PIFU: A novel framework to evaluate the interpretability of synthetic free-text explanations in digital mental health

Y.H.P.P Priyadarshana
Kyoto University of Advanced Science
Kyoto, Japan
2022md05@kuas.ac.jp

Ashala Senanayake
Kyoto University of Advanced Science
Kyoto, Japan
2024mm25@kuas.ac.jp

Zilu Liang
Kyoto University of Advanced Science
Kyoto, Japan
liang.zilu@kuas.ac.jp

ABSTRACT

In-context learning (ICL) in large language models (LLMs) improves performance across a wide range of tasks by utilizing a small number of in-context examples. A recent phenomenon has been reported in detecting depressive symptoms using social media postings, leveraging suitable examples from labeled depression corpora. Prompt-based explanations, on the other hand, has been recognized as an effective approach to better understand predictions derived from ICL. However, the interpretability of these synthetic explanations and their contribution to digital mental health (DMH) screening remain unclear. To address this, we propose *PIFU*, a novel theoretical framework for diverse stakeholders, including domain experts, to evaluate the interpretability of LLM-generated free-text explanations in DMH. We gather a cohort of clinicians (N=12), including certified psychologists and psychiatrists, to assess free-text explanations generated for depression and anxiety samples using a Likert scale rating system, followed by qualitative feedback in free-text form to develop human-interpretability metrics for plausibility, informativeness, and utility. We further reformulate faithfulness by focusing on predicting model decisions using explanations rather than gold labels. These human-centered metrics pave the way for promising advancements in interpretable LLM-based screening tools within the context of DMH.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence • **Human-centered computing** → Human computer interaction • **Applied computing** → Health informatics

KEYWORDS

Large language models, Free-text explanations, Human-interpretability, Human-centered AI, Digital mental health

1 Introduction

Few-shot prompting of large language models (LLMs), which involves learning from a few in-context examples in prompts, has led to significant improvements across various natural language processing (NLP) tasks, including classification, multi-step reasoning, and summarization. These in-context examples, also called demonstrations, cast downstream tasks together with task-specific prompts into a frozen LLM format to achieve state-of-the-art (SOTA) in-context learning (ICL) performance in both in-domain and previously unseen tasks [4, 6]. A few recent approaches have been introduced to detect depressive symptoms in social media text conversations by leveraging suitable demonstrations from labeled depression data [11], and to support mental health interventions through the use of conversational agents such as chatbots [32, 33]. However, this entire process remains poorly understood, as models are influenced by factors such as the number, order, and diversity of demonstrations and may not utilize instructions or labels in the expected manner [5].

Prompt-based explanations has been recognized as an effective approach to better understand predictions derived from ICL [2]. Free-text explanations, a prominent category, have received increasing attention by providing detailed reasoning behind an LLM's decisions, unlike extractive methods such as LIME and SHAP, which focus solely on input tokens [3]. Unlike *predict-then-explain* (P-E) approach, which generates the explanation after making a prediction, as seen in methods like SHAP, free-text explanations follow an *explain-then-predict* (E-P) process, where explanations are generated before the model makes its prediction. However, the interpretability of these synthetic explanations and how they contribute to ICL downstream performance, including depression screening, remains unclear.

Interpretability refers to how well the inner workings of a black-box model can be presented in understandable terms to a human [7]. Prior research has identified four key measures for evaluating the interpretability of explanations: *plausibility*, *informativeness*, *utility*, and *faithfulness*. Plausibility measures how convincing the explanation is to humans [1], while informativeness assesses how much new information is provided by the explanation to justify the prediction [8]. Utility measures the usefulness of the explanation for the target audience [9]. Faithfulness, on the other hand, explains the reasoning process behind the model's prediction, meaning that human judgment should not be involved in evaluating faithfulness [10]. Although evaluating the interpretability of explanations is an active area of research, its adoption in

domains such as digital mental health (DMH) has not been thoroughly explored.

In this study, we propose a novel framework to evaluate the interpretability of LLM-generated free-text explanations in DMH. Our objective is to employ a cohort of experts, including certified psychologists and psychiatrists, to evaluate synthetic free-text explanations generated for depression and anxiety samples on a Likert-scale (1-5) for consistency, reliability, and professionalism, while also prompting them to provide qualitative feedback in a free-text format. The rating values on consistency, reliability, and professionalism and free-text feedback are utilized to develop interpretability metrics for plausibility, informativeness, and utility. To the best of our knowledge, we are the first to construct a human interpretability framework for evaluating LLM-generated explanations in the context of DMH. We further reformulate faithfulness by applying the criteria of predictive power evaluation (PPE) [1, 10], focusing on predicting model decisions using explanations instead of gold labels.

