nature portfolio

Peer Review File

Natural Language Processing of Youth Speech Predicts Psychopathology Across Adolescence

Corresponding Author: Mr Chase Antonacci

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

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 paper investigates several NLP techniques—including LIWC, TF-IDF, LDA, and transformer-based models—to predict Internalizing Problems scores and future diagnostic outcomes using an in-house dataset of approximately 200 participants followed over four years. The authors identified intriguing patterns in word usage and discussion topics that may signal potential mental disorders.

The sample size is reasonable, and the four-year longitudinal design makes the dataset particularly valuable, although it is unfortunate that the data are not publicly available. The NLP methods employed are relatively basic but appropriate for the study's goals.

The analysis is thorough, and the conclusions appear well-supported by the results.

Overall, I found this paper enjoyable to read. It is clear, well-structured, and presents interesting analyses and findings.

(Remarks on code availability)
Code is not available yet.

Reviewer #2

(Remarks to the Author)
Summary
This longitudinal study applies multiple NLP techniques (LIWC, TF-IDF, LDA, RoBERTa) to stress interview transcripts from 204 youth (ages 9-13 at baseline) to predict internalizing psychopathology up to 6 years later. Key findings demonstrate that linguistic features substantially outperform expert stress ratings, explaining 16.2% of variance versus 7.0% for baseline models. Across methods, linguistic style proved more predictive than explicit emotional content. A novel interpretability technique for transformer embeddings revealed clinically intuitive risk markers (physical violence, social exclusion) and protective factors (structured activities, healthcare access). Notably, data-driven semantic dimensions significantly predicted future clinical diagnoses while expert-rated cumulative stress severity did not.

Originality and Significance
This work makes several highly original contributions, presenting a new and important conceptualization of the common phenomenon of heterogeneous mental health outcomes following early adversity. It represents the first application of comprehensive NLP methods to in-depth stress narratives for predicting adolescent psychopathology onset, introducing a novel interpretability technique that successfully "opens the black box" of transformer embeddings. The finding that linguistic style outperforms explicit content extends foundational psycholinguistic theory to longitudinal clinical prediction in youth populations. This study addresses a critical "prediction gap" in identifying at-risk youth and provides compelling proof-of-concept for scalable, objective risk assessment tools that could transform early intervention targeting.

Positive Findings

The study demonstrates methodological rigor through its longitudinal design, high-quality clinical assessments (validated measures, expert consensus ratings), sophisticated audio processing pipeline, and proper cross-validated modeling. The methodology section provides a comprehensive description of all procedures. The complementary multimodal NLP approach provides robust convergent validity, with multiple independent approaches identifying function words as key predictors. The results are presented clearly and logically. The discussion appropriately addresses main topics, clearly articulating both the strengths and limitations of the work. The clinical interpretability of findings makes the work relevant to practitioners and demonstrates thoughtful engagement with the real-world implications of this research.

Justification for Revision
Minor revisions are needed to address several concerns: (0) figure referring inconsistency and refinement to improve clarity, (1) ethical considerations regarding implementation of predictive tools must be discussed in the introduction given the sensitive nature of identifying at-risk youth, (2) consent/assent procedures need clarification since the majority of baseline participants were minors, and (3) statistical rigor requires enhancement through addition of confidence intervals.

1. Figures: Systematic review of the manuscript reveals inconsistent figure citations throughout. Lines 285-299 incorrectly cite 'Figure 1' panels when referring to Figure 2 (LIWC results). Lines 318 and 322 incorrectly cite 'Figure 2' (TF-IDF results). Please perform a complete systematic check of all figure citations against figure legends and correct all instances. Additionally, line 326 is missing a figure citation for the LDA performance plot (should be Figure 4a). This affects manuscript readability and must be addressed.
• Figure 2e (scree plot): Could be removed from main figure and mention only in text. The elbow is clear and the figure takes valuable space
• Figure 3d (scree plot): Similarly, remove from main figure and reference in text only
• Figure 3b: Check for text cutoff below Y-axis title when zoomed out. Ensure all axis labels are fully visible and properly formatted

2. Ethical considerations discussion (Introduction)
• Add a dedicated paragraph in the introduction discussing ethical implications of implementing psychopathology prediction tools in youth:
o Concerns about stigmatization and labelling effects
o Potential for discrimination or misuse
o Importance of human clinical oversight
o Balance between early intervention benefits and potential harms
• Consider citing relevant bioethics literature on predictive algorithms in healthcare

3. Consent/assent procedures clarification (Methods)
• Lines 110-111 state "All participants provided written informed consent" but majority were minors (ages 9-13)
• Clarify the consent/assent procedure:
o Did minors provide written assent while parents/guardians provided consent?
o Were participants re-consented when reaching age of majority during follow-up?

4. Add confidence intervals
• Could be important to report bootstrapped 95% CIs for all R^2 estimates, regression coefficients, and odds ratios in text and figures, this would help assess stability of performance estimates
• Same with cross-validation folds, for the same objective

5. Comprehensive diagnostic performance metrics
• For all logistic regression models predicting diagnostic outcomes, I would suggest reporting: (1) ROC-AUC with 95% confidence intervals; (2) sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) at the threshold that maximizes Youden's index; and (3) consider adding an ROC curve figure comparing the baseline model to linguistic models. These metrics are essential for evaluating the clinical utility of linguistic markers for risk stratification and are standard in prognostic model evaluation. If authors decide otherwise, please add a justification, address it as a limitation, or encourage future work to do those analyses.

