Artificial Allies: Validation of Synthetic Text for Peer Support Tools through Data Augmentation in NLP Model Development

Josué Godeme
Research Computing and Data Services, Information, Technology & Consulting, Dartmouth College
Hanover, NH 03784, USA
E-mail: josue.f.godeme.26@dartmouth.edu
www.dartmouth.edu

Julia Hill
Department of Psychiatry, Geisel School of Medicine, Dartmouth College
Hanover, NH 03784, USA
E-mail:Julia.clark.hill@gmail.com

Stephen P. Gaughan, Wade J. Hirschbuhl, Amanda J. Emerson, Christian Darabos, Carly A. Bobak
Research Computing and Data Services, Information, Technology & Consulting, Dartmouth College
Hanover, NH 03784, USA
Email: Stephen.P.Gaughan, Wade.J.Hirschbuhl, Amanda.J.Emerson, Christian.Darabos, Carly.A. Bobak@dartmouth.edu

Karen L. Fortuna
Department of Psychiatry, Geisel School of Medicine, Dartmouth College
Hanover, NH 03784, USA
E-mail: Karen.L.Fortuna@dartmouth.edu

This study investigates the potential of using synthetic text to augment training data for Natural Language Processing (NLP) models, specifically within the context of peer support tools. We surveyed 22 participants—13 professional peer supporters and 9 AI-proficient individuals—tasked with distinguishing between AI-generated and human-written sentences. Using signal detection theory and confidence-based metrics, we evaluated the accuracy and confidence levels of both groups. The results show no significant differences in rater agreement between the two groups (p = 0.116), with overall classification accuracy falling below chance levels (mean accuracy = 43.10%, p < 0.001). Both groups exhibited a tendency to misclassify low-fidelity sentences as AI-generated, with peer supporters showing a significant bias (p = 0.007). Further analysis revealed a significant negative correlation between errors and confidence among AI-proficient raters (r = -0.429, p < 0.001), suggesting that as their confidence increased, their error rates decreased. Our findings support the feasibility of using synthetic text to mimic human communication, with important implications for improving the fidelity of peer support interventions through NLP model development.

Keywords: Synthetic text generation; Natural language processing; Peer support; Signal detection theory; AI-generated content; Rater agreement; Fidelity classification.

1. Introduction

Peer support specialists play a crucial role in the mental health care system.[1–4] These individuals, who have lived experiences of mental health conditions, provide emotional, social, and practical assistance to others facing similar challenges. The peer support movement has grown significantly, with peer support specialists becoming an integral part of mental health services due to their ability to engage and support individuals in ways that complement traditional clinical interventions.[1–4] This form of support is particularly important for adults with serious mental illnesses, who often face high rates of morbidity and reduced life expectancy due to poorly managed health conditions.[5]

Despite the proven benefits of peer support,[3,4,6] there is a gap in tools that can assist peer supporters in delivering consistent and high-quality care. An ideal tool would not only aid in real-time fidelity monitoring but also enhance the delivery of evidence-based practices.[7,8] Kadakia et al's[9,10] prior work has shown promise in this area, utilizing a deep learning model trained on data from both recorded peer support conversations and Reddit to classify high-fidelity peer support techniques. This approach demonstrated that natural language processing (NLP) could be used to scale and ensure the fidelity of digital peer support interventions.

It has previously been established that improving data quality and quantity is a critical step in improving deep learning model accuracy, particularly for NLP models.[11,12] However, in our application, accessing mental health data is often difficult, and transcription of interactions can be labor-intensive and prone to errors.[13–15] Furthermore, deep learning NLP algorithms typically require large amounts of high-quality data to perform optimally.[16,17] Previously, researchers have demonstrated that LLM generated text can be used to improve the performance of NLP-related tasks, including text classification.[18–20] Hence, we hypothesize that large language models (LLMs) can be used to generate synthetic data that closely mimic real peer support mental health sessions, thereby enhancing the fidelity classification of peer support interventions.

In this study, we seek to demonstrate the feasibility of using synthetic data to mimic human-written content effectively in the peer-supporter context. We also aim to validate that peer supporters, as well as individuals who are professionally engaged in working with LLMs (dubbed AI-proficient non-peer supporters) struggle to reliably distinguish between LLM-generated sentences and real human sentences. This research will contribute to the understanding of synthetic data validation and its potential to support the development of robust tools for peer supporters, ultimately enhancing the quality of mental health care.

2. Methods

2.1. *Original Data Collection*

Collection and transcription of the original human generated conversations are described in Kadakia et al.[9] In short, anonymized records of peer support conversations from the PeerTECH platform where manually recorded verbatim.[9] High-fidelity and low-fidelity sentences are defined in the context of adherence to best practices for peer support in mental health.[9,10,21] High-fidelity sentences refers to interactions that strictly follow established protocols and best practices, ensuring comprehensive and consistent delivery of peer support. These interactions typically include elements such as active listening, empathy, validation of experiences, and appropriate use of self-disclosure.[9,10,21] Low-fidelity sentences, on the other hand, denote interactions that deviate from these best practices, potentially lacking in one or more critical aspects of effective peer support.[9,10,21] Such deviations might include inadequate listening, insufficient emotional engagement, or inappropriate self-disclosure, which can undermine the effectiveness of the support provided.

2.2. *Synthetic Text Generation*

To generate the synthetic text, we utilized OpenAI's Application programming interface (API), specifically the GPT-4 Turbo model,[22] which was the most advanced model available at the time of the study. The process aimed to produce 10,000 sentences, which should provide a robust training set for downstream NLP modeling.[23] The GPT-4 Turbo model was configured with a temperature setting of 0.9.

