Journal of High School Science

# Cognitive-congestion language profiling in ADHD: lexical signal auditing and negative-manifold specificity testing

Ju S

Submitted: June 6, 2026, Revised: version 1, July 15, 2026, version 2, August 21, 2026, version 3, October 2, 2026, version 4, October 5, 2026
Accepted: October 5, 2026

**Abstract**
High machine-learning accuracy in classifying ADHD-related writing does not necessarily establish an ADHD-specific linguistic phenotype, as predictions may reflect diagnostic vocabulary, medication references, or community-specific topics rather than cognitive organization. This study evaluated an interpretable cognitive-congestion language profile (LCCI) and a *negative-signature* framework for detecting disorder-specific linguistic distinctiveness. A dataset of 25,000 Reddit-derived posts from ADHD, OCD, depression, PTSD, and Asperger/autism-related communities was analyzed using condition-name masking and author-grouped cross-validation. Twenty-five features measuring discourse structure, uncertainty, urgency, lexical diversity, and coherence were compared with TF-IDF classification. The equal-weight LCCI composite performed near chance (ROC-AUC 0.506), while the multivariable profile achieved ROC-AUC 0.654, increasing to 0.681 with learned weights. Individual feature differences were small. TF-IDF substantially outperformed LCCI (ROC-AUC 0.940), with negligible improvement from combining them. An audit classified 59 of 60 high-weight TF-IDF coefficients (98.3%) as topic-related, without quantifying their contribution to predictive performance. To test specificity independently of supervised classification, the negative-signature framework excluded each psychiatric community in turn, constructed a reference language manifold from the remaining communities, and measured residual distinctiveness using PCA reconstruction, shrinkage Mahalanobis distance, and isolation forest. ADHD exhibited no positive residual signal (ROC-AUC 0.451, 0.419, and 0.429, respectively), while OCD showed the greatest residual separability. These findings distinguish predictive discrimination, interpretable linguistic association, and disorder specificity. The principal contribution is a reproducible, generalizable falsification framework demonstrating that strong supervised classification does not establish an ADHD-specific linguistic phenotype and providing a means to evaluate candidate linguistic biomarkers before clinical translation.

**Keywords**
Attention-deficit/hyperactivity disorder (ADHD), Computational psychiatry, Natural language processing (NLP), Digital biomarkers, Linguistic biomarkers, Machine learning, Disorder-specific language signatures, Anomaly detection, One-class classification, Psychiatric disorder classification

Sieun Ju, North London Collegiate School Jeju33 Globaleduro 145 beongil, Daejeongeup Seogwiposi, Code 63644 Zip, South Korea, Jeju-do, South Korea, irisot5190@gmail.com


https://doi.org/10.64336/001c.172399

## 1. Introduction

ADHD is frequently described through symptom checklists, but cognitive accounts also emphasize instability: variable response times, inconsistent goal maintenance, lapses of attention, and fluctuating control. A meta-analysis of reaction-time variability found robust ADHD-related variability across hundreds of studies (1). This motivates the present language question. Writing is not only word choice. It is also sequencing, monitoring, repair, referential control, topic maintenance, sentence construction, and regulation of urgency. If these processes become congested, the resulting language may show small but measurable shifts in structure even when the text remains readable.

Natural-language processing offers scale, but it also creates a scientific trap. A model may classify ADHD-labeled posts while leaving uncertain how much performance reflects disorder-relevant organization versus correlated medication, diagnosis, school/work, or subreddit-specific vocabulary. This is why we did not treat TF-IDF dominance as a final biomarker result. Instead, TF-IDF is used as a strong comparator and signal-source for audit target.

The proposed cognitive-congestion language profile was developed as a transparent alternative to purely lexical classification. The first scalar prototype collapsed multiple feature families into one equal-weight score. Reviewer feedback and the present results showed that this collapse was too severe. The approach therefore treats the same theoretical feature families as a profile and learns regularized weights under nested author-grouped cross-validation. This revision preserves interpretability while allowing the data to estimate which transparent features contribute most.

The central claim is deliberately narrow. The proposed profile is not a diagnostic biomarker. It is an interpretable digital biomarker candidate for ADHD-labeled public writing: a reproducible measurement framework that can be tested, audited, rejected, refined, and eventually transferred to clinically confirmed transcript data. The feature set is theory-motivated and does not directly encode diagnosis names, medication names, or community labels; the present manuscript does not claim prospective preregistration of the feature set.

## 2. Prior work and rationale

Three lines of work are most relevant. First, ADHD cognitive research provides the instability hypothesis. Reaction-time variability is one of the most replicated performance-level correlates of ADHD, but it is measured in structured tasks rather than natural language (1). The present study asks whether a related instability concept can be operationalized in written discourse.

Second, language studies in ADHD and adjacent neurodevelopmental conditions show that communication can be quantified beyond vocabulary. For example, work on ambiguous pronoun production has examined whether children with ASD and ADHD adapt referential choices to the listener (2). This supports the idea that language can reflect control, monitoring, and discourse organization.

Third, computational mental-health studies have shown that social-media text contains strong group-level signal (3-6). TF-IDF and n-gram models are especially important because they often produce strong classification performance with minimal feature engineering. Coppersmith and colleagues used social-media language to quantify mental-health signals and analyze self-reported diagnoses on Twitter, including ADHD and related categories (4, 5). The CLPsych 2015 shared task further demonstrated that lexical representations can classify depression and PTSD-related language (6). These studies justify TF-IDF as a serious benchmark rather than a strawman.

Non-language ADHD biomarker work also clarifies the standard of evidence. The ADHD-200 Global Competition showed that personal characteristic data could outperform resting-state fMRI for ADHD classification (7). More recent multimodal virtual-reality-assisted work has used eye tracking, EEG, actigraphy, and behavioral indices with independent train-test samples (8). Compared with these designs, subreddit-labeled public writing is a weaker endpoint. Therefore, this paper frames its result as a candidate language measure, not a clinical test.

## 3. Study questions

The study addresses four questions. First, does an equal-weight scalar version of the proposed profile separate ADHD-labeled posts from other mental-health-community posts after condition-name masking? Second, does a profile-level representation perform better than scalar collapse? Third, can learnable but still transparent regularized weights improve the interpretable-feature approach? Fourth, if TF-IDF remains stronger, which lexical features receive high weights in a fitted comparator model, and does ADHD-labeled writing appear as a distinctive residual outlier relative to the non-ADHD psychiatric-language manifold?

These questions intentionally separate predictive utility from biological interpretation. A classifier can predict a label without discovering a biomarker. Conversely, an interpretable profile can be scientifically useful even if it is not the top-performing predictor, especially when the top predictor appears dominated by lexical artifacts.

## 4. Methods

**4.1** Data source and sampling

The primary dataset was quocanh34/mental_health_dataset_1, a public Hugging Face dataset containing Reddit-derived text and metadata (9). The deterministic analytic subset used five subreddit labels: ADHD, OCD, depression, PTSD, and aspergers. After filtering invalid and deleted text, the pipeline sampled up to 5,000 posts per label with random seed 20260505.

The comparison group is not a healthy-control group. It consists of other mental-health-community writing, which makes the task harder and more relevant to specificity than ADHD versus general Reddit text. Table 1 shows that post counts were class-balanced by design, but unique-author counts differed across subreddits. This imbalance is important. The primary mitigation was author-grouped cross-validation: known same-author posts were kept within the same fold so that model estimates did not depend on seeing the same identified author in both training and test sets.

Remaining author ambiguity is treated as a limitation of public-corpus research.

TalkBank was evaluated as a future validation target because it provides structured spoken-language corpora and documented language-analysis infrastructure (10). The Dutch Asymmetries corpus is particularly relevant because it contains structured language data from ADHD and typically developing participants (11). In the present environment, TBDBpy metadata probing was possible, but authenticated access was required for transcript bodies (12). Therefore, no TalkBank transcript text was analyzed in the reported experiments.

**Table 1.** Deterministic analytic sample after filtering.

| subreddit | n | authors | lcci_mean | lcci_sd | median_tokens | median_chars |
|---|---|---|---|---|---|---|
| ADHD | 5000 | 4481 | 0.370 | 0.031 | 160.000 | 847.000 |
| OCD | 5000 | 3826 | 0.373 | 0.039 | 126.000 | 666.000 |
| depression | 5000 | 4457 | 0.375 | 0.033 | 138.000 | 692.000 |
| ptsd | 5000 | 3981 | 0.370 | 0.032 | 159.000 | 827.500 |
| aspergers | 5000 | 3339 | 0.365 | 0.034 | 129.000 | 690.000 |

**4.2** Preprocessing and feature engineering

Preprocessing lowercased text for lexical masking and replaced direct condition-name terms such as ADHD, OCD, PTSD, depression, and Asperger/autism-related labels. Masking was applied before deterministic feature extraction and model fitting. This control removes the most obvious label leakage but does not remove medication names, symptom words, school and work topics, or community-specific vocabulary. The manuscript therefore treats condition-name masking as a minimum leakage control, not a complete semantic decontamination procedure.

The proposed profile features were grouped into interpretable families. Repair and uncertainty features count surface markers such as uncertainty terms, negation, discourse markers, and repair-like language. Affective and urgency features measure question marks, exclamation marks, all-caps usage, repeated punctuation, and temporal urgency markers. Length and structure features summarize token counts, character counts, sentence counts, sentence-length variability, and related structural variability. Lexical diversity features include type-token ratio, moving-average type-token ratio, and hapax ratio. Local coherence and topic-shift features estimate adjacent sentence similarity and topical transitions using deterministic lexical representations. These families are theory-motivated measurement choices; they were not defined as a prospective preregistered feature set, and they do not directly encode diagnosis names, medication names, or community labels.

The scalar prototype was an equal-weight standardized combination of selected feature families. This scalar was useful as a conservative baseline but too rigid for the current corpus. The weighted profile preserves the transparent feature families while learning

regularized coefficients inside nested cross-validation. This matters because cognitive congestion is unlikely to be represented by a single fixed direction across all text lengths, communities, and discourse contexts.

