

Repurposing Large Language Models as Instruments to Detect Transdiagnostic Speech Deviations across the Psychosis–Affective Spectrum

Esra Lenz

esra.lenz@gmail.com

Central Institute of Mental Health, Heidelberg University

Svenja Seuffert

Marburg University

Julius Wiegert

Central Institute of Mental Health, Heidelberg University

Rieke Roxanne Mülfarth

Marburg University

Agata Mrozek

Marburg University

Tilo Kircher

Marburg University

Emmanuel Schwarz

Central Institute of Mental Health, Heidelberg University

Frederike Stein

Marburg University

Article

Keywords:

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10942665/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Repurposing Large Language Models as Instruments to Detect Transdiagnostic Speech Deviations across the Psychosis–Affective Spectrum

Esra Lenz¹, Svenja Seuffert^{2,3}, Julius Wiegert^{1,4}, Rieke Roxanne Mülfarth^{2,3}, Agata Mrozek^{2,3}, Tilo Kircher^{2,3}, *Emmanuel Schwarz^{1,4,5} and *Frederike Stein^{2,3}

¹Hector Institute for Artificial Intelligence in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

²Marburg University, School of Medicine, Department of Psychiatry and Psychotherapy, Marburg, Germany

³Center for Mind, Brain and Behavior, Marburg University, Marburg, Germany

⁴Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

⁵German Center for Mental Health (DZPG), Partner Site Mannheim - Heidelberg - Ulm, Germany

Corresponding Authors:

Esra Lenz (esra.lenz@gmail.com)

Frederike Stein (frederike.stein@staff.uni-marburg.de)

Emanuel Schwarz (emanuel.schwarz@zi-mannheim.de)

*Emanuel Schwarz and Frederike Stein contributed equally as senior authors.

Abstract

Trained on vast text corpora, large language models (LLMs) encode statistical expectations about language. We tested whether deviations from these expectations, quantified as LLM surprisal, capture clinically relevant speech alterations. Surprisal was defined as median sentence-level cross-entropy of Qwen2.5-32B-Instruct conditioned on a healthy-speaker persona. In this cross-sectional German sample with schizophrenia-spectrum (SSD, n=48), bipolar (BD, n=26) or depressive disorders (MDD, n=119) and healthy controls (n=178), surprisal was elevated in SSD ($\beta=0.49^{**}$) and BD ($\beta=0.39^{*}$), but not MDD, after adjustment for demographic confounders and average sentence length. Surprisal correlated with eight of 18 symptom scales, including positive formal thought disorder ($p=0.23^{**}$), largely reflecting between-group rather than within-group differences. It improved model fit beyond sentence length, confounders and semantic coherence, although discrimination remained modest (AUC 0.74 SSD, 0.62 BD), and was robust across prompts and models. LLM surprisal may provide a scalable marker of language alterations across psychiatric disorders, requiring further validation.

Introduction

Psychiatry is, to a large extent, a clinical science of language and speech^{1,2}. In the absence of established biological markers³, clinicians infer diagnosis², symptom severity⁴ and deviations of healthy functioning primarily from patients' verbal self-report. For this purpose, they use mainly speech, supplemented by psychomotor⁵ and expressive features observed during the clinical psychiatric interview⁶. As specific dimensions of this speech behavior, clinicians rate syntactic structure, semantic connectedness, and discourse flow of speech. Disruptions across these domains characterize formal thought disorder (FTD)—a disturbance in the organization of language^{7,8} and thought, and a hallmark feature of schizophrenia (SZ)⁷⁻⁹. Unlike symptoms inferred primarily from patients' self-reported inner experience, many manifestations of FTD are directly expressed in the form and organization of speech and are therefore, in principle, observable and measurable. In practice, however, speech assessment relies on trained clinical raters, whose judgments show heterogeneous inter-rater reliability despite extensive standardization efforts¹⁰ across rating scales^{11,12} and classification of diagnosis¹³⁻¹⁵.

To make such speech particularities scalable and objective, computational psychiatry has applied natural language processing (NLP) to psychiatric speech¹⁶⁻²². Approaches like latent semantic analysis first quantified incoherence in SZ²³, and contextual representations from transformer-based architectures^{24,25} later extended this to automated FTD detection, in some cases distinguishing patients from controls even when clinician-rated thought-language disturbance did not differ²⁶, suggesting these metrics capture clinically relevant variation below the threshold of human observation or may even reflect dimensions of language pathology beyond traditional rating scales²⁷. However, these earlier approaches primarily measured semantic relatedness, which largely neglects one fundamental property of human language: its predictability based on preceding context^{28,29}.

Autoregressive large language models (LLMs) possess the needed inherent ability and scale to extend the idea from *relatedness* to *predictability*^{24,30}. These LLMs are trained on vast amounts of text to predict the most probable next token given the preceding context. This training objective can be turned on its head and repurposed: the same probability estimates that drive text generation also provide a direct computational measure of linguistic predictability. This can be operationalized via the LLM's prediction loss, quantified as cross-entropy loss and its so-called corresponding perplexity^{31,32}. Because the model evaluates semantic and contextual likelihood alongside syntax, elevated deviation stems from both grammatical breakdown and grammatically fluent yet derailed speech (Figure 1), capturing and potentially extending the FTD construct. These LLM-derived measures have already yielded initial positive results in SZ^{32,33}, first evidence in bipolar disorder (BD) in a Korean cohort³⁴ and findings in dementia^{35,36}. They carry important caveats: they are susceptible to social, educational, and demographic confounds^{37,38}, and fundamental methodological questions remain, regarding optimal model architecture and scale^{31,39}, as well as the mitigation of measurement-inherent text-length artifacts⁴⁰.