The generation process involved three key components: a system prompt, a specific prompt, and user profiles. Two distinct system prompts were employed to generate transcripts demonstrating both high- and low-fidelity practices in peer support conversations. The specific prompt was constructed using characteristics of both the peer supporter and the patient. For the peer supporter, the prompt included their age, gender, personality traits, mental health history, and the topic of the support session. For the patient, the prompt specified their age, gender, personality traits, and their diagnosed mental health condition. This structured approach ensured that each generated conversation was contextualized with specific demographic and psychological information for both participants.

> **Example Prompt 1**
> Peer Supporter - Age: 35, Gender: female, Traits: compassionate, insightful, Mental Health History: post-traumatic stress disorder, Session Topic: coping with trauma
> Patient - Age: 29, Gender: male, Traits: distrustful, struggling, Mental Health Issue: trauma recovery
>
> **Example Prompt 2**
> Peer Supporter - Age: 41, Gender: female, Traits: calm, reassuring, Mental Health History: post-traumatic stress disorder, Session Topic: managing triggers
> Patient - Age: 30, Gender: female, Traits: jumpy, anxious, Mental Health Issue: post-traumatic stress disorder

All data manipulation and analysis were conducted in R version 4.3.2,[24] with extensive use of the `tidyverse` suite of packages[25] for data manipulation, cleaning, and visualization. A total of 154 API calls were executed, resulting in the generation of 10,736 sentences, exceeding the initial target of 10,000 sentences.

2.3. *Synthetic Text Validation*

To evaluate the accuracy and confidence of human raters in distinguishing between human-generated and synthetic text, we randomly selected 100 sentences. These sentences were categorized based on their origin and fidelity: 17 high-fidelity and 14 low-fidelity sentences were human-generated, while 43 high-fidelity and 26 low-fidelity sentences were synthetic. High-fidelity refers to adherence to best practices for peer support in mental health providing, while low-fidelity indicates lesser adherence.

We recruited two types of raters: AI-proficient non-peer supporters from Information, Technology, and Consulting at Dartmouth College, and peer professionals recruited through social media calls and email lists. Raters rated their confidence in how each sentence was generated, using the options: Definitely Human, Maybe Human, I Don't Know, Maybe AI, and Definitely AI. Responses were collected using the *Qualtrics* survey platform (Qualtrics, Provo, UT).

The survey data, which included ratings from AI-proficient non-peer supporters and peer supporters was rated on a scale from 1 to 5, where 1 represented "Definitely AI" and 5 represented "Definitely Human." Confidence ratings were assigned numerical values: 100 for "Definitely," 60 for "Maybe," and 0 for "I don't know."

To evaluate rater performance, several metrics were calculated:

1. **Percentage Agreement**: For each sentence, the percentage agreement among all raters was calculated by determining the proportion of ratings that matched the most common rating.

$$\text{Agreement} = \left(\frac{\sum_{i=1}^{n} I(r_i = \text{mode}(r))}{n} \right) \times 100$$

where r_i represents the rating of the i-th rater, mode(r) is the most common rating among all raters, and n is the total number of raters.

2. **Group-Specific Agreement**: The percentage agreement was calculated separately for AI-proficient non-peer supporters and peer supporters to understand agreement within each group.

3. **Weighted Accuracy**: Weighted accuracy was determined by comparing each rating to the true origin of the sentence and adjusting for confidence levels.

$$\text{Weighted Accuracy} = \left(\frac{\sum_{i=1}^{n} w_i \cdot I(r_i = \text{true origin})}{n} \right) \times 100$$

where w_i is the weight assigned based on the confidence level of the i-th rating, r_i is the rating of the i-th rater, and true origin indicates whether the sentence is human or AI-generated.

4. **Percentage of Errors**: The percentage of incorrect ratings was calculated by determining the proportion of ratings that deviated from the true origin of the sentence.

5. **Average Confidence**: The average confidence level for each sentence was calculated by averaging the confidence scores provided by the raters.

Summary statistics were generated to provide an overview of the overall agreement, accuracy, error rates, and confidence levels among the different rater groups.

We also calculated accuracy, errors, and confidence at the rater level. For each rater, the percentage of sentences judged as human that were actually human was calculated:

$$\text{Percentage Judged Human (Actual Human)} = \left(\frac{\sum_{i=1}^{n} I(r_i \in \{4,5\} \wedge t_i = \text{Human})}{\sum_{i=1}^{n} I(r_i \in \{4,5\})} \right) \times 100$$

and the percentage of sentences judged as human that were actually AI was calculated:

$$\text{Percentage Judged Human (Actual AI)} = \left(\frac{\sum_{i=1}^{n} I(r_i \in \{4,5\} \wedge t_i = \text{Synthetic})}{\sum_{i=1}^{n} I(r_i \in \{4,5\})} \right) \times 100$$

2.4. *Statistical Analysis*

We performed various statistical tests to evaluate differences in rater performance and confidence. Paired t-tests[26,27] compared the accuracy, confidence, and agreement between peer supporters and AI-proficient non-peer supporters. A one-sample t-test assessed whether the overall accuracy differed significantly from 50%. An independent t-test evaluated the overall agreement among all raters.

Correlation tests[28] examined the relationship between errors and confidence levels for both rater groups and for sentences labeled with low-fidelity. Specifically, correlations between errors and confidence for AI-proficient non-peer supporters and peer supporters were assessed, as well as for low-fidelity sentences.

To compare the proportion of AI judgments between low and high-fidelity sentences, paired t-tests were performed separately for AI-proficient non-peer supporters and peer supporters. Paired t-tests[26,27] were also conducted to compare the percentage of judgments that were actually human versus AI for each rater type.

