nature portfolio

Peer Review File

A Data-Centric Approach to Detecting and Mitigating Demographic Bias in Pediatric Mental Health Text: A Case Study in Anxiety Detection

Corresponding Author: Dr Julia Ive

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This study targets a timely and impactful research question. The issue of biases in applications of AI models for mental health is critical and potentially interesting to many readers. While the overall structure of the current paper seems rationale and fairly straightforward, there are some major and minor concerns that are listed below.

Major Weaknesses
- The primary concern about the study design is the choice of one specific method (i.e., Clinical-BigBird) and basing almost all analysis on the method choice. The listed Objectives 1 and 3 seem to be directly dependent on the method choice. The identified biased patterns and the proposed method are not shown to be "generalizable" beyond the current method. Additionally, this model is referred to as a "state-of-the-art" model; while it dates back to 2022 and before offering many newer variations of NLP (decoder-, encoder-, or transformer-based) methods. This method only has a context limit of ~4K tokens (vs. million-sized ones in 2025). This issue may be less of a concern for Objective 2, but that part relates to descriptive analysis and is not related to the main theme of the paper (AI-based methods).

- The overall formatting, organization, and flow of the paper make it hard to follow the ideas easily. There are many short paragraphs, typos, missing punctuation, and undefined acronyms. For instance, the first sentence below Objective 1 is unfinished, or "Unc" is not clearly defined in the text.

Minor Weaknesses
- The text uses subjective language. Examples include "highest-ranked pediatric institutions in US" and "state-of-the-art NLP". A 2018 paper is considered the best in the vibrant field of NLP.

- LIME is fine, but the choice should be explained given other options for local explainability.

- Unclear why .25 below and above the threshold of 1 is considered significant. Relatedly, the study used the term "significant" loosely in many places without offering an objective statistical measure.

- The choice of 20% random filtering as a baseline (sole baseline?) seems unfair. Could the study use other mitigation techniques as a baseline?

- It is not discussed clearly why the experiments start with both gender and race and then primarily focus on gender. Also, the implications of the study for the mental health or pediatric patients are less clear. The study remains almost agnostic to the disease of interest.

Reviewer #2

(Remarks to the Author)
The study on pediatric mental health, particularly its focus on bias and gender disparity, is both compelling and valuable. The author develops a data-centric de-biasing method that neutralizes biased language while preserving essential clinical

information. Tested on an automatic anxiety detection model for pediatric patients, the approach significantly reduced diagnostic bias, particularly mitigating the under-diagnosis of female adolescents.

Strengths:
This work addresses the often-overlooked issue of bias in unstructured clinical data, especially in pediatric mental health;
Demonstrates a substantial reduction in diagnostic bias, enhancing fairness in AI-driven healthcare solutions.
The overall evaluation and conclusion, including detailed analysis is very comprehensive, and the paper is easy to follow.

Suggestions:
1. Figure 1 has to be re-checked. Here, what is the "overamplification" points to?
2. Well the main method is still focused on Tranformer-based retrained LMs, and the de-biasing methods are somehow very basic. It is crucial to see how LLMs would perform in this scenario.
3. When selecting the patient EHRs, if a patient has visit multiple times, would the latest record be selected only?
4. From Tab 6, it maybe concluded that both de-biasing methods have limited impact to the performance, this also may indicate that more advanced models maybe tried out.

Reviewer #3

(Remarks to the Author)
I think this paper is very interesting but there are a number of things that should be addressed before it is ready for publication. As such, I recommend a revision. I've done my best to group my comments into major topics to ease the revision process.
1. More specific details are required in some places throughout the manuscript. In the introduction, it mentions anxiety rates have doubled, but not from what rate or to what rate. Top of page 2, what constitutes "mental health help"? Why is "comprehensive screening for pediatric mental health concerns is challenging"? Middle of page 2, "their selection process" needs to be clarified. I could think of multiple sources of selection bias which are separate from process, such as who has access to care and who seeks care. Fig. 1 caption mentions "we consider biases of four types" but then also "Our framework focuses on the textual bias", which on the surface seems contradictory. In "observed differences between demographic groups can generally be characterized as follows", which category does different presentations/manifestations of disease fall into? "This hierarchy of disparities", but unclear which list was a hierarchy. "anxiety symptoms among underrepresented groups, particularly females", need to explain how females are an underrepresented group with regards to anxiety. "In the challenging pediatric primary care setting" is an important setup and should be more fully discussed rather than relegated to a list in parentheses. The sentences "Our data come from one of the highest-ranked pediatric institutions in US. We apply a range of state-of-the-art NLP methods to these data" are so vague as to be pointless to include. "one encounter in the EHR", what does encounter mean in this context? Would like more detail about the model such as how similar was the training data and how does it handle longer sequences better than other similar models? "Those methods could be used separately or combined" should instead mention that you test them both separately and combined. "average length" vs "concatenation" on page 6 is unclear. "~540 words longer (~180 words", why not report on exact amounts?
2. The manuscript would benefit more thorough explanation of how you imagine AI being implemented within the healthcare system. For example, "AI constitutes a promising solution for expanding mental health diagnostics". However, many believe AI should only be used for screening as it's important to retain clinicians-in-the-loop, especially given the propensity for AI to exacerbate bias, as mentioned later in the introduction.
3. Sometimes unclear why specific studies are referenced. For example, "A recent study by Yates Coley et al.8", is that the only relevant study or most relevant study?
4. "For instance, techniques like swapping gender words between sentences may introduce unrealistic symptom profiles". However, it is my understanding, what your de-biasing method consists of is swapping out all gendered names and pronouns for gender neutral alternatives. This needs to be better explained why the existing method would be ineffective and how your method differs, or that you are testing whether the existing method would be effective. When you mention "we propose two following bias mitigation methods", does that mean you created the methods? If so, how do they differ from existing methods? Why not test them against existing methods?
5. A motivation for this work in the introduction is that "healthcare data is highly heterogeneous, as clinical records often come from various care sites", but then it seems like you used EHRs from a single institution. Please explain this as well as why all anxiety diagnoses were considered together.
6. More details are needed about the data, especially given no data will be shared and the statistical analysis is highly subject to dataset attributes. Specifically, a dataset table is needed with the number of participants and cases for different data subsets (train vs test, anxiety vs control) and demographics. Right now, the manuscript is missing how many EHR and participants after cleaning and binning. Why would there be duplicate notes in the dataset? Did the data that not fit into a bin get discarded? Why those specific age bins? In the modeling, you mention using 1,000 tokens (words?) but not what the average and standard deviation of the inputs. For the matched controls, I assume they had clinical notes for reasons other than anxiety, so could their reason for being treated introduce bias into the models as different sexes may be more likely to get treatment for different medical concerns?
7. Given that a pretrained model was used for anxiety classification, please explain how your "gender-word debiasing" method does not introduce different bias with the use of gender-neutral pronouns. Please explain how it does not "introduce unrealistic symptom profiles".
8. While you mention it is above random, a diagnostic model with an accuracy of 0.61 is VERY low. Please better explain the usefulness of your analysis given that the classification model performance is so poor.
9. "rnd_filt" is first mentioned in Table 1, and not clear what it refers to. I don't particularly like the green vs red color schema.

