Rethinking Large Language Models in Mental Health Applications

Shaoxiong Ji † , Tianlin Zhang *, Kailai Yang *, Sophia Ananiadou *, and Erik Cambria $^{\alpha}$

[†]University of Helsinki, Finland

*The University of Manchester, UK

^a Nanyang Technological University, Singapore

Email: shaoxiong.ji@helsinki.fi; {kailai.yang,tianlin.zhang}@postgrad.manchester.ac.uk;,

sophia.ananiadou@manchester.ac.uk, cambria@ntu.edu.sg

Abstract

Large Language Models (LLMs) have become valuable assets in mental health, showing promise in both classification tasks and counseling applications. This paper offers a perspective on using LLMs in mental health applications. It discusses the instability of generative models for prediction and the potential for generating hallucinatory outputs, underscoring the need for ongoing audits and evaluations to maintain their reliability and dependability. The paper also distinguishes between the often interchangeable terms "explainability" and "interpretability", advocating for developing inherently interpretable methods instead of relying on potentially hallucinated self-explanations generated by LLMs. Despite the advancements in LLMs, human counselors' empathetic understanding, nuanced interpretation, and contextual awareness remain irreplaceable in the sensitive and complex realm of mental health counseling. The use of LLMs should be approached with a judicious and considerate mindset, viewing them as tools that complement human expertise rather than seeking to replace it.

1 Introduction

Mental health is important and has been studied by natural language processing (NLP) using text (e.g., social posts and doctor-patient conversations) as the data sources, leading to the development of automatic methods for various applications, including early detection of mental disorders [49] and mental health counseling [1]. Researchers have employed techniques, ranging from traditional feature engineering to automatic feature learning, such as convolutional neural networks, recurrent neural networks, and transformer networks, for mental illness detection and classification [49]. Recent advances utilize pretrained language models (PLMs). PLMs trained with the masked language modeling objective have become popular for training classification models in this domain. Domainspecific continual pre-training has also undergone intensive development to acquire domain knowledge with representative discriminative models including PsychBERT [41], MentalBERT [18], PHS-BERT [31], and MentalLongformer [19]. A recent shift as in Figure 1 has occurred towards prompt learning, where generative large language models (LLMs) such as SmileChat [33], Psy-LLM [22], Mental-LLM [46], MentalLLaMA [48], ChatCounselor [26], and MindWatch [4], are used to generate predictions or counseling based on input prompts related to mental health conditions. This shift signifies a growing interest in leveraging generative LLMs and prompt learning for mental health-related tasks. However, one question looms large: is this a mere hype? This paper delves into the recent developments and concerns surrounding the use of LLMs for early prediction of mental health conditions, generating explanations for mental health conditions, and generating responses in mental health counseling.

The landscape of large language models has undergone substantial transformation in recent years. Current LLMs boast hundreds of billions of parameters, a stark contrast to the relatively modest sizes seen in the early 2010s, typically ranging from millions to tens of millions of parameters. Notably, models such as BERT, with 110 million parameters, and GPT-2, with 1.5 billion

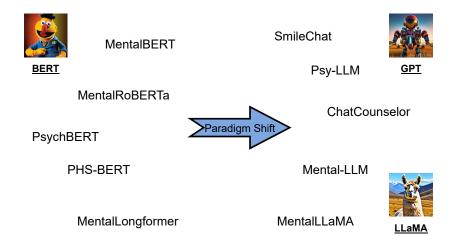


Figure 1: A paradigm shift in NLP for mental health applications from masked language models such as BERT to generative language models such as GPT and LLaMA. Images of BERT, GPT, LLaMA are generated by Midjourney AI Art Generator.

parameters, which were once considered large, now fall into the category of medium-sized language models by standards at the time of writing. It is important to note that the size of language models is not the only factor determining their performance. Other factors, such as model architecture, training data, and fine-tuning, also play significant roles in their capabilities. This growth in model size reflects the ongoing evolution of AI language models. This paper focuses on the recent use of generative LLMs in mental health applications. For the purpose of this paper, the term "LLMs" refers to generative models trained with the causal language modeling objective, often called next-word prediction in a simpler term.