The overall organization of the manuscript is as follows: Section 2 critically reviews related work. Section 3 elaborates the proposed framework and then discussed in Section 4.

2 Related work

Multiple attempts have recently been made to enhance the interpretability of E-P-based synthetic explanation generation utilizing automatic and human evaluation techniques. Herman (2017) formulated functional interpretability correlated with cognitive functions and user preferences as functional metrics to measure plausibility of explanations [12]. Wiegreffe et al. [3] empirically demonstrated that LLMs, including GPT-3, can generate plausible explanations using few-shot human-written explanations and high-quality prompts for question answering (QA) and natural language inference (NLI) tasks. However, a few other studies have shown that the plausibility of synthetic explanations was not acceptable with the absence of gold examples since the existing plausibility metrics were conflated with faithfulness [13, 14].

In a related vein, faithfulness in synthetic explanations has been both theoretically and empirically evaluated, considering the fact that plausibility does not entail faithfulness. Jacovi and Goldberg were the first to establish a criterion for evaluating faithfulness of post-hoc explanations [1]. These assumptions have subsequently been used to formulate four types of faithfulness evaluation methods, including PPE, which incorporates generated explanations for model predictions [10]. Chen et al. introduced precision- and generality-based criteria to evaluate the counterfactual simulatability of post-hoc explanations using PPE [15]. On a different line of work, Lanham et al. demonstrated that corrupted, unfaithful explanations would lead an LLM to a different prediction [16]. Recently, MentaLLaMA was introduced for interpretable mental health analysis, focusing on the quality and consistency of generated content, but lacking evidence for the human-interpretability criteria of plausibility and faithfulness [17].

The impact of informativeness and utility in evaluating the interpretability of free-text explanations has been minimally explored. Chen et al. assessed the amount of new information in generated explanations using automatic metrics such as rationale quality for QA tasks [8] while Sun et al. measured the additional knowledge provided by LLM-generated rationales through a human study [18]. Since a proxy model was used to evaluate explanations in the aforementioned approaches, an explanation could potentially reveal the corresponding labels, making it uninformative. The label leakage was mitigated in a later approach, RORA by Jiang et al. [19], but human reliability and consistency were not assessed. Joshi et al., on the other hand, assessed the human utility of free-text explanations for QA tasks using accuracy [20]. However, they concluded that a reliable metric is necessary to estimate the human utility, considering the shortcomings of automatic metrics, including accuracy.

To this end, we propose a theoretical framework, *PIFU*, to evaluate the human-interpretability of free-text explanations reformulating metrics, such as Plausibility, Informativeness, Faithfulness, and Utility. The proposed framework can be used to demonstrate how synthetic explanations contribute to ICL downstream performance, including depression detection.

3 Interpretability framework

3.1 Study overview

Our aim is to detect depression and anxiety of user postings P extracted from Reddit and Twitter incorporating task-specific demonstrations $D = \{d_1, d_2, \ldots, d_n\}$ and their explanations E generated by multiple LLMs. Due to its effectiveness in retrieving D from unseen datasets across various ICL tasks, we choose unified demonstration retriever (UDR) [21] encoder as our demonstration retrieval method. UDR selects the most relevant samples from the Reddit Self-reported Depression Diagnosis (RSDD) [22] and Self-Reported Mental Health Diagnoses (SMHD) [23] corpora to serve as task-specific D for depression and anxiety detection. The P and their D are used to generate free-text E. The E, D, and P are further processed to form downstream tasks. Subsequently, the generated E are assessed by clinicians for interpretability.

3.2 *E* generation and depression/anxiety classification

Multiple LLMs, including Mistral-7B-Instruct and Gemma-7B, are utilized to generate E, which are then ranked using E ranking method introduced by Ye et al. [25] to select the top 2 E. These LLMs are known for their reliable E generation in decision support systems, including emotion analysis [24]. The following prompt is crafted to emphasize the semantic

aspects of *D* and *P*, directing the LLM to focus on identifying and expressing relevant content when generating *E*.

