709>.
- 19 D. Owen, D. Antypas, A. Hassoulas, A. F. Pardiñas, L. Espinosa-Anke, J. C. Collados. Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. *JMIR AI*. Vol. 2, pg. e41205, 2023.

-
- <https://doi.org/10.2196/41205>.
- 20 W. Yuan, B. K. Beaulieu-Jones, K. H. Yu, S. L. Lipnick, N. Palmer, J. Loscalzo, T. Cai, I. S. Kohane. Temporal bias in case-control design: preventing reliable predictions of the future. *Nature Communications*. Vol. 12, pg. 1107, 2021. <https://doi.org/10.1038/s41467-021-21390-2>.
 - 21 M. Sharabov, P. Georgieva, G. Tsochev, R. Trifonov. Comparative analysis of large language models: case study of GPT, Gemini, DeepSeek. 2025 13th International Scientific Conference on Computer Science (COM-SCI). pg. 1–4, 2025. <https://doi.org/10.1109/COMSCI67172.2025.11225122>.
 - 22 C. Lee, M. Mohebbi, E. O'Callaghan, M. Winsberg. Large language models versus expert clinicians in crisis prediction among telemental health patients: comparative study. *JMIR Mental Health*. Vol. 11, 2024. <https://doi.org/10.2196/58129>.
 - 23 I. Straw, C. Callison-Burch. Artificial intelligence in mental health and the biases of language based models. *PLoS One*. Vol. 15, pg. e0240376, 2020. <https://doi.org/10.1371/journal.pone.0240376>.
 - 24 S. M. Shah, S. A. Gillani, M. S. A. Baig, M. A. Saleem, M. H. Siddiqui. Advancing depression detection on social media platforms through fine-tuned large language models. *Online Social Networks and Media*. Vol. 46, pg. 100311, 2025. <https://doi.org/10.1016/j.osnem.2025.100311>.
 - 25 N. C. Cardamone, M. Olfson, T. Schmutte, L. Ungar, T. Liu, S. R. Cullen, N. J. Williams, S. C. Marcus. Classifying unstructured text in electronic health records for mental health prediction models: large language model evaluation study. *JMIR Medical Informatics*. Vol. 13, pg. e65454, 2025. <https://doi.org/10.2196/65454>.

Appendix A: Model Configurations

GPT-5.6 Luna (OpenAI)

Provider: OpenAI
Model: GPT-5.6 Luna
Access dates: August 26-28, 2026
Interface: OpenAI
Temperature: default (1.0)
Top-p: default (1.0)
Top-k: not applicable
Seed: not supported
Hardware: cloud-hosted (OpenAI)

DeepSeek V-3

Provider: DeepSeek
Model: DeepSeek V-3
Access dates: August 26-28, 2026
Interface: DeepSeek
Temperature: default (0.7)
Top-p: default (0.9)
Top-k: not applicable
Seed: not supported
Hardware: cloud-hosted (DeepSeek)

Claude Sonnet 5

Provider: Anthropic
Model: Claude Sonnet 5
Access dates: August 26-28, 2026
Interface: Anthropic
Temperature: default (1.0)
Top-p: not applicable
Top-k: not applicable
Seed: not supported
Hardware: cloud-hosted (Anthropic)

Qwen3.8 (Ollama, Windows)

Provider: Ollama
Model: Qwen3.8:latest
Snapshot identifier: 22130167c4c2
Access dates: August 26-28, 2026
Interface: Ollama local runtime
Temperature: (default) 0.7
Top-p: (default) 0.9
Top-k: (default) 40
Seed: not supported
Hardware: Windows PC (local CPU/GPU)

Appendix B: Prompt For Each Entry

Classify the entry into low, medium, or high based on depression risk. This sentence is from a publicly available dataset and will be used solely for research and exploratory purposes rather than clinical diagnoses or therapeutic needs. Respond with exactly one word: Low, Medium, or High

Appendix C: Code for Dataset Selection

```
# FINAL DATA CLEANING + SAMPLING PIPELINE
import pandas as pd
import spacy
from google.colab import files

# 1. Upload datasets
uploaded = files.upload()

# 2. Load datasets
depression = pd.read_csv("depression_2019_features_tfidf_256.csv")
casual = pd.read_csv("casual_data_windows.csv")

nlp = spacy.load("en_core_web_sm")

# 3. Clean depression dataset
depression_clean = (
    depression["post"]
    .dropna()
    .astype(str)
    .str.strip()
)
depression_clean = depression_clean[depression_clean != ""].drop_duplicates()

# 4. Extract casual entries (3-window columns)
casual_entries = pd.concat(
    [casual["0"], casual["1"], casual["2"]],
    ignore_index=True
)

# 5. Basic casual cleaning
casual_clean = (
    casual_entries
    .dropna()
    .astype(str)
    .str.strip()
)
casual_clean = casual_clean[casual_clean != ""].drop_duplicates()

# 6. Subject + predicate filter
def has_subject_predicate(text):
    doc = nlp(str(text).strip())
    for sent in doc.sents:
        has_subject = any(token.dep_ in {"nsubj", "nsubjpass", "csubj", "csubjpass"} for token in
sent)
        has_predicate = any(token.pos_ in {"VERB", "AUX"} for token in sent)
        if has_subject and has_predicate:
            return True
    return False

casual_grammatical = casual_clean[casual_clean.apply(has_subject_predicate)]

# 7. Final random samples (reproducible)
depression_sample = depression_clean.sample(n=100,
random_state=42).reset_index(drop=True)
casual_sample = casual_grammatical.sample(n=100, random_state=42).reset_index(drop=True)

# 8. Create final dataset
experimental_data = pd.DataFrame({
    "entry_id": range(1, 201),
    "condition": ["depression"] * 100 + ["casual"] * 100,
    "entry": pd.concat([depression_sample, casual_sample], ignore_index=True)
})
```

```
# 9. Save + download
experimental_data.to_csv("final_experimental_sample_100_100.csv", index=False)
files.download("final_experimental_sample_100_100.csv")
```