nature portfolio

Corresponding author(s): Chase Antonacci

Last updated by author(s): May 17, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Audio was transcribed using WhisperX (v3.4.2; large-v2 model), with forced alignment via wav2vec 2.0 and speaker diarization using pyannote-audio (v3.1). Speaker-label corrections were performed using a secure GPT-4o deployment configured for PHI with no external data retention
Data analysis	LIWC variables were extracted using the LIWC-22 toolbox (version 1.11.0). TF-IDF variables were extracted using scikit-learn (version 1.7.2) and statsmodels v0.14.6. LDA topics were extracted using the DLATK package v1.3.0 in python3-mysqldb with MALLET v2.1.0 backend. Sentence embeddings were generated in Python v3.11.7 using PyTorch (v2.9.1), sentence-transformers (v5.5.0), and HuggingFace Transformers (v4.57.1) with the pretrained RoBERTa-base checkpoint (roberta-base). Statistical analyses used R v4.4.2 (with glmnet v4.1.10, uwot v0.2.3, dbscan v1.2.3, and ggplot2 v 4.0.0 for figure generation).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex was recorded at baseline (86 male; 118 female). Sex was included as a covariate in all predictive models. Supplemental analyses tested sex-by-language interaction terms; none were statistically significant.
Reporting on race, ethnicity, or other socially relevant groupings	The sample was racially and ethnically diverse (45.6% White, 11.3% Asian, 8.8% African American, 8.3% Hispanic/Latino, 19.6% biracial, 6.4% other). Race was included as a covariate in predictive models.
Population characteristics	Participants were youth from the San Francisco Bay Area enrolled in a longitudinal study of early life stress. Baseline age range was 9-13 years. Household income-to-needs ratio and parent education were assessed and included as covariates (see Supplement for details).
Recruitment	Participants were recruited via community-based outreach (flyers, school newsletters, community events, and online advertisements) across the San Francisco Bay Area as part of a larger longitudinal study of early life stress and adolescent psychobiological development; this is a convenience sample, with recruitment targeted to ensure variability in stress exposure. Inclusion and exclusion criteria are described in the Methods section. No formal a priori power analysis was conducted for the present NLP analyses, as this work was a secondary analysis of an existing longitudinal cohort. Baseline assessments were conducted between 09/2013 and 04/2017; the 4-year follow-up assessments occurred between 02/2018 and 10/2021; and the 6-year follow-up assessments occurred between 08/19 and 03/2025.
Ethics oversight	All procedures were approved by the Stanford University Institutional Review Board. Written parental consent and child assent were obtained for minors; participants provided informed consent upon reaching age 18. Participants were compensated for all study activities. The study was conducted in accordance with the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study is a longitudinal observational cohort study examining whether natural language features derived from structured stress interviews in youth predict internalizing symptoms and mood/anxiety diagnoses across adolescence. Participants completed a baseline structured clinical interview assessing lifetime stress exposure (TESI) at ages 9-13 years and were followed approximately 4 to 6 years later for self-reported internalizing symptoms (YSR) and interviewer-administered diagnostic outcomes (KSADS-PL). Multimodal natural language processing methods were applied to baseline speech transcripts, and predictive models were evaluated using nested cross-validation.
Research sample	The analytic sample consisted of 204 youth (86 male, 118 female) from the San Francisco Bay Area who were participating in a longitudinal study of early life stress and psychobiological development. At baseline, participants were aged 9-13 years (M = 11.38, SD = 1.05). Participants were included in the analytic sample if they completed the baseline stress interview and had follow-up behavioral and/or diagnostic data at 4 or 6 years.
Sampling strategy	Participants were recruited as part of an ongoing longitudinal study examining the effects of early life stress on development and

Sampling strategy	mental health. Recruitment was community-based in the San Francisco Bay Area and did not use random population sampling. Participants were recruited via community-based outreach (flyers, school newsletters, community events, and online advertisements). No a priori power calculation was conducted for the NLP analyses, as this study represents a secondary analysis of an existing longitudinal cohort. Sample size was determined by cohort enrollment and follow-up retention. Given the predictive modeling framework, model performance was evaluated using 5-fold nested cross-validation to estimate out-of-sample generalizability and mitigate overfitting. Data saturation was not applicable, as this was not a qualitative sampling design.
Data collection	At baseline, participants completed the Traumatic Events Screening Inventory for Children (TESI), a structured clinical interview assessing exposure to stress and trauma across 30 domains. Interviews were conducted by trained research staff and audio-recorded using digital recording equipment. For participants under age 18, a parent or guardian was present in the laboratory but not during the interview itself. Follow-up assessments included the Youth Self Report (YSR) and the Kiddie Schedule for Affective Disorders and Schizophrenia (KSADS-PL), administered by trained interviewers under clinical supervision. Expert raters assigning objective stress severity scores were blind to participant identity and subjective ratings. Diagnostic interviewers were blind to language-derived predictors. NLP analyses were conducted after data collection was complete and did not influence assessment procedures.
Timing	Baseline data collection occurred when participants were aged 9-13 years. The 4-year follow-up assessment occurred approximately four years after baseline (mean age ≈ 15.6 years), and the 6-year follow-up occurred approximately six years after baseline (mean age ≈ 17.4 years). Baseline TESI interviews were conducted between 09/2013 and 04/2017; the 4-year follow-up assessments occurred between 02/2018 and 10/2021; and the 6-year follow-up assessments occurred between 08/19 and 03/2025. Data were collected on a rolling basis as part of the ongoing longitudinal study; no experimental manipulation occurred between timepoints.
Data exclusions	Of the 225 enrolled participants, 21 were excluded from the present analyses due to missing baseline TESI interview data, missing 4-year follow-up behavioral data, or missing diagnostic data at 4- or 6-year follow-up. Exclusion criteria were pre-specified based on availability of required predictor and outcome variables. No additional exclusions were made based on transcript content, model performance, or outcome values.
Non-participation	The initial cohort included 225 participants. The analytic sample included 204 participants with complete baseline and follow-up data, reflecting attrition and missing data across waves (n = 21 excluded due to incomplete data). Participants who were excluded did not differ significantly from the analytic sample on key demographic or baseline stress variables. Reasons for attrition were typical of longitudinal research (e.g., loss to follow-up, incomplete assessments). No participants were excluded due to refusal after baseline enrollment.
Randomization	This was an observational longitudinal cohort study. No randomization was performed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N/A
Study protocol	The full study protocol (#27671, approved 03/2025) is available upon request to the study team (see corresponding author) or via the Stanford University Institutional Review Board (https://irb.stanford.edu/)
Data collection	All study visits were conducted in person by trained research staff at the Department of Psychology at Stanford University. Baseline data were collected from 2013-2017, and follow-up assessments occurred ~4 and ~6 years later.
Outcomes	Primary outcomes were YSR Internalizing Problems (continuous) at the 4-year follow-up and mood/anxiety disorder diagnosis (binary) assessed via KSADS-PL at 4 and 6 years. Diagnosis was coded as present if any mood/anxiety disorder was rated probable, remitted, or current at either follow-up assessment.

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A