Our paper offers perspectives on rethinking large language models in the context of mental healthcare. When using generation to predict mental health conditions based on a prompt and post, it is worth noting that the generation-as-prediction process can exhibit instability and unpredictability, even with minor changes to the input prompt. We discuss empirical results related to this instability and explore theoretical studies on meta-optimization that may underlie this unpredictability. Consequently, we advocate carefully auditing generative LLMs when they are used to predict mental health conditions.

When employing LLMs for mental health prediction, a significant concern revolves around the interpretability of the generated output or the so-called explanations. LLMs often operate as blackbox neural networks, making it challenging to discern how they arrive at their conclusions. Therefore, it is essential to emphasize that claims of interpretable mental health analysis should not be taken at face value but substantiated with rigorous proof and verification. One fundamental concern when using LLMs for mental health prediction is that the generated explanations may not necessarily imply true interpretability.

In the context of early prediction of mental disorders, providing explanations for mental health conditions, and counseling for mental health-related queries, LLMs have the potential to produce incorrect information akin to hallucinations. This risk underscores the necessity for further research to assess the reliability and accuracy of these models. Developing safeguards and validation mechanisms is essential to minimize the potential for misinformation in mental health applications. Notably, some LLMs, like LLaMA and BLOOM, have explicitly stated that their use in high-stakes settings is either out of scope or prohibited. This underscores the recognition within the AI community of the ethical and practical concerns surrounding the application of LLMs in situations where human well-being and mental health are at stake.

In conclusion, these limitations and guidelines should serve as a reminder that LLMs are not universally suitable for all mental health applications and should be used judiciously with a full understanding of their strengths and limitations.

2 Early Prediction Through LLMs' Generation

Social media platforms have become a rich data source for studying and potentially detecting mental health issues [32; 15; 49]. Early detection of mental health concerns on social media involves using models to identify signs and patterns that may indicate emotional distress, mental health issues, or potential crises. The emergent prompt-based learning follows the steps of pretraining, prompting, and predicting. For example, an LLM, such as GPT-3 [6] and its successors, generates the predicted mental disorder label given a prompt and a social post as the input. The generation-as-prediction paradigm has many advantages in many NLP tasks. In the mental health analysis, an early evaluation on ChatGPT [2] and other LLMs [47] indicate LLMs are good generalist models but not as good as specialized discriminative classification models trained specifically for downstream tasks. Recently, Mental-LLM [46] and MentalLLaMA [48] show that instructional fine-tuning can improve the prediction performance. However, finetuning generative large models with billions of model parameters still did not outperform discriminative models with millions of parameters.

Instablity of Generation-as-Prediction The dynamic nature of generative models means that small alterations in the input prompt can lead to significantly different outputs. In the context of mental health prediction, this unpredictability poses a serious concern. A minor modification in the wording or framing of a prompt could yield varying and potentially incorrect assessments of an individual's mental health condition. For example, Yang et al. [47] reported that the model's performance is highly sensitive to variations in adjectives describing condition severity while mentioning that few-shot learning could be a possible way to mitigate it. Specifically, altering the adjectives of severity from *any*, *some* to *very severe* can result in fluctuations in predictive accuracy without a discernible pattern. This instability underscores the necessity of thorough audits of the model's performance and response to different inputs.

From the View of Meta Optimization Some recent research in machine learning has provided insights into the in-context learning behavior of LLMs. For example, Dai et al. [10] suggested that LLMs perform implicit gradient descent at inference time. There are some similar views, such as the concept of learning in-context through gradient descent [42] and the mechanism of causal language modeling through meta-learning [45]. Meta optimization seems a quite reasonable explanation for the "learning" process of LLMs' generation given a prompt. However, there is no definitive consensus on this matter. For example, Min et al. [28] showed that ground truth demonstrations are not required for in-context learning, raising the question of whether LLMs might rely on a form of hard memorization. The debate continues, and a conclusive answer remains elusive. In the context of unpredictable prediction of LLMs' generation, the optimization process, when viewed as a form of meta-optimization, can appear arbitrary without a certain optimization objective, especially when prompted with free-form inputs as illustrated in Figure 2. In the case study conducted by Yang et al. [47], the adjectives of severity affect the inference time optimization. This underscores the challenges in adapting LLMs to complex human mental states and the nuances present in self-reported mental health posts. Overall, the nature of LLMs' in-context learning and the design of prompts remain subjects of ongoing research and debate, given the unique characteristics and challenges posed by LLMs in their generation of text.