Developing separately from these measurement advances, a growing corpus of studies indicates that speech disturbances such as FTD are not confined to SZ^{8,41-46}.

Figure 1 Visualization of the concept of LLM surprisal as a measure of deviation of statistical expectation

① Fluent / Normal Speech

Grammatically fluent, coherent, and highly typical picture description.

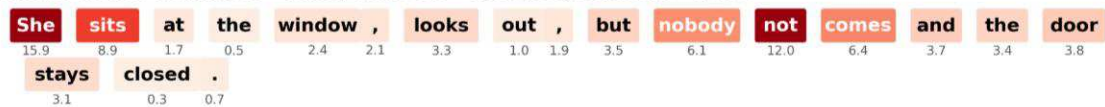
Sentence Metrics: Median Loss = 2.259 | Mean Loss = 3.373 | 95th Percentile = 11.083



② Syntactic Disorganization

Characterized by severe grammatical violations, missing auxiliary verbs, and poor inflection.

Sentence Metrics: Median Loss = 3.293 | Mean Loss = 4.253 | 95th Percentile = 12.389



③ Semantic / Content Disorganization

Grammatically correct, but introduces highly bizarre, non-sequitur delusional content.

Sentence Metrics: Median Loss = 3.500 | Mean Loss = 5.450 | 95th Percentile = 12.796



④ Pathological Content (Auditory Hallucinations / Suicidal Ideation)

Structurally cohesive, but contains explicit pathological content and auditory voices.

Sentence Metrics: Median Loss = 2.288 | Mean Loss = 3.638 | 95th Percentile = 12.690



low loss  high loss

Token Loss (Surprisal)

Token-level surprisal from an autoregressive LLM (Qwen-2.5-32B-Instruct) primed with a healthy persona (Supplementary Table 9 - English translated). Each box is one model token; the number indicates how "surprised" the model was by the token given the context preceding it. Because the LLM was trained on large amounts of text and instruction-tuned to produce coherent text, its statistical expectations also reflect typical language. Surprisal therefore indexes how far speech departs from the inherently expected next token.

Panel 1 — Fluent/Normal Speech: low surprisal throughout.

Panel 2 — Syntactic Disorganization: spikes at the grammatical errors.

Panel 3 — Semantic/Content Disorganization: high surprisal on perfectly grammatical words, showing the measure captures meaning, not just grammar.

Panel 4 — Pathological Content (Suicidal/Depressive): one local spike, where content derails to pathological content.

Two patterns are visible in every panel: surprisal is high at the start of a sentence, where the model has little context to predict from, and falls as the sentence continues and context builds up. Longer sentences therefore look more predictable on average. Therefore sentence-length emerges as an important confounder.

Given examples are synthetic and no patient data.

FTD dimensions vary continuously across MDD, BD, and SSD, and model-based clustering has identified transdiagnostic FTD profiles whose members are distributed across clinical diagnoses rather than aligning with them^{41,44,45,47,48}. Additionally dimensional frameworks — the Hierarchical Taxonomy of Psychopathology⁴⁹, Research Domain Criteria⁵⁰, and the proposed language ("L") factor⁵¹ — recast language and thought disturbance as a continuum crossing diagnostic boundaries⁵². For example, they place schizophrenia spectrum disorders (SSD) and BD along a shared psychosis spectrum⁵³, and disturbed predictive language processing has been proposed as a mechanism relevant across conditions^{54–56}. Yet the existing evidence derives largely from small, mostly English-language cohorts⁵⁷, examining single diagnoses (primarily SZ) and focusing on limited constructs, such as a thought-disorder rating or diagnosis discrimination⁵⁸, rather than the broader symptom and cognitive dimensions.

The present study is motivated by the convergence of these two developments: a language-model signal, trained on vast amounts of text, that indexes predictability rather than mere relatedness, and a dimensional view of language pathology that crosses diagnostic boundaries. How such LLM-derived measures behave across a transdiagnostic sample, and whether they withstand the methodological limitations noted above, therefore remains unknown.

To address these gaps we quantified the predictability of spontaneous narrative speech as the LLM surprisal, computed as the median cross-entropy LLM loss, in a deeply phenotyped German sample⁵⁹ of patients with SSD, BD, and major depressive disorder (MDD) and healthy controls (HC), using the Thematic Apperception Test (TAT)⁶⁰ as a standardized narrative stimulus across participants. We tested three primary clinical hypotheses: (i) that speech is less predictable - scoring higher for LLM surprisal - in patients than in controls, most pronounced along the psychosis spectrum; (ii) that this LLM surprisal reflects clinically meaningful variation, associating with symptom and cognitive dimensions across diagnoses, in particular FTD; and (iii) that these associations are not reducible to demographic or syntactic confounds. Complementing these clinical objectives, and given the lack of established best practices for LLM-derived metrics, our secondary analyses focused on key methodological questions: specifically, whether the signal provides incremental value beyond semantic coherence and whether it remains robust across model architectures, prompting techniques, and sensitivity analyses.

By situating a language-model-derived signal within this broader clinical picture, we take a step toward objective tools that could complement clinical assessment in measuring and

monitoring language pathology, and that could supplement existing rating scales with a more automated, cross-diagnostic description of psychiatric patients.