The green vs red vs bold is confusing (higher vs reduction vs best). Are different metrics reported for Org than other methods (higher values vs reduction)? Make sure all abbreviations like org, unc, and av fr (Table 3) are defined. In Table 2, missing green in bin 12 rows. Also, what does “non-privileged” group mean in caption, any reason not to be more specific?
10. There are some sentences that are missing citations. For example, the first sentences in the introduction and conclusions. “Additionally, clinicians often have concerns about the interpretability of predictions…” also needs a citation. Further, citations are sometimes after et al. and sometimes at end of sentence.
11. There are some typos, as well as excessive paragraph division. Paragraphs should generally be more than two lines/sentences. The typos I noticed: “This is particularly challenging in mental health care the primary source of information in mental health care” doesn’t work as a sentence, “selection bias” is the only type of bias not italicized in the text, switching from using entire number to k to represent a thousand (ie “7,810,849 notes for these 73k patients”), “This cohort contains demographic data (sex and race)” and age according to the next subsection, “model authors” lacks possession, “(see section Descriptive Analysis in Results)” on page 4 would be better as “refer to” and lacks a period at the end of the sentence, bin capitalization is inconsistent (noticed on page 5), av vs Av. In Table 3, and “representations16,are” missing space after comma.
12. Please explain how the code can is confidential and can not be shared. The lack of code greatly reduces the reproducibility and contribution of the work.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The raised concerns by this reviewer are addressed for the most part, and this reviewer appreciates the authors' efforts. However, the one major concern seems to be still there. The concern was about demonstrating the generalizability of the methods beyond one specific model. At this point, the paper only justifies why that one certain model was chosen, but falls short of demonstrating the generalizability beyond that.

The new text reads "Nevertheless, our proposed de-biasing framework is model-agnostic and can be integrated with a wide range of AI systems, including those with larger context windows." This is exactly the claim that needed strong evidence (beyond only the model used, which is arguably less capable than the modern ones).

As for the authors' response to R3's comments, they have mostly addressed the raised concerns. They seemed fine to me. The only minor note is about the very last comment about not sharing the code publicly. Even if their code is "tightly linked to confidential data pipelines," they may want to release the portions that are not directly related to the internal data. Admittedly, this could sometimes be challenging, but without that, the impact could be very limited.

Dear Editor,

We sincerely appreciate the thoughtful feedback and constructive comments provided by you and the reviewers on our manuscript entitled "A Data-Centric Approach to Detecting and Mitigating Demographic Bias in Pediatric Mental Health Text: A Case Study in Anxiety Detection". The suggestions have been invaluable in enhancing the clarity and overall quality of our work. We are encouraged by the positive feedback and have carefully revised the manuscript in accordance with the reviewers' recommendations.

For your convenience, we have attached a detailed, point-by-point response to each comment. All changes made in the revised manuscript have been highlighted in red.

We hope that our responses will be positively received. Should there be any further questions or comments, we would be very happy to address them.

Sincerely,
Julia Ive

We sincerely thank all reviewers for their thoughtful remarks and constructive feedback. We have carefully addressed each point, revisited parts of the analysis, and integrated suggested improvements. We believe this process has significantly enhanced the quality of the manuscript and hope it now meets your expectations.

Point-by-point responses to Reviewers' Comments.

Reviewer #1 (Remarks to the Author):

This study targets a timely and impactful research question. The issue of biases in applications of AI models for mental health is critical and potentially interesting to many readers. While the overall structure of the current paper seems rationale and fairly straightforward, there are some major and minor concerns that are listed below.

Major Weaknesses
- The primary concern about the study design is the choice of one specific method (i.e., Clinical-BigBird) and basing almost all analysis on the method choice. The listed Objectives 1 and 3 seem to be directly dependent on the method choice. The identified biased patterns and the proposed method are not shown to be "generalizable" beyond the current method. Additionally, this model is referred to as a "state-of-the-art" model; while it dates back to 2022 and before offering many newer variations of NLP (decoder-, encoder-, or transformer-based) methods. This method only has a context limit of ~4K tokens (vs. million-sized ones in 2025). This issue may be less of a concern for Objective

2, but that part relates to descriptive analysis and is not related to the main theme of the paper (AI-based methods).

We thank the reviewer for raising this important point. We fully acknowledge that our study relies primarily on Clinical-BigBird, and that referring to it as "state-of-the-art" overstates its novelty. We now describe it more appropriately as a representative clinical NLP model. Our motivation for selecting Clinical-BigBird was threefold: (i) it belongs to a family of compact, discriminative pretrained models that remain widely used in mental health contexts, where efficiency and robustness are essential (they are less to hallucinations) [1,2]; (ii) such models have been shown to perform on par with, or in some cases outperform, much larger generative models on in-domain classification tasks [3]; and (iii) its low computational requirements allowed us to run systematic experiments in the limited computational environment. We have clarified this in the manuscript (p. 5, ll. 154-158).

Moreover, Clinical-BigBird is also representative in that its performance is modest (diagnostic accuracy of 0.61). This reflects the reality of many clinical NLP applications. Also, as prior work suggests [4,5], models with lower absolute performance can sometimes make subgroup disparities more apparent, since their predictions are more sensitive to input features. We therefore used Clinical-BigBird as a diagnostic bias tool. We also experimented with other traditional machine learning approaches (e.g., Logistic Regression using word frequencies) and observed broadly similar performance levels (p. 13, ll. 421-424).

At the same time, we agree that our contribution should not rely on this specific model. To address concerns about generalisability, we clarify that our proposed de-biasing methods are model-agnostic and can be applied across architectures (p. 13, ll. 425-427).

1. Chopra, S. et al. (2024). Roberta and BERT: Revolutionizing mental healthcare through natural language. *SN computer science*, 5 (7), pp.1–12.
2. Hossain, M. M. et al. (2025). Multi task opinion enhanced hybrid BERT model for mental health analysis. *Scientific reports*, 15 (1), pp.3332.
3. Chen, Q. et al. (2025). Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16 (1), pp.3280.
4. Blanzeisky, W. and Cunningham, P. (2021). Algorithmic factors influencing bias in Machine Learning. *arXiv [cs.LG]*.
5. Ding, X., Xi, R., & Akoglu, L. (2024). Outlier Detection Bias Busted: Understanding Sources of Algorithmic Bias through Data-centric Factors. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 384-395.