 # Prompt: Below are in-context examples with depressive/anxiety elements and their detailed explanations:
 # Demonstrations: [examples]
 # Prompt: Example: [examples] Explanation: This example shows signs of depression because…
 # Input: [postings]
 # Prompt: Now, analyze the [postings] for depressive/anxiety elements and provide a detailed explanation

ProDepDet, the prompt framework for LLM transferability to previously unseen tasks, is used for depression and anxiety classification [11]. The prompt manager logic of ProDepDet integrates *P*, *D*, and top-ranked *E*, incorporating task-specific prompt templates and verbalizers while keeping the LLM frozen. Multiple downstream tasks, such as depressed post classification, are formulated for depression and anxiety classification purposes using 100M-300M-parameter LLMs, including MentalBERT and DisorBERT, which are specifically designed for detecting mental disorders [27, 28]. Twitter depression 2022 [26] corpus and SMHD are used to construct task-specific verbalizers, as well as to evaluate downstream tasks.

3.3 Clinician evaluation protocol

We design the clinician evaluation protocol to measure the interpretability of the generated *E* and answer the question: *How can the interpretability of human distress detection in social media text-based postings be further improved?* A cohort of clinical experts, including psychologists and psychiatrists (*N*=12) over the age of 30, is recruited based on their qualifications and experience in the field of mental health. Informed consent is obtained from each subject, ensuring they understand their rights and the nature of the experiment. The study protocol is submitted to the IRB at the host institution for the study. We follow all IRB procedures to avoid potential conflict of interest.

Each subject is sent out an online form, shown in Appendix A, detailing the instructions to be followed for evaluating a set of 200 free-text explanations generated in Section 3.2. This includes an overview of the context surrounding the depression- and anxiety-related examples. For each example, there are two generated explanations. The evaluation protocol consists of a rating assessment followed by feedback evaluation. The rating assessment criteria (1–5, with 1 being the minimum and 5 being the maximum)—consistency, reliability, and professionalism, shown in Appendix B, are based on similar human evaluation methods used for mental health-related tasks [17].

- **Consistency (α)**: the LLM-generated explanations should provide clues and analyses that align with their associated depression and anxiety examples.
- **Reliability (β)**: the trustworthiness of the evidence to support the classification results (depressed, anxiety, or normal) in the generated explanations.
- **Professionalism (γ)**: the rationality of the evidence in generated explanations from the perspective of psychology.

Following the rating process, the subjects are prompted to provide qualitative feedback δ in a free-text format of about 100 words, highlighting the strengths of the synthetic explanations and suggesting areas for improvement. The 100-word limitation is determined based on the constraints of the model refinement process, which is intended to enhance the *E* generation performance. A sample of generated explanations is shown in Appendix C.

3.4 Clinician evaluation metrics

The rating values and free-text expert feedback are used to develop the interpretability metrics. Inspired by cognitive functional metrics and user preferences [10], we incorporate the expert ratings and free-text feedback as input to develop interpretability metrics for plausibility, informativeness, and utility.

Plausibility incorporates reliability β and professionalism γ ratings to assesses how convincing the explanation is to clinicians. The evidence supporting the downstream classification results in the generated *E*, along with the rationality of those results from the perspective of experts, collectively measures the degree of plausibility using β and γ. The plausibility of explanations for a given distress example $\mathcal{I}_p$ can be presented as

$$\mathcal{I}_p = \frac{1}{C}\left(\frac{\sum_{i=1}^{C}\beta \oplus \sum_{i=1}^{C}\gamma}{\sum_{i=1}^{C}(\beta.\gamma)}\right), \quad (1)$$

where *C* denotes the number of *E* per example. Here, $\mathcal{I}_p \in [0,1]$, indicating any value within the range of 0 to 1, inclusive.

Informativeness, on the other hand, comprises consistency α and reliability β human ratings to measure how much new information is provided by the number of explanations, *C*, to justify the downstream prediction. α assesses the clues provided by *E*, while β evaluates the trustworthiness of supporting the model classification results, making them suitable for determining the degree of informativeness. The informativeness of *E* for a given distress example $\mathcal{I}_i$ can be presented as:

$$\mathcal{I}_i = \frac{1}{C}\left(\frac{\sum_{i=1}^{C}\alpha \oplus \sum_{i=1}^{C}\beta}{\sum_{i=1}^{C}(\alpha.\beta)}\right) \quad (2)$$

Here, $\mathcal{I}_i \in [0,1]$, indicating any value within the range of 0 to 1, inclusive. Incorporating reliability as a measure of informativeness allows us to bridge the gap identified in the prior work by Jiang et al. [19].