Results

Participants

The initial sample comprised 372 participants: 178 HC and 194 patients (27 BD, 119 MDD, 48 SSD). Years of education (YoE) were missing for 15 participants and were imputed by k-nearest-neighbour imputation (KNN). For one BD patient, the variables used to identify neighbours were themselves incomplete, and this participant was excluded. All analyses were therefore based on 371 participants, including 26 BD patients. SSD, BD and MDD did not differ significantly in age or verbal IQ (Table 1). Regarding clinical presentation, patients with SSD exhibited the highest overall symptom severity across multiple domains, showing the highest scores for positive symptoms (Scale for the Assessment of Positive Symptoms - SAPS⁶¹), negative symptoms (Scale for the Assessment of Negative Symptoms - SANS⁶²), manic symptoms (Young Mania Rating Scale - YMRS⁶³), and depressive symptoms (Hamilton Depression Rating Scale - HAMD-17⁶⁴), while demonstrating the lowest global functioning scores (Global Assessment of Functioning - GAF⁶⁵) (Table 1). Further sample characteristics for this cohort have been reported in previous publications^{43,59,66}.

Confounding Variables

To ensure demographic and syntactic variance did not drive our findings, we residualized raw LLM surprisal against age, sex, YoE, and average sentence length (ASL); throughout, we refer to raw LLM surprisal as *unadjusted* and corrected surprisal as *residualized*. Unadjusted LLM surprisal was most strongly associated with ASL ($\rho = -0.556$, $P < 0.001$), and weakly but significantly with YoE, age, and sex. Residualization removed any significant association with all tested covariates (Supplementary Table 1; Supplementary Figure 1).

We additionally controlled for medication influences. Residualized surprisal showed no significant correlation with chlorpromazine-equivalent antipsychotic dosage within the SSD or with general polypharmacy load transdiagnostically, and all primary symptom associations remained robust, and in some cases strengthened, when controlling for medication dosage (Supplementary Figures 2-3; Supplementary Table 2). Given the absence of strong pharmacological influences we did not include these as primary covariates.

Table 1 Demographic, linguistic, and clinical characteristics by diagnostic group

Variable	HC	BD	MDD	SSD	H / χ^2	df	P
n	178	26	119	48			
Age, years	44.4 $\pm$ 13.5	44.4 $\pm$ 12.4	44.3 $\pm$ 13.8	41.2 $\pm$ 12.2	2.51	3	0.47
Education, years	14.5 $\pm$ 3.1	13.2 $\pm$ 3.0	13.4 $\pm$ 3.0	12.3 $\pm$ 2.3	21.24	3	< 0.001
Verbal IQ	116.1 $\pm$ 12.8	111.5 $\pm$ 13.4	116.4 $\pm$ 14.2	111.9 $\pm$ 15.1	4.21	3	0.24
Female, n (%)	114 (64.0%)	17 (65.4%)	74 (62.2%)	18 (37.5%)	11.90	3	0.008
Comorbidity, n (%)	0 (0.0%)	16 (61.5%)	44 (37.0%)	26 (54.2%)	113.79	3	< 0.001
Words per sentence	20.3 $\pm$ 11.1	19.6 $\pm$ 13.3	18.9 $\pm$ 9.8	15.6 $\pm$ 6.4	12.64	3	0.005
Clinical measures							
HAMD-17 total	1.3 $\pm$ 2.2	5.9 $\pm$ 5.9	5.6 $\pm$ 5.8	9.2 $\pm$ 8.4	8.56	2	0.01
YMRS total	0.4 $\pm$ 1.3	4.5 $\pm$ 5.7	0.4 $\pm$ 1.0	6.0 $\pm$ 7.2	55.86	2	< 0.001
GAF	85.4 $\pm$ 9.0	60.4 $\pm$ 13.0	67.4 $\pm$ 12.6	51.9 $\pm$ 15.7	31.80	2	< 0.001
SAPS total	0.5 $\pm$ 1.3	4.6 $\pm$ 5.6	1.5 $\pm$ 2.9	13.2 $\pm$ 13.8	47.49	2	< 0.001
SANS alogia	0.2 $\pm$ 0.8	1.5 $\pm$ 2.6	1.1 $\pm$ 1.9	2.7 $\pm$ 2.9	15.85	2	< 0.001
SANS affective flattening	0.6 $\pm$ 1.8	3.9 $\pm$ 5.5	2.5 $\pm$ 3.3	6.7 $\pm$ 5.8	14.62	2	< 0.001
SANS abulia/apathy	0.2 $\pm$ 0.7	2.6 $\pm$ 3.0	1.3 $\pm$ 2.0	4.4 $\pm$ 3.7	31.60	2	< 0.001

SANS anhedonia	0.2 ± 0.9	2.4 ± 3.9	2.4 ± 4.0	4.6 ± 4.8	11.34	2	0.003
SAPS positive FTD	0.5 ± 1.5	3.6 ± 4.9	1.2 ± 2.5	6.1 ± 7.6	25.79	2	< 0.001
SAPS delusions	0.0 ± 0.2	0.5 ± 1.3	0.1 ± 0.3	4.6 ± 5.9	65.62	2	< 0.001
SAPS hallucinations	0.0 ± 0.1	0.2 ± 0.7	0.0 ± 0.2	2.5 ± 4.6	36.94	2	< 0.001
SAPS bizarre behaviour	0.0 ± 0.2	0.2 ± 0.7	0.2 ± 0.7	0.6 ± 0.9	15.87	2	< 0.001
TLI impoverished	0.3 ± 0.6	0.6 ± 0.7	0.5 ± 1.1	0.7 ± 1.1	4.94	2	0.08
TLI disorganization	0.6 ± 0.8	0.9 ± 1.2	1.0 ± 1.2	0.8 ± 0.9	0.49	2	0.78
TLI non-specific	0.2 ± 0.3	0.2 ± 0.3	0.3 ± 0.5	0.2 ± 0.3	0.79	2	0.67
Lexical verbal fluency	11.8 ± 4.1	9.2 ± 3.2	11.1 ± 4.3	9.2 ± 4.2	8.68	2	0.01
Semantic verbal fluency	23.4 ± 5.0	20.8 ± 5.1	23.5 ± 5.3	18.5 ± 5.7	25.67	2	< 0.001
Antipsychotic dose (CPZ equivalent)	0.0 ± 0.0	119.8 ± 312.3	9.4 ± 47.2	245.5 ± 331.9	60.02	2	< 0.001
Antidepressant load	0.0 ± 0.1	1.6 ± 2.0	1.1 ± 1.8	0.7 ± 1.2	3.54	2	0.17
General medication index	0.0 ± 0.1	2.6 ± 1.9	0.7 ± 1.2	1.7 ± 1.7	33.46	2	< 0.001