6. Feature stability analysis
• Report how frequently each feature is selected across CV folds in elastic net models (particularly for Figures 2c, 3a). This could be presented as "selection frequency" percentages in the coefficient plots or mentioned in text. This information supports the robustness of feature importance claims and helps distinguish stable predictors from those selected inconsistently.

7. Effect size contextualization
• Provide clinical context for 16.2% variance explained
• Compare to other clinical screeners or typical effect sizes in developmental psychopathology
• This contextualization would help readers interpret whether linguistic prediction represents a meaningful advance over existing approaches

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have thoughtfully and thoroughly addressed all of my concerns, significantly strengthening the paper's statistical rigor, methodological transparency, and ethical framing. This is an outstanding, highly original study that will make a valuable contribution to the field.

(Remarks on code availability)
The authors have done a commendable job making their analytical pipeline publicly available, which greatly enhances the methodological transparency of the study.

Response to Referees

February 25, 2026

Rebecca Cooney, PhD
Editor-in-Chief
Nature Mental Health

RE: NATMH-25-1964

Dear Dr. Cooney,

We appreciate the thoughtful comments from the reviewers and are grateful for the time and effort they have devoted to our manuscript. We believe that our manuscript is stronger for the revisions we have made in response to their comments and suggestions. We provide responses to their comments below. In this letter, reviewer comments are presented in plain text, our responses are indented, and quoted sections of the manuscript are in blue.

Reviewer 1

1. This paper investigates several NLP techniques—including LIWC, TF-IDF, LDA, and transformer-based models—to predict Internalizing Problems scores and future diagnostic outcomes using an in-house dataset of approximately 200 participants followed over four years. The authors identified intriguing patterns in word usage and discussion topics that may signal potential mental disorders.

   The sample size is reasonable, and the four-year longitudinal design makes the dataset particularly valuable, although it is unfortunate that the data are not publicly available. The NLP methods employed are relatively basic but appropriate for the study's goals. The analysis is thorough, and the conclusions appear well-supported by the results. Overall, I found this paper enjoyable to read. It is clear, well-structured, and presents interesting analyses and findings.

   (Remarks on code availability): Code is not available yet.

   We thank the reviewer for their thoughtful and positive evaluation of our manuscript, and for noting the strengths of the longitudinal design and analytic approach. We apologize that the code was not available at the time of the initial review. Since we submitted the manuscript, however, we have fully documented the transcription pipeline and all analytic procedures and have made the complete codebase publicly available. Consistent with *NMH's* reproducibility guidelines, the code is now hosted on both GitHub and Code Ocean:

   - GitHub: https://github.com/eugiampetruzzi/nlp_stress_antonacci
   - Code Ocean: https://codeocean.com/capsule/7267047/tree

   With respect to data availability, we should note that the dataset contains sensitive information about clinical interviews with minors in which we assess lifetime stress and trauma exposure, including substantial protected health information (PHI); accordingly, IRB approval and participant consent do not permit public sharing of speech-derived

data. We report this constraint in the Data Sharing and Availability section of the manuscript (page 27):

Code and Data Availability: All code used for transcription, preprocessing, and analysis is publicly available at https://github.com/eugiampetruzzi/nlp_stress_antonacci. The full dataset contains information about sensitive clinical interviews with minors assessing lifetime history of stress and trauma exposure and, therefore, includes protected health information (PHI). Consequently, we cannot de-identify the data, and neither IRB approval nor participant consent permit public sharing of the speech-derived data.

Reviewer 2

Summary: This longitudinal study applies multiple NLP techniques (LIWC, TF-IDF, LDA, RoBERTa) to stress interview transcripts from 204 youth (ages 9-13 at baseline) to predict internalizing psychopathology up to 6 years later. Key findings demonstrate that linguistic features substantially outperform expert stress ratings, explaining 16.2% of variance versus 7.0% for baseline models. Across methods, linguistic style proved more predictive than explicit emotional content. A novel interpretability technique for transformer embeddings revealed clinically intuitive risk markers (physical violence, social exclusion) and protective factors (structured activities, healthcare access). Notably, data-driven semantic dimensions significantly predicted future clinical diagnoses while expert-rated cumulative stress severity did not.

Originality and Significance: This work makes several highly original contributions, presenting a new and important conceptualization of the common phenomenon of heterogeneous mental health outcomes following early adversity. It represents the first application of comprehensive NLP methods to in-depth stress narratives for predicting adolescent psychopathology onset, introducing a novel interpretability technique that successfully "opens the black box" of transformer embeddings. The finding that linguistic style outperforms explicit content extends foundational psycholinguistic theory to longitudinal clinical prediction in youth populations. This study addresses a critical "prediction gap" in identifying at-risk youth and provides compelling proof-of-concept for scalable, objective risk assessment tools that could transform early intervention targeting.

Positive Findings: The study demonstrates methodological rigor through its longitudinal design, high-quality clinical assessments (validated measures, expert consensus ratings), sophisticated audio processing pipeline, and proper cross-validated modeling. The methodology section provides a comprehensive description of all procedures. The complementary multimodal NLP approach provides robust convergent validity, with multiple independent approaches identifying function words as key predictors. The results are presented clearly and logically. The discussion appropriately addresses main topics, clearly articulating both the strengths and limitations of the work. The clinical interpretability of findings makes the work relevant to practitioners and demonstrates thoughtful engagement with the real-world implications of this research.