**4.3** Classification models and validation

The first model group reproduced the original comparison: prevalence baseline, scalar prototype, interpretable profile model, sanitized TF-IDF, and TF-IDF plus transparent profile features. All models were implemented in Python with scikit-learn (13) and used author-grouped five-fold validation.

The classification endpoint was binary, distinguishing ADHD-labeled posts from posts originating in the four comparison communities (OCD, depression, PTSD, and aspergers). Five model configurations were evaluated: a prevalence baseline, an equal-weight scalar LCCI prototype, a multivariable interpretable profile, a sanitized TF-IDF classifier, and a combined TF-IDF plus interpretable-feature classifier. The prevalence baseline provided a reference for performance without informative linguistic predictors, while the scalar and multivariable models tested whether retaining the separate dimensions of cognitive congestion improved discrimination relative to their aggregation into a single score.

Model evaluation used five-fold author-grouped cross-validation, with posts from the same identified author assigned exclusively to either training or validation data within each fold. Performance was summarized using balanced accuracy, ROC-AUC, average precision, and F1 score. These metrics were calculated at the post level rather than the author level, so authors contributing multiple posts received proportionally greater weight. Fold-to-fold standard deviations were used to describe variation across the five validation partitions rather than to estimate formal confidence intervals.

**4.4** Weighted profile and feature ablation

The weighted profile was designed to answer a specific weakness of the scalar prototype: equal weights force heterogeneous discourse features into one direction. The weighted-profile analysis used nested author-grouped validation. In each outer fold, regularized logistic models were tuned in inner grouped folds, then evaluated on held-out authors. This design prevents choosing weights on the same fold used for performance reporting. The original profile ablation helps interpret which feature families mattered before learnable weighting.Feature-family ablation was performed by removing each predefined linguistic feature family in turn and comparing the resulting balanced accuracy and ROC-AUC against the complete 25-feature interpretable profile.

Four transparent-feature configurations were compared: the core profile, the core profile augmented with length-related features, the weighted profile, and the weighted profile augmented with interaction terms. For the weighted models, regularized logistic regression coefficients were selected within three-fold inner author-grouped cross-validation, nested within five outer author-grouped validation folds. The outer folds were reserved for evaluating generalization to previously unseen identified authors. This separation reduced the risk of optimistic performance estimates arising from hyperparameter selection on evaluation data.

Coefficient stability was assessed across the five outer folds by summarizing the mean signed coefficient, mean absolute coefficient, number of folds in which each feature was selected, and corresponding selection rate. Feature-family ablation was evaluated separately using the original 25-feature interpretable profile. Each predefined feature family was removed in turn, and changes in balanced accuracy and ROC-AUC were measured relative to the complete profile. The ablation analysis therefore evaluated the contribution of feature families to the original profile rather than the incremental contribution of those families to the subsequently optimized weighted model.

**4.5** TF-IDF signal-source audit

Because TF-IDF strongly outperformed all transparent-profile variants, the key question here is which lexical features received the largest weights in a fitted comparator. A representative fold audit was conducted on exactly 60 fold-1 TF-IDF coefficients: the 30 most positive coefficients and the 30 most negative coefficients after fitting the sanitized TF-IDF classifier. Each coefficient was categorized as medication/treatment vocabulary, diagnosis/symptom vocabulary, function-school-work-attention vocabulary, community-identity vocabulary, trauma/self-harm/affect vocabulary, or general lexical affect.

The coefficient audit was restricted to one representative fitted model from the first validation fold and was intended to characterize the vocabulary receiving the largest positive and negative classifier weights. Positive coefficients indicated association with the ADHD-labeled class, whereas negative coefficients indicated association with the comparison class. The 60 selected terms were assigned to the predefined semantic categories, and the number of terms and mean absolute coefficient magnitude within each category were summarized. The complete term-level audit and categorization rules are provided in Supplementary Table S1.

This procedure was descriptive and did not estimate the fraction of predictive performance attributable to each vocabulary category. In particular, the proportion of audited terms assigned to diagnosis-, treatment-, symptom-, or community-related categories should not be interpreted as the proportion of classifier accuracy explained by those categories. Establishing such attribution would require additional experiments, such as category-specific vocabulary removal followed by model refitting and evaluation on held-out data.

**4.6** Negative-manifold specificity testing

The negative-manifold analysis reversed the conventional supervised classification problem. Instead of learning features that positively identified a target psychiatric community, the analysis tested whether that community exhibited linguistic structure that could not be adequately represented by a reference distribution constructed without its observations. Each of the five communities was designated as the excluded target in turn, while the remaining four communities supplied the reference population. This leave-one-community-out design allowed ADHD residual distinctiveness to be compared directly with the corresponding distinctiveness of OCD, depression, PTSD, and aspergers.

Three complementary unsupervised methods were evaluated. Principal component analysis (PCA) quantified residual variation through reconstruction error, measuring how poorly an observation could be represented by a lower-dimensional reference structure. Shrinkage Mahalanobis distance measured multivariate deviation from the reference feature distribution while regularizing covariance estimation. Isolation forest provided an alternative anomaly score based on the relative ease with which observations could be isolated within the learned reference distribution. These approaches evaluated different forms of deviation rather than assuming that disorder-specific language must manifest as a single type of statistical outlier.

For each excluded community, residual scores were compared between the excluded target observations and held-out observations from the four-community reference population. ROC-AUC quantified whether higher anomaly scores preferentially identified the excluded community, with 0.5 representing chance-level ranking discrimination. Standardized differences between target and reference residual scores were additionally summarized using Cohen's d. The 95th percentile of training-reference residual scores defined an anomaly threshold independent of the excluded target observations. This framework evaluated relative residual distinctiveness within the selected linguistic feature space; it did not establish clinical specificity or the existence or absence of a disorder-specific biological phenotype.

## 5. Results

### **5.1** Feature level differences

Table 2 summarizes the differences in individual linguistic features between ADHD-labeled posts and the combined comparison communities following condition-name masking. Overall, the differences were small, with the composite LCCI score nearly identical between groups (0.370 versus 0.371; Cohen's d = −0.015). The largest standardized differences were observed for negation density (d = −0.174), first-person density (d = −0.170), and exclamation density (d = 0.147). ADHD-labeled posts also showed slightly greater sentence-length variability (d = 0.046) and topic-shift rate (d = 0.081). Although several features exhibited statistically significant differences after false-discovery-rate correction, all reported effect sizes were small ($|d| \leq 0.174$). Thus, no individual feature demonstrated substantial separation between ADHD-labeled and comparison-community writing, motivating the subsequent evaluation of multivariable feature combinations.

**Table 2.** Largest ADHD-minus-comparison feature contrasts after masking.

| Feature | ADHD mean | Comparison mean | Mean diff | Cohen’s d | FDR q |
|---|---|---|---|---|---|
| lcci_nlp | 0.370 | 0.371 | -0.001 | -0.015 | 0.198 |
| sentence_len_cv | 0.620 | 0.611 | 0.009 | 0.046 | 0.008 |
| repair_density | 0.010 | 0.010 | -0.001 | -0.059 | 0.296 |
| uncertainty_density | 0.003 | 0.003 | 0.000 | 0.013 | <0.001 |

| Feature | ADHD mean | Comparison mean | Mean diff | Cohen's d | FDR q |
|---|---|---|---|---|---|
| discourse_marker_density | 0.053 | 0.053 | 0.001 | 0.034 | 0.039 |
| negation_density | 0.013 | 0.016 | -0.002 | -0.174 | <0.001 |
| first_person_density | 0.095 | 0.102 | -0.006 | -0.170 | <0.001 |
| question_density | 0.002 | 0.002 | 0.000 | 0.020 | <0.001 |
| exclamation_density | 0.001 | 0.000 | 0.000 | 0.147 | <0.001 |
| repeated_punct_density | 0.001 | 0.001 | -0.000 | -0.109 | 0.163 |
| allcaps_density | 0.014 | 0.013 | 0.001 | 0.041 | <0.001 |
| adjacent_jaccard_sd | 0.046 | 0.047 | -0.000 | -0.004 | 0.021 |
| topic_shift_rate | 0.826 | 0.810 | 0.016 | 0.081 | 0.549 |
| low_adjacent_coherence | 0.961 | 0.958 | 0.003 | 0.080 | 0.198 |

**5.2** Model comparisons

Table 3 and Figure 1 compare the predictive performance of five classification models under author-grouped five-fold cross-validation.

**Table 3.** Original model comparison under author-grouped five-fold validation.

| Model | Balanced accuracy | SD | ROC-AUC | Avg precision | F1 |
|---|---|---|---|---|---|
| TF-IDF + profile features | 0.859 | 0.007 | 0.940 | 0.836 | 0.743 |
| TF-IDF only | 0.858 | 0.006 | 0.940 | 0.836 | 0.743 |
| Transparent profile model | 0.613 | 0.004 | 0.654 | 0.292 | 0.388 |
| Scalar prototype | 0.506 | 0.006 | 0.506 | 0.195 | 0.295 |
| Prevalence baseline | 0.500 | 0.000 | 0.500 | 0.200 | 0.000 |

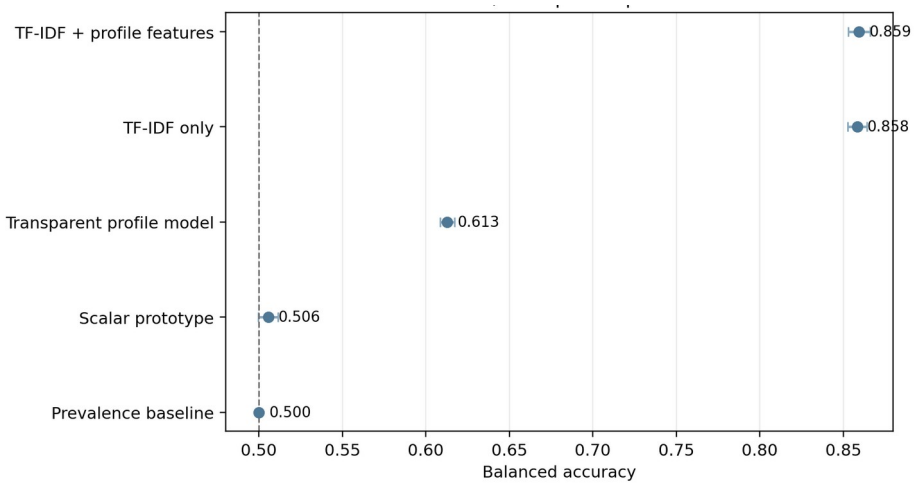



**Figure 1.** Original model comparison. TF-IDF is strongest, whereas the scalar prototype is near baseline and the full profile carries modest signal.