Data are mean ± S.D. for continuous variables and n (%) for categorical variables. For demographic and linguistic variables, the four groups (HC, BD, MDD, SSD) were compared using Kruskal–Wallis tests (H statistic, $df = 3$), except sex, which was compared using a χ^2 test ($df = 3$). Clinical measures were compared across the three patient groups (BD, MDD, SSD) using Kruskal–Wallis tests (H statistic, $df = 2$); healthy-control values are shown for reference only and were not included in these tests. H / χ^2 column reports H for Kruskal–Wallis tests and χ^2 for the sex comparison. HC, healthy controls; BD, bipolar disorder; MDD, major depressive disorder; SSD, schizophrenia spectrum disorder GAF, Global Assessment of Functioning; HAMD-17, 17-item Hamilton Depression Rating Scale; YMRS, Young Mania Rating Scale; SAPS, Scale for the Assessment of Positive Symptoms; SANS, Scale for the Assessment of Negative Symptoms; FTD, formal thought disorder; TLI, Thought and Language Index; CPZ, Chlorpromazine equivalent doses.

Group Differences in LLM surprisal

After deconfounding for age, sex, YoE, and ASL, the overall patient–control difference in LLM surprisal was small and did not reach significance (Figure 2; Table 2 and Supplementary Table 3).

Table 2 Group differences in median LLM surprisal across language models

Comparison	Qwen-2.5-32B-Instruct	Qwen-2.5-7B-Instruct	LeoMistral-Hessianai-7b
HC vs. All Patients	$\beta=0.15$ [-0.02, 0.33], P=0.09	$\beta=0.14$ [-0.04, 0.33], P=0.13	$\beta=0.12$ [-0.01, 0.25], P=0.08
HC vs. SSD	$\beta=0.49$ [0.21, 0.76], P=0.002	$\beta=0.61$ [0.32, 0.90], P=< 0.001	$\beta=0.32$ [0.11, 0.53], P=0.009
HC vs. MDD	$\beta=-0.02$ [-0.21, 0.18], P=0.87	$\beta=-0.05$ [-0.26, 0.15], P=0.61	$\beta=0.00$ [-0.14, 0.15], P=0.97
HC vs. BD	$\beta=0.39$ [0.04, 0.73], P=0.04	$\beta=0.30$ [-0.05, 0.66], P=0.14	$\beta=0.34$ [0.08, 0.60], P=0.01
MDD vs BD	$\beta=0.40$ [0.05, 0.75], P=0.04	$\beta=0.36$ [-0.01, 0.73], P=0.08	$\beta=0.34$ [0.07, 0.61], P=0.02
MDD vs SSD	$\beta=0.50$ [0.22, 0.79], P=0.002	$\beta=0.66$ [0.37, 0.96], P=< 0.001	$\beta=0.32$ [0.10, 0.53], P=0.01
BD vs SSD	$\beta=0.10$ [-0.30, 0.50], P=0.62	$\beta=0.31$ [-0.11, 0.73], P=0.15	$\beta=-0.02$ [-0.33, 0.28], P=0.89

Standardized beta β coefficients, 95% confidence intervals [in brackets], and P-values derived from OLS regression models comparing group differences in residualized median surprisal. Values represent residualized median surprisal after regressing out age, sex, YoE, and ASL via OLS. Positive β values denote higher residualized surprisal (lower text predictability). Estimates are shown for the primary model (Qwen2.5-32B-Instruct) and two validation models: a smaller same-family model (Qwen2.5-7B) and a German-language model (LeoMistral-Hessianai-7b). Significant results ($P < 0.05$) are indicated in bold. $n = 178$ HC, 48 SSD, 26 BD, 119 MDD. HC, healthy controls; SSD, schizophrenia spectrum disorder; BD, bipolar disorder; MDD, major depressive disorder.

In contrast to this null finding, patients with SSD showed the strongest elevation over HC ($\beta = 0.49$ [0.21, 0.76], $P = 0.002$), followed by patients with BD ($\beta = 0.39$ [0.04, 0.73], $P = 0.04$), whereas patients with MDD were indistinguishable from HC.

In the in-between group comparison LLM surprisal was significantly higher in SSD than in MDD and significantly higher in BD than in MDD, while SSD and BD did not differ significantly.

Additionally controlling for Verbal IQ, group elevations relative to controls remained significant for SSD ($\beta = 0.18$, $P = 0.001$) and non-significant for MDD ($\beta = 0.00$, $P = 0.912$), but became non-significant for BD ($\beta = 0.12$, $P = 0.068$).