Justification for Revision: Minor revisions are needed to address several concerns: (0) figure referring inconsistency and refinement to improve clarity, (1) ethical considerations regarding implementation of predictive tools must be discussed in the introduction given the sensitive nature of identifying at-risk youth, (2) consent/assent procedures need clarification since the majority of baseline participants were minors, and (3) statistical rigor requires enhancement through addition of confidence intervals.

We thank the reviewer for their thoughtful and detailed evaluation of our manuscript. We appreciate their recognition of the originality and significance of the work, including the longitudinal design, the multimodal NLP framework, and the contribution of our interpretability approach for transformer-based models. We are also grateful for the reviewer's careful engagement with the clinical and methodological implications of the findings. We have addressed each of the points that the reviewer raised to improve clarity of presentation, strengthen the ethical aspects of the study, and enhance statistical reporting. Below, we provide point-by-point responses to the specific requests for revision.

1. Figures: Systematic review of the manuscript reveals inconsistent figure citations throughout. Lines 285-299 incorrectly cite 'Figure 1' panels when referring to Figure 2 (LIWC results). Lines 318 and 322 incorrectly cite 'Figure 2' (TF-IDF results). Please perform a complete systematic check of all figure citations against figure legends and correct all instances. Additionally, line 326 is missing a figure citation for the LDA performance plot (should be Figure 4a). This affects manuscript readability and must be addressed.

We appreciate the reviewer identifying these inconsistencies. We have reviewed the entire manuscript and have corrected all figure citations to ensure they align with the corresponding figure legends and panels. Specifically, we have updated citations in the dictionary-based analysis section to correctly refer to Figure 2 panels; corrected citations in the TF-IDF section to accurately reference Figure 3; and inserted the missing citation for the LDA performance plot to reference Figure 4a.

• Figure 2e (scree plot): Could be removed from main figure and mention only in text. The elbow is clear and the figure takes valuable space.
• Figure 3d (scree plot): Similarly, remove from main figure and reference in text only

We appreciate this comment and agree that the scree plots occupy valuable space in the main figures. We have moved Figures 2e and 3d from the main manuscript to the Supplemental Information as Supplementary Figures S2a and S2b, respectively. We have updated the text in the Results section to reflect this change, referencing their new location in the Supplement. Changes to the manuscript (pages 12-13):

Results:

To identify broader linguistic dimensions, we conducted a PCA on the LIWC features, retaining two components based on the scree plot (Figure S2a)… To consolidate this stylistic signal into a more stable dimension, we conducted a PCA on the TF-IDF features, retaining three components based on the scree plot (Figure S2b).

Figure 2: *Prediction of Youth Internalizing Outcomes from Dictionary-Based Linguistic Features*

Redacted

Note. (a) Participant-level LIWC feature usage, organized by hierarchical clustering. Rows represent participants, columns are LIWC variables grouped by theme, and cell color indicates standard usage (*z*-score). **(b-d)** Performance and interpretation of an elastic net model predicting continuous internalizing scores from LIWC features. **(b)** Model performance on held-out data, plotting predicted versus observed scores from the pooled out-of-fold predictions in 5-fold nested cross-validation. **(c)** Standardized coefficients of the top risk and protective features from the full-sample model used for interpretation. **(d)** Incremental variance explained (cross-validated R^2), showing the added predictive contribution of LIWC features beyond demographics and cumulative stress. **(e-g)** Principal component analysis (PCA) identifying latent linguistic dimensions. **(e)** Biplot of the first two components, revealing PC1 as a dimension of structured to dynamic/social language and PC2 a dimension of cognitive to motivational language. **(f)** Association between PC1 and internalizing scores. **(g)** Predicted probability of a future mood or anxiety diagnosis as a function of PC1 score (covariate adjusted); rug marks show participants with (red) and without (gray) a diagnosis. For panels with fitted lines **(b, f, g)**, shaded bands indicate 95% confidence intervals around the fitted line.

Figure 3: *Prediction of Youth Internalizing Outcomes from Specific and Latent Lexical Features*

Redacted

Note. (a-b) Prediction of internalizing problems from individual word TF-IDF scores. **(a)** Standardized coefficients of the top word-level predictors from the full-sample elastic net model used for interpretation. **(b)** Model performance on held-out data, plotting predicted versus observed scores from pooled out-of-fold predictions in 5-fold nested cross-validation. **(c)** Top 15 words whose TF-IDF scores were significantly associated with internalizing problems in an exploratory bivariate analysis (FDR q<.05). **(d-f)** Principal component analysis (PCA) identifying latent lexical dimensions as predictors. **(d)** Top 10 word loadings for the first three principal components, interpreting PC1 as "Function Words," PC2 as "Interpersonal Focus," and PC3 as "Visceral Experience." **(e)** Association between PC1 ("Function Words") scores and YSR internalizing scores. **(f)** Predicted probability of future mood or anxiety diagnosis as a function of PC1 score (covariate-adjusted); rug marks indicate participants with (red) and without (gray) a diagnosis. For panels with fitted lines **(b, e, f)**, shaded bands indicate 95% confidence intervals around the fitted line.