Reliablity and Auditing Generative language models provide a more flexible and accessible way through API than the preceding pre-training and fine-tuning paradigm, largely due to their rapid development. OpenAI's ChatGPT, for example, stands as a prominent model accessible via an API, facilitating the creation of generation-as-prediction pipelines for a wide user base. At the time of writing, despite their larger model size, LLMs utilized for generation-as-prediction still exhibit lower predictive performance than previous models trained with task-specific classification heads [46; 48]. Additional (instruction) fine-tuning mitigates this performance gap. The significance of fine-tuning on diverse and representative datasets cannot be overstated. Addressing biases in the training data and optimizing model hyperparameters to achieve improved performance in mental health classification tasks remain less explored, primarily due to the extensive computing requirements for training large models and searching hyperparameters. These factors collectively contribute to the responsible and effective utilization of LLMs in mental health assessment. Further-

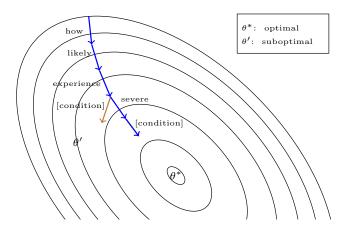


Figure 2: An illustration of prompting from the view of meta update. The change in the prompt might lead to suboptimal, possibly explaining the unpredictable LLMs' generation-as-prediction.

more, the instability of LLMs' generation-as-prediction remains a challenging problem. The view-point of meta-optimization can probably potentially shed light on the early prediction of mental disorder through LLMs' generation, for example, quantitatively evaluating the equivalent parameter contributions [23] of prompts tailored for mental health applications. Besides, it is crucial to establish auditing processes [30] that assess the model's reliability, sensitivity to input variations, and potential biases to ensure the responsible and accurate use of generative models in mental health applications. Such audits can help identify and mitigate issues related to unpredictability and instability, ultimately improving the model's suitability for assisting in mental health prediction and ensuring that its outputs are consistent and dependable.

3 LLM-generated Explanation ≠ Interpretablity

While deep learning models are often considered opaque, recent research has unveiled that these hidden representations can, to some extent, offer explanations. For instance, there has been ongoing discussion regarding whether attention mechanisms serve as explanations [5], and we acknowledge that there is no definitive consensus on this matter. In the context of mental health applications, our stance aligns with the perspective put forth in these publications regarding explainability. LLMs have the ability of self-explanation to provide explanations for their responses or generate text that clarifies the reasoning behind their answers, which is a form of step-by-step reasoning as referred to chain-of-thought [44]. However, such explanations can be unfaithful [46] and require targeted efforts for improvements [40]. Assessing the robustness and faithfulness of LLM-generated explanations in the context of mental health is crucial. LLMs may sometimes produce explanations that are overly simplistic or misleading, potentially impacting the quality of mental health interventions. It is essential to rigorously evaluate the explanations generated by LLMs to ensure they align with established clinical knowledge and guidelines. Besides, it is crucial to exercise caution and prudence when making claims about the explainability and interpretability of LLMs-based methods applied to mental health applications. LLMs' generated explanations do not imply LLMs for mental health analysis are inherently interpretable. Research works must refrain from using "interpretability" and "explainability" interchangeably to avoid misconceptions and ensure clarity in discussions surrounding LLMs in mental health applications.