Clinical Correlations

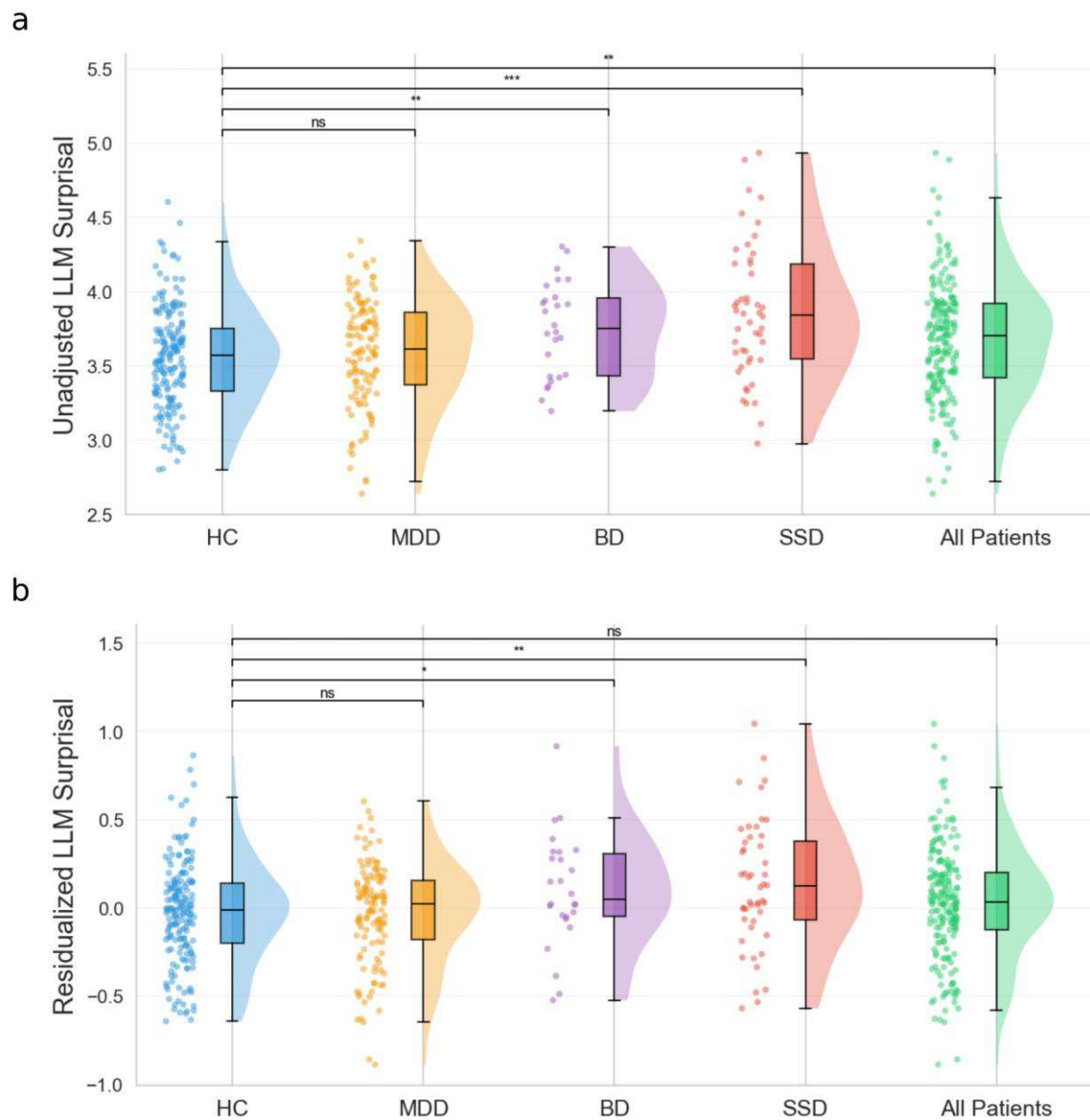
Initially, unadjusted LLM surprisal was significantly correlated (Spearman) with 13 of 18 clinical variables after FDR correction (Supplementary Table 4; Supplementary Figure 4), reduced to eight following residualization (Figure 3). The surviving associations spanned manic symptoms, positive symptoms, negative symptoms, and cognition, with effects for verbal IQ ($\rho = -0.25$, $q = 0.009$) and FTD ($\rho = 0.23$, $q = 0.009$) among the strongest. Associations with global functioning, structural speech impoverishment, and semantic fluency were attenuated after deconfounding and no longer survived correction, whereas bizarre behavior reached significance only after residualization. When post hoc controlling for verbal IQ, correlations with manic, positive, and SANS abulia/apathy survived FDR correction (Supplementary Table 5)

However, all symptom-based correlations were attenuated below the significance threshold when accounting for diagnostic group membership. Within the MDD cohort, higher residualized LLM surprisal remained significantly associated with lower verbal IQ, reduced lexical verbal fluency, and SANS abulia/apathy. Within the smaller SSD and BD cohorts, no clinical or cognitive variables survived FDR correction for the diagnostic subgroups (Supplementary Table 6 and accompanying text).

Incremental Predictive Value of LLM surprisal

To determine whether median LLM surprisal provides diagnostic utility beyond basic speech and demographic features, we evaluated nested predictive models, using Akaike Information Criterion (AIC)⁶⁷ for model fit, leave-one-out cross-validated Area Under the Curve for predictive accuracy, likelihood ratio tests (LRT) for improvement in fit, and DeLong tests for improvement in cross-validated discrimination⁶⁸. We tested whether adding LLM surprisal improved on a semantic baseline model comprising the confounders and semantic coherence between sentences.

Figure 2 Group differences in median LLM surprisal, unadjusted and residualized



In each panel, individual points denote participants, the box marks the median and interquartile range (whiskers at $1.5 \times$ IQR), and the half-violin shows the kernel density estimate. Group differences were tested by OLS regression with HC as the reference; standardized betas, confidence intervals, and p-values are reported in Table 2.