While plausibility and informativeness offer valuable interpretability insights for expert clinicians, their benefits are less significant without human reasoning. **Utility**, therefore, incorporates consistency α and reliability β ratings along with free-text expert feedback δ to measure the usefulness of E for the target audience. To quantify the feedback, a similarity model, such as SBERT [29], is used to measure the similarity between the contextual representations of E and encodings derived from expert feedback. According to Joshi et al., the similarity between synthetic rationales and the corresponding expert feedback is an effective indicator of human utility [20]. The utility $\mathcal{I}_u$ can be presented as

$$\mathcal{I}_u = \frac{1}{C}\sum_{i=1}^{C} sim(\phi_\varepsilon, \phi_\delta).\left(\frac{\sum_{i=1}^{C}\alpha \oplus \sum_{i=1}^{C}\beta}{\sum_{i=1}^{C}(\alpha.\beta)}\right), \quad (3)$$

where ϕ_ε and ϕ_δ denote contextual representations of E and encodings derived from the expert feedback for the i-th explanation. Here, $\mathcal{I}_u \in [0,1]$, indicating any value within the range of 0 to 1, inclusive. In contrast to automatic metrics, such as accuracy, used to evaluate the usefulness of explanations, $\mathcal{I}_u$ provides a more interpretable metric for measuring human utility.

Faithfulness indicates how accurately an E represents the reasoning process of a model. E that lack faithfulness can be risky, as they may appear plausible and convincing to humans, leading users to over-trust the model despite its potential biases. To reformulate the faithfulness of free-text E, we use PPE, a widely used methodology for assessing the faithfulness of post-hoc E [10]. We use PPE under the assumption that an E is unfaithful if it leads to a different prediction than the one made by the model it is meant to explain. Drawing inspiration from the approach of Ye et al. [30], we adopt the feature importance score as the PPE technique for predicting model decisions. For the downstream prediction tasks performed by LLMs, such as MentalBERT in Section 3.2, feature importance scores are obtained using postings and their demonstrations as input. These scores are derived by integrating the gradients of the model's output with respect to input features, both with (w/) and without (w/o) E. The Euclidean Distance $\mathcal{D}$ is used to quantify the semantic distance between the feature importance scores.

$$\mathcal{D}(\phi_x^{w/\,E}, \phi_x^{w/o\,E}) = \sqrt{\sum_{i=1}^{n}\left(\phi_{x,i}^{w/\,E} - \phi_{x,i}^{w/o\,E}\right)^2}, \quad (4)$$

where n, x, ϕ_x, and $\phi_{x,i}$ denote number of input features, the input feature vector, the feature importance vector, and the importance score for the i-th feature in the vector, respectively.

If $\mathcal{D}$ is larger than a threshold $\mathcal{T}$ determined empirically, it indicates that E alters model's decision-making process, suggesting that E may not faithfully represent the original model's reasoning. The faithfulness $\mathcal{I}_f$ can be presented as:

$$\mathcal{I}_f = \begin{cases} Faithful \leftarrow \mathcal{D}(\phi_x^{w/\,E}, \phi_x^{w/o\,E}) \le \mathcal{T} \\ Unfaithful \leftarrow \mathcal{D}(\phi_x^{w/\,E}, \phi_x^{w/o\,E}) > \mathcal{T} \end{cases} \quad (5)$$

4 Discussion

The proposed framework establishes more sophisticated interpretability metrics to assess LLM-generated free-text explanations, potentially addressing limitations of automatic evaluation metrics used for ICL-based DMH tasks. Selecting expert reliability and professionalism is effective for measuring the plausibility of explanations, as reliability provides evidence to support the model's outcome, while professionalism evaluates the rationality of the evidence in the generated explanations from a clinical perspective. In this context, the target audience consists of expert clinicians, where the combination of reliability and professionalism determines the plausibility of synthetic explanations that are interpretable to them. An explanation may appear entirely plausible if it fully aligns with human reasoning, yet it could be completely unfaithful if it does not reflect the model's actual decision-making process. Hence, plausibility does not ensure faithfulness, and vice versa. Since we only consider expert ratings for the plausibility metric, we can confidently state that our proposed plausibility metric is not conflated with the metric of faithfulness.