2.5. *Signal Detection Analysis*

To evaluate the ability of raters to distinguish between human-generated and synthetic text, we calculated signal detection measures. Weights for definite and maybe confidence levels were defined, assigning a weight of 1 for definite judgments and 0.6 for maybe judgments.

For each rater, we calculated the signal detection theory (SDT) measures,[29] including the signal detection score (d’), beta, and criterion (c).

For each sentence, the counts of hits (true positives), false alarms (false positives), misses (false negatives), and correct rejections (true negatives) were determined based on the ratings and true origin. Specifically, for sentences with a true origin of human, hits were defined as ratings of “Definitely Human” or “Maybe Human,” and false alarms were defined as ratings of “Definitely AI” or “Maybe AI.” For sentences with a true origin of synthetic, hits were defined as ratings of “Definitely AI” or “Maybe AI,” and false alarms were defined as ratings of “Definitely Human” or “Maybe Human.”

The SDT measures were calculated using the `psycho` package in R.[30] The d’ score was computed as:

$$d' = \Phi^{-1}(\text{hit rate}) - \Phi^{-1}(\text{false alarm rate})$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

For each rater, the weighted SDT measures (d’, beta, and c) were calculated separately for human and synthetic origins. The combined measures for each rater were used to perform t-tests to compare against a hypothesized mean of 0. T-tests[26,27] for the combined d' scores, beta, and c values were conducted to determine if there was a significant ability to distinguish between human and synthetic text.

2.6. *Insight Calculation (Meta d')*

To evaluate rater insight, we calculated the meta d' score, which measures a rater's metacognitive ability to discriminate between their correct and incorrect judgments.[31,32]

The meta d' score was calculated using the negative log-likelihood optimization approach. Specifically, we minimized the negative log-likelihood to find the meta d' value that best describes the observed data. The optimization was performed using the L-BFGS-B method,[33] ensuring the parameter estimates stayed within reasonable bounds.

The steps to calculate meta d' included:

(1) Aggregating the ratings data for each rater to count the occurrences of each confidence level (0, 60, 100) for human and synthetic sentences.
(2) Defining the negative log-likelihood function based on the signal detection theory model parameters.
(3) Using the `optimx` package[34] to optimize the parameters and calculate the meta d' score.
(4) Extracting and summarizing the meta d' scores for each rater.

2.7. *Sensitivity and Specificity Analysis*

We calculated the sensitivity and specificity for each rater to evaluate their ability to correctly identify human-generated text (with human as the positive case). The analysis was performed using R with the `dplyr`,[35] `tidyr`,[25] `purrr`,[36] `ggplot2`,[37] and `pROC`[38] packages.

Area under the receiver operating characteristic (AUROC) curves were calculated for each rater. AUROC curves were plotted for the best, worst, and median raters.[39,40]

2.8. *Code Availability Statement*

The code used in this study is publicly available on GitHub at [https://github.com/FrejusGdm/Synthetic-Text-Validation-Karen-Fortuna].

3. Results

The age of peer supporters ranged from 26 to 45 years (M = 32.5, SD = 4.2), while patients' ages ranged from 19 to 45 years (M = 29.3, SD = 5.7). The most common mental health issues addressed were depression (22.4%), social anxiety (14.9%), and obsessive-compulsive disorder (13.0%). These are shown in Figure 1.

We recruited 9 AI-proficient non-peer supporters professionals and 13 professional peer supporters to complete the survey (n=22). The mean agreement across all raters was 27.97 (95% CI: 25.55, 30.39). There was no significant difference in the levels of agreement between AI-proficient non-peer supporters and peer supporters ($p = 0.12$). The overall accuracy of raters was lower than what would be expected by random chance, with a mean accuracy of 43.10