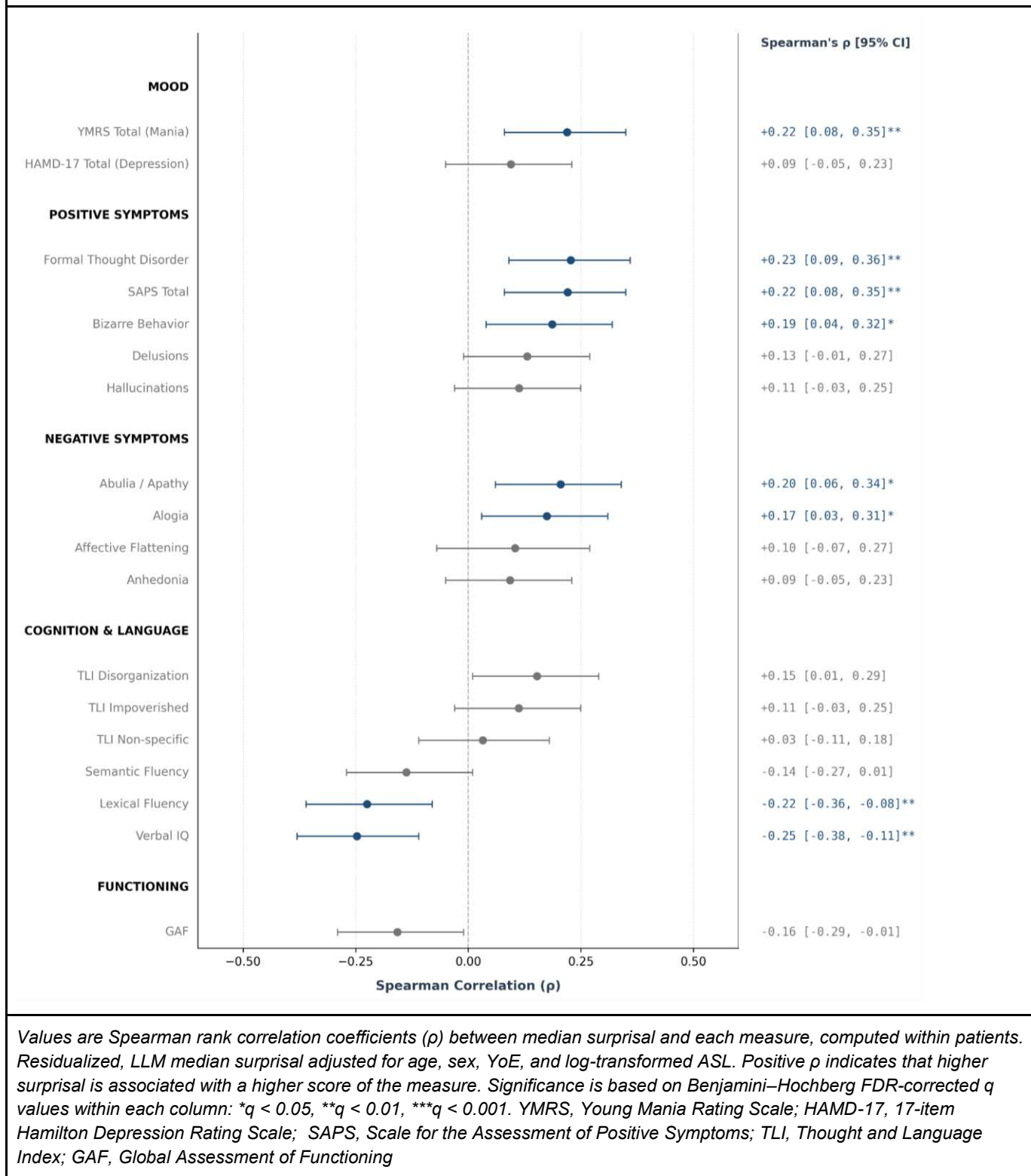
(a) Unadjusted LLM surprisal across the individual diagnostic groups (MDD, BD, SSD) and the combined patient cohort (All Patients) relative to HC. Overall, patients showed higher unadjusted surprisal than controls (HC: 3.56 ± 0.34 ; patients: 3.68 ± 0.39), with significant elevations visible in the SSD, BD groups and all patients, whereas MDD did not differ.

(b) Residualized LLM surprisal across the same groups after residualizing for age, sex, years of education, and average sentence length. Under this deconfounded model, SSD ($\beta = 0.49$ [0.21, 0.76], $P = 0.002$) and BD ($\beta = 0.39$ [0.04, 0.73], $P = 0.04$) remained significantly elevated relative to HC, whereas MDD ($\beta = -0.02$ [-0.21, 0.18], $P = 0.87$) and the overall combined patient cohort ($\beta = 0.15$ [-0.02, 0.33], $P = 0.09$) were non-significant.

Horizontal brackets denote pairwise significance contrasts against HC;

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. MDD, Major Depressive Disorder; BD, Bipolar Disorder; HC, Healthy Controls.

Figure 3 Association of median LLM surprisal with clinical metrics



In the full, mixed-diagnosis cohort, adding LLM surprisal to the semantic model improved fit (LRT - $P = 0.033$) but did not yield a significant gain in cross-validated discrimination (DeLong $P = 0.133$; Table 3). This pattern held in disease-specific subanalyses: LLM surprisal consistently improved model fit in SSD and BD, but the corresponding DeLong tests were non-significant, with CIs that crossed zero. For MDD, LLM surprisal added no value in fit and worsened performance (Supplementary Table 7; Supplementary Figure 5). When evaluated against clinician-rated scales in SSD (Supplementary Table 8), adding median LLM surprisal significantly improved cross-validated classification over SAPS positive FTD alone ($\Delta = 0.115$, DeLong $P = 0.003$), but did not yield significant gains over the TLI baseline or the combined SAPS and TLI clinician-rated baseline.