Interpretability and Explanability Interpretability and/or explainability are frequently employed in many mental health publications [20]. It is crucial to recognize the distinct contrast between "interpretability", which pertains to a model's inherent characteristics, and "explainability", which refers to the methods used to explain a model or make a model interpretable, while "explanations" encompass the actual insights or justifications provided by the model to facilitate users' comprehension of its predictions [35]. Despite this, it is worth noting that some literature within the field of LLM uses interpretability and explainability interchangeably, such as Zhao et al. [50]. Recent

work such as Yang et al. [47] explores how LLMs generate text to explain the prediction of mental disorders. It is important to recognize that these explanations may lack interpretability. In other words, LLMs may provide detailed explanations (putting aside the faithfulness aspect for now), but these explanations may not be straightforward or easily comprehensible to human users who seek to understand why the model generates such textual explanations. It is crucial to understand that LLMs' explanations, as post-hoc generated text, do not guarantee that the model will be inherently interpretable. Users may need to exert additional effort or engage in further processing to make sense (or nonsense) of these explanations and render them readily digestible.

Call for Interpretability LLMs are getting more performant in many applications and improving in mental health applications. While LLM-based methods are employed for mental health analysis, the claim of interpretability should be considered cautiously. Our intention is not to dismiss the value of ongoing research on self-explanation. Instead, we aim to clarify definitions and claims, particularly within critical applications like mental healthcare. One avenue of research in the realm of LLM self-explanation involves engineering techniques or experimental testing that explains the significance of the model's representations and draws intuitive conclusions about the performance of these generated explanations or representations. In mental health, relying solely on the modelgenerated explanation is insufficient. Human judgment and clinical expertise should be integral in explaining and validating the results. When explaining the causes behind mental disorders, it is crucial to verify the accuracy and evidence-based nature of the explanations provided by LLMs. Additionally, it is essential to carefully monitor and mitigate the potential for LLMs to generate stigmatizing or harmful explanations. Interpretability is a critical factor, especially in fields where decisions can profoundly affect individuals' well-being [20]. More importantly, we call upon the computational research community in the field of mental health to focus on developing techniques that make these models more inherently interpretable, rigorously define the knowledge being modeled or applied within the mathematical theory, and adhere to proof or analysis that has been done through conceptual representation capacity, generalization, and robustness of neural networks in theory. The black-box nature of neural networks in LLMs underscores the need for transparency and validation in mental health applications, allowing clinicians and experts to trust and validate the results. Although the trade-off between interpretability and accuracy is still a matter of debate, the emphasis on interpretability can help ensure that LLMs become valuable tools in mental health while mitigating the risks associated with their black-box nature. The future trajectory in integrating LLMs into mental health applications could be ensuring that the LLMs' outputs align with clinical perspectives in interpreting and validating the model's prediction and developing specialized tools tailored for mental health professionals to comprehend the model. In this context, data-driven methods like LLMs serve as a user interface to improve the overall usability of the mental health support system and interpretable methods are used for certain aspects of decisionmaking (Figure 3). An analogy is the well-established diagnostic tools like the nine-item Patient Health Questionnaire (PHQ-9) for assessing depressive symptoms [21]. Interpretable methods that foster an understanding of their inner workings enable users, especially mental health professionals, to grasp the rationale behind the model's prediction.

4 LLMs in Mental Health Counseling

Chatbots for mental health, such as Woebot, have been developed to provide emotional support and aid in Cognitive Behavioral Therapy (CBT) through conversation with people living in mental health conditions [12]. Sarkar et al. [37] reviewed conversational agents for mental health and emphasized clinical knowledge and clinical practice guidelines in making them explainable and safe. Developments such as MindWatch [4] may play a role in monitoring such risks, but their use should be guided by best practices and ethical considerations.

Here, we discuss whether LLM-generated text is good for counseling in recent studies on empathy, user intention, emotion cause, and beyond. We expect reinforcement learning from human feedback to enable helpful and harmless generation, which could possibly enhance LLM-based mental health counseling. LLM-generated text can have potential applications in counseling, especially in psychological therapies like cognitive behavioral therapy (CBT). However, it's essential to approach this with caution. The recent development of LLMs might assist in providing information

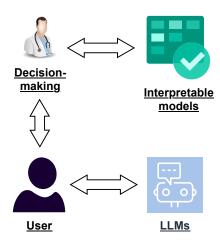


Figure 3: LLMs serve as the user interface to facilitate service quality, while interpretable models are critical for decision-making.

or exercises but should not replace the human element of counseling to offer personalized guidance and adapt interventions to the individual's unique needs, especially in sensitive mental health contexts.