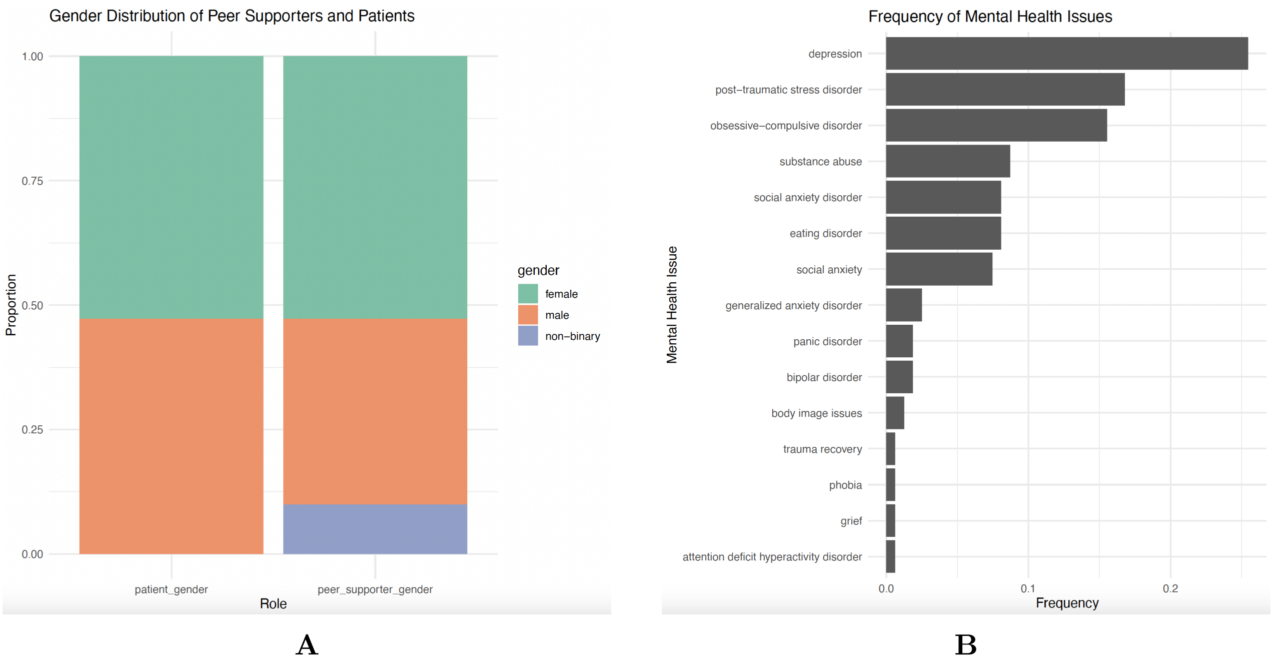

Fig. 1: Demographic Distributions of Peer Supporters and Patients. (A) Gender distribution among peer supporters and patients. (B) Frequency of various mental health issues reported by patients.

(95% CI: 41.11, 45.09; $p < 0.001$ for a two-sided t-test). Within this, AI-proficient non-peer supporters demonstrated higher accuracy (48.62%) compared to peer supporters (36.41%; mean difference -12.21$p < 0.001$) and reported higher confidence levels (mean difference -13.30$p < 0.001$), although the overall confidence was generally low, with a mean confidence score of 47.75 (95% CI: 45.61, 49.89). These relationships are illustrated in Figure 2 (A)-(C).

Overall, we found that errors and confidence were not significantly correlated ($p = 0.08$). However, this overall trend masks important differences between groups and conditions. Among AI-proficient non-peer supporters, there was a significant negative correlation between errors and confidence ($r = -0.43$, 95% CI: -0.55, -0.25; $p < 0.001$), indicating that as confidence increased, errors decreased. In contrast, for peer supporters, the correlation between errors and confidence was not significant ($r = -0.19$, 95% CI: -0.37, 0.01; $p = 0.06$). These results are shown in Figure 2 (D).

When examining sentences labeled with low-fidelity, the correlation between errors and confidence for peer supporters was not significant ($r = -0.03$, 95% CI: -0.34, 0.29; $p = 0.87$). However, for AI-proficient non-peer supporters, there was a significant negative correlation ($r = -0.33$, 95% CI: -0.58, -0.02; $p = 0.04$) in low-fidelity sentences. These results are shown in Figures 2 (E) and (F).

In high-fidelity sentences, peer supporters exhibited a significant negative correlation between errors and confidence ($r = -0.38$, 95% CI: -0.58, -0.14; $p = 0.003$). Similarly, AI-proficient non-peer supporters showed a significant negative correlation ($r = -0.51$, 95% CI: -0.68, -0.30; $p < 0.001$).