The equal-weight LCCI scalar performed near chance, with balanced accuracy and ROC-AUC both 0.506, compared with 0.500 for the prevalence baseline. Retaining the individual linguistic features in the transparent profile improved balanced accuracy to 0.613 and ROC-AUC to 0.654, although average precision remained low (0.292). The TF-IDF classifier substantially outperformed both interpretable approaches, achieving balanced accuracy of 0.858, ROC-AUC of 0.940, average precision of 0.836, and F1 of 0.743. Combining TF-IDF with the transparent profile produced virtually identical performance, increasing balanced accuracy by only 0.001 (0.858 to 0.859), with no improvement in ROC-AUC, average precision, or F1. These results demonstrate that retaining the individual LCCI features improves discrimination relative to equal-weight aggregation, but that the resulting profile provides substantially weaker classification performance than conventional lexical features and adds no measurable incremental predictive benefit when combined with TF-IDF.

### **5.3** Weighted profile: learnable transparent weighting

Table 4 and Figure 2 compare four interpretable-feature configurations evaluated using nested author-grouped cross-validation.

**Table 4.** Weighted-profile nested author-grouped validation performance.

| Model | Balanced accuracy | SD | ROC-AUC | ROC SD | Avg precision | F1 |
|---|---|---|---|---|---|---|
| Weighted profile | 0.629 | 0.005 | 0.681 | 0.008 | 0.312 | 0.404 |
| Weighted profile + interactions | 0.626 | 0.013 | 0.675 | 0.015 | 0.316 | 0.401 |
| Core profile + length | 0.615 | 0.010 | 0.656 | 0.012 | 0.297 | 0.390 |
| Core profile | 0.593 | 0.015 | 0.632 | 0.013 | 0.280 | 0.370 |

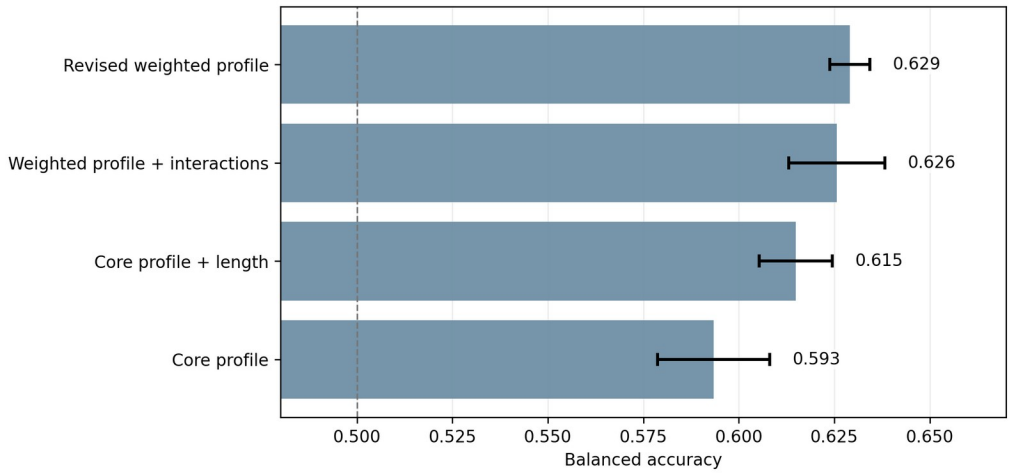



**Figure 2.** Weighted-profile performance by transparent feature set. Engineered linear weighting provides the strongest interpretable-feature result.

The core profile achieved balanced accuracy of 0.593 and ROC-AUC of 0.632, increasing to 0.615 and 0.656, respectively, when length-related features were added. The regularized weighted profile achieved the highest balanced accuracy (0.629), ROC-AUC (0.681), and F1 score (0.404), representing improvements of 0.036, 0.049, and 0.034, respectively, over the core profile. Adding interaction terms to the weighted profile did not improve performance, yielding balanced accuracy of 0.626 and ROC-AUC of 0.675, although average precision increased slightly from 0.312 to 0.316. The weighted profile also exhibited relatively low fold-to-fold variability (SD = 0.005 for balanced accuracy and 0.008 for ROC-AUC). Overall, learned feature weighting improved discrimination compared with the unweighted core profile, while the addition of interaction terms provided no further improvement in balanced accuracy or ROC-AUC.

Table 5 and Figure 3 summarize the magnitude, direction, and stability of the weighted-profile coefficients across the five outer validation folds. Token count (mean coefficient = −1.097), urgency per content token (−1.049), and all-caps density (0.988) exhibited the largest absolute coefficients, with all three selected in every fold. Character count also showed a relatively large positive coefficient (0.843), although it was selected in four of five folds. Other consistently selected features included the first-person/negation composite (−0.487), type-token ratio (−0.479), sentence count (−0.413), question density (0.330), exclamation density (0.311), and sentence-length standard deviation (−0.238). Positive coefficients indicated greater association with ADHD-labeled writing, whereas negative coefficients indicated association with the comparison communities. Most of the highest-weighted features were retained across all five folds, although first-person density, urgency surface sum, and the topic-shift/coherence composite were selected in only three folds. These results identify the linguistic features contributing most consistently to the weighted classifier, with both positive and negative associations distributed across structural, lexical-diversity, and urgency-related measures.

**Table 5.** Most stable weighted-profile coefficients across outer folds.

| Feature | Mean coefficient | Mean \|coef\| | Selected folds | Selection rate |
|---|---|---|---|---|
| n_tokens | -1.097 | 1.097 | 5 | 1.000 |
| urgency_per_content_token | -1.049 | 1.049 | 5 | 1.000 |
| allcaps_density | 0.988 | 0.988 | 5 | 1.000 |
| n_chars | 0.843 | 0.843 | 4 | 0.800 |
| self_negation_sum | -0.487 | 0.487 | 5 | 1.000 |
| type_token_ratio | -0.479 | 0.479 | 5 | 1.000 |
| n_sentences | -0.413 | 0.413 | 5 | 1.000 |
| question_density | 0.330 | 0.330 | 5 | 1.000 |
| first_person_density | 0.322 | 0.322 | 3 | 0.600 |

| Feature | Mean coefficient | Mean \|coef\| | Selected folds | Selection rate |
|---|---|---|---|---|
| urgency_surface_sum | -0.319 | 0.319 | 3 | 0.600 |
| exclamation_density | 0.311 | 0.311 | 5 | 1.000 |
| n_content_tokens | 0.254 | 0.254 | 4 | 0.800 |
| sentence_len_sd | -0.238 | 0.238 | 5 | 1.000 |
| topic_shift_minus_coherence | -0.226 | 0.226 | 3 | 0.600 |

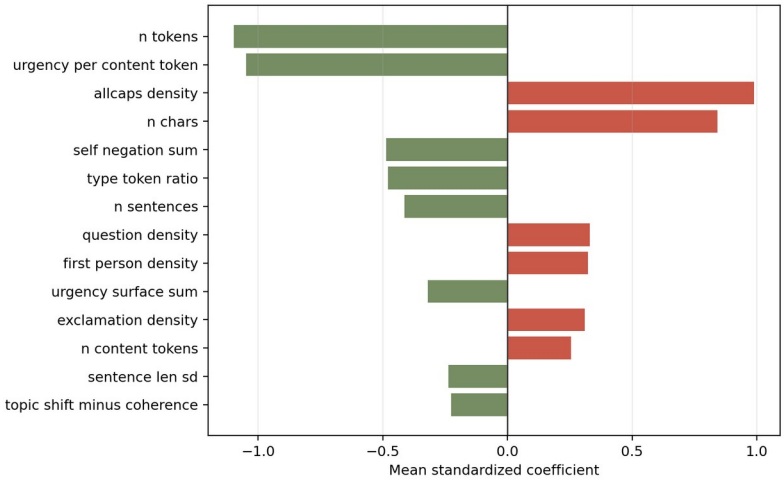



**Figure 3.** Stable weighted-profile coefficients. The largest repeated weights are discourse-structural rather than direct diagnosis-name terms.

**5.4** Ablation and feature-family interpretation

Table 6 and Figure 4 present the effects of removing individual linguistic feature families from the original 25-feature transparent profile.

**Table 6.** Original feature-family ablation relative to the full interpretable profile.

| Removed family | Features left | Balanced accuracy | ROC-AUC | Delta vs full |
|---|---|---|---|---|
| none | 25 | 0.613 | 0.654 | 0.000 |
| length_structure | 18 | 0.600 | 0.640 | -0.013 |
| lexical_diversity | 21 | 0.591 | 0.629 | -0.021 |
| repair_uncertainty | 22 | 0.613 | 0.655 | -0.000 |
| affective_urgency_surface | 19 | 0.590 | 0.627 | -0.023 |
| local_coherence | 21 | 0.612 | 0.654 | -0.001 |

The complete profile achieved balanced accuracy of 0.613 and ROC-AUC of 0.654. Removing affective-urgency features produced the largest reduction in balanced accuracy (−0.023), followed by lexical-diversity features (−0.021) and length/structure features (−0.013). Corresponding ROC-AUC values declined to 0.627, 0.629, and 0.640, respectively. In

contrast, removing repair/uncertainty features produced virtually no change in performance (balanced accuracy = 0.613; ROC-AUC = 0.655), while removing local-coherence features resulted in a negligible reduction in balanced accuracy (−0.001) and no change in ROC-AUC (0.654). These findings indicate that affective-urgency, lexical-diversity, and length/structure features contributed most to the transparent profile's discrimination, whereas repair/uncertainty and local-coherence features provided little measurable incremental predictive information.