- The overall formatting, organization, and flow of the paper make it hard to follow the ideas easily. There are many short paragraphs, typos, missing punctuation, and undefined acronyms. For instance, the first sentence below Objective 1 is unfinished, or "Unc" is not clearly defined in the text.

Thank you for your valuable feedback. We have carefully revised the text throughout to improve readability and flow. Specifically, we addressed the issues you highlighted by merging short paragraphs, correcting typos and punctuation, and ensuring that all acronyms (including "Unc") are clearly defined upon first use. The unfinished sentence under Objective 1 has also been completed (p. 7, ll. 221-223).

Minor Weaknesses

- The text uses subjective language. Examples include "highest-ranked pediatric institutions in US" and "state-of-the-art NLP". A 2018 paper is considered the best in the vibrant field of NLP.

Thank you for pointing out the use of subjective language in the manuscript. We have carefully revised the text to ensure a more objective and neutral tone. Specifically, we replaced phrases such as “highest-ranked pediatric institutions in the US” with the name of the institution, and “state-of-the-art NLP” with “representative NLP methods” (p. 4, ll. 111-112).

- LIME is fine, but the choice should be explained given other options for local explainability.

We selected LIME for local explainability due to its speed and intuitive output, which makes it particularly suitable for communicating model behavior to non-technical stakeholders in clinical settings. While alternative methods such as SHAP [1] offer more theoretically grounded attributions, they are computationally more intensive and conceptually complex hence less practical for our use case (see p. 6 of the manuscript, ll. 176-179).

1. Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*.

- Unclear why .25 below and above the threshold of 1 is considered significant. Relatedly, the study used the term "significant" loosely in many places without offering an objective statistical measure.

We thank you for this insightful comment. The ratio thresholds of >1.25 and <0.80 are used to indicate meaningful bias. In our study, we adopt the fairness ratio bounds of 0.80 and 1.25 as defined by Feldman et al. (2015) [1], which are commonly accepted in algorithmic

fairness literature to denote meaningful differences in outcomes between groups [2]. These thresholds are not derived from statistical hypothesis testing but rather serve as practical cutoffs to identify disparities in model performance.

We amended the manuscript accordingly and removed the unjustified usage of the term significant (p. 7, ll. 229-233).

1. Feldman, M. et al. (2015). Certifying and Removing Disparate Impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM. Doi:10.1145/2783258.2783311.
2. Bellamy, R. K. E. et al. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv [cs.AI]*.

- The choice of 20% random filtering as a baseline (sole baseline?) seems unfair. Could the study use other mitigation techniques as a baseline?

We thank the reviewer for highlighting this important point. We agree that 20% random filtering is a relatively simple baseline. Our rationale for selecting it was that it represents the most straightforward, low-cost strategy that could be realistically applied in practice without requiring additional annotations, compute resources, or complex modeling infrastructure.
We also acknowledge that incorporating other established mitigation strategies, such as reweighting, adversarial de-biasing [1,2], or representation learning [3], would have further strengthened the comparative analysis. Unfortunately, due to the expiration of our project-specific data access, we were unable to implement additional baselines within the scope of this study (see p.6, ll. 190-195).

1. Sun, Z. et al. (2024). Causal-guided active learning for debiasing large language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Pp.14455–14469.
2. Gallegos, I. O. et al. (2025). Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Pp.873–888.
3. Vaidya, A. et al. (2024). Demographic bias in misdiagnosis by computational pathology models. *Nature medicine*, 30 (4), pp.1174–1190.

- It is not discussed clearly why the experiments start with both gender and race and then primarily focus on gender. Also, the implications of the study for the mental health or

pediatric patients are less clear. The study remains almost agnostic to the disease of interest.

We thank the reviewer for this thoughtful observation. The study indeed initially considered both sex and race to reflect key dimensions of bias in clinical data. However, we focused primarily on sex in the later stages due to more pronounced differences in subgroup distributions and clearer clinical relevance. For example, females have a higher lifetime prevalence of anxiety disorders compared to males, a disparity that becomes particularly evident during adolescence [1,2]. These patterns are consistently reflected in our dataset, allowing for more robust analysis: percentages of female patients grow from 36% to 69% in age bins from 5 to 15. In contrast, ethnicity-related patterns remain understudied and, in our case, could not be reliably assessed due to data sparsity (the percentage of the other races is stable around 30% across age bins).

We use gender and sex interchangeably in the manuscript because our data record biological sex, but our analysis often relates to linguistic gendered patterns as a social construct. We clarify this distinction in the revised text (p. 3, ll. 94-96).

Regarding disease specificity, our aim was indeed to develop a generalizable mitigation framework. However, the reasons behind gender-related anxiety patterns remain understudied and are likely the result of a complex interplay of biological, psychological, and social factors. De-biasing for gender is hence particularly important in pediatric anxiety research, as it helps ensure that AI-driven tools do not reinforce existing disparities. Moreover, fair AI models can support abductive reasoning and the generation of new hypotheses, potentially uncovering causal mechanisms for these gender patterns.

We clarified this in the Discussion section of the manuscript (p. 13, ll. 409-416).

1. Dalsgaard, S. et al. (2020). Incidence rates and cumulative incidences of the full spectrum of diagnosed mental disorders in childhood and adolescence. *JAMA psychiatry (Chicago, Ill.)*, 77 (2), pp.155–164.
2. Warner, E. N. et al. (2023). Developmental epidemiology of pediatric anxiety disorders. *Child and adolescent psychiatric clinics of North America*, 32 (3), pp.511–530.

Reviewer #2 (Remarks to the Author):

The study on pediatric mental health, particularly its focus on bias and gender disparity, is both compelling and valuable. The author develops a data-centric de-biasing method that neutralizes biased language while preserving essential clinical information. Tested on an automatic anxiety detection model for pediatric patients, the approach significantly

reduced diagnostic bias, particularly mitigating the under-diagnosis of female adolescents.

Strengths:
This work addresses the often-overlooked issue of bias in unstructured clinical data, especially in pediatric mental health;
Demonstrates a substantial reduction in diagnostic bias, enhancing fairness in AI-driven healthcare solutions.
The overall evaluation and conclusion, including detailed analysis is very comprehensive, and the paper is easy to follow.

Thank you very much for highlighting the strengths of our work. We are especially grateful for your recognition of the focus on bias in unstructured pediatric mental health data, and the emphasis on fairness in AI-driven healthcare. Your comments on the clarity and comprehensiveness of our work are deeply appreciated.