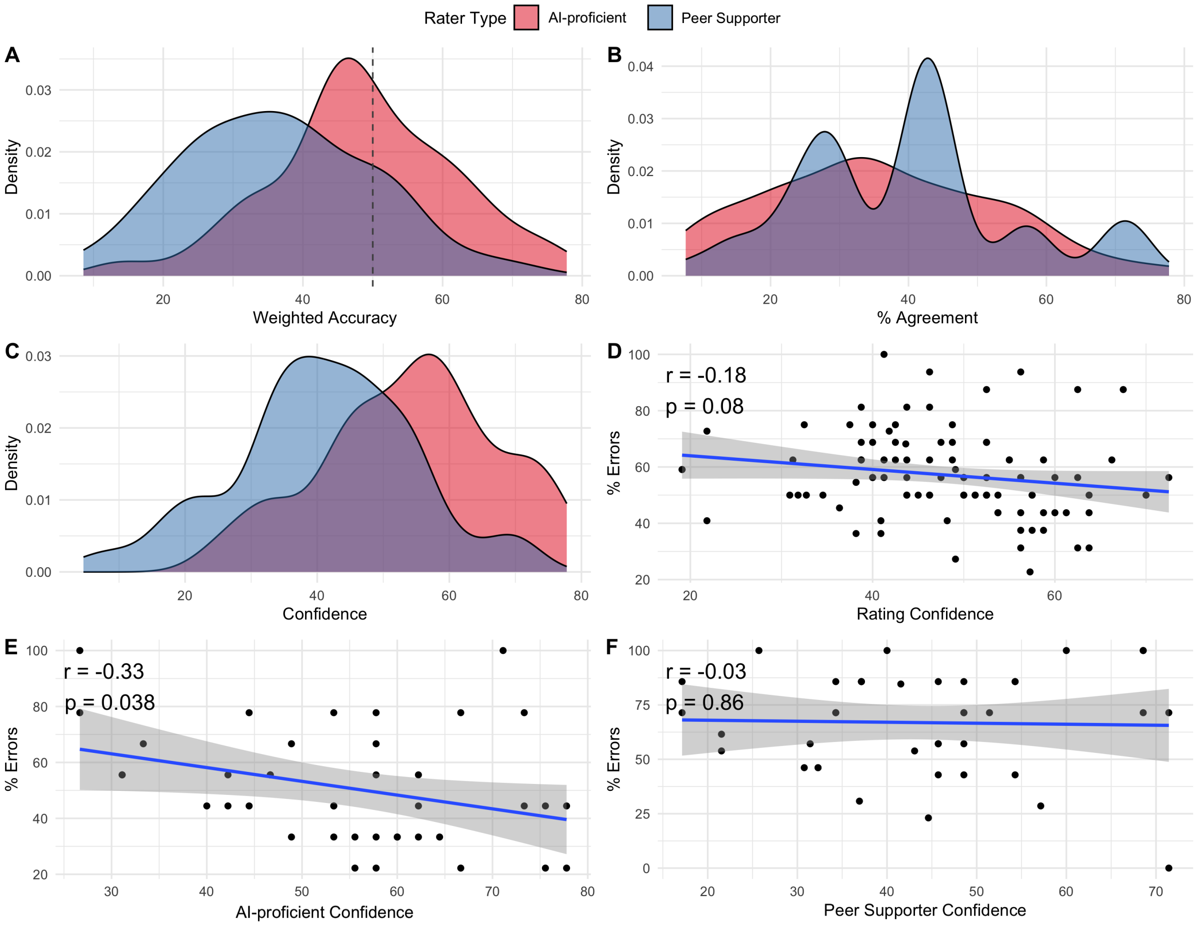

Fig. 2: Analysis of rater performance and confidence. (A) Weighted accuracy for AI-proficient non-peer supporters and peer supporters, with a 50% accuracy line indicated. (B) Percentage agreement among AI-proficient non-peer supporters and peer supporters. (C) Rating confidence. (D) Scatter plot with fitted line and 95% confidence intervals showing rating confidence by percentage errors. (E) Same as (D) for AI-proficient non-peer supporters with low-fidelity ratings. (F) Same as (D) for peer supporters with low-fidelity ratings.

Peer support raters were more likely to assume that low-fidelity sentences were AI-generated compared to high-fidelity sentences. This difference in proportions was statistically significant ($p = 0.007$), with a difference in proportions ranging from 0.02 to 0.10. For AI-proficient non-peer supporters, the tendency to assume low-fidelity sentences were AI-generated was also observed, although the difference was only borderline significant ($p = 0.05$), with a difference in proportions ranging from 0.00 to 0.09. These findings indicate that both peer supporters and AI-proficient non-peer supporters are more inclined to classify low-fidelity sentences as AI-generated, though this tendency is more pronounced and statistically significant among peer supporters.

The bar plot in Figure 3 (A) visualizes the percentage of sentences judged as "Human" by two different groups of raters: AI-proficient non-peer supporters and peer supporters, for sentences that were actually AI-generated (AI). Of all sentences rated as “Human” by AI-proficient non-peer supporters, 66.7% were AI-generated. This percentage was higher in the peer supporter group, with 83.3% of sentences rated as “Human” being AI-generated. Statistical tests revealed that for AI-proficient non-peer supporters, the tendency to judge AI sentences as “Human” was borderline statistically significant ($p = 0.05$). In contrast, Peer Supporters showed a statistically significant tendency to judge AI-generated sentences as “Human” ($p < 0.001$). These findings indicate a tendency for both AI-proficient non-peer supporters and Peer Supporters to be deceived by AI-generated content, with Peer Supporters being particularly susceptible. This could perhaps indicate AI hyperrealism[31]— where even trained individuals are frequently unable to distinguish AI from human-generated text.

Signal detection theory was applied to evaluate the ability of raters to distinguish between human-generated and AI-generated text. The d’ (d-prime) score is a measure of a rater’s ability to discriminate between signal (human-generated text) and noise (AI-generated text), where a higher d’ indicates better discrimination ability. Our analysis revealed that the combined mean d’ score was significantly greater than zero (mean = 0.39, 95% CI: 0.22, 0.55; $p < 0.001$), indicating that detection is occurring among raters.

In addition to d’, we also evaluated beta (β) and criterion (c), which provide insights into the decision-making biases of the raters. A positive beta (β) indicates a conservative response bias, meaning raters are less likely to label sentences as human. The combined mean beta (β) was significantly greater than zero (mean = 1.37, 95% CI: 1.17, 1.57; $p < 0.001$), suggesting a strong bias towards not labeling sentences as human. Similarly, the combined mean criterion (c) was significantly greater than zero (mean = 0.62, 95% CI: 0.44, 0.80; $p < 0.001$), reinforcing the notion of a reluctance to label sentences as human.

We also calculated the meta d’ score, referred to as insight, based on the raters’ confidence levels. The meta d’ score measures a rater’s metacognitive ability to discriminate between their correct and incorrect judgments. Our results showed that the combined mean meta d’ score was not significantly different from zero (mean = -0.39, 95% CI: -1.37, 0.59; $p = 0.42$). This near-zero insight score indicates that raters do not have a reliable metacognitive awareness of their accuracy in distinguishing between human and AI-generated sentences, suggesting that their confidence levels do not effectively reflect their true performance.

Plotting the signal detection score against the insight score allows us to identify how detection and insight are interrelated. As shown in Figure 3 (B), we observe only two raters (9%) with both good insight and good detection in the top-left quadrant of the plot.

Sensitivity and specificity were calculated with regards to the raters’ ability to discern human-written sentences and are displayed in Figure 2(C). No rater achieved both a sensitivity and specificity greater than 0.7, indicating that none of the raters were highly proficient at correctly identifying human-written sentences while also correctly rejecting AI-generated ones.

We calculated the area under the receiver operating characteristic curve (AUROC) for each rater, and a boxplot of AUROC scores across AI-proficient non-peer supporters and peer supporters is shown in Figure 2(D). The mean AUROC for peer supporters was 0.59 (95% CI: 0.52, 0.67), while the mean AUROC for AI-proficient non-peer supporters was 0.61 (95% CI: 0.56, 0.66).