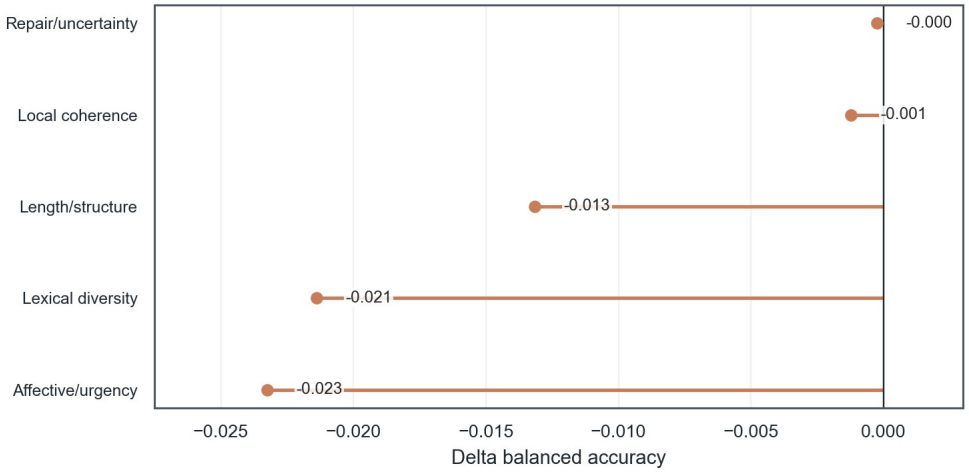



**Figure 4.** Feature-family ablation. Lexical-diversity and affective/urgency surface families contributed more than repair/uncertainty or local-coherence proxies in this corpus.

**5.5** TF-IDF signal-source audit

Table 7 and Figure 5 summarize the thematic classification of the 60 highest-magnitude TF-IDF coefficients, comprising 30 positively and 30 negatively associated terms. Of these, 59 (98.3%) were assigned to topic-related categories, while only one (1.7%) was classified as general lexical affect. Diagnosis/symptom terminology constituted the largest category (19 terms), followed by medication/treatment (13), trauma /self-harm /affect (12), school /work / attention (8), and community identity (7). Thus, the overwhelming majority of the strongest lexical coefficients were associated with clinical, treatment-related, or community-specific subject matter rather than general linguistic expression. This distribution was observed among the selected high-magnitude coefficients and does not quantify the proportion of overall classifier performance attributable to these categories.

**Table 7.** Representative fold signal-source comparison between TF-IDF and the proposed profile.

| Signal category | Source | Count | Mean \|weight\| |
|---|---|---|---|
| diagnosis/symptom | TF-IDF top coefficients | 19 | 3.194 |

| Signal category | Source | Count | Mean \|weight\| |
|---|---|---|---|
| medication/treatment | TF-IDF top coefficients | 13 | 4.608 |
| trauma/self-harm affect | TF-IDF top coefficients | 12 | 3.185 |
| school/work attention | TF-IDF top coefficients | 8 | 3.631 |
| community identity | TF-IDF top coefficients | 7 | 3.525 |
| general affect | TF-IDF top coefficients | 1 | 2.598 |
| coherence/topic | Weighted profile | 6 | 0.090 |
| length/structure | Weighted profile | 6 | 0.488 |
| urgency surface | Weighted profile | 6 | 0.511 |
| lexical diversity | Weighted profile | 4 | 0.202 |
| repair/uncertainty | Weighted profile | 4 | 0.062 |
| self/negation | Weighted profile | 3 | 0.300 |
| discourse markers | Weighted profile | 2 | 0.083 |
| other profile feature | Weighted profile | 1 | 0.072 |

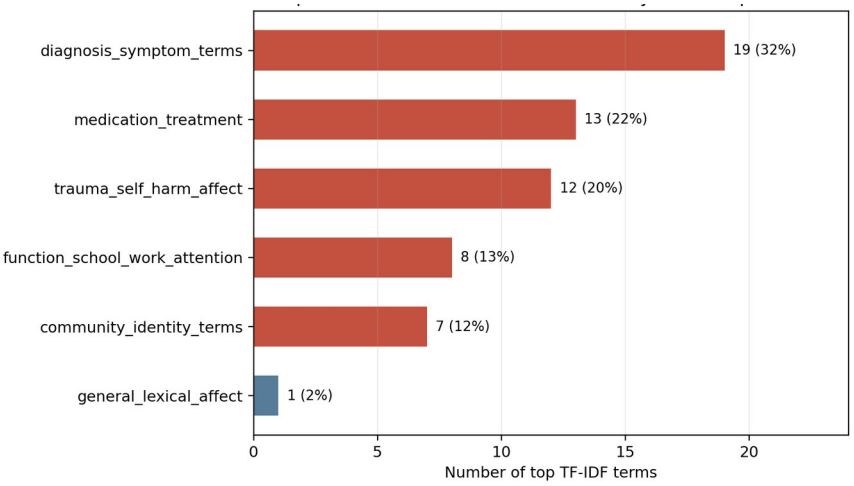



**Figure 5.** Representative fold TF-IDF top-coefficient audit. Many high-weight lexical coefficients correspond to medication, diagnosis, symptom, community, or topic-related categories.

Table 8 and Figure 6 present representative high-magnitude TF-IDF coefficients associated with ADHD-labeled and comparison-community posts. Terms positively associated with ADHD-labeled writing included "adderall" (+7.645), "medication" (+6.857), "meds" (+6.801), "vyvanse" (+6.584), and "focus" (+4.978), reflecting predominantly medication- and attention-related vocabulary. Conversely, terms associated with the comparison communities included references to trauma, intrusive thoughts, flashbacks, autism spectrum terminology, compulsions, and suicide. The opposing coefficient directions demonstrate that the classifier distinguished communities partly through differences in the clinical conditions, treatments, and experiences discussed. These findings complement the thematic distribution in Table 7 by identifying specific lexical

features contributing to the observed classification performance.

**Table 8.** Selected representative fold TF-IDF coefficients and audit categories. The complete audit is presented in Supplementary Table S1.

| Term | Direction | Audit category | Coefficient |
|---|---|---|---|
| adderall | ADHD | medication/treatment | 7.645 |
| medication | ADHD | medication/treatment | 6.857 |
| meds | ADHD | medication/treatment | 6.801 |
| vyvanse | ADHD | medication/treatment | 6.584 |
| focus | ADHD | school/work attention | 4.978 |
| concerta | ADHD | medication/treatment | 4.959 |
| ritalin | ADHD | medication/treatment | 4.438 |
| diagnosed | ADHD | diagnosis/symptom | 4.198 |
| forget | ADHD | school/work attention | 4.166 |
| task | ADHD | school/work attention | 3.912 |
| add | ADHD | community identity | 3.789 |
| stimulants | ADHD | medication/treatment | 3.702 |
| doctor | ADHD | medication/treatment | 3.649 |
| bored | ADHD | school/work attention | 3.631 |
| dopamine | ADHD | diagnosis/symptom | 3.623 |
| hyperfocus | ADHD | diagnosis/symptom | 3.527 |
| strattera | ADHD | medication/treatment | 3.518 |
| medicated | ADHD | medication/treatment | 3.418 |

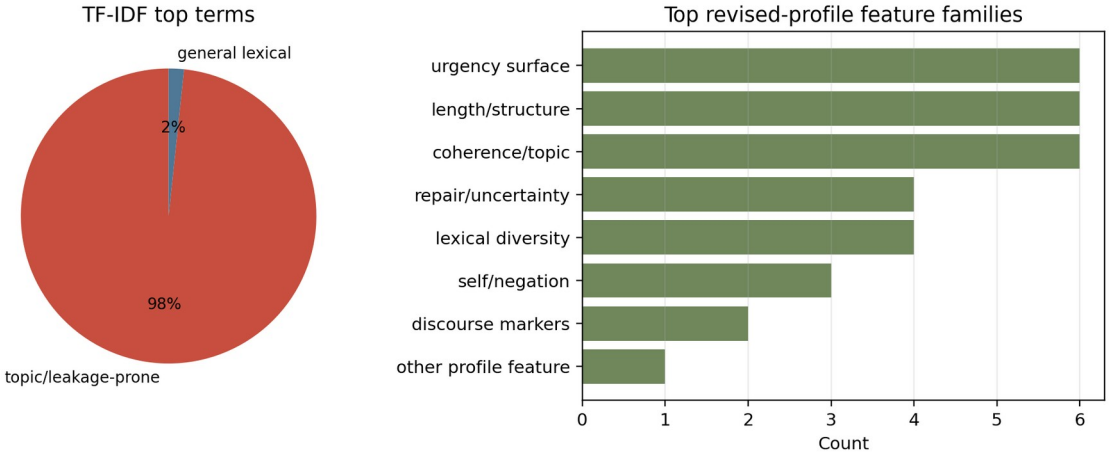



Figure 6. TF-IDF versus the proposed profile signal source. In the audited fold, many high-weight TF-IDF terms correspond to lexical topic categories, whereas the proposed profile uses transparent discourse and structure features.

**5.6** Manifold specificity test

Table 9 and Figure 7 present the results of the negative-manifold specificity test using three independent anomaly-detection methods. ADHD-labeled posts exhibited ROC-AUC values below chance across all three methods: 0.451 for PCA reconstruction error, 0.419 for shrinkage Mahalanobis distance, and 0.429 for isolation forest. Corresponding standardized effect sizes were also negative (Cohen's d = −0.170, −0.263, and −0.298, respectively), indicating that ADHD-labeled posts did not exhibit greater anomaly scores than the comparison communities. Among the evaluated communities, OCD showed the highest ROC-AUC values across all three methods (0.545, 0.568, and 0.568, respectively), although these remained only modestly above chance. The consistency of the ADHD results across three distinct methods indicates that ADHD-labeled writing was not distinguishable as an outlying linguistic profile relative to the modeled negative manifold within the selected feature space.