Human Intent, Touch and Empathy While LLMs can provide automated responses and information and process long context [27], the generation mainly relies on learned model parameters from the pretraining corpus and the calculation of the likelihood of the next word. They can be distracted by irrelevant context [39] and may not fully understand the nuances of individual experiences, especially when there is insufficient individual training data, making their advice less tailored. These models lack the empathetic and contextual understanding that human counselors possess, which are crucial in counseling, especially in mental health contexts [38]. LLMs are required to have the human touch, empathy, and comprehension that human counselors can provide to enable more effective counseling. Reinforcement learning is adopted to facilitate empathic conversations [38], generate motivational and empathetic responses with long-term reward maximization [36], and promote polite and empathetic counseling [29]. Reinforcement learning in combination with LLMs can enhance the potential for a better dialogue system and reinforce counseling strategies in mental health. Ji [17] showcased that language models struggle to comprehend user intentions and can inadvertently generate harmful or hateful content. In such cases, it becomes essential to employ contextual intent understanding, model the intention awareness [8], and reason to identify the root causes of mental conditions, which could be used to enable empathetic conversational chatbot [25], and generate responses with human-consistent empathetic intents [9]. These strategies highlight the importance of understanding why users turn to LLM-based counseling and what they anticipate, offering insights for the design and deployment of these systems and making them more humane and responsive.

Promises and Caveats The use of LLMs in mental health counseling brings both potential benefits and perils. Chatbots engage in complex conversations with mental health consumers but struggle to identify and respond effectively to signs of distress, and consumers react negatively to unhelpful or risky chatbot responses [11]. The responses of some publicly available LLM-based chatbots, when presented with prompts of increasing depression severity and suicidality, failed to recognize the risk progression appropriately [16]. Generative LLMs such as ChatGPT struggled to detect unsafe responses in mental health support dialogues [34]. Cabrera et al. [7] examined ethical issues using chatbots for mental health, identifying 24 moral dilemmas that cut across bioethical principles. LLMs as chatbots require regulation, but the unreliability stops them from applying to the real world [14]. ChatCounselor [26] conducted supervised fine-tuning of base LLMs on real-world conversations between consulting clients and professional psychologists, although only evaluated the performance with GPT-4 without human grounds. Deploying LLM-based technology at scale for mental health may pose risks of misuse and require careful development and ongoing evaluated

ation more systematically. [24] studied an annotation framework for understanding counselors' strategies and client reactions. Interacting with LLMs can provide insights into tailoring the use of these models in mental health counseling and make LLMs more professional virtual counselors, leading to the need for interactive language processing [43]. Human preference data tailored for mental health scenarios can also be used to train reinforcement learning models to enable helpful and harmless dialogue agents [3]. LLM-based methods facilitate psychological intervention and educational outreach for non-professionals [13] and meanwhile posit some potential risks such as unreliable generation [14] and weakness in assessment of risk progression [16]. To ensure safe usage, more rigorous improvements and tests are needed.

5 Conclusion

AI indeed has the potential to be a valuable tool for identifying and supporting individuals who may be facing mental health challenges on social media platforms. However, it is essential to acknowledge and address the current challenges and concerns associated with its use. This paper discusses the problems associated with large language models in mental health applications and emphasizes the importance of conducting further research to enhance the safety and reliability of LLM-based methods. While this study raises concerns regarding their effectiveness and explainability, it is essential to clarify that its intention is not to discredit the existing efforts made in this field. Instead, the research works discussed in this paper are regarded as essential steps toward exploring LLMs' real-world applications. Our perspectives on rethinking LLMs in mental health applications aim to encourage the research community to reflect more deeply on LLMs' applicability, accountability, trustworthiness, and reliability [8].

Firstly, it is noteworthy to mention that the application of LLMs for generative prediction in mental health has made significant progress, albeit without achieving a breakthrough. Nevertheless, several crucial issues, such as the instability in generated predictions and the performance of the generation-as-prediction paradigm, continue to persist and remain unresolved.