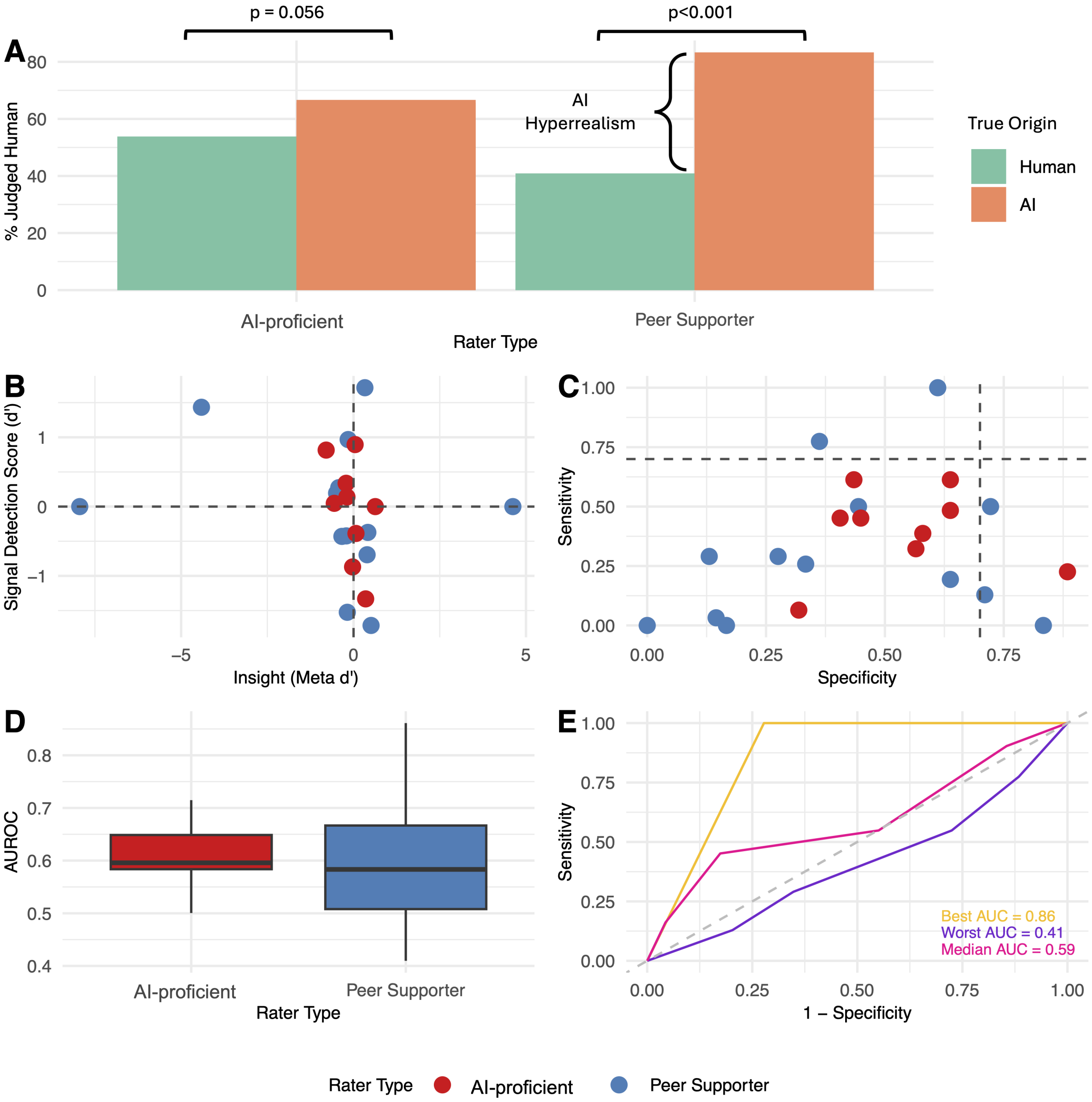

Fig. 3: Comparison of rater performance and insights. (A) Bar chart showing the percentage of ratings within each rater type judged as human, grouped by the true origin (AI vs Human generation). Low and high fidelity sentences are pooled. (B) Plot of rater insight, calculated as Meta d' (as described by[31]), and signal detection score (d') using the `psycho`[30] package. (C) Scatter plot of sensitivity versus specificity of a human rating for a true human label by rater. (D) Boxplot of AUROC scores by rater type. (E) AUROC for the best, worst, and median rater.

The top AUROC calculated (Peer Supporter), the bottom AUROC calculated (also Peer Supporter), and the rater with an AUROC closest to the median (AI-proficient non-peer supporters) are shown in Figure 2(E). These results highlight the variability in rater performance and suggest that, overall, raters struggled to consistently distinguish between human and AI-generated text.

4. Discussion

There is a clear need for targeted training programs to enhance peer supporters' ability to critically evaluate their performance during peer-support calls, allowing them to improve the quality of care they are producing.[9,10] This aligns with recommendations by Naslund et al.[41] philosophies on the importance of digital literacy in mental health support contexts. To support peer supporters in self-evaluating and improving their job performance, digital tools can play a crucial role, but such tools require access to large amounts of high-quality data.

Given the difficulty in obtaining sufficient real-world data, large language models (LLMs) offer a promising solution by generating synthetic data, as evidenced by the low detection accuracy (43.10%) in our study, where the AI-generated text closely mimicked human-created content.

The significant difference in accuracy between AI-proficient non-peer supporters and peer supporters, with the AI-proficient group demonstrating higher accuracy, is expected. However, the overall low accuracy for both groups underscores the challenges in reliably detecting AI-generated content, even for those with technical expertise.

This raises the possibility that exposure to AI in professional settings may confer some advantage in detecting synthetic content. However, the performance gap was small, which suggests that even AI-exposure may not be sufficient to reliably distinguish between human and AI-generated text in all cases. This brings into question how evaluators are selected for similar studies, as familiarity with AI might not always correlate with better performance in validation tasks. Future research should consider how varying levels of AI-exposure might impact evaluators' ability to assess synthetic text, and whether additional training or background knowledge is required for optimal evaluation.