Supplement:

Figure S2: LIWC and TF-IDF Scree Plots

Redacted

Note. Scree plots show eigenvalues across principal components for the **(a)** LIWC and **(b)** TF-IDF feature sets; the elbow criterion supported retaining two components for LIWC and three components for TF-IDF, consistent with the PCA solutions used in the main analyses. No scree plot is shown for the LDA pipeline because topics were reduced *a priori* by applying PCA to the 500-topic participant-topic matrix and retaining 20 orthogonal components to yield a stable, interpretable thematic space for prediction (see Methods). No scree plot is shown for sentence embeddings because PCA was not applied to the RoBERTa embeddings; instead, contextualized embeddings were used directly in predictive models, and UMAP was applied for low-dimensional visualization and semantic-structure analyses.

• Figure 3b: Check for text cutoff below Y-axis title when zoomed out. Ensure all axis labels are fully visible and properly formatted

> We have carefully inspected Figure 3b and all other performance plots (Figures 2b, 4a, and 5b) for potential text cutoff or formatting issues. While we did not see any cutoff issues in our high-resolution source files, we appreciate that problems can arise during PDF rendering or when viewing figures at lower levels of resolution. To address this, we have (a) increased the left-hand margin between the y-axis title and the edge of the plot across all performance figures to ensure that "Observed Score" remains fully legible regardless of zoom level or compression; and (b) verified that all axis labels are properly formatted and fully visible in the revised high-resolution figures.

2. Ethical considerations discussion (Introduction): Add a dedicated paragraph in the introduction discussing ethical implications of implementing psychopathology prediction tools in youth:
   - Concerns about stigmatization and labelling effects
   - Potential for discrimination or misuse
   - Importance of human clinical oversight
   - Balance between early intervention benefits and potential harms
   - Consider citing relevant bioethics literature on predictive algorithms in healthcare

We appreciate the reviewer raising these important ethical considerations. We agree that it is important to address the ethical implications of using language-based predictive tools in youth mental health. Because these issues most directly involve interpretation, limitations, and potential translation of our findings, we felt they fit most naturally in the Discussion rather than the Introduction. Accordingly, in the revised manuscript we have added a dedicated paragraph to the Discussion framing language-based prediction not as a tool for deterministic labeling, but as a probabilistic signal intended to support (rather than replace) clinical judgment. We emphasize that, although our models outperform traditional risk markers, they nevertheless explain only a portion of the variance in future outcomes and, therefore, cannot and should not be interpreted as diagnostic or deterministic. We further state that the value of early risk identification lies in informing the delivery of timely, preventive, and resilience-focused support, rather than in labeling or pathologizing youth. When used with appropriate clinical oversight, scalable behavioral markers such as naturalistic speech may help direct limited preventive resources to those most likely to benefit, including youth who may be overlooked when using conventional screening approaches. At the same time, we acknowledge that there are risks related to stigmatization, discrimination, and misuse of language-based tools and emphasize the necessity of safeguards and human oversight in this research. Finally, we have incorporated discussion of several relevant bioethics publications addressing predictive algorithms in healthcare and precision psychiatry. Changes to the Discussion (page 19):

Given the potential of data-driven tools to improve risk stratification for youth mental health outcomes, it is also important to consider the ethical implications of any future translation of our findings. We want to emphasize here our position that language-based prediction should be treated as a probabilistic risk signal intended to support (rather than replace) clinical judgment; consistent with this, even our best-performing models explain only a portion of variance in later outcomes and, therefore, are not diagnostic or deterministic. This level of predictive performance is broadly consistent with other data-driven approaches to forecasting youth mental health (e.g., neurobiological risk models[59,60]), suggesting that such tools are best positioned to support early, preventive care by helping direct limited resources to youth most likely to benefit from them. Any clinical deployment would require safeguards to minimize unintended labeling or discrimination and to prevent misuse in higher-stakes contexts. These considerations are consistent with bioethics frameworks for precision psychiatry and child-centered medical AI, which emphasize transparent communication of uncertainty, privacy and data protection, fairness, and accountable human oversight when using predictive algorithms in healthcare[61,62].

3. Consent/assent procedures clarification (Methods): Lines 110-111 state "All participants provided written informed consent" but majority were minors (ages 9-13). Clarify the consent/assent procedure.
   - Did minors provide written assent while parents/guardians provided consent?
   - Were participants re-consented when reaching age of majority during follow-up?

   We appreciate the opportunity to clarify the consent and assent procedures in this study. For all study visits conducted when participants were under 18 years of age, written informed consent was obtained from a parent or legal guardian, and written assent was obtained from the participant. For follow-up assessments that were conducted after a participant reached the age of majority, participants provided their own written informed consent. We have revised the Methods section (Participants and Study Design) to

describe these procedures (page 5):

The study was conducted in compliance with all relevant laws and guidelines and was approved by the Stanford University Institutional Review Board. For study visits conducted when participants were under 18 years of age, written informed consent was obtained from a parent or legal guardian, and written assent was obtained from the participant. For follow-up assessments conducted after participants reached the age of majority (18 years), participants provided their own written informed consent. Participants were compensated for all study activities.