Suggestions:
1. Figure 1 has to be re-checked. Here, what is the "overamplification" points to?

Regarding your suggestion on Figure 1, we appreciate your attention to detail. We have amended the Figure to improve clarity.

Each component is annotated with the type of bias it may introduce. Specifically, the term “overamplification" refers to the phenomenon where biases are intensified by the NLP model during training (see p. 2).

2. Well the main method is still focused on Tranformer-based retrained LMs, and the de-biasing methods are somehow very basic. It is crucial to see how LLMs would perform in this scenario.

Thank you for raising this important point. Our model choice was motivated by three factors: (i) compact discriminative models remain widely used in clinical contexts for their efficiency and robustness [1,2]; (ii) they often perform on par with much larger generative models on in-domain classification tasks [3]; and (iii) their low computational cost enabled systematic experimentation in the limited computational environment (see p. 5, ll. 154-158; p. 13, ll. 417-420).

Clinical-BigBird is also representative in its modest performance (accuracy ~0.61), reflecting typical outcomes in clinical NLP. Prior work shows that such models may better reveal subgroup disparities [4,5]. We clarify that our contribution is not tied to this model.

The proposed de-biasing methods are model-agnostic and applicable across architectures (see p. 13, ll. 421-424).

We also acknowledge that the de-biasing techniques explored here are relatively basic. This was a deliberate choice to prioritise transparent, low-cost, and easily reproducible strategies.

We fully agree that evaluating large language models (LLMs) in this context is crucial, however, this was not possible due to the computational limitations when conducting the study.

1. Chopra, S. et al. (2024). Roberta and BERT: Revolutionizing mental healthcare through natural language. *SN computer science*, 5 (7), pp.1–12.
2. Hossain, M. M. et al. (2025). Multi task opinion enhanced hybrid BERT model for mental health analysis. *Scientific reports*, 15 (1), p.3332.
3. Chen, Q. et al. (2025). Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16 (1), p.3280.
4. Blanzeisky, W. and Cunningham, P. (2021). Algorithmic factors influencing bias in Machine Learning. *arXiv [cs.LG]*.
5. Ding, X., Xi, R., & Akoglu, L. (2024). Outlier Detection Bias Busted: Understanding Sources of Algorithmic Bias through Data-centric Factors. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 384-395.

3. When selecting the patient EHRs, if a patient has visit multiple times, would the latest record be selected only?

Thank you for the question. No, we do not select only the latest record for each patient.
Instead, we include the 25 most recent clinical notes per patient, concatenated in temporal order. This results in an average input length of approximately 4,000 tokens. However, the model processes only the last 1,000 tokens, as we empirically observed no performance improvement beyond this threshold. We have clarified this in the manuscript (p. 5, ll. 159-163).

4. From Tab 6, it maybe concluded that both de-biasing methods have limited impact to the performance, this also may indicate that more advanced models maybe tried out.

From Table 6, it is correct that both de-biasing methods have only a limited effect on predictive performance. We would like to stress, however, that this was expected and in line with our primary objective: to preserve baseline accuracy while reducing bias, rather than to

optimise overall performance. Our de-biasing methods represent accessible and cost-effective mitigation strategies that could be implemented without additional data, heavy computation, or complex modelling (see p. 13, ll. 426-427).

We acknowledge that more advanced approaches (e.g., reweighting, adversarial de-biasing [1,2], or representation learning techniques [3]) may yield stronger mitigation effects. These approaches are typically more resource-intensive.

1. Sun, Z. et al. (2024). Causal-guided active learning for debiasing large language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Pp.14455–14469.
2. Gallegos, I. O. et al. (2025). Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Pp.873–888.
3. Vaidya, A. et al. (2024). Demographic bias in misdiagnosis by computational pathology models. *Nature medicine*, 30 (4), pp.1174–1190.

Reviewer #3 (Remarks to the Author):

I think this paper is very interesting but there are a number of things that should be addressed before it is ready for publication. As such, I recommend a revision. I've done my best to group my comments into major topics to ease the revision process.

Thank you very much for your thoughtful feedback. We appreciate the organisation of the comments into major topics which helped us navigate our revision.

1. More specific details are required in some places throughout the manuscript. In the introduction, it mentions anxiety rates have doubled, but not from what rate or to what rate.

Clinically elevated anxiety symptoms during the COVID-19 pandemic was 21%, compared to a pre-pandemic estimate of 12% [1]. We have included this into the manuscript (pp. 1-2, ll. 25-26).

1. Racine, N. et al. (2021). Global Prevalence of Depressive and Anxiety Symptoms in Children and Adolescents During COVID-19: A Meta-analysis. JAMA pediatrics, 175, pp.1142–1150.

Top of page 2, what constitutes "mental health help"? Why is "comprehensive screening for pediatric mental health concerns is challenging"?

By "mental health help," we refer to a continuum of clinical services, including outpatient care such as psychiatric consultations, diagnostic assessments, psychotherapy, psychopharmacology, and neuropsychological testing. In addition, more intensive support options include hospitalization (see manuscript p. 2, ll. 27-29).

Comprehensive screening for pediatric mental health concerns is challenging due to several factors inherent to the pediatric primary care setting. These include the need to integrate input from multiple informants (e.g., parents, teachers, and the child), variability in how risk factors and symptoms are interpreted, and the presence of overlapping symptoms across different mental health conditions, which can complicate accurate identification and diagnosis. These complexities, combined with time constraints and limited mental health training in many primary care environments, make consistent and equitable screening difficult to implement at scale [1,2,3] (see manuscript p. 2, ll. 30-32).

1. Behrens, B. et al. (2019). The screen for Child Anxiety Related Emotional Disorders (SCARED): Informant discrepancy, measurement invariance, and test-retest reliability. *Child psychiatry and human development*, 50 (3), pp.473–482.

2. Strawn, J. R. et al. (2021). Research Review: Pediatric anxiety disorders - what have we learnt in the last 10 years? *Journal of child psychology and psychiatry, and allied disciplines*, 62 (2), pp.114–139.

3. Tulisiak, A. K. et al. (2017). Antidepressant prescribing by pediatricians: A mixed-methods analysis. *Current problems in pediatric and adolescent health care*, 47 (1), pp.15–24.

Middle of page 2, "their selection process" needs to be clarified. I could think of multiple sources of selection bias which are separate from process, such as who has access to care and who seeks care.

Thank you for this insightful comment. We agree that the phrase “their selection process” was ambiguous and have revised the sentence for clarity.

The updated passage now reads:
“However, the available data are often sparse and biased because the process by which individuals enter the healthcare system often reflects underlying inequities. These include structural barriers such as who has access to care, who seeks care, and how symptoms are recognized and documented, all of which contribute to selection bias.” (p. 2, ll. 44-46)

Fig. 1 caption mentions “we consider biases of four types” but then also “Our framework focuses on the textual bias”, which on the surface seems contradictory.