Conversely, faithfulness has been used to evaluate post-hoc explanations, but it has yet to be adapted for assessing free-text explanations. Although our framework assesses the human interpretability of LLM-generated output, it remains incomplete without a metric to evaluate model faithfulness. Given the shortcomings of other LLM faithfulness evaluation methods, such as axiomatic-based evaluation, and robustness evaluation, PPE is used in this study due to its sensitivity to faithfulness [10]. We reformulate the metric of faithfulness by incorporating feature importance scores as the PPE technique for predicting model decisions. Feature importance scores, derived from methods like integrated gradients, aim to highlight the contribution of input features to the model's output. This directly reflects the model's internal decision-making process, making it a good proxy for evaluating faithfulness. These scores are obtained solely by considering the model's output, without incorporating expert ratings for free-text explanations, to distinguish model reasoning from human reasoning. This allows us to evaluate the model's predictions with and without explanations while preserving the fundamentals, including the fact that faithfulness should exclude human judgment regarding the quality of the explanations. Moreover, faithfulness evaluation ensures that explanations accurately reflect the model's reasoning, helping

stakeholders avoid blindly trusting outputs that might be based on misleading or irrelevant factors.

Evaluating informativeness and utility in the context of DMH remains unexplored. Incorporating expert ratings on consistency and reliability to evaluate informativeness fulfill the need for an interpretability metric that measures how much new information the explanation provides to justify a model's prediction, surpassing automatic metrics such as rationale quality. Although rationale quality provides an average score for the generated explanations, it lacks detailed insights into the important clues offered by the explanations and how they align with associated depression or anxiety examples. More importantly, an automatic metric, such as an average score, does not adequately address the interpretability of LLM-generated content in the context of DMH. Utility, on the other hand, targets a wide range of the stakeholders, providing valuable interpretability insights. We incorporate human feedback alongside ratings, as expert input can serve as a rubric for evaluating generated explanations, with clinicians as the primary target audience. The semantic similarity between synthetic rationales and the corresponding free-text feedback, along with consistency and reliability ratings, demonstrates the usefulness of generated content in DMH. This includes the provided clues and trustworthiness of evidence to support for downstream predictions, helping audience understand the utility of the generated content. The proposed interpretability metric for human utility addresses the limitations of existing automatic metrics, such as accuracy, to assess the usefulness of LLM-generated content for the target audience. Therefore, the question – *How can the interpretability of human distress detection in social media text-based postings be further improved?* – can be considered answered.

The proposed design of *PIFU* outlines a robust framework that goes beyond the criteria of plausibility, informativeness, and utility by incorporating additional factors such as accountability and credibility [31]. While one could argue that experts can directly assess plausibility, informativeness, and utility, this generalization broadens the scope of evaluation. To validate the proposed framework, we will adopt key strategies, including expert agreement validation, thematic validation for feedback analysis, and ground truth comparison. Expert agreement validation measures the inter-rater reliability among the experts to ensure consistency in human ratings. Metrics such as Fleiss' Kappa or Cohen's Kappa will be employed to quantify agreement. Clinician-provided free-text feedback will undergo thematic analysis to identify recurring themes and key insights about generated explanations. Two independent researchers will analyze the feedback to ensure reliability and resolve discrepancies through discussion. In ground truth comparison, a benchmark dataset with explanations pre-labeled by expert consensus will serve as the ground truth. The IMHI corpus is one such benchmark for DMH tasks [17]. The framework's outputs will be compared against this dataset to validate its ability to differentiate between high- and low-quality explanations. This multi-faceted validation ensures rigor and alignment with clinical expectations.