4. Add confidence intervals
   - Could be important to report bootstrapped 95% CIs for all R^2 estimates, regression coefficients, and odds ratios in text and figures, this would help assess stability of performance estimates
   - Same with cross-validation folds, for the same objective

We agree that reporting uncertainty is important for evaluating the stability and precision of model estimates; therefore, we have added uncertainty metrics for cross-validated performance throughout. Specifically, we computed nonparametric bootstrap 95% confidence intervals for pooled out-of-fold R^2 (1,000 iterations) by resampling participants' paired observed outcomes and nested-CV out-of-fold predictions; in addition, we report outer-fold performance (mean ± SD) to illustrate stability across folds. For consistency with the bootstrap procedure (which targets standard R^2, not adjusted R^2), we now report standard R^2 values with bootstrap 95% CIs throughout the manuscript.

For elastic net feature-importance models, coefficients are presented descriptively because these models were fit on the full dataset for interpretability (as noted in the Methods). Because elastic net involves shrinkage and variable selection, conventional coefficient CIs are not straightforward. Instead, we quantified stability via bootstrap refits (500 iterations), reporting selection frequencies (proportion of refits with non-zero coefficients) and coefficient distributions across resamples (Supplemental Fig. S1; see also Reviewer 2, Point 6).

We have also updated figure captions to clarify that, in panels with fitted relationships (e.g., Figs. 2b, 2f, 2g; 3b, 3e, 3f; 4a; 5b), shaded bands represent 95% confidence intervals around the fitted line. For panels summarizing variance partitioning or elastic net coefficients (e.g., Figs. 2c-d, 3a, 4b), we report uncertainty in the Results text and the Supplement (i.e., bootstrap CIs for pooled out-of-fold R^2; bootstrap feature stability/selection frequencies), given that adding more intervals directly to these panels would reduce interpretability.

Finally, we now report 95% confidence intervals for regression coefficients and odds ratios in the baseline (non-regularized) models and in downstream inferential models built on derived linguistic dimensions (e.g., PCA/UMAP-based regressions and logistic regressions). We have added these intervals to the Results text. Changes to the Results section (pages 12-16):

Baseline Models:

A baseline linear regression conducted using only demographic variables and SumSev to predict Internalizing Problems explained 20.73% of the variance (R^2=0.207, bootstrap 95% CI [0.107, 0.287]), with significant effects for SumSev (B=0.361, SE=0.076, p<.001, 95% CI [0.212, 0.511]) and female sex (B=5.854, SE=1.269, p<.001, 95% CI [3.352, 8.356]); however, when evaluated via 5-fold cross validation, the model's performance on the held-out data was substantially lower, accounting for 7.0% of the variance (pooled out-of-fold R^2=0.070, bootstrap 95% CI [-0.041, 0.161], outer fold R^2 mean ± SD=0.023±0.163). Sex (F) was the only significant variable to predict mood or anxiety diagnoses (B=1.119, SE=0.446, p=.012, 95% CI [0.277, 2.042]).

LIWC:

An elastic net model incorporating LIWC features significantly improved prediction, accounting for 16.2% of the variance in held-out data (pooled out-of-fold R^2=0.162; outer-fold R^2 mean ± SD=0.157±0.040; bootstrap 95% CI [0.047, 0.258]; Figure 2b), more than double that of the baseline model (Figure 2d)… In a linear regression controlling for covariates, only PC1 significantly predicted higher Internalizing Problems (B=0.405, SE=0.139, p=.004, 95% CI [0.130, 0.679]; Figure 2f), and its inclusion significantly improved model fit over the baseline (F(1, 192)=7.984, p=.005), increasing the explained variance to 23.9% (R^2=0.239, bootstrap 95% CI [0.114, 0.336]). Further, this dynamic/social dimension predicted diagnostic outcomes (Figure 2g); each one-standard-deviation increase in PC1 was associated with a 12.1% higher odds of receiving a future diagnosis (OR=1.121, 95% CI: [1.014, 1.251], p=.032); in contrast, the expert-rated SumSev score was not a significant predictor (p=.816).

TF-IDF:

An initial elastic net model performed comparably to the baseline model (pooled out-of-fold R^2=0.076; bootstrap 95% CI [-0.026, 0.157]; outer-fold R^2 mean ± SD=0.076±0.103; Figure 3b) but yielded a set of selected predictors that did not converge on a clear psychological theme (Figure 3a)… This data-driven "Function Words" dimension was a robust predictor of Internalizing Problems scores (B=0.258, SE=0.101, p=.011, 95% CI [0.060, 0.457]; Figure 3e) and improved model fit over the baseline (F(1, 192)=6.324, p=.013). PC2 (“Interpersonal Focus”) and PC3 (“Visceral Experience”) were not significant predictors (ps>.077). Further, in a logistic regression, each one-standard-deviation increase in PC1 was associated with a 7.6% increase in the odds of receiving a future diagnosis (OR=1.076, 95% CI [1.007, 1.155], p=.034; Figure 3f).

LDA:

Evaluated via 5-fold nested cross-validation, the model accounted for 13.3% of the variance in held-out Internalizing Problems scores (pooled out-of-fold R^2=0.133, bootstrap 95% CI [0.053, 0.204]; outer-fold R^2 mean ± SD=0.115±0.090; Figure 4a) and significantly improved model fit over baseline (F(5, 188)=3.602, p=.004)… For interpretability, we refit the elastic net model on the full sample and summarized the non-zero standardized coefficients as indicators of feature importance (Figure 4b; See Figure S1 for feature-selection stability analysis).