Table 3 Incremental predictive value of median LLM surprisal for patient–control classification

Cohort	Model Specification	AIC	CV-AUC [95% DeLong CI]	LRT p-value	DeLong p-value
All Patients vs. HC					
	Median LLM surprisal Only	508.2	0.580 [0.521, 0.638]	—	—
	Confounders + Semantic Coherence	499.3	0.610 [0.553, 0.667]	—	—
	Confounders + Semantic Coherence + Median LLM surprisal	496.7	0.627 [0.571, 0.684]	0.033	0.133
SSD vs. HC					
	Median LLM surprisal Only	215.2	0.676 [0.584, 0.767]	—	—
	Confounders + Semantic Coherence	212.4	0.706 [0.622, 0.789]	—	—
	Confounders + Semantic Coherence + Median LLM surprisal	203.6	0.744 [0.665, 0.824]	< 0.001	0.059
BD vs. HC					
	Median LLM surprisal Only	152.8	0.622 [0.502, 0.741]	—	—

	Confounders + Semantic Coherence	158.5	0.590 [0.464, 0.715]	—	—
	Confounders + Semantic Coherence + Median LLM surprisal	155.1	0.622 [0.495, 0.749]	0.021	0.252
MDD vs. HC					
	Median LLM surprisal Only	403.4	0.476 [0.409, 0.544]	—	—
	Confounders + Semantic Coherence	401.6	0.558 [0.493, 0.624]	—	—
	Confounders + Semantic Coherence + Median LLM surprisal	403.5	0.550 [0.484, 0.615]	0.730	0.011

Nested logistic regression models predicting patient status versus healthy controls (HC), fitted separately for each cohort. The likelihood-ratio test (LRT) P-value and DeLong P-value evaluate whether adding Median LLM surprisal significantly improves model fit and discrimination (CV-AUC), respectively, over the baseline model containing Confounders + Semantic Coherence. Confounders include age, sex, YoE, and ASL. Semantic Coherence represents the mean cosine similarity between embeddings of utterances (see Methods). AIC, Akaike information criterion (lower indicates better fit); CV-AUC, leave-one-out cross-validated area under the receiver operating characteristic curve; CI, confidence interval; LRT, likelihood-ratio test; HC, healthy controls; SSD, schizophrenia spectrum disorder; BD, bipolar disorder; MDD, major depressive disorder.

Cross-Model and Cross-Prompt Stability

We compared the primary Qwen2.5-72B-Instruct model with a smaller same-family model (Qwen2.5-7B-Instruct) and a German-language model (LeoMistral-Hessianai-7b). Subject-level LLM surprisal correlated strongly with the primary model, with a uniform upward shift for the smaller models (Supplementary Figure 6). Group and clinical effects were stable: SSD was robustly elevated in every model, BD elevations were significant in the large model and German Model but attenuated only in the smallest multilanguage model, and MDD never differed from controls (Table 2). Clinical correlations were likewise consistent (Supplementary Figure 7). Comparing the instruction-tuned Qwen-7B against its base counterpart, subject-level estimates agreed near-perfectly, indicating the metric reflects the underlying probability distribution rather than an artifact of instruction tuning (Supplementary Figure 8).

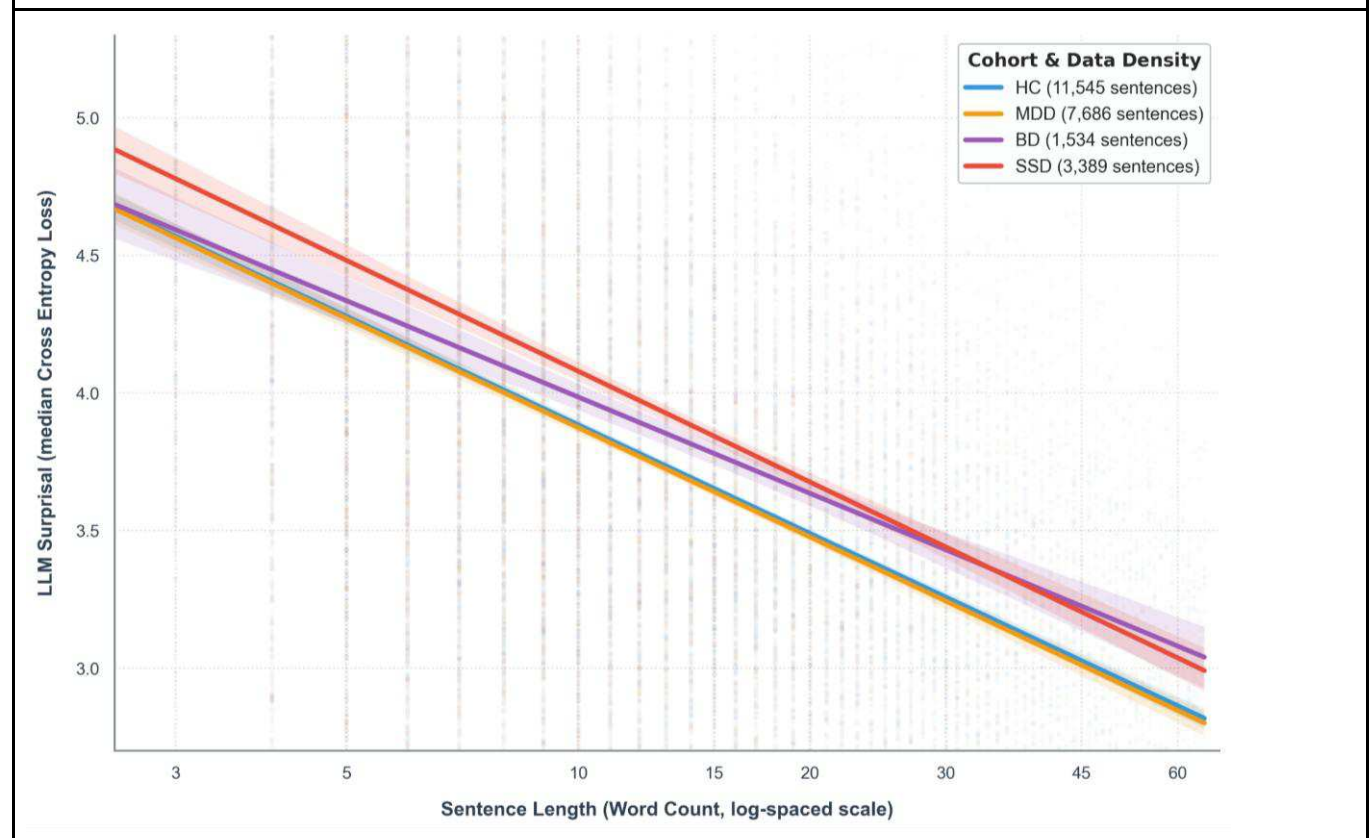
Additionally we repeated the pipeline across four system-prompt configurations (healthy persona, patient persona, neutral, and none; Supplementary Table 9). Subject-level LLM surprisal agreed almost perfectly (Supplementary Figure 9), and effects were essentially invariant: SSD remained robustly elevated throughout, BD elevations held under all prompts