Thank you for pointing out the inconsistency. We have revised the passage to clarify the distinction between the types of bias considered and the specific focus of our framework:

“Types of predictive bias and their origin. We consider four types of bias in clinical NLP models, as outlined by Shah et al. (2020): (a) *selection bias*, arising from training or testing data that are not representative; (b) *label bias*, introduced through human annotation; (c) *textual bias*, stemming from differences in word distributions across subpopulations; and (d) *overamplification bias*, where statistical models intensify discrepancies present in the training data. While our framework acknowledges all four, it specifically targets textual bias, as this is the most directly addressable through data-centric interventions.” (p. 2)

In “observed differences between demographic groups can generally be characterized as follows”, which category does different presentations/manifestations of disease fall into? “This hierarchy of disparities”, but unclear which list was a hierarchy. “anxiety symptoms among underrepresented groups, particularly females”, need to explain how females are an underrepresented group with regards to anxiety. “In the challenging pediatric primary care setting” is an important setup and should be more fully discussed rather than relegated to a list in parentheses.

Thank you for these thoughtful observations. We have revised the passage to clarify the categorization of observed differences, the meaning of hierarchy, and the context around pediatric care and gender representation:

“This categorization establishes a hierarchy of disparities, ranging from acceptable biological variation to deeply problematic systemic bias, and serves as a framework for guiding how bias should be identified and addressed in clinical AI. For example, different presentations or manifestations of disease may fall under genuine differences if biologically driven, or under contextual disparities if shaped by access or reporting

practices. While females are not underrepresented in terms of prevalence of anxiety, they may be underrepresented in fair representation within data, particularly when symptoms are documented differently. In the pediatric primary care setting, these challenges are amplified by the need to integrate input from multiple informants (e.g., parents, teachers, and the child), variability in symptom interpretation, and overlapping diagnostic criteria. These challenges complicate consistent and equitable screening [1,2,3]. These complexities also limit the applicability of many existing de-biasing methods: for instance, swapping gendered terms or removing gender components from embeddings can distort clinical meaning, while algorithmic modifications may lead to overfitting." (p. 3, ll. 84-93)

1. Behrens, B. et al. (2019). The screen for Child Anxiety Related Emotional Disorders (SCARED): Informant discrepancy, measurement invariance, and test-retest reliability. *Child psychiatry and human development*, 50 (3), pp.473–482.
2. Strawn, J. R. et al. (2021). Research Review: Pediatric anxiety disorders - what have we learnt in the last 10 years? *Journal of child psychology and psychiatry, and allied disciplines*, 62 (2), pp.114–139.
3. Tulisiak, A. K. et al. (2017). Antidepressant prescribing by pediatricians: A mixed-methods analysis. *Current problems in pediatric and adolescent health care*, 47 (1), pp.15–24.

The sentences "Our data come from one of the highest-ranked pediatric institutions in US. We apply a range of state-of-the-art NLP methods to these data" are so vague as to be pointless to include. "one encounter in the EHR", what does encounter mean in this context?

We have rephrased the indicated sentences into "Our dataset originates from the Cincinnati Children's Hospital Medical Center (CCHMC). We apply a range of representative NLP methods to analyze these data." (pp. 4, ll. 111-112) and "Additional selection criteria required that each patient had never received one of the anxiety diagnoses at the time of the matched case's anxiety diagnosis and had at least one documented healthcare contact in the EHR within the 18 months preceding their anxiety diagnosis, ensuring sufficient clinical history for analysis." (pp. 4, ll. 118, ll. 131-133)

Would like more detail about the model such as how similar was the training data and how does it handle longer sequences better than other similar models?

We built our anxiety prediction models by fine-tuning the popular Transformer-based Clinical-BigBird model [1]. Clinical-BigBird is pre-trained on MIMIC-III [2], a large, publicly available dataset of de-identified electronic health records (EHRs) aligned with our task of

analysing mental health notes. One of its key advantages is the ability to process long input sequences of up to 4,096 tokens, which is particularly beneficial in clinical NLP, where patient histories often span multiple notes. This capability stands in contrast to models like BERT [3], which are limited to 512 tokens and may truncate important context. Clinical-BigBird uses a computationally-efficient sparse attention mechanism. We have amended the manuscript accordingly. While our fine-tuning data is narrower in scope than MIMIC-III, it remains domain-consistent, consisting of mental health text notes from pediatric settings. (see p. 5, ll. 148-158 of the manuscript)

1. Li, Y. et al. (2022). Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv [cs.CL]*.
2. Johnson, A. E. W. et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific data, 10 (1), p.1.
3. Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C. and Solorio, T. (Eds). *Proceedings of the 2019 Conference of the North*. pp.4171–4186.

“Those methods could be used separately or combined” should instead mention that you test them both separately and combined.

We have revised the manuscript to state that both de-biasing methods were tested independently and in combination, allowing us to evaluate their individual and joint effects on model performance. (p. 6, ll. 213)

“average length” vs “concatenation” on page 6 is unclear. “

We have revised the indicated paragraph into the following: “*Average length of* a patient note in words (without tokenization, i.e., separation of punctuation). To characterize the typical volume of clinical documentation per patient, we assessed the average length of patient notes by aggregating all valid entries into a single text string per individual.” (p. 10, ll. 275-276)

~540 words longer (~180 words”, why not report on exact amounts?

We addressed the reviewer’s concern regarding the use of approximate values by reporting the exact differences in note length: on average, notes for the male subgroup were 543 words longer, and notes for white patients were 161 words longer. (p. 11, ll. 297-298)

2. The manuscript would benefit more thorough explanation of how you imagine AI being implemented within the healthcare system. For example, “AI constitutes a promising

solution for expanding mental health diagnostics". However, many believe AI should only be used for screening as it's important to retain clinicians-in-the-loop, especially given the propensity for AI to exacerbate bias, as mentioned later in the introduction.

In the revised manuscript, we clarify that while AI holds promise for expanding access to mental health diagnostics, its role is envisioned primarily as a supportive tool for screening and triage, not as a replacement for clinical judgment. We emphasize the importance of keeping clinicians in the loop, particularly given the risk of bias amplification and the nuanced nature of psychiatric assessment. (p. 2, ll. 32-36)

3. Sometimes unclear why specific studies are referenced. For example, "A recent study by Yates Coley et al.8", is that the only relevant study or most relevant study?