**Table 9.** Leave-one-community-out residual-manifold specificity test.

| Residual method | ADHD ROC-AUC | ADHD Cohen's d | Highest residual label | Highest ROC-AUC | Specificity result |
|---|---|---|---|---|---|
| PCA reconstruction | 0.451 | -0.170 | OCD | 0.545 | ADHD not residual outlier |
| Shrinkage Mahalanobis | 0.419 | -0.263 | OCD | 0.568 | ADHD not residual outlier |
| Isolation forest | 0.429 | -0.298 | OCD | 0.568 | ADHD not residual outlier |

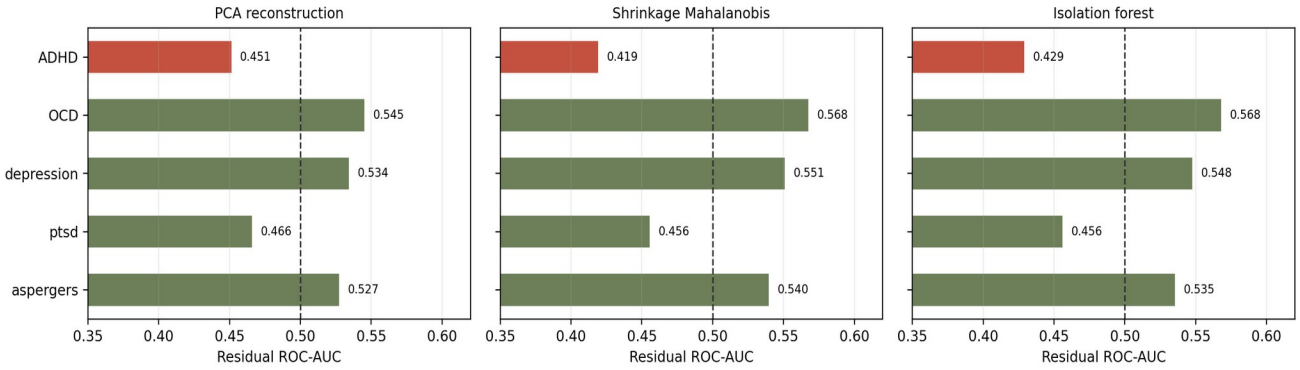



**Figure 7.** Residual-manifold ROC-AUC by community. ADHD is not the most residual label under the proposed profile features.

Figure 8 contrasts supervised classification performance with unsupervised residual discrimination. Although TF-IDF achieved ROC-AUC of 0.940, the three ADHD negative-manifold methods produced ROC-AUC values below 0.500, illustrating that

supervised discrimination and residual distinctiveness measure different properties.

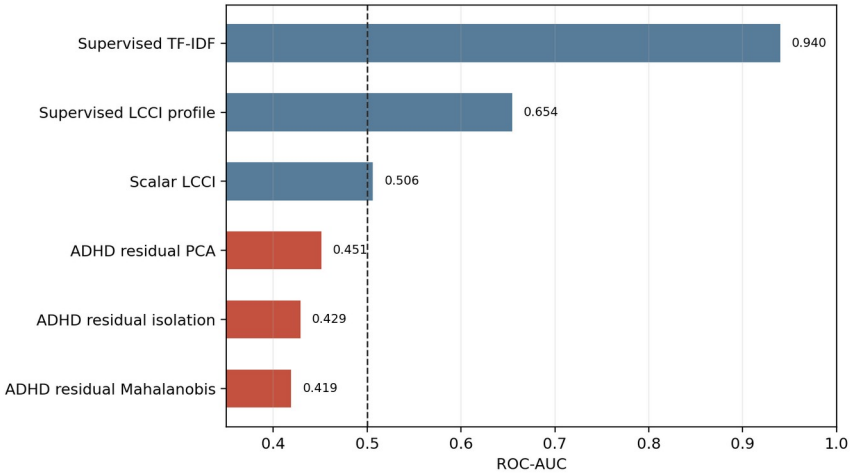



**Figure 8.** Supervised discrimination and unsupervised residual specificity answer different questions. The proposed profile can classify modestly without proving ADHD-specific outlier structure.

## 6. Discussion

### 6.1 Predictive accuracy, linguistic measurement, and diagnostic specificity are distinct scientific questions

The central finding of this study is that the ability to classify ADHD-labeled online writing does not necessarily imply the identification of an ADHD-specific linguistic phenotype. Across 25,000 posts drawn from five mental-health communities, three complementary analytical approaches produced different but mutually informative results. Conventional lexical classification distinguished ADHD-labeled posts from comparison-community posts with high accuracy; interpretable cognitive-congestion features demonstrated considerably weaker but measurable discrimination; and unsupervised negative-manifold testing failed to identify ADHD-labeled writing as a distinctive linguistic outlier. Rather than representing contradictory outcomes, these findings expose three separate questions that are frequently conflated in computational language research: whether group labels can be predicted, which linguistic properties account for that prediction, and whether those properties are specific to the condition under investigation.

This distinction constitutes the principal scientific contribution of the manuscript. A classifier can achieve high predictive accuracy by learning vocabulary associated with a diagnostic community without necessarily identifying linguistic characteristics that arise from the cognitive processes associated with that diagnosis. Conversely, a transparent linguistic profile can capture modest group differences without demonstrating either clinical validity or diagnostic specificity. Finally, a feature set may support supervised classification while failing to establish that the target population occupies a distinct region of

linguistic feature space. The present study evaluates these possibilities within a common dataset rather than treating predictive accuracy as sufficient evidence of biomarker validity.

The results therefore support a qualified conclusion. ADHD-labeled Reddit writing is strongly distinguishable from the selected comparison communities using lexical features, and modestly distinguishable using interpretable discourse and structural features. However, the current evidence does not establish that the latter constitute an ADHD-specific cognitive-congestion signature. This is a scientifically consequential negative finding because it identifies the precise inferential boundary between community-label prediction and condition-specific linguistic measurement.

**6.2** Weak individual feature effects challenge the concept of a unidirectional cognitive-congestion signature

The feature-level analysis in Table 2 provides the first indication that cognitive-congestion language cannot be adequately represented as a simple accumulation of linguistic irregularities. The composite LCCI score was nearly identical between ADHD-labeled and comparison-community posts (0.370 versus 0.371), with a negligible standardized difference (Cohen's d = −0.015). Even the largest individual feature contrasts remained small, including negation density (d = −0.174), first-person density (d = −0.170), and exclamation density (d = 0.147). Sentence-length variability (d = 0.046) and topic-shift rate (d = 0.081) showed similarly limited separation.

These findings have two implications. First, the absence of a substantial composite difference suggests that ADHD-labeled writing does not exhibit a uniformly elevated cognitive-congestion profile as operationalized by the present index. Second, the opposing directions of individual feature effects indicate that the constituent measures cannot be assumed to vary together or contribute equally to classification. Features theoretically associated with discourse disruption, urgency, self-reference, and structural variability may differ in direction, magnitude, and relevance across communities.

The distinction between statistical significance and substantive discrimination is particularly important. Several features exhibited statistically significant group differences after false-discovery-rate correction, yet all reported standardized effects were small. With thousands of observations, small differences may be detected reliably without being sufficiently large to support meaningful classification or interpretation at the individual-post level. Accordingly, the feature-level results do not justify treating any single measure as a robust ADHD marker.

Hence, a theoretically motivated collection of linguistic features does not automatically constitute a coherent measurement scale. Construct validity requires evidence that the features jointly represent the proposed underlying phenomenon. In the present study, the near-identical composite scores and heterogeneous individual effects challenge the assumption that cognitive congestion can be measured by summing features with equal positive contributions. This does not invalidate the individual features, but it shifts the relevant measurement unit from an assumed unidirectional index toward a potentially

multidimensional profile whose components must be evaluated separately.

**6.3** The failure of equal-weight aggregation and the value of retaining feature structure

The model comparison in Table 3 and Figure 1 demonstrates the practical consequences of collapsing heterogeneous linguistic features into a single scalar. The equal-weight LCCI prototype achieved balanced accuracy and ROC-AUC of only 0.506, effectively indistinguishable from the prevalence baseline of 0.500. By contrast, retaining the individual features in a transparent multivariable classifier increased balanced accuracy to 0.613 and ROC-AUC to 0.654.

The difference between these models is more informative than either performance estimate considered independently. Both models draw on interpretable linguistic measurements, but they differ in whether the relationships among those measurements are preserved. The near-chance scalar result indicates that equal-weight aggregation discarded much of the discriminatory information available in the underlying feature set. The multivariable result indicates that some of this information can be recovered when individual features are allowed to contribute separately.

One plausible explanation is that the features have opposing associations with community labels, causing their contributions to cancel during scalar aggregation. Another is that some features are informative only when considered jointly with others, or that the same observed feature can reflect different communicative processes across communities. These possibilities are consistent with the observed coefficient directions but have not been independently established as cognitive mechanisms.

The distinction has broader relevance to interpretable biomarker construction. Interpretability is not synonymous with reducing a multidimensional measurement to one number. A scalar may be easy to communicate while concealing the heterogeneous structure of the underlying data. Conversely, a multivariable profile can remain transparent if its component definitions, coefficients, and validation procedures are accessible. The present findings therefore favor retaining feature-level information rather than prematurely compressing linguistic complexity into a single cognitive-congestion score.

Nevertheless, the transparent profile's balanced accuracy of 0.613 and ROC-AUC of 0.654 indicate only modest discrimination. Its average precision of 0.292, compared with a baseline of 0.200, further illustrates the limited predictive value of the original profile in this dataset. These results support continued investigation of the individual linguistic measurements but do not support deployment of the profile as a diagnostic classifier.