Our findings are currently limited to evaluate LLM-generated free-text explanations. Improving these metrics for evaluating synthetic explanations generated by other types, such as structured explanations, is a potential direction. Although we critically consider the potential overlap between plausibility and faithfulness, we did not examine such relationships between faithfulness and informativeness or faithfulness and utility. The expert natural language feedback is restricted to a 100-word limit; therefore, the semantic similarity used for the utility metric is constrained by the context length. Furthermore, we limited ourselves to using SBERT for quantifying the expert feedback, and more constructive methods, such as thematic analysis, should be incorporated to effectively extract key insights from the free-text clinician feedback. While task-specific instructions are essential for certain ICL few-shot reasoning tasks, the present study does not incorporate them alongside the in-context examples and their corresponding explanations. The present study is limited to 7B-parameter LLMs for generating explanations. Further evaluations should include larger models, such as LLaMA-3-70B, to assess interpretability at an increased scale.

Although the proposed framework marks a significant step forward in evaluating human interpretability of synthetic content, several challenges need to be addressed. Different clinicians might interpret consistency, reliability, and professionalism differently, potentially impacting the robustness of the interpretability metrics. Restricting expert feedback to 100 words could limit the depth of insights provided, especially when evaluating complex explanations or diverse DMH contexts. The interpretability metrics may perform well for specific tasks like depression or anxiety detection but may need extensive re-tuning to generalize to other DMH tasks. Faithfulness evaluations based on PPE require threshold values for metrics like semantic distance. Determining and validating these thresholds across different DMH tasks can be challenging and might affect the consistency of results. Evaluating utility for a diverse range of stakeholders, including both expert clinicians and laypersons, may require distinct criteria, leading to complexity in defining a single utility metric that satisfies both audiences.

5 Conclusion

In this study we propose *PIFU*, a novel theoretical human-in-the-loop framework designed to evaluate interpretability of synthetic explanations within the context of DMH. First, we utilize SOTA LLMs to generate free-text explanations for the retrieved ICL examples sourced from RSDD and SMHD, Reddit-based corpora used for DMH tasks, including

depression and anxiety detection. A prompt-based framework is then employed for downstream depression and anxiety classification tasks, incorporating task-specific prompt templates and verbalizers, along with input user postings, ICL examples, and their explanations. Second, a cohort of clinical experts is recruited to rate the LLM-generated explanations on a Likert-scale (1-5) based on consistency, reliability, and professionalism. This is followed by expert-written natural language feedback to develop interpretability metrics for plausibility, informativeness, and utility. We further refine the metric of faithfulness by focusing on predicting model decisions using explanations rather than gold labels.

In the next phase of the study, we will utilize the proposed framework, along with appropriate automatic metrics, to conduct experiments on multiple downstream DHM screening tasks, including depression and anxiety classification. Enhancing the human-in-the-loop aspect by incorporating real-time clinician feedback into the model refinement process could lead to more effective and contextually relevant explanations. By systematically applying thematic insights to modify the prompt design, the model explanation generation process becomes more aligned with expert expectations. Future work could expand the interpretability metrics beyond plausibility, informativeness, faithfulness, and utility, incorporating additional dimensions like fairness, accountability, or actionability, particularly in high-stakes fields like DMH. Ethical considerations in data privacy, anonymization, and consent need to be deeply integrated into the framework. Given the global nature of mental health issues, future work could evaluate how the framework performs across different languages and cultural contexts, ensuring that the interpretability metrics are robust and applicable worldwide.

REFERENCES

[1] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. DOI: 10.18653/v1/2020.acl-main.386.

[2] Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, et al. 2022. Can language models learn from explanations in context?. In *Findings of the Association for Computational Linguistics:*, pages 537–563. DOI: 10.18653/v1/2022.findings-emnlp.38.

[3] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658. Doi: 10.18653/v1/2022.naacl-main.47.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

[5] Xi Ye and Greg Durrett. 2024. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Article 2202, 30378–30392.

[6] Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, et al. 2022. Knowledge Inheritance for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3921–3937, Seattle, United States. Association for Computational Linguistics. DOI: 10.18653/v1/2022.naacl-main.288.

[7] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57. DOI: 10.1145/3236386.3241340.

[8] Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: Information-Theoretic Evaluation of Free-Text Rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030. DOI: 10.18653/v1/2023.acl-long.112.

[9] Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, et al. 2023. Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-text Rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128. DOI: 10.18653/v1/2023.acl-long.392.

[10] Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, 50(2):657–723. DOI: 10.1162/coli_a_00511.

[11] Yapa Priyadarshana, Zilu Liang, and Ian Piumarta. 2024. ProDepDet: Out-of-domain Knowledge Transfer of Pre-trained Large Language Models for Depression Detection in Text-Based Multi-Party Conversations. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. pages. 1-8. doi: 10.1109/IJCNN60899.2024.10650774.