Sentence Embeddings:

The RoBERTa-based elastic net model achieved strong predictive performance for internalizing symptoms (pooled out-of-fold R^2=.156, bootstrap 95% CI [0.053, 0.242]; outer-fold R^2 mean ± SD=0.156±0.047, range [0.082, 0.200]; Figure 5b), underscoring the value added by capturing deep semantic content… A participant's position along this gradient was a robust predictor of their future mental health, correlating with UMAP Dimensions 1 (r=.316, p<.001, 95% CI [0.186, 0.434]) and 2 (r=.300, p<.001, 95% CI [0.169, 0.420]). Further, HDBSCAN clustering identified two distinct linguistic styles along this gradient, separating participants into "High-Risk" or "Low-Risk" style clusters that differed in the severity of their future symptoms (t(195)=4.433, p<.001, 95% CI for mean difference [3.002, 7.813]; Figure 5d). To avoid collinearity, we focused on the primary axis (UMAP 1), which remained a significant predictor of higher Internalizing Problems (B=0.435, SE=0.180, p=.017, 95% CI [0.080, 0.790]) and significantly improved model fit (F(1, 192)=5.838, p=.017) over baseline covariates. Finally, this latent semantic gradient was especially powerful in predicting clinical outcomes; each one-standard-deviation increase along the UMAP 1 axis was associated with a 15.1% increase in the odds of a participant receiving a future diagnosis (OR=1.151, 95% CI [1.023, 1.303], p=.023).

Figure Captions:

Figure 2: *Prediction of Youth Internalizing Outcomes from Dictionary-Based Linguistic Features*

Note. (a) Participant-level LIWC feature usage, organized by hierarchical clustering. Rows represent participants, columns are LIWC variables grouped by theme, and cell color indicates standard usage (z-score). **(b-d)** Performance and interpretation of an elastic net model predicting continuous internalizing scores from LIWC features. **(b)** Model performance on held-out data, plotting predicted versus observed scores from the pooled out-of-fold predictions in 5-fold nested cross-validation. **(c)** Standardized coefficients of the top risk and protective features from the full-sample model used for interpretation. **(d)** Incremental variance explained (cross-validated R^2), showing the added predictive contribution of LIWC features beyond demographics and cumulative stress. **(e-g)** Principal component analysis (PCA) identifying latent linguistic dimensions. **(e)** Biplot of the first two components, revealing PC1 as a dimension of structured to dynamic/social language and PC2 a dimension of cognitive to motivational language. **(f)** Association between PC1 and internalizing scores. **(g)** Predicted probability of a future mood or anxiety diagnosis as a function of PC1 score (covariate adjusted); rug marks show participants with (red) and without (gray) a diagnosis. For panels with fitted lines **(b, f, g)**, shaded bands indicate 95% confidence intervals around the fitted line.

Figure 3: *Prediction of Youth Internalizing Outcomes from Specific and Latent Lexical Features*

Note. (a-b) Prediction of internalizing problems from individual word TF-IDF scores. **(a)** Standardized coefficients of the top word-level predictors from the full-sample elastic net model used for interpretation. **(b)** Model performance on held-out data, plotting predicted versus observed scores from pooled out-of-fold predictions in 5-fold nested cross-validation. **(c)** Top 15 words whose TF-IDF scores were significantly associated with internalizing problems in an exploratory bivariate analysis (FDR q<.05). **(d-f)** Principal component analysis (PCA) identifying latent lexical dimensions as predictors. **(d)** Top 10 word loadings for the first three principal components, interpreting PC1 as "Function Words," PC2 as "Interpersonal Focus," and PC3 as "Visceral Experience." **(e)** Association between PC1 ("Function Words") scores and YSR internalizing scores. **(f)** Predicted probability of future mood or anxiety diagnosis as a function of PC1 score (covariate-adjusted); rug marks indicate participants with (red) and without (gray) a

diagnosis. For panels with fitted lines **(b, e, f)**, shaded bands indicate 95% confidence intervals around the fitted line.

Figure 4: *Predicting Internalizing Outcomes from Latent Thematic Content*

Note. (a) Performance of the elastic net model on held-out data, predicting continuous internalizing scores from 20 PCA-reduced topic dimensions; plots show pooled out-of-fold predictions from 5-fold nested cross-validation; shaded band indicates the 95% confidence interval around the fitted line. **(b)** Standardized coefficients of the predictors retained by the elastic net model refit on the full sample for interpretability, yielding five topic components and key demographic and stress covariates; bootstrap CIs for pooled out-of-fold R^2 are reported in the Results and bootstrap feature-selection stability is reported in Figure S1. **(c)** Word clouds providing qualitative interpretation of the five retained topic components; titles report the corresponding standardized coefficient (*B*) from the interpretive elastic net refit. Within each cloud, word size is proportional to loading strength on the component and color reflects the direction of association with internalizing problems (red=positive; green=negative).