**6.4** Learned feature weighting improves measurement but does not resolve specificity

The weighted-profile experiments in Table 4 and Figure 2 provide a direct test of whether the limited performance of the original profile arose partly from its weighting assumptions. The core profile achieved balanced accuracy of 0.593 and ROC-AUC of 0.632. Adding length-related features improved these values to 0.615 and 0.656, respectively. The regularized weighted profile performed best, achieving

balanced accuracy of 0.629 and ROC-AUC of 0.681.

These results show that learnable weighting can extract additional discriminatory information from the transparent feature set. The improvement from the core profile to the weighted configuration is consistent with the hypothesis that linguistic dimensions differ in their relevance and should not be treated as interchangeable manifestations of a single underlying process.

The nested author-grouped validation design is important in interpreting this result. Feature weighting and model selection were conducted within inner validation folds, with evaluation reserved for held-out authors in the outer folds. This provides a stronger assessment of generalization than selecting feature weights and reporting performance on the same observations. The relatively small reported fold-to-fold variability of the weighted model is encouraging, although it does not establish stability across platforms, populations, or independently collected datasets.

Adding interaction terms did not improve balanced accuracy or ROC-AUC, which declined slightly to 0.626 and 0.675, respectively. Although average precision increased marginally from 0.312 to 0.316, the overall pattern does not support a meaningful advantage for the interaction-augmented configuration. Within the tested model family, regularized linear weighting appears sufficient to capture most of the available discriminatory information. More elaborate interactions may add complexity without corresponding gains in generalization.

Importantly, the improvement achieved through learned weighting remains limited in absolute terms. A ROC-AUC of 0.681 is better than the original transparent profile's 0.654 but substantially below the TF-IDF classifier's 0.940. The weighted model therefore strengthens the argument that interpretable linguistic features contain some reproducible information while simultaneously demonstrating that optimization alone does not transform this feature set into a highly discriminative or condition-specific marker.

**6.5** Coefficient directionality reveals a heterogeneous rather than uniformly elevated profile

Table 5 and Figure 3 provide further insight into why a simple additive cognitive-congestion score performed poorly. The strongest weighted-profile coefficients were not uniformly positive. Token count (−1.097) and urgency per content token (−1.049) showed large negative coefficients, whereas all-caps density (0.988) and character count (0.843) showed large positive coefficients. Other consistently selected features included the first-person/negation composite, type-token ratio, sentence count, question density, and exclamation density.

The opposing coefficient signs indicate that the fitted model distinguishes ADHD-labeled posts through a combination of relatively elevated and reduced linguistic measurements. In particular, urgency-related features did not behave as a single homogeneous family: urgency per content token showed a strong negative association, while exclamation density and all-caps density showed positive associations. This divergence suggests that different operational definitions of apparent

urgency capture different aspects of written expression, rather than interchangeable measurements of a single construct.

The simultaneous importance of token count, character count, and sentence count also warrants careful interpretation. These variables are related but not identical, and their coefficients may reflect differences in message length, word length, sentence segmentation, or correlations among predictors. Their signed values should not be interpreted as independent causal effects. In a regularized multivariable model, a coefficient describes the association of a feature conditional on the other included variables and the model's preprocessing. Correlated predictors can acquire opposing coefficients even when their marginal relationships are similar.

Coefficient stability across outer folds adds useful information but should also be interpreted narrowly. Selection in all five folds indicates that a feature was repeatedly retained under the study's validation procedure; it does not establish that the feature is clinically stable, causally related to ADHD, or invariant across different writing environments. Features selected in only three folds, including first-person density and the topic-shift/coherence composite, provide weaker evidence of stable model contribution.

Taken together, these results support a multidimensional interpretation of the measured linguistic differences. The weighted classifier does not identify a general increase in cognitive congestion. Instead, it identifies a combination of structural, lexical-diversity, and surface-expression variables with different directional associations. This is potentially more informative than a scalar index, but the psychological meaning of the coefficient pattern remains an empirical question requiring independent validation.

**6.6** Feature-family ablation identifies predictive contributions, not cognitive mechanisms

The ablation analysis in Table 6 and Figure 4 examines the contribution of predefined feature families to the original 25-feature transparent profile. Removing affective-urgency features reduced balanced accuracy by 0.023, while removing lexical-diversity features reduced it by 0.021. Removing length/structure features produced a smaller decline of 0.013. In contrast, removing repair/uncertainty features had virtually no effect, and removing local-coherence features reduced balanced accuracy by only 0.001.

These results indicate that predictive information was distributed unevenly across the proposed cognitive-congestion dimensions. Affective-urgency, lexical-diversity, and length/structure features contributed more to the original classifier's performance than the repair/uncertainty and local-coherence families. The relatively weak contribution of local-coherence measures is particularly relevant because disrupted discourse organization is an intuitively plausible manifestation of cognitive congestion. The results do not demonstrate that discourse coherence is unrelated to ADHD. Rather, they indicate that the particular local-coherence measurements used here contributed little incremental predictive information under the study's sampling and modeling conditions. Their limited contribution may reflect weak construct alignment, measurement noise, redundancy with other features, or genuine

overlap in discourse characteristics across the selected communities.

Similarly, the negligible effect of removing repair/uncertainty features does not establish that uncertainty or self-correction is psychologically unimportant. Such behaviors may be highly context-dependent, and their surface manifestations in informal online posts may not adequately represent the cognitive processes of interest.

Ablation measures the effect of removing a feature family from a fitted predictive framework; it does not directly measure the causal importance of that family or the underlying psychological mechanism. Correlated features may compensate for removed variables, while weakly measured constructs may appear unimportant despite theoretical relevance. The analysis also applies specifically to the original transparent profile, not to the subsequently optimized weighted model. Its results therefore should not be treated as a direct decomposition of the weighted model's performance.

The broader implication is that theoretical plausibility and predictive utility must be evaluated separately. Features should not be retained solely because they correspond to an appealing cognitive hypothesis, nor discarded as theoretically irrelevant solely because they contribute little to one classifier. The ablation results instead provide an empirical basis for prioritizing future refinement of the measurement framework.

**6.7** Strong lexical classification does not establish a cognitive-linguistic biomarker

The most striking performance contrast appears in Table 3 and Figure 1. The TF-IDF classifier achieved balanced accuracy of 0.858, ROC-AUC of 0.940, average precision of 0.836, and F1 of 0.743, substantially outperforming every transparent-profile configuration.

Viewed in isolation, this performance might appear to provide strong evidence that ADHD can be identified from language. However, the actual prediction target was membership in an ADHD-labeled Reddit community rather than independently verified ADHD diagnosis. The classifier therefore demonstrates that the writing associated with the sampled community is highly distinguishable from writing associated with the four comparison communities. It does not establish that the learned distinctions represent intrinsic linguistic manifestations of ADHD.

The difference is not semantic but methodological. Online mental-health communities are organized around conditions, experiences, treatments, and shared identities. Their members are likely to discuss different medications, symptoms, concerns, and daily-life problems. A lexical classifier can exploit these differences without learning anything about the cognitive processes governing ADHD associated sentence construction, discourse organization, or language production.

Condition-name masking reduces the most direct form of label leakage but does not eliminate topic-related information. Medication names, associated symptoms, treatment discussions, and community-specific terminology can remain highly informative even when explicit condition labels are removed. Consequently, strong TF-IDF performance following masking is not

sufficient evidence that the classifier has learned a condition-specific linguistic phenotype.

The near-identical performance of TF-IDF alone and TF-IDF combined with the transparent profile provides a second important result. Adding the LCCI features increased balanced accuracy by only 0.001, from 0.858 to 0.859, without improving ROC-AUC, average precision, or F1. Thus, the interpretable features did not provide a measurable incremental predictive advantage over the lexical representation under the tested conditions.

This finding admits several possible explanations. The transparent features may contain little additional information beyond the lexical representation. Alternatively, TF-IDF may indirectly capture some of the same linguistic regularities through correlated words and expressions. It is also possible that the transparent features contain independent information too weak to improve aggregate predictive metrics in this sample. The current experiments do not distinguish these explanations.

What can be concluded is narrower but important: the proposed cognitive-congestion profile has not demonstrated incremental predictive utility when added to the stronger lexical baseline. This does not render the transparent measurements scientifically irrelevant, because their intended interpretability differs from that of TF-IDF. It does, however, prevent the claim that they improve the best-performing classifier.

**6.8** Lexical signal-source auditing exposes the distinction between topic recognition and language phenotype

The TF-IDF signal-source audit in Tables 7 and 8 and Figures 5 and 6 addresses the critical question of what the lexical classifier appears to be using. Among the 60 highest-magnitude coefficients examined from one fitted validation fold, 59 (98.3%) were categorized as topic-related. Diagnosis/symptom terminology accounted for 19 terms, medication/treatment for 13, trauma/self-harm/affect for 12, school/work/attention for eight, and community identity for seven. Only one term was categorized as general lexical affect.

The representative coefficients reinforce this pattern. ADHD-associated terms included “adderall,” “medication,” “meds,” “vyvanse,” and “focus,” whereas comparison-associated vocabulary included references to trauma, intrusive thoughts, flashbacks, autism spectrum conditions, compulsions, and suicide. These are readily interpretable differences in the subjects discussed by the communities. They are not, by themselves, measurements of cognitive congestion.

The audit therefore offers a plausible explanation for the divergence between lexical and transparent-profile performance. The TF-IDF model has access to highly discriminative topical information, whereas the interpretable profile is deliberately constrained to more general properties of written expression. A high-performing lexical classifier and a modestly performing discourse classifier may consequently be measuring substantially different phenomena.

Nevertheless, the 98.3% figure must not be interpreted as indicating that 98.3% of predictive performance is attributable to topical vocabulary. It represents the proportion of selected high-magnitude coefficients assigned to topic-related categories. Coefficient magnitude does not directly equal contribution to overall predictive accuracy, and the audit covers only 60 terms from one fitted fold rather than the entire lexical representation across all folds.