[12] Herman, Bernease. 2017. The promise and peril of human evaluation for model interpretability. *ArXiv preprint*, abs/1711.07414.

[13] Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness. In *Proceedings. of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086. DOI: 10.18653/v1/2023.emnlp-main.622.

[14] Hangfeng He, Hongming Zhang, and Dan Roth. 2024. SocREval: Large Language Models with the Socratic Method for Reference-free Reasoning Evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2736–2764. DOI: 10.18653/v1/2024.findings-naacl.175.

[15] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, et al. 2025. Do models explain themselves? counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Vol. 235. JMLR.org, Article 310, 7880–7904.

[16] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *ArXiv preprint*, abs/2307.13702.

[17] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. In *Proceedings of the ACM Web Conference 2024* 4489–4500. DOI: 10.1145/3589334.3648137.

[18] Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the Benefits of Free-Form Rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5867–5882. DOI: 10.18653/v1/2022.findings-emnlp.432.

[19] Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024. RORA: Robust Free-Text Rationale Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1070–1087. DOI: 10.18653/v1/2024.acl-long.60.

[20] Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, et al. 2023. Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-text Rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128. DOI: 10.18653/v1/2023.acl-long.392.

[21] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, et al. 2023. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668. DOI: 10.18653/v1/2023.acl-long.256.

[22] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978. DOI: 10.18653/v1/D17-1322.

[23] Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497.

[24] Marco Siino. 2024. TransMistral at SemEval-2024 Task 10: Using Mistral 7B for Emotion Discovery and Reasoning its Flip in Conversation. In

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 298–304. DOI: 10.18653/v1/2024.semeval-1.46.

[25] Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary Explanations for Effective In-Context Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484. DOI: 10.18653/v1/2023.findings-acl.273.

[26] Junyeop Cha, Seoyun Kim, and Eunil Park. 2022. A lexicon-based approach to examine depression detection in social media: The case of Twitter and university community. *Humanities and Social Sciences Communications* 9, 1 (2022), pages 1–10. DOI: 10.1057/s41599-022-01313-2.

[27] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190.

[28] Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318. DOI: 10.18653/v1/2023.acl-long.853.

[29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. DOI: 10.18653/v1/D19-1410.

[30] Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting Attributions and QA Model Behavior on Realistic Counterfactuals. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5496–5512. DOI: 10.18653/v1/2021.emnlp-main.447.

[31] Jiaying Liu, Yan Zhang, and Yeolib Kim. 2023. Consumer Health Information Quality, Credibility, and Trust: An Analysis of Definitions, Measures, and Conceptual Dimensions. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR '23)*, pages 197–210. DOI: 10.1145/3576840.3578331.

[32] van der Schyff EL, Ridout B, Amon KL, Forsyth R, Campbell AJ. 2023. Providing Self-Led Mental Health Support Through an Artificial Intelligence-Powered Chat Bot (Leora) to Meet the Demand of Mental Health Care. *J Med Internet Res*. Jun 19;25:e46448. DOI: 10.2196/46448.

[33] Qi Yue. 2025. Pilot quasi-experimental research on the effectiveness of the Woebot AI Chatbot for reducing mild depression symptoms among athletes. *International Journal of Human-Computer Interaction*, 41(1), pages 452–459. DOI: 10.1080/10447318.2023.2301256.

Appendix A – Instructions for clinicians

Enhancing the interpretability of large language models (LLMs) for depression and anxiety detection.

Human Rating Evaluation

In the Google sheet provided to you, the first two columns display the depression/anxiety example and it's LLM-generated explanation. Please read both the examples and explanations. For each example, there are two generated explanations. You are subjected to review a set of 200 LLM-generated explanations related to 100 depression/anxiety examples. Each explanation is rated on a scale of 1 to 5, where 1 indicates the lowest and 5 the highest, for consistency, reliability, and professionalism.

- **Consistency**: the LLM-generated explanations should provide clues and analyses that align with their associated depression/anxiety examples.
- **Reliability**: the trustworthiness of the evidence to support the classification results (depressed, anxiety or normal) in the generated explanations.
- **Professionalism**: the rationality of the evidence in generated explanations from the perspective of psychology.

You may indicate the rating for each of the 3 criteria under the respective columns.