except the no-prompt condition, and MDD remained non-significant (Supplementary Table 10). Clinical correlations were stable (Supplementary Figure 10).

Sentence-Level Moderation and Mediation Analyses

To confirm that group differences in surprisal were not artifacts of ASL, we ran two complementary analyses on the unadjusted LLM surprisal. The first asked whether the ASL–surprisal *relationship* itself differed by group (moderation); the second asked whether each group's elevation was *carried by* differences in sentence length (mediation). The first modeled surprisal at the individual sentence level, estimating each group's difference from controls at the ASL while allowing the length–surprisal relationship to differ by group. At the ASL, SSD patients scored higher surprisal than controls ($\beta = 0.19$, $P < 0.001$), with no group difference in how length related to surprisal (interaction $P = 0.60$). BD showed a nuanced pattern: the elevation at average length was non-significant ($\beta = 0.11$, $P = 0.056$) but the length–surprisal slope differed from controls (interaction $\beta = 0.07$, $P = 0.012$). MDD differed from controls on neither (both $P > 0.70$). Across the sample, longer sentences showed lower surprisal ($\beta = -0.54$, $P < 0.001$) (Figure 4).

Figure 4 Interaction between sentence length and sentence-level LLM surprisal across diagnostic cohorts.



Bold lines depict fitted linear regression curves with shaded 95% confidence intervals for Healthy Controls (HC, blue), Major Depressive Disorder (MDD, orange), Bipolar Disorder (BD, purple), and Schizophrenia Spectrum Disorder (SSD, red). Individual sentence-level observations are plotted as ultra-light background dots to depict sample density. Sentence length (x-axis) is plotted in log-transformed space but labeled with raw word counts. While the MDD group (orange) is virtually indistinguishable from controls in both slope and baseline elevation ($p > 0.70$), the SSD group (red) displays a significant, constant vertical elevation across all sentence lengths ($\beta = 0.19$, $p < 0.001$, no interaction). In contrast, the BD group (purple) exhibits a significant slope interaction with controls ($\beta = 0.07$, $p = 0.012$), where its LLM surprisal elevation is marginal for short utterances but diverges sharply as sentences grow longer.

The second analysis partitioned each group's total difference into a length-dependent (indirect) and a length-independent (direct) component. For SSD, the direct component was significant ($c' = 0.19$, $P < 0.001$) while the indirect path did not reach significance ($ab = 0.057$, 95% CI $[-0.002, 0.120]$), indicating the elevation was largely independent of ASL. For BD, the direct component was likewise significant ($c' = 0.15$, $P = 0.015$) and the indirect path was near zero ($ab = 0.008$, 95% CI $[-0.064, 0.077]$). For MDD, the total effect was negligible.

Exploratory Distributional Analysis

Because some speech disturbances are intermittent and may concentrate in occasional, highly unpredictable sentences, we examined whether distributional features of the LLM surprisal captured associations the median missed. In patients, the rate of high-surprisal sentences, defined as the proportion of a participant's sentences whose surprisal exceeded the healthy-control 95th percentile, was associated with TLI disorganization (Supplementary Table 11) and remained significant after partialling out the median (partial $\rho = 0.19$, $q = 0.009$; Supplementary Table 12). The association was specific to disorganization and absent for FTD, impoverishment, and non-specific subscales (Supplementary Table 11-12). Comparing tail metrics across the three diagnostic groups, only the 90th-percentile surprisal was elevated in SSD relative to controls (Supplementary Table 13). Trimming the contextless first word as baseline noise resulted in largely stable correlations and showed no robust changes with clinical metrics (Supplementary Figure 11-12).

Sensitivity to the Residualization Reference

Our primary analyses deconfounded LLM surprisal across the full sample. Because our framework treats surprisal as deviation from healthy speech, we repeated the analysis with the covariate model estimated in HC only and applied it to