The audit is therefore best understood as a qualitative investigation of signal sources, not a quantitative decomposition of model performance. Establishing the actual dependence of TF-IDF performance on topic-related vocabulary would require additional experiments, such as removing predefined lexical categories, refitting the model, and measuring the resulting change in held-out performance. Similar analyses across all validation folds would help establish whether the observed thematic pattern is stable.

Even with these limitations, the audit contributes an important methodological safeguard. It shows that high classification performance should trigger an examination of the vocabulary driving that performance before biological or psychological meaning is assigned to the model. In this dataset, the strongest audited lexical associations provide substantial reason to question whether the classifier's success reflects a condition-specific linguistic phenotype rather than recognition of community subject matter.

**6.9** Negative-manifold testing directly challenges the ADHD-specific outlier hypothesis

The negative-manifold analysis in Table 9 and Figures 7 and 8 provides a distinct test of specificity. Whereas supervised classification asks whether a model can distinguish labeled groups using their known examples, the negative-manifold approach asks whether a held-out community exhibits unusually high anomaly scores relative to a reference distribution learned without that community.

This distinction is central to the manuscript's novelty. A supervised model is explicitly optimized to separate ADHD-labeled posts from comparison posts. It can therefore exploit even small, distributed differences between the groups. An unsupervised residual model is not trained to maximize that particular separation. Instead, it assesses whether the excluded community is unusually distant from the structure represented by the other communities.

Across all three methods, the ADHD-specific outlier hypothesis was unsupported. ADHD ROC-AUC values were 0.451 for PCA reconstruction error, 0.419 for shrinkage Mahalanobis distance, and 0.429 for isolation forest. The corresponding Cohen's d values were −0.170, −0.263, and −0.298. Thus, ADHD-labeled posts did not exhibit the elevated anomaly scores expected if they formed a distinct residual population relative to the modeled comparison-community manifold. The consistency of this result across methods is noteworthy because the methods characterize deviations differently. PCA reconstruction error measures how poorly observations can be represented within a learned lower-dimensional structure. Shrinkage Mahalanobis distance evaluates deviations relative to the estimated covariance structure of the reference distribution. Isolation forest estimates anomaly

through a different, tree-based mechanism. Their agreement reduces the likelihood that the negative finding is merely an artifact of one particular residual-scoring procedure.

OCD exhibited the highest ROC-AUC values among the evaluated communities, ranging from 0.545 to 0.568. These values were only modestly above chance and should not be interpreted as demonstrating a strong or clinically meaningful OCD-specific linguistic phenotype. They establish only that OCD was the most residual community under the reported comparisons.

The below-chance ADHD ROC-AUC values also require careful interpretation. They indicate that, under the chosen scoring direction, ADHD-labeled observations tended not to receive higher anomaly scores than held-out reference observations. They do not demonstrate that ADHD language is the inverse of a distinctive phenotype, nor do they prove that ADHD is linguistically indistinguishable from all other conditions.

The manifold analysis tests a particular proposition: whether the selected features position ADHD-labeled writing as unusually distant from the modeled psychiatric-language reference distribution. A condition-specific signature could instead appear as a subtle shift within a shared manifold, a different covariance structure, a context-dependent pattern, or a nonlinear combination not adequately captured by the selected features and methods. Failure to detect residual distinctiveness therefore rejects the tested outlier hypothesis, not every possible form of ADHD-associated linguistic specificity.

**6.10** Supervised separability and unsupervised distinctiveness need not coincide

The apparent disagreement between supervised and negative-manifold results is one of the most informative aspects of the study. The transparent profile achieved modest supervised discrimination, particularly after learned weighting, yet ADHD-labeled writing was not a positive residual outlier under any of the three manifold methods.

These findings are compatible because classification and anomaly detection optimize different objectives. Supervised discrimination depends on differences between labeled distributions, which may involve modest shifts in their centers or combinations of features. Anomaly detection depends on whether observations are unusually distant or poorly represented relative to a learned reference distribution. A group can therefore be partially predictable without being anomalous.

For example, two linguistic communities could occupy extensively overlapping regions of feature space while differing slightly in the relative frequencies of several features. A supervised classifier might exploit these shifts to obtain above-chance performance. An unsupervised anomaly detector, however, might assign similar residual scores to both groups because neither lies outside the broader distribution.

This distinction clarifies why the modest performance of the transparent profile does not contradict the negative-manifold finding. The former establishes some predictive association with community labels; the latter fails to establish that ADHD-labeled posts constitute a distinct outlying linguistic population.

More broadly, the results suggest that linguistic biomarker evaluation should distinguish at least three levels of evidence: predictive discrimination, construct-related measurement, and condition specificity. Predictive discrimination is necessary for many classification applications but does not establish the other two. Construct-related measurement requires evidence that features correspond to the proposed cognitive processes. Condition specificity requires evidence that those measurements distinguish the target condition from relevant alternatives, rather than reflecting shared distress, general psychopathology, topic selection, or communication context.

The present study supplies evidence of community-label discrimination and limited evidence of stable multivariable linguistic differences. It does not yet establish cognitive construct validity or ADHD-specificity. Making this distinction explicit converts an otherwise ambiguous mixture of positive and negative findings into a coherent evaluation of competing interpretations.

**6.11** A shared psychiatric-language manifold as a hypothesis for future investigation

The absence of ADHD-specific residual distinctiveness raises a further hypothesis: some of the measured discourse characteristics may be shared across mental-health communities rather than unique to ADHD. Self-reference, uncertainty, urgency, lexical repetition, variable sentence structure, and local coherence changes may arise in multiple psychological or communicative contexts.

Under this interpretation, the selected communities may occupy overlapping regions of a broader psychiatric-language feature space. Their differences may involve shifts in feature prevalence or combinations rather than discrete, diagnosis-specific linguistic boundaries.

This interpretation is consistent with the modest effect sizes in Table 2, the limited performance of the transparent classifiers, and the failure of ADHD to emerge as a residual outlier. However, the present study does not establish a common psychiatric-language phenotype. The reference distribution was constructed from selected Reddit communities, not from an independently characterized population with verified diagnoses and standardized writing conditions. Shared features may reflect platform conventions, emotional self-disclosure, demographic differences, or other unmeasured factors.

A direct test of the shared-manifold hypothesis would require a broader reference design incorporating nonclinical writing, additional psychiatric conditions, and clinically characterized participants. It would also require determining whether the same feature relationships persist when writing topic, task, and communication setting are controlled.

The hypothesis nevertheless provides a constructive direction for interpreting the negative result. Rather than assuming that every psychiatric condition must possess a unique linguistic signature, future research could first identify general dimensions of linguistic variation and then test whether individual conditions produce reproducible deviations within those dimensions. This would shift the emphasis from searching for diagnostic uniqueness by default toward

experimentally establishing which aspects of language are shared, condition-associated, or condition-specific.

**6.12** Implications for digital biomarker development

The combined findings suggest that digital linguistic biomarker development should proceed through a sequence of increasingly demanding evidentiary tests.

The first question is whether a proposed feature or model can distinguish groups under appropriate validation. The present study shows that TF-IDF performs strongly on community labels, while the transparent features perform modestly.

The second question is whether the observed predictive signal arises from the intended construct. The lexical audit shows why this cannot be assumed: a model may primarily exploit vocabulary associated with diagnoses, treatments, or community discussions. The interpretable profile is more closely aligned with discourse measurement by design, but that alignment alone does not establish that the features measure cognitive congestion.

The third question is whether the proposed signal is specific to the target condition. The negative-manifold results show that ADHD-labeled posts were not unusually distant from the comparison-community reference distribution within the tested feature space.

A fourth question, not resolved here, is whether the signal generalizes beyond the original sampling environment. This requires testing on independently collected writing, ideally from clinically characterized participants under comparable elicitation conditions. A fifth question is whether the measurements remain stable within individuals and respond meaningfully to relevant changes in cognitive state or clinical status.

These distinctions are particularly important when the term *biomarker* is used. The current LCCI framework is best described as an exploratory, interpretable linguistic measurement candidate. It has not demonstrated the clinical validity, diagnostic specificity, or external generalizability required to support a clinical biomarker claim.

The study's value lies partly in showing how such claims can be evaluated before translation. A negative specificity result should not be regarded as a failed analysis if it prevents an otherwise attractive but unsupported interpretation of predictive accuracy. In this sense, specificity testing is not merely an optional supplement to classifier evaluation; it is an important safeguard against mistaking dataset separability for clinically meaningful measurement.

**6.13** A targeted experimental path toward establishing linguistic specificity

The present findings identify several concrete experiments that would distinguish competing explanations for the observed results.

First, topic-controlled lexical experiments could determine how much of the TF-IDF classifier's performance depends on clinical and community-specific vocabulary. The audited categories provide a starting point for predefined vocabulary removal or masking, followed by complete model refitting and evaluation on held-out authors. Comparing

performance before and after these interventions would provide a more direct estimate of the predictive contribution of topical information than coefficient inspection alone.

Second, the transparent profile could be evaluated under matched writing conditions. Participants from different diagnostic groups could complete identical writing prompts, with comparable text lengths and communication settings. This would reduce differences arising simply because communities discuss different topics. If the interpretable features retained their discriminatory performance under these conditions, the case for a condition-associated linguistic signal would become stronger.

Third, the measurement framework could be refined through construct-specific experiments. Features intended to represent repair, uncertainty, coherence, urgency, and discourse organization could be evaluated against independently measured cognitive or behavioral variables. Such studies would test whether the linguistic features correspond to the proposed cognitive processes rather than merely correlating with community membership.

Fourth, negative-manifold analysis could be repeated with broader and independently collected reference populations. In particular, comparing ADHD against both nonclinical writing and clinically characterized psychiatric groups would help distinguish general departures from typical writing from differences specific to ADHD. The same residual methods could then be evaluated across alternative feature spaces and controlled elicitation conditions.

Fifth, longitudinal measurements could assess whether the proposed features exhibit within-person stability and meaningful state dependence. A useful cognitive-language measure should ideally distinguish relatively stable individual characteristics from fluctuations caused by task demands, emotional state, medication, fatigue, or conversational context. Such experiments would be particularly valuable because the present cross-sectional Reddit analysis cannot establish whether observed group differences are trait-related or situational.