Figure 5: *Contextual Sentence Embeddings Predict Youth Internalizing Problems*

Note. (a) Illustrative sentences with the highest (risk) and lowest (protective) predicted YSR scores (representing the top 100 (0.3%) and bottom 100 (0.3%) of all sentences; see Supplement for full list). Predicted YSR score column reflects the sentence-level risk score, which was calculated via a dot product projection of the elastic net model's coefficients onto each sentence embedding. Thematic content was qualitatively analyzed and is displayed for each sentence. **(b)** Predictive performance of the elastic net model on test data using participant-level RoBERTa embeddings, plotting predicted versus actual scores from a 5-fold nested cross-validation. **(c)** A UMAP projection of the participant-level embeddings, revealing two distinct, data-driven linguistic style clusters ("Style A" and "Style B") identified via HDBSCAN. **(d)** Violin plot demonstrating that internalizing scores differ significantly between the two linguistic style clusters.

5. Comprehensive diagnostic performance metrics

For all logistic regression models predicting diagnostic outcomes, I would suggest reporting: (1) ROC-AUC with 95% confidence intervals; (2) sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) at the threshold that maximizes Youden's index; and (3) consider adding an ROC curve figure comparing the baseline model to linguistic models. These metrics are essential for evaluating the clinical utility of linguistic markers for risk stratification and are standard in prognostic model evaluation. If authors decide otherwise, please add a justification, address it as a limitation, or encourage future work to do those analyses.

We appreciate this comment and agree that ROC-AUC and related operating characteristics can help readers gauge diagnostic discrimination. Therefore, we followed the reviewer's suggestions and, for each logistic regression model predicting mood/anxiety diagnoses, now report ROC-AUC with nonparametric bootstrap 95% confidence intervals, as well as sensitivity, specificity, PPV, and NPV at the threshold maximizing Youden's J. We also added ROC curve comparisons between the baseline model (demographics + expert-rated cumulative stress severity) and language-augmented models across LIWC, TF-IDF, and RoBERTa approaches. Because we included these logistic regressions primarily to test whether linguistic dimensions predict diagnostic outcomes beyond conventional risk markers (rather than to develop or validate a deployable clinical classifier), and because they were not cross-validated or externally validated, we present these diagnostic performance summaries in the

Supplement as descriptive/exploratory metrics (Table S3; Fig. S3), while keeping the main Results focused on inferential estimates (odds ratios and confidence intervals for linguistic predictors and SumSev). Across models, discrimination gains from adding language were modest, which is expected given the community-based nature of our sample, the limited number of diagnostic events, and our emphasis on incremental prediction over covariates rather than on classifier optimization and validated clinical deployment. Changes: Supplement Table S3 and Fig. S3; Results text updated to reference these materials (page 13).

Results: Complementary diagnostic operating characteristics (e.g., ROC-AUC, sensitivity, and specificity) are reported in the Supplement (Table S3; Fig. S3).

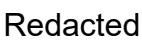
Supplement:

Table S3: Diagnostic Operating Characteristics of Baseline and Language-Augmented Logistic Regression Models

	AUC	Bootstrap 95% CI	Youden Threshold	Sensitivity	Specificity	PPV	NPV
Baseline Model (Demographics + SumSev)	0.704	[0.616, 0.785]	0.240	0.634	0.712	0.356	0.885
LIWC Model (Demographics + SumSev + LIWC PCs)	0.732	[0.648, 0.812]	0.238	0.683	0.718	0.378	0.900
TF-IDF Model (Demographics + SumSev + TF-IDF PCs)	0.735	[0.649, 0.814]	0.182	0.780	0.620	0.340	0.918
LDA Model (Demographics + SumSev + LDA PCs)	0.709	[0.623, 0.792]	0.238	0.634	0.706	0.351	0.885
RoBERTa Model (Demographics + SumSev + RoBERTa PCs)	0.729	[0.641, 0.805]	0.226	0.683	0.724	0.384	0.901

Note. Diagnostic operating characteristics of baseline and language-augmented logistic regression models. Area under the ROC curve (AUC) is reported with nonparametric bootstrap 95% confidence intervals. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are computed at the probability threshold maximizing Youden's J statistic. Models compare a baseline covariate model (demographics + expert-rated cumulative stress severity) against models augmented with linguistic dimensions derived from LIWC, TF-IDF, LDA, and RoBERTa features. Prevalence of mood/anxiety disorders in the sample = 20.1%. SumSev = expert-rated cumulative stress severity.

Figure S3: ROC Curves for Diagnostic Prediction: Baseline vs Language-Augmented Models



Note. Receiver operating characteristic (ROC) curves comparing a baseline logistic regression model (demographics + expert-rated cumulative stress severity, SumSev) to models augmented with language-derived predictors for mood/anxiety diagnosis (prevalence = 20.1%). Across methods, language features yielded modest improvements in overall discrimination relative to baseline (AUCs shown in-panel; operating characteristics and bootstrap 95% CIs reported in Table S3). Importantly, consistent with the inferential aim of these analyses, language-derived dimensions significantly predicted diagnostic outcomes in three of the four NLP approaches (LIWC PCs, TF-IDF PCs, and RoBERTa/UMAP1), whereas SumSev did not significantly predict diagnosis in these models; only the LDA-derived topic components did not show a significant association. Dashed diagonal indicates chance performance.

6. Feature stability analysis

Report how frequently each feature is selected across CV folds in elastic net models (particularly for Figures 2c, 3a). This could be presented as "selection frequency" percentages in the coefficient plots or mentioned in text. This information supports the robustness of feature importance claims and helps distinguish stable predictors from those selected inconsistently.