We thank the reviewer for their helpful comment regarding the clarity of cited studies. To address this, we have rephrased the indicated paragraph to clarify that the referenced works are examples of studies within the mental health domain, rather than the only or most relevant ones. The revised paragraph now reads:

"In the context of mental health, AI models are especially vulnerable to amplifying existing biases due to the subjective nature of psychiatric assessment, historical underrepresentation and the complexity of symptom expression across diverse populations. When trained on such data, models often underperform for marginalized groups, leading to inequitable outcomes and reinforcing existing disparities [1,2]. Several studies have highlighted this issue within mental health applications [2,3,4]. For example, Yates Coley et al. (2021) found that suicide risk prediction models performed poorly for underrepresented demographic groups, including Black, Native American, and Alaskan patients [4]." (p. 2, ll. 47-52)

1. Ji, Y. et al. (2025). Mitigating the risk of health inequity exacerbated by large language models. *npj digital medicine*, 8 (1), p.246.
2. Timmons, A. C. et al. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 18 (5), pp.1062–1096.
3. Chen, I. Y., Szolovits, P. and Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21 (2), pp.E167-179.
4. Coley, R. Y. et al. (2021). Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA psychiatry (Chicago, Ill.)*, 78 (7), pp.726–734.

4. “For instance, techniques like swapping gender words between sentences may introduce unrealistic symptom profiles”. However, it is my understanding, what your de-biasing method consists of is swapping out all gendered names and pronouns for gender neutral alternatives. This needs to be better explained why the existing method would be ineffective and how your method differs, or that you are testing whether the existing method would be effective. When you mention “we propose two following bias mitigation methods”, does that mean you created the methods? If so, how do they differ from existing methods? Why not test them against existing methods?

Thank you for your thoughtful feedback. We appreciate your request for greater clarity regarding our gender-word debiasing method and its distinction from existing approaches.

We agree that the sentence “techniques like swapping gender words between sentences may introduce unrealistic symptom profiles” could be more precise. Indeed, our method does not involve gender-swapping (e.g., replacing “she” with “he” [1]) but instead replaces gendered names and pronouns with neutral alternatives (e.g., “person1”, “they”). This distinction is critical, as gender-swapping can result in biologically implausible narratives (for example, “He reported experiencing abdominal pain and mood changes around his period”) which compromise clinical validity. We have revised the manuscript to make this distinction clearer and to better explain why we chose not to benchmark against gender-swapping techniques (p. 6., ll. 208-212). While such methods may yield similar bias mitigation outcomes, they risk distorting symptom descriptions.

Regarding your question about novelty: yes, when we state that “we propose two following bias mitigation methods,” we mean that these are adaptations of existing techniques, developed by us to address the challenges of mental health clinical text. We have now clarified this in the manuscript (p.6, ll. 182-183).

Information Density Filtering builds on the established technique of information filtering, which is widely used across domains [2]. However, its effectiveness depends heavily on how it is adapted to the specific context. In our case, we apply sentence-level TF-IDF scoring to clinical notes to retain the most informative content. Importantly, there is no universally accepted “go-to” benchmark for information filtering in clinical text. Because filtering strategies must be customized to the data, comparisons across methods are often not straightforward. To provide a meaningful baseline, we benchmark our approach against random information filtering (see p. 6, ll. 190-195 of the manuscript).

For Gender-Word Debiasing, we have expanded our rationale for avoiding gender-swapping and included references to prior work that highlights the risks of introducing clinically implausible narratives above.

1. Kiritchenko, S. & Mohammad, S. (2018) Examining gender and race bias in two hundred sentiment analysis systems. In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Pp. 43–53.
2. Rousseau, A. (2013). XenC: An Open-Source Tool for Data Selection in Natural Language Processing. *Prague Bull. Math. Linguistics*, 100 (1), pp.73–82.

5. A motivation for this work in the introduction is that "healthcare data is highly heterogeneous, as clinical records often come from various care sites", but then it seems like you used EHRs from a single institution. Please explain this as well as why all anxiety diagnoses were considered together.

Thank you for your insightful comment. Our introduction highlights the heterogeneity of healthcare data across care sites and our dataset reflects internal diversity. Cincinnati Children's operates across approximately 600 care sites, including a wide range of settings such as primary care clinics, specialty centers, urgent care facilities, school-based health centers, and regional outreach locations (e.g., Liberty Campus, College Hill Campus, Anderson Primary Care, etc.).

We have added a summary to the manuscript accordingly (p. 13, ll. 395-396).

6. More details are needed about the data, especially given no data will be shared and the statistical analysis is highly subject to dataset attributes. Specifically, a dataset table is needed with the number of participants and cases for different data subsets (train vs test, anxiety vs control) and demographics. Right now, the manuscript is missing how many EHR and participants after cleaning and binning. Why would there be duplicate notes in the dataset? Did the data that not fit into a bin get discarded? Why those specific age bins? In the modeling, you mention using 1,000 tokens (words?) but not what the average and standard deviation of the inputs. For the matched controls, I assume they had clinical notes for reasons other than anxiety, so could their reason for being treated introduce bias into the models as different sexes may be more likely to get treatment for different medical concerns?

We thank the reviewer for requesting more detail on the dataset. To address this, we provide Table 1 (in the manuscript p. 5 and below) and the accompanying Figure 1 below (added Figure 2 to the manuscript p. 4). All 73,288 patients who matched the selection criteria were successfully binned by age (0 to 24 years) at first anxiety diagnosis using a data-driven approach, ensuring that no eligible data was discarded during the process. Age bins of 5, 8, 10, 12, and 15 years were chosen to ensure diverse female representation: percentages of

female patients grow from 36% to 69% in age bins from 5 to 15. The percentage of the other races is stable around 30% across the selected age bins. On average, 25 notes per patient were retained (average minimum count of notes per patient across the Bins). Duplicates are common due to clinical copy-paste practices. While input length was capped at 1,000 tokens, some concatenated notes were still shorter than this limit, resulting in a mean input length of 728 tokens per patient.

Each age bin included from 1,828 to 2,532 cases in the training set and from 426 to 639 cases in the test set (sampled as 20% of each bin cases). We maintained a consistent 1:1 ratio of cases to matched controls (drawn from the same patient pool, matched by age and sex) across all age bins. Controls had never received one of the anxiety diagnoses at the time of the matched case's anxiety diagnosis and had at least one documented healthcare contact in the electronic health record (EHR) within the 18 months preceding their anxiety diagnosis, ensuring sufficient clinical history for analysis. As a result, the total number of patients per bin is double the number of cases.

We have clarified this in the manuscript text (pp. 4-5, ll. 127-138).

Thank you for this thoughtful observation regarding the bias in the cohort selection. While it is true that matched controls had clinical notes for reasons other than anxiety, we minimized potential bias by drawing them from the same patient pool as the cases. This ensured that both groups shared a similar clinical context. The purpose of our model was to distinguish the early clinical presentation of anxiety from the formal diagnosis stage (p. 13, ll. 404-408).