The tendency of both peer supporters and AI-proficient non-peer supporters to classify low-fidelity sentences as AI-generated more often than high-fidelity sentences is particularly interesting. This suggests that the quality or adherence to best practices in peer support conversations might be a key factor in how text is perceived.

The promising results of this study, reflected in the low detection accuracy, suggest that synthetic text could be effectively integrated into training data for automatic feedback algorithms designed for peer supporters. However, it is essential to carefully consider the ethical implications and ensure that the human element, which is crucial to peer support, is maintained.[42]

4.1. *Limitations*

There are several limitations to this analysis. Firstly, the small sample size ($n = 22$) limits further generalizing our findings. However, the effects observed achieve statistical significance, which suggests that the findings are robust despite the sample size. Secondly, our analysis was based on the classification of individual sentences without additional context. While this serves our goal of creating technologies to highlight sentences of high- and low-fidelity, it is reasonable to expect that providing more context around each sentence might yield different results, as raters could potentially make more accurate judgments with more information.

Another limitation is the potential bias introduced by the specific demographic and professional backgrounds of our raters, which may not be representative of the broader population of peer supporters and AI-proficient individuals. Additionally, the inherent variability in individual raters' experiences and familiarity with AI-generated content could influence their performance and confidence levels.

Testing was conducted on a single model, which restricts our ability to generalize the findings across LLMs that may perform differently. Furthermore, we did not conduct a qualitative analysis of the synthetic data, which could have provided deeper insights into its linguistic quality, semantic accuracy, stylistic consistency, and realism. A more detailed assessment, such as annotation by professional peer supporters, could offer valuable perspectives on the text's quality and its alignment with human communication in similar contexts.

We did not evaluate the synthetic data for downstream tasks, leaving its practical application in real-world settings unexplored. This remains an important area for future work, and in our next follow-up study, we plan to investigate how synthetic data can be integrated into various downstream tasks, including its potential to enhance peer-support tools and other applications in similar domains.

Despite these limitations, our findings support the hypothesis that synthetic data generation for augmentation is feasible. The validation efforts indicate that both AI-proficient non-peer supporters and peer supporters struggle to reliably distinguish between human and AI-generated text, suggesting that AI-generated synthetic data can effectively mimic human-written content. This finding has promising implications for the use of synthetic data to augment training datasets and improve the performance of fidelity classification algorithms.

5. Conclusion

This study demonstrates the potential for using LLMs in synthetic text generation to create diverse datasets of peer support conversations, encompassing both high- and low-fidelity examples. Our findings reveal that both our test groups had below 50% in distinguishing synthetic text from human-created content, underscoring the sophisticated nature of current AI language models. Importantly, this work does not aim to replace human peer supporters with AI chatbots, but instead lays the groundwork for developing an automated feedback system to enhance peer support training and quality assurance. The synthetic sentences generated provide a rich dataset for training AI models to classify the quality of peer support interactions, potentially offering real-time feedback to supporters.

Future work might focus on developing and validating an AI-based feedback algorithm using our synthetic dataset, exploring its ethical implications, and investigating the long-term impacts of AI-assisted training on peer support outcomes. Ultimately, this study represents a significant step towards leveraging AI to enhance, rather than replace, human-delivered peer support, contributing to improved mental health support services.

Acknowledgements

We acknowledge Information, Technology, and Consulting (ITC) at Dartmouth College and Research Computing and Data Services (RC) team for providing the high-performance computational expertise and resources crucial to our research project's success. Their support and dedication significantly contributed to our progress and outcomes.

References

1. R. Cooper, K. Saunders, A. Greenburgh *et al.*, The effectiveness, implementation, and experiences of peer support approaches for mental health: a systematic umbrella review, *BMC Medicine* **22**, p. 72 (2024).
2. F. Schneider, M. Erhart, W. Hewer, L. AK Loeffler and F. Jacobi, Mortality and Medical Comorbidity in the Severely Mentally Ill, *Deutsches Ärzteblatt International* **116**, 405 (2019).
3. S. White, R. Foster, J. Marks *et al.*, The effectiveness of one-to-one peer support in mental health services: a systematic review and meta-analysis, *BMC Psychiatry* **20**, p. 534 (2020).
4. C. S. T. Yim, J. H. L. Chieng, X. R. Tang, J. X. Tan, V. K. F. Kwok and S. M. Tan, Umbrella review on peer support in mental disorders, *International Journal of Mental Health* **52**, 379 (2023).
5. K. Barnhouse, S. Clark and J. Waters Davis, Special population: Adults with severe and persistent mental health disorders, in *Chronic Illness Care*, eds. T. Daaleman and M. Helton (Springer, Cham, 2023)
6. D. Smit, C. Miguel, J. Vrijsen, B. Groeneweg, J. Spijker and P. Cuijpers, The effectiveness of peer support for individuals with mental illness: systematic review and meta-analysis, *Psychological Medicine* **53**, 5332 (2023).
7. C. Collins-Pisano, M. Johnson, G. Mois, J. Brooks, A. Myers, A. Muralidharan, M. Storm, M. Wright, N. Berger, A. Kasper *et al.*, Core competencies to promote consistency and standardization of best practices for digital peer support: focus group study, *JMIR Mental Health* **8**, p. e30221 (2021).
8. K. L. Fortuna, M. Venegas, E. Umucu, G. Mois, R. Walker and J. M. Brooks, The future of peer support in digital psychiatry: promise, progress, and opportunities, *Current treatment options in psychiatry* **6**, 221 (2019).
9. A. Kadakia, S. Preum, A. Bohm and K. Fortuna, Investigating the fidelity of digital peer support: A preliminary approach using natural language processing to scale high-fidelity digital peer support, in *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, (SCITEPRESS – Science and Technology Publications, Lda., 2023).
10. K. Fortuna, A. Wright, G. Mois *et al.*, Feasibility, acceptability, and potential utility of peer-supported ecological momentary assessment among people with serious mental illness: a pilot study, *Psychiatric Quarterly* **93**, 717 (2022).
11. Y. Roh, G. Heo and S. E. Whang, A survey on data collection for machine learning: a big data-ai integration perspective, *IEEE Transactions on Knowledge and Data Engineering* **33**, 1328 (2019).
12. Y. Cui, W. Che, T. Liu, B. Qin, S. Wang and G. Hu, Revisiting pre-trained models for chinese natural language processing, *arXiv preprint arXiv:2004.13922* (2020).
13. E. Watson, S. Fletcher-Watson and E. J. Kirkham, Views on sharing mental health data for research purposes: qualitative analysis of interviews with people with mental illness, *BMC Medical Ethics* **24**, p. 99 (2023).
14. F. Bernardi, D. Alves, N. Crepaldi, D. Yamada, V. Lima and R. Rijo, Data quality in health