We thank the reviewer for suggesting that we demonstrate the robustness of the features identified by the elastic net models. Because elastic net performs shrinkage and variable selection, we assessed feature stability using a bootstrap-based selection analysis rather than classical coefficient confidence intervals. Specifically, for each method we refit the full-sample elastic net model across 500 nonparametric bootstrap resamples and computed each predictor's selection frequency (proportion of refits with a non-zero coefficient). We report these results in the Supplement (Fig. S2), alongside the full-sample standardized coefficients used for interpretability in the main text.

Bootstrap resampling indicated that feature selection was highly stable for the LIWC, TF-IDF, and LDA models (selection frequencies ≥0.75 for nearly all top predictors / all retained topic components), supporting the robustness of the primary linguistic markers discussed in the Results. For RoBERTa sentence embeddings, selection frequencies for individual embedding dimensions were more moderate, which is expected given the high-dimensional, highly collinear embedding space. In this setting, many embedding dimensions carry overlapping information, so different bootstrap refits may select different (but redundant) subsets of embeddings while yielding similar out-of-sample performance, reflecting stability of the predictive signal even when the exact selected coefficients vary. Changes to the manuscript (pages 12-15; Supplement Fig. S1):

Figure S1: Bootstrap Stability of Elastic Net Feature Selection

Redacted

Note. Bootstrap stability of elastic net feature selection across NLP methods (B=500 refits). Bars show the selection frequency (proportion of bootstrap refits with non-zero coefficients) for the top 15 predictors from each full-sample elastic net model; within each panel, predictors are ordered by selection frequency. Colors indicate the direction of association in the full-sample model (risk factor: positive coefficient; protective: negative coefficient). The dashed vertical line marks a 75% selection-frequency threshold used to denote highly stable features. Feature selection was highly stable for LIWC, TF-IDF, and LDA (LIWC: 13/15 predictors; TF-IDF: 14/15 predictors; LDA: 15/15 predictors) with selection frequencies ≥75% across refits. In contrast, for RoBERTa sentence embeddings, the top-ranked embedding dimensions showed more moderate selection frequencies, consistent with a high-dimensional, collinear embedding space, predictive signal is distributed across many correlated dimensions such that the specific embedding coefficients selected can vary between refits, even as overall model performance remains robust.

Results:

LIWC: The final elastic net model fit on the full sample for feature interpretation (Figure 2c) confirmed that female sex (B=1.299) and SumSev (B=0.795) predicted Internalizing Problems, but also indicated that language features were comparably strong (or stronger) predictors (see Supplemental Figure S1 for bootstrap feature-selection stability)…For interpretability, we refit the elastic net model on the full sample and summarized the non-zero standardized coefficients as indicators of feature importance (Figure 4b; See Figure S1 for feature-selection stability analysis).

7. Effect size contextualization
   - Provide clinical context for 16.2% variance explained
   - Compare to other clinical screeners or typical effect sizes in developmental psychopathology
   - This contextualization would help readers interpret whether linguistic prediction represents a meaningful advance over existing approaches

   We agree that contextualizing the variance explained is important for interpreting the clinical significance of our findings. While 16.2% variance explained may appear modest in absolute terms, it represents more than a two-fold improvement over a baseline model that included expert-rated cumulative severity of stress and demographic risk factors (7.0%), which themselves are widely used indicators of risk in developmental research. Importantly, it is clear from research in developmental psychopathology that long-term mental health outcomes are shaped by numerous interacting biological, social, and environmental influences, making large effect sizes from any single baseline measure uncommon, particularly in longitudinal cross-validated prediction models that span several years. In this context, a single behavioral phenotype derived from a brief, one-time speech sample that accounts for over 16% of the variance in future symptoms represents a meaningful and substantial effect.

   In the revised Discussion, we have provided this context and compare our results to performance estimates reported for other prognostic approaches, including digital/clinical phenotyping and neurobiological risk models risk forecasting. We note that the predictive performance observed here is broadly consistent with that reported in prior work, despite the added challenge in this study of predicting outcomes up to six years later from a single baseline assessment conducted in childhood. This contextualization clarifies that linguistic style and semantic features provide a sensitive, scalable signal for long-term

risk stratification that represents a meaningful advance over existing approaches, including experts' ratings of stress severity. Changes to the Discussion (page 17):

Moreover, given that the LIWC model demonstrated the highest predictive performance across methods (R^2=16.2%), our findings suggest that style-linked linguistic markers are detectable even in less emotionally charged or domain-specific speech contexts such as routine healthcare interactions, potentially enhancing the scalability of language-based risk assessment. This level of performance also helps contextualize the potential utility of language as a prognostic signal; because multi-year symptom trajectories are shaped by many interacting biological and environmental factors, models that are based on a single baseline assessment typically show modest out-of-sample explanatory power. In our data, LIWC features (R^2=0.162) and RoBERTa-derived sentence embeddings (R^2=0.156) each more than doubled the variance explained relative to a baseline model with demographics and expert-rated cumulative severity of stress (R^2=0.070), indicating that language captures prognostic information beyond conventional risk markers. This magnitude is broadly consistent with other data-driven forecasting approaches used in youth mental health, including scalable behavioral/digital signals[58,59] and, in some cases, neurobiological risk models[60].