Table 1. Demographic statistics and properties of textual notes across Age Bins (measured for the training data). We report average values per concatenated patient note.

	Bin 5	**Bin 8**	**Bin 10**	**Bin 12**	**Bin 15**
count, total	4188	3884	3656	3662	5064
%, cases		50			
%, Male	64	60	56	45	31
%, Female	36	40	44	55	69
%, White	68	72	74	74	76
%, Other	32	28	26	26	24
Average note length, words	3887	4162	3957	3783	3916
Average model input length, tokens	724	733	726	729	735

Average terms, %	21	20	20	21	21
Average biased words, %	3.0	3.1	3.2	3.2	3.0

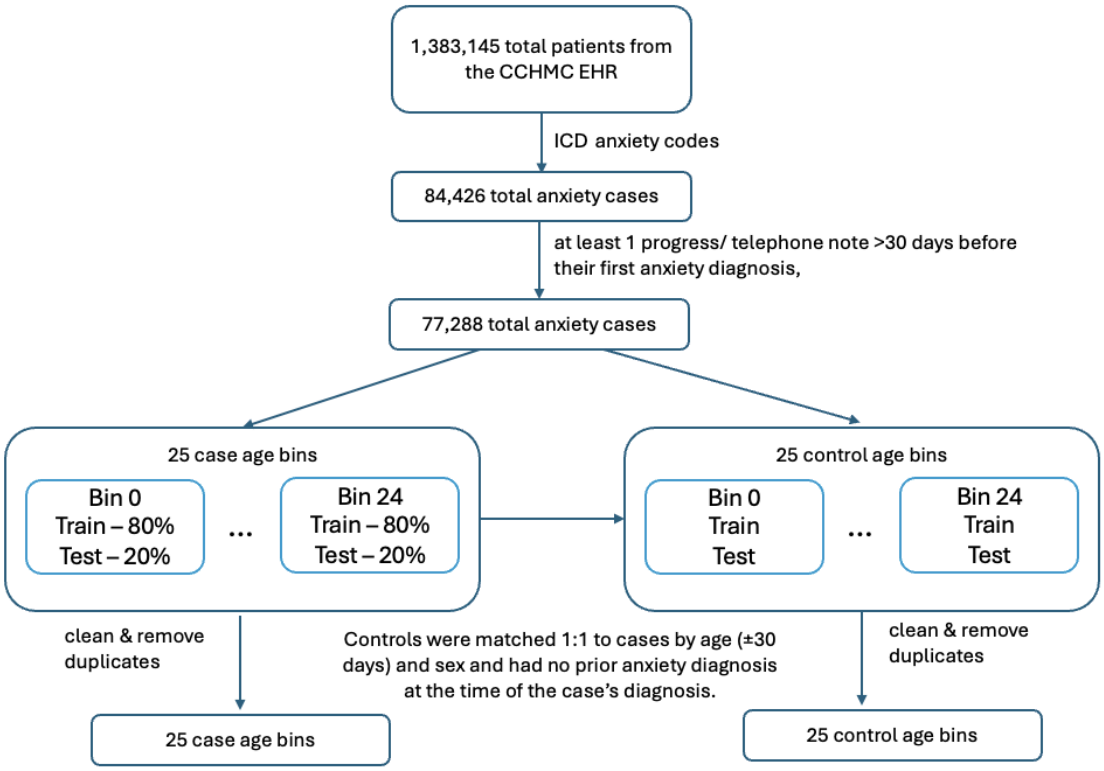

Figure 1. Flowchart illustrating the selection and binning of 77,288 anxiety cases from the Cincinnati Children's EHR. Cases were identified using ICD codes and filtered for patients with at least one note >30 days before diagnosis. All the anxiety cases were divided into 25 age bins and split into 80% training and 20% test sets. Controls (drawn from the same patient pool) were matched 1:1 by age (±30 days) and sex, with no prior anxiety diagnosis. On average, 25 notes per patient were retained, duplicates were removed.

7. Given that a pretrained model was used for anxiety classification, please explain how your "gender-word debiasing" method does not introduce different bias with the use of gender-neutral pronouns. Please explain how it does not "introduce unrealistic symptom profiles".

Thank you for this insightful question. We agree that any intervention on clinical text must be carefully designed to avoid introducing new biases.

Our gender-word debiasing method was developed with this concern in mind. Rather than performing gender-swapping (e.g., replacing “he” with “she”), which can result in biologically implausible or clinically misleading narratives (e.g., “He reported mood changes around his period”), we use gender-neutral substitutions that preserve the semantic integrity of the original text. Specifically, we replace gendered pronouns (“he”/“she”) with “they” and substitute names with neutral placeholders such as “person1” or “person2.” These replacements are designed to avoid altering symptom descriptions or introducing additional differences between gender subgroups. We conducted manual reviews of the debiased text to confirm that symptom profiles remained clinically coherent (p. 6, ll. 208-212).

Regarding the model pipeline: the pretrained BERT model is initially exposed to the original, unaltered clinical text. Debiasing is applied only during the fine-tuning phase, where the model learns task-specific representations (in this case, anxiety classification). This approach allows us to assess the impact of gender-neutral language on downstream predictions (see p.11, ll. 326-329).

8. While you mention it is above random, a diagnostic model with an accuracy of 0.61 is VERY low. Please better explain the usefulness of your analysis given that the classification model performance is so poor.

Thank you for highlighting this point. We agree that a diagnostic model accuracy of 0.61 is modest. However, in this study our intent was not to optimise predictive performance, but rather to employ the model as a diagnostic probe for examining bias in clinical text. In this context, the absolute performance is less critical than the model’s ability to reveal relative differences across demographic groups. Prior work has shown that lower-performance models can in fact make bias more visible, since their predictions are more influenced by spurious features of the inputs [1,2]. Thus, while the 0.61 accuracy highlights the limitations of the diagnostic model, it does not diminish its utility for the specific aim of probing demographic bias.
We have clarified that in the manuscript text (p. 13, ll. 421-424).

1. Blanzeisky, W. and Cunningham, P. (2021). Algorithmic factors influencing bias in Machine Learning. *arXiv [cs.LG]*.
2. Ding, X., Xi, R., & Akoglu, L. (2024). Outlier Detection Bias Busted: Understanding Sources of Algorithmic Bias through Data-centric Factors. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 384-395.