Human Feedback Evaluation

Following the rating process, you are prompted to provide qualitative feedback in a free-text format. You are instructed to write approximately 100 words, focusing on your insights regarding the explanations' strengths and areas for improvement. Generally speaking, the feedback should be a critique of what is good and bad about an explanation and how an explanation can be improved. It should be as precise as possible and mention the concrete ways in which the explanation can be improved.

Here is an example of how feedback could look: *"The explanation is good/bad/lacks detail because..... Concretely, the explanation should elaborate..... The explanation should also convey that....."*

You may access the Google sheet here.

Appendix B – Rating assessment criteria

Consistency

1 (Very Inconsistent): The explanation is completely disconnected from the associated depression or anxiety example. It provides conflicting clues and analyses that do not align with the model's output.
2 (Inconsistent): The explanation is mostly inconsistent with the depression or anxiety example. It contains several clues or analyses that don't align well with the example, causing confusion.
3 (Moderately Consistent): The explanation is somewhat consistent with the depression or anxiety example. Some clues and analyses align with the example, but there are noticeable discrepancies that may hinder understanding.
4 (Consistent): The explanation is mostly consistent with the depression or anxiety example, and most clues and analyses align well with the example, providing a clear understanding of the reasoning process.
5 (Highly Consistent): The explanation is entirely consistent with the depression or anxiety example. All clues and analyses align perfectly with the example, providing a coherent and well-supported reasoning process.

Reliability

1 (Very Unreliable): The evidence presented in the explanation is completely untrustworthy and does not support the classification results (depressed, anxiety, or normal). The reasoning is flawed or misleading.
2 (Unreliable): The explanation provides some evidence, but it is not fully trustworthy. The evidence may be weak, inconsistent, or irrelevant, making the classification results questionable.
3 (Moderately Reliable): The evidence in the explanation is somewhat reliable, but it may lack strong support for the classification results. It generally aligns with the model's decision, but with some areas of uncertainty.
4 (Reliable): The explanation provides trustworthy evidence that strongly supports the classification results. The reasoning is clear, and the evidence aligns well with the model's output.
5 (Highly Reliable): The evidence in the explanation is highly reliable, providing strong support for the classification results. The reasoning is clear, trustworthy, and fully aligned with the model's decision-making.

Professionalism

1 (Very Unprofessional): The explanation is highly unprofessional, using inappropriate or irrational evidence that would not be accepted in a psychological or clinical context. It lacks critical reasoning.
2 (Unprofessional): The explanation is unprofessional in some aspects. It may use language or reasoning that is not appropriate from a psychological perspective or contains flaws that reduce its credibility.
3 (Moderately Professional): The explanation is mostly professional but may include some slight lapses in rationality or tone that could benefit from refinement. It is generally appropriate from a psychological perspective but not perfect.
4 (Professional): The explanation is professional, using rational and reasonable evidence that aligns with psychological standards. It maintains a tone and style appropriate for clinical or academic settings.
5 (Highly Professional): The explanation is highly professional, demonstrating clear, rational, and evidence-based reasoning. It aligns perfectly with psychological standards and is appropriate for clinical settings.

Appendix C – Sample of explanations to be used for evaluations

Post_Id	Example	Explanation	Consistency (1 to 5)	Reliability (1 to 5)	Professionalism (1 to 5)	Feedback (100 words max)
1.26E+18	my biggest problem be overthinking everything	Overthinking is often associated with anxiety disorders, but it can also be a symptom of major depressive disorder (MDD). Persistent overthinking can lead to rumination, which is a persistent focus on a negative thought, feeling, or experience. Rumination can result in increased sadness, hopelessness, and feelings of helplessness.				
		The repeated emphasis on "overthinking everything" is indicative of rumination, which is a common symptom of Major Depressive Disorder (MDD) as per the DSM-5. Rumination includes dwelling on negative emotions and thoughts.				
1.26E+18	the worst sadness be the sadness you have teach yourself to hide	This statement reflects the person's inability to express or acknowledge their emotions, which is a common symptom of depression. Hidden sadness can contribute to feelings of isolation and alienation, worsening the overall mood and increasing the likelihood of depression.				
		Statement 3 implies the presence of "depressed mood most of the day, nearly every day" (Criterion A). Additionally, the need to hide sadness highlights the possibility of feelings of worthlessness (Criterion A).				