These experiments represent a progression from identifying statistical associations toward establishing what the associations mean. They also provide explicit criteria by which the cognitive-congestion hypothesis could be strengthened, modified, or rejected.

**6.14** Scientific novelty and overall interpretation

The novelty of this study lies less in proposing another language-based ADHD classifier than in examining the evidentiary relationship between classification, interpretability, lexical signal sources, and diagnostic specificity within a unified framework.

Four findings are particularly consequential. First, the near-chance equal-weight LCCI scalar demonstrates that theoretically motivated linguistic features cannot automatically be combined into a meaningful unidirectional score. Second, the improvement obtained through multivariable modeling and learned weighting shows that retaining feature structure can recover modest predictive information that scalar aggregation obscures. Third, the TF-IDF audit reveals that the strongest inspected lexical

predictors are overwhelmingly associated with clinical topics, treatments, and community-specific subject matter, complicating the interpretation of high classification accuracy as evidence of an intrinsic language phenotype. Fourth, the negative-manifold analysis demonstrates that above-chance supervised classification does not necessarily imply that the target community constitutes a distinctive linguistic outlier.

These findings are complementary rather than mutually exclusive. They suggest that ADHD-labeled writing contains detectable linguistic and lexical associations, but that the most powerful associations in the present dataset are not established as cognitive-congestion markers, while the interpretable features have not demonstrated ADHD-specific residual distinctiveness.

The study therefore advances a methodological distinction with implications beyond ADHD: a model may predict a diagnostic-community label accurately without measuring a condition-specific phenotype, and a transparent feature set may capture reproducible group differences without constituting a validated biomarker.

The most defensible conclusion is consequently neither that the proposed cognitive-congestion framework has been validated nor that linguistic investigation of ADHD is unproductive. Rather, the framework has undergone a series of increasingly demanding tests that identify where its present evidentiary support ends. The results favor a multidimensional approach to linguistic measurement, expose the interpretive limitations of topic-driven classification, and demonstrate the importance of evaluating specificity independently of supervised performance.

By identifying these boundaries explicitly, the manuscript offers a more rigorous foundation for future language-based research than a classification study reporting predictive accuracy alone.

## 7. Limitations

The most important limitation is label validity. Subreddit membership is not clinical diagnosis. The findings therefore cannot be used for individual-level ADHD screening, triage, diagnosis, or clinical decision-making. They are corpus-level associations between ADHD-labeled public writing and other mental-health-community writing.

The second limitation is residual leakage. Direct condition names were masked, but medication names, treatment discussions, school and work topics, symptom descriptions, and community vocabulary remain. The TF-IDF audit indicates that many high-weight coefficients correspond to residual semantic vocabulary, but attribution of overall classifier performance would require ablation or stronger causal tests. We believe future work should test stronger semantic decontamination, held-out topic controls, and external corpora where labels are not community names.

The third limitation is the evaluation unit and author contribution imbalance. Performance metrics are computed at the post level, while author-grouped folds prevent posts from the same known author from appearing in both training and test folds. Consequently, authors who contribute multiple posts receive proportionally greater weight in the reported

post-level metrics. Author grouping reduces dependence caused by author reuse but does not make the estimand author-level performance. Public datasets may also contain deleted, missing, or ambiguous author identifiers. The reported fold variability is descriptive split variability, not a formal clinical confidence interval.

The fourth limitation is modality. The theory of cognitive congestion may be better tested in spoken transcripts, narrative tasks, or experimentally constrained writing than in Reddit posts. TalkBank remains an appropriate future validation pathway once authenticated transcript access and corpus-specific governance requirements are satisfied.

## 8. Reproducibility and data availability

The analysis is designed to be reproducible from scripts and saved artifacts. The main original experiment is implemented in run_real_text_lcci.py. The residual-manifold analysis is implemented in run_pivot_manifold_study.py. The weighted-profile analysis is implemented in run_lcci_v2_biomarker_candidate.py. The signal-source audit is implemented in analyze_signal_source_tfidf_vs_lcci.py. The present manuscript is generated by build_lcci_v2_pivot_submission.py. The accompanying reproducibility archive contains the scripts, dependency specification, dataset and sampling manifests, aggregate tables, figures, the complete 60-coefficient audit, categorization rules, and reproduction instructions. The executable code and minimal reproduction instructions are publicly available at https://github.com/GitSE1/LCCI; raw Reddit text and generated result files are not redistributed in that repository. Raw social-media text is not reproduced in the manuscript tables. Exported feature files use deterministic features and row hashes. The weighted-profile manifest records the endpoint as "ADHD-labeled subreddit posts versus other mental-health-community posts" and the validation design as "Nested author-grouped validation: 5 outer GroupKFold folds; 3 inner GroupKFold folds for hyperparameter selection." using the no-raw-text feature source. Public release should include code, manifests, aggregate tables, figures, and instructions for obtaining the original Hugging Face dataset under its own terms.

## 9. Conclusion

This study demonstrates that the ability to classify ADHD-related writing is fundamentally different from identifying an ADHD-specific linguistic phenotype. Although conventional TF-IDF classification achieved high predictive performance (ROC-AUC 0.940), the proposed interpretable cognitive-congestion language profile (LCCI) showed only modest discrimination (ROC-AUC 0.654), improving to 0.681 with learned feature weights. Individual linguistic differences were uniformly small, the equal-weight composite performed near chance, and combining LCCI with TF-IDF produced virtually no additional predictive benefit. Furthermore, a lexical audit identified medication, diagnostic, symptom, and community-related vocabulary among 59 of the 60 highest-weighted TF-IDF coefficients, suggesting substantial dependence on contextual language rather than demonstrating a distinctive cognitive phenotype. These findings establish that predictive accuracy, statistical association, and mechanistic interpretability cannot be treated as equivalent evidence of a linguistic biomarker.

The principal methodological contribution is the introduction of a complementary *negative-signature* framework that reverses the conventional classification question. Rather than training a model to recognize ADHD, each psychiatric community was excluded in turn while a reference language manifold was constructed exclusively from the remaining communities. The excluded condition was then evaluated for residual distinctiveness using three independent approaches: PCA reconstruction error, shrinkage Mahalanobis distance, and isolation forest scoring. This approach tests whether a candidate disorder exhibits linguistic characteristics that cannot be adequately represented by the language structure learned without that disorder. It therefore provides an additional, falsifiable test of disorder specificity that does not depend on learning the target disorder's positive signature.

Contrary to the original ADHD-specificity hypothesis, ADHD did not exhibit a distinctive positive residual signal under any of the three methods (ROC-AUC 0.451, 0.419, and 0.429), whereas OCD showed the greatest residual separability. Thus, although ADHD-related writing was readily distinguishable using conventional lexical classification, it was not an unusually distant outlier within the psychiatric-language feature space examined. This negative result does not establish that ADHD lacks distinctive linguistic characteristics, but demonstrates that the proposed cognitive-congestion features do not establish such distinctiveness in the present corpus.

The broader scientific contribution is therefore a shift from *detecting a disorder* to *testing whether the disorder possesses an independently distinguishable signal*. The study establishes a reproducible framework combining interpretable feature engineering, lexical signal auditing, supervised classification, and leave-one-disorder-out residual analysis. Importantly, the negative-signature approach can be extended to other psychiatric conditions and candidate digital biomarkers, providing a means of challenging apparent disorder specificity before clinical translation. The central finding is that strong classification performance can coexist with weak interpretable feature differences and an absence of positive residual distinctiveness. Consequently, the scientific value of a candidate linguistic biomarker should be judged not merely by how accurately it predicts a diagnostic label, but also by whether it reveals information that remains distinctive when the target disorder is excluded from model construction. This distinction, rather than the identification of an ADHD biomarker, represents the principal contribution of the present study.

## 10. References

1. Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (2013). Reaction time variability in ADHD: A meta-analytic review of 319 studies. *Clinical Psychology Review, 33(6)*, 795–811. https://doi.org/10.1016/j.cpr.2013.06.001

2. Kuijper, S. J. M., Hartman, C. A., & Hendriks, P. (2015). Who is he? Children with ASD and ADHD take the listener into account in their production of ambiguous pronouns. *PLOS ONE, 10*, e0132408. https://doi.org/10.1371/journal.pone.0132408

3. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1485–1497.

4. Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 51–60. https://doi.org/10.3115/v1/W14-3207

5. Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 1–10. https://doi.org/10.3115/v1/W15-1201

6. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 31–39. https://doi.org/10.3115/v1/W15-1204

7. Brown, M. R. G., Sidhu, G. S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P. H., Greenshaw, A. J., & Dursun, S. M. (2012). ADHD-200 global competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Frontiers in Systems Neuroscience, 6*, Article 69. https://doi.org/10.3389/fnsys.2012.00069

8. Wiebe, A., Selaskowski, B., Paskin, M., Asché, L., Pakos, J., Aslan, B., Lux, S., Philipsen, A., & Braun, N. (2024). Virtual reality-assisted prediction of adult ADHD based on eye tracking, EEG, actigraphy and behavioral indices: A machine learning analysis of independent training and test samples. *Translational Psychiatry, 14*, Article 508. https://doi.org/10.1038/s41398-024-03217-y

9. quocanh34. (n.d.). *mental_health_dataset_1* [Data set]. Hugging Face. https://huggingface.co/datasets/quocanh34/mental_health_dataset_1

10. MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods*, 51, 1919–1927. https://doi.org/10.3758/s13428-018-1174-9

11. Hendriks, P., Koster, C., & Kuijper, S. J. M. (n.d.). *ASDBank Dutch asymmetries corpus* [Data set]. TalkBank. https://doi.org/10.21415/T5911K

12. TalkBank. (n.d.). *TBDBpy: Python API to TalkBankDB* [Computer software]. GitHub. https://github.com/TalkBank/TBDBpy

13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html