9. “rnd_filt” is first mentioned in Table 1, and not clear what it refers to. I don’t particularly like the green vs red color schema. The green vs red vs bold is confusing (higher vs reduction vs best). Are different metrics reported for Org than other methods (higher values vs reduction)? Make sure all abbreviations like org, unc, and av fr (Table 3) are defined. In Table 2, missing green in bin 12 rows. Also, what does “non-privileged” group mean in caption, any reason not to be more specific?

Thank you for your detailed and helpful feedback. We have addressed each of your points as follows:

“rnd_filt” refers to random filtering, a baseline method where a random subset of uncertain predictions is removed. We have now defined this term explicitly throughout the manuscript.

We appreciate your concern regarding the clarity of the visual encoding. We have added a new column to Tables 2 and 3 that reports the difference (Δ) between female and male FNR values. A downward arrow indicates a reduction in this difference compared to the Original method, highlighting improvements in fairness. Bold formatting is used to emphasize the lowest FNR differences within each bin. Accordingly, results for Bin 12 in Table 3 have been updated to reflect this markup. All methods, including Orig (Original), report the same set of metrics.

We have now fully rewritten the abbreviations: Orig to Original model, Av fr to average frequency. Unc is defined as the percentage of uncertain predictions.

We have clarified this in the captions (see p. 8 and p.9).

We agree that the term “non-privileged” could be more specific. In our context, the “non-privileged” is female. We have revised all the captions to specify this explicitly.

10. There are some sentences that are missing citations. For example, the first sentences in the introduction and conclusions. “Additionally, clinicians often have concerns about the interpretability of predictions…” also needs a citation. Further, citations are sometimes after et al. and sometimes at end of sentence.

Thank you for your helpful observations. We have reviewed the manuscript and added citations to general claims in both the introduction and conclusion [1,2] (p.1, ll. 23-24; p.13, ll. 429-430). For example, the sentence regarding clinicians’ concerns about interpretability now includes a citation to recent work on explainable AI in healthcare [3] (p. 2, ll. 39-40).

In addition, we have standardized citation formatting throughout the manuscript. Citations are now consistently placed at the end of the relevant logical unit (typically the sentence), also for the sentences directly referencing a named author (e.g., “Ribeiro et al. (2016)”).

1. COVID-19 Mental Disorders Collaborators. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *Lancet*, 398 (10312), pp.1700–1712.
2. Warner, E. N. et al. (2023). Developmental epidemiology of pediatric anxiety disorders. *Child and adolescent psychiatric clinics of North America*, 32 (3), pp.511–530.
3. Hou, J. and Wang, L. L. (2025). Explainable AI for clinical outcome prediction: A survey of clinician perceptions and preferences. *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2025, pp.215–224.

11. There are some typos, as well as excessive paragraph division. Paragraphs should generally be more than two lines/sentences.

Thank you for your careful reading and detailed feedback.

The typos I noticed: “This is particularly challenging in mental health care the primary source of information in mental health care” doesn’t work as a sentence, “selection bias” is the only type of bias not italicized in the text, switching from using entire number to k to represent a thousand (ie “7,810,849 notes for these 73k patients”), “This cohort contains demographic data (sex and race)” and age according to the next subsection, “model authors” lacks possession, “(see section Descriptive Analysis in Results)” on page 4 would be better as “refer to” and lacks a period at the end of the sentence, bin capitalization is inconsistent (noticed on page 5), av vs Av. In Table 3, and “representations16,are” missing space after comma.

Thank you for your careful reading and detailed feedback. We have addressed the issues you raised in the manuscript text (i.a., p. 2, ll. 42-43, l. 46; p. 4, l. 121, l. 122; p. 5, l. 163, l. 138).

12. Please explain how the code can is confidential and can not be shared. The lack of code greatly reduces the reproducibility and contribution of the work.

In this case, the code cannot be shared due to data governance restrictions, as it is tightly linked to confidential data pipelines. We are however committed to supporting reproducibility as much as possible within these constraints. To that end, we have provided

detailed descriptions of the model architecture, data preprocessing steps, and evaluation procedures in the manuscript. We are also happy to provide pseudocode upon request.

We have provided the revised code statement accordingly (p.13, ll. 436-439).

Dear Editor,

Thank you for the thoughtful feedback and constructive comments provided by you and the reviewers on our manuscript entitled *“A Data-Centric Approach to Detecting and Mitigating Demographic Bias in Pediatric Mental Health Text: A Case Study in Anxiety Detection.”* We greatly appreciate the time and effort invested in reviewing our work. The suggestions have been invaluable in improving our manuscript.

In response to the reviewer's concerns, we have amended our contribution to not overclaim regarding generalizability of our model-agnostic framework, and we have publicly released the cleaned code with synthetic examples to facilitate reproducibility.

For your convenience, we have included a detailed response to the reviewer's comment below. We have also outlined all revisions in the attached table, as requested, with detailed descriptions provided in the right-hand column. The completed table is included as a Related Manuscript file alongside the updated manuscript. All changes in the revised manuscript are highlighted in red for easy reference.

We hope these revisions satisfactorily address all concerns. Please let us know if any further clarification or adjustments are needed.

Sincerely,
Julia Ive

We thank the reviewer for the insightful comment. Below, we provide a detailed response to the point raised.

Point-by-point responses to Reviewers’ Comments.

Reviewers' comments:
Reviewer #1 (Remarks to the Author):

The raised concerns by this reviewer are addressed for the most part, and this reviewer appreciates the authors' efforts. However, the one major concern seems to be still there. The concern was about demonstrating the generalizability of the methods beyond one specific model. At this point, the paper only justifies why that one certain model was chosen, but falls short of demonstrating the generalizability beyond that.

The new text reads "Nevertheless, our proposed de-biasing framework is model-agnostic and can be integrated with a wide range of AI systems, including those with larger context windows." This is exactly the claim that needed strong evidence (beyond only the model used, which is arguably less capable than the modern ones).

As for the authors' response to R3's comments, they have mostly addressed the raised concerns. They seemed fine to me. The only minor note is about the very last comment about not sharing the code publicly. Even if their code is "tightly linked to confidential data pipelines," they may want to release the portions that are not directly related to the internal data. Admittedly, this could sometimes be challenging, but without that, the impact could be very limited.

We really appreciate the reviewer's acknowledgment of our efforts and the feedback regarding the claim of generalizability. We agree that the original wording may have overstated the generalizability claim. In the revised manuscript, we have toned down the claim to state that *"our proposed de-biasing framework has the potential to be generalizable across different model types. However, future work is needed to empirically validate this across ML models."*

Regarding code availability, we recognize the importance of reproducibility and impact. To address this, we have released a public version of the code for our models and bias detection metrics, along with synthetic data, in a dedicated repository: https://github.com/julia-ive/bias-pediatric-anxiety. This code version has been carefully sanitized to remove sensitive components while closely replicating the functionality used in this study.