research: Integrative literature review, *Journal of Medical Internet Research* **25**, p. e41446 (2023).

15. R. Syed, R. Eden, T. Makasi, I. Chukwudi, A. Mamudu, M. Kamalpour, D. Kapugama Geeganage, S. Sadeghianasl, S. Leemans, K. Goel, R. Andrews, M. Wynn, A. ter Hofstede and T. Myers, Digital health data quality issues: Systematic review, *Journal of Medical Internet Research* **25**, p. e42615 (2023).
16. A. R. Luca, T. F. Ursuleanu, L. Gheorghe, R. Grigorovici, S. Iancu, M. Hlusneac and A. Grigorovici, Impact of quality, type and volume of data used by deep learning models in the analysis of medical images, *Informatics in Medicine Unlocked* **31**, p. 100911 (2022).
17. Y. K. A. Baqraf, P. Keikhosrokiani and M. Al-Rawashdeh, Evaluating online health information quality using machine learning and deep learning: A systematic literature review, *Digital Health* **9** (2023).
18. A. Rosenbaum, S. Soltan, W. Hamza, A. Saffari, M. Damonte and I. Groves, Clasp: Few-shot cross-lingual data augmentation for semantic parsing, in *AACL-IJCNLP 2022*, 2022.
19. A. Rosenbaum, S. Soltan, W. Hamza, Y. Versley and M. Boese, Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging, in *COLING 2022*, 2022.
20. H. Zhao, H. Chen, T. A. Ruggles, Y. Feng, D. Singh and H.-J. Yoon, Improving text classification with large language model-based data augmentation, *Electronics* **13**, p. 2535 (2024).
21. M. Chinman, S. McCarthy, C. Mitchell-Miland, K. Daniels, A. Youk and M. Edelen, Early stages of development of a peer specialist fidelity measure, *Psychiatric Rehabilitation Journal* **39**, 256 (2016).
22. OpenAI, ChatGPT API `https://openai.com/blog/chatgpt`, (2022).
23. S. Gholizadeh and N. Zhou, Model explainability in deep learning based natural language processing (2021), arXiv preprint.
24. R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, (2021).
25. H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo and H. Yutani, *Welcome to the tidyverse*, (2019). R package version 1.3.1.
26. D. W. Zimmerman, A note on interpretation of the paired-samples t test, *Journal of Educational and Behavioral Statistics* **22**, 349 (1997).
27. Student, The probable error of a mean, *Biometrika* **6**, 1 (1908).
28. K. Pearson, Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London* **58**, 240 (1895).
29. H. Stanislaw and N. Todorov, Calculation of signal detection theory measures, *Behavior Research Methods, Instruments, & Computers* **31**, 137 (1999).
30. D. Makowski, *psycho: Procedures for Psychological, Psychometric, and Personality Research*, (2018). R package version 0.6.0.
31. E. J. Miller, B. A. Steward, Z. Witkower, C. A. M. Sutherland, E. G. Krumhuber and A. Dawel, AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones, *Psychological Science* **34**, 1390 (2023).
32. B. Maniscalco and H. Lau, A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings, *Consciousness and Cognition* (2012), journal homepage: www.elsevier.com/locate/concog.
33. D. Liu and J. Nocedal, On the limited memory bfgs method for large scale optimization, *Mathematical Programming* **45**, 503 (1989).
34. J. C. Nash and R. Varadhan, Unifying optimization algorithms to aid software system users: optimx for r, *Journal of Statistical Software* **43**, 1 (2011).

35. H. Wickham, R. François, L. Henry and K. Müller, *dplyr: A Grammar of Data Manipulation*, (2021). R package version 1.0.7.
36. L. Henry and H. Wickham, *purrr: Functional Programming Tools*, (2022). R package version 0.3.5.
37. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, (2016). R package version 3.3.5.
38. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez and M. Müller, *pROC: Display and Analyze ROC Curves*, (2022). R package version 1.18.0.
39. J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, *Radiology* **143**, 29 (1982).
40. E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* **44**, 837 (1988).
41. J. A. Naslund, P. P. Gonsalves, O. Gruebner, S. R. Pendse, S. L. Smith, A. Sharma and G. Raviola, Digital innovations for global mental health: opportunities for data science, task sharing, and early intervention, *Current treatment options in psychiatry* **6**, 337 (2019).
42. K. L. Fortuna, A. L. Myers, J. Ferron, A. Kadakia, C. Bianco, M. L. Bruce and S. J. Bartels, Assessing a digital peer support self-management intervention for adults with serious mental illness: feasibility, acceptability, and preliminary effectiveness, *Journal of Mental Health* **31**, 812 (2021).