Secondly, LLMs possess the capability to generate explanations during generative predictions. This feature provides supplementary information to support the predictions made by LLMs. However, it is important to note that this does not necessarily imply that the model is inherently interpretable when applied to mental health analysis.

Thirdly, LLMs have shown considerable promise in generating coherent and fluent textual content. This quality makes them viable candidates for automatic mental health counseling. Nonetheless, it is important to emphasize the need for further research and development to ensure that the generated content is genuinely helpful and harmless for safe and effective application in mental health scenarios.

In conclusion, LLMs represent a promising frontier in mental health, but they must be approached with caution and respect for ethical principles. Their role should be one of support for human experts, emphasizing their unique abilities in the field. Our perspectives serve an important but not exhaustive view of applying LLMs in mental health. Notably, there are other important aspects and considerations, such as cultural sensitivity, privacy, and data security. Robustness, ethical guidelines, and careful monitoring are essential components of deploying LLMs in the crucial task of addressing challenges in computational methods for mental health.

Ethical Considerations

Ethical considerations are undeniably pivotal in deploying LLMs in mental health applications. Guaranteeing user safety, preserving privacy, and mitigating biases in responses are critical aspects that demand careful attention. This paper underscores the exclusive reliance on publicly available publications for its research. Furthermore, it is essential to emphasize that there is no engagement in efforts to identify or directly interact with the individuals behind the social media posts. For research involving LLMs engaging in interactions with human beings, it is important to adhere to the user guidelines provided by the corresponding LLMs and to uphold the principles of research ethics rigorously. This ensures that ethical standards are maintained in all aspects of LLM-based research, especially in sensitive contexts like mental health.

References

- [1] T. Althoff, K. Clark, and J. Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.
- [2] M. M. Amin, E. Cambria, and B. W. Schuller. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of ChatGPT. *IEEE Intelligent Systems*, 38(2):15–23, 2023.
- [3] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* preprint arXiv:2204.05862, 2022.
- [4] R. Bhaumik, V. Srivastava, A. Jalali, S. Ghosh, and R. Chandrasekharan. MindWatch: A smart cloud-based AI solution for suicide ideation detection leveraging large language models. *medRxiv*, 2023. doi: 10.1101/2023.09.25.23296062. URL https://www.medrxiv.org/content/early/2023/09/26/2023.09.25.23296062.
- [5] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.269. URL https://aclanthology.org/2022.acl-long.269.
- [6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.
- [7] J. Cabrera, M. S. Loyola, I. Magaña, and R. Rojas. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 313–326. Springer, 2023.
- [8] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6), 2023.
- [9] M. Y. Chen, S. Li, and Y. Yang. EmpHi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074, 2022.
- [10] D. Dai, Y. Sun, L. Dong, Y. Hao, S. Ma, Z. Sui, and F. Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL* 2023, pages 4005–4019, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247. URL https://aclanthology.org/2023.findings-acl.247.
- [11] J. De Freitas, A. K. Uğuralp, Z. Oğuz-Uğuralp, and S. Puntoni. Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology*, 2022.
- [12] K. K. Fitzpatrick, A. Darcy, and M. Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.
- [13] G. Fu, Q. Zhao, J. Li, D. Luo, C. Song, W. Zhai, S. Liu, F. Wang, Y. Wang, L. Cheng, J. Zhang, and B. X. Yang. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals, 2023.

- [14] S. Gilbert, H. Harvey, T. Melvin, E. Vollebregt, and P. Wicks. Large language model AI chatbots require approval as medical devices. *Nature Medicine*, pages 1–3, 2023.
- [15] K. Harrigian, C. Aguirre, and M. Dredze. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24. ACL, 2021.
- [16] T. F. Heston. Evaluating risk progression in mental health chatbots using escalating prompts. medRxiv, 2023. doi: 10.1101/2023.09.10.23295321. URL https://www.medrxiv.org/content/early/2023/09/12/2023.09.10.23295321.
- [17] S. Ji. Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of EMNLP*, pages 4028–4038. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.findings-emnlp.297.
- [18] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*, pages 7184–7190, Marseille, France, 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.778.
- [19] S. Ji, T. Zhang, K. Yang, S. Ananiadou, E. Cambria, and J. Tiedemann. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv* preprint *arXiv*:2304.10447, 2023. URL https://arxiv.org/abs/2304.10447.
- [20] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, 6(1):6, 2023.
- [21] K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.
- [22] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang. Psy-LLM: Scaling up global mental health psychological services with AI-based large language models, 2023.
- [23] J. Lan, R. Liu, H. Zhou, and J. Yosinski. LCA: Loss change allocation for neural network training. *Advances in Neural Information Processing Systems*, 32, 2019.
- [24] A. Li, L. Ma, Y. Mei, H. He, S. Zhang, H. Qiu, and Z. Lan. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, 2023.
- [25] Y. Li, K. Li, H. Ning, X. Xia, Y. Guo, C. Wei, J. Cui, and B. Wang. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2041–2045, 2021.
- [26] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu. ChatCounselor: A large language models for mental health support. arXiv preprint arXiv:2309.15461, 2023.
- [27] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.
- [28] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759.
- [29] K. Mishra, P. Priya, and A. Ekbal. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14408–14416, Jun. 2023. doi: 10.1609/aaai.v37i12.26685. URL https://ojs.aaai.org/index.php/AAAI/article/view/26685.

- [30] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, pages 1–31, 2023.
- [31] U. Naseem, B. C. Lee, M. Khushi, J. Kim, and A. Dunn. Benchmarking for public health surveil-lance tasks on social media with a domain-specific pretrained language model. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 22–31, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlppower-1.3. URL https://aclanthology.org/2022.nlppower-1.3.
- [32] U. Pavalanathan and M. De Choudhury. Identity management and mental health discourse in social media. In *WWW*, pages 315–321. ACM, 2015.
- [33] H. Qiu, H. He, S. Zhang, A. Li, and Z. Lan. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. *arXiv* preprint arXiv:2305.00450, 2023.
- [34] H. Qiu, T. Zhao, A. Li, S. Zhang, H. He, and Z. Lan. A benchmark for understanding dialogue safety in mental health support. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 1–13. Springer, 2023.
- [35] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [36] T. Saha, V. Gakhreja, A. S. Das, S. Chakraborty, and S. Saha. Towards motivational and empathetic response generation in online mental health support. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2650–2656, 2022.
- [37] S. Sarkar, M. Gaur, L. K. Chen, M. Garg, and B. Srivastava. A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement. *Frontiers in Artificial Intelligence*, 6, 2023.
- [38] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference*, pages 194–205, 2021.
- [39] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023.
- [40] M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv* preprint *arXiv*:2305.04388, 2023.
- [41] V. Vajre, M. Naylor, U. Kamath, and A. Shehu. PsychBERT: a mental health language model for social media mental health behavioral analysis. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1077–1082. IEEE, 2021.
- [42] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.
- [43] Z. Wang, G. Zhang, K. Yang, N. Shi, W. Zhou, S. Hao, G. Xiong, Y. Li, M. Y. Sim, X. Chen, Q. Zhu, Z. Yang, A. Nik, Q. Liu, C. Lin, S. Wang, R. Liu, W. Chen, K. Xu, D. Liu, Y. Guo, and J. Fu. Interactive natural language processing. *arXiv* preprint arXiv:2305.13246, 2023.
- [44] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [45] X. Wu and L. R. Varshney. A meta-learning perspective on transformers for causal language modeling. arXiv preprint arXiv:2310.05884, 2023.

- [46] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang. Mental-LLM: Leveraging large language models for mental health prediction via online text data. arXiv preprint arXiv:2307.14385, 2023.
- [47] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou. Towards interpretable mental health analysis with large language models. In *Proceedings of EMNLP*, 2023. URL https://arxiv.org/abs/2304.03347.
- [48] K. Yang, T. Zhang, Z. Kuang, Q. Xie, and S. Ananiadou. MentalLLaMA: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023.
- [49] T. Zhang, A. Schoene, S. Ji, and S. Ananiadou. Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, 5, 2022.
- [50] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. Explainability for large language models: A survey. *arXiv* preprint arXiv:2309.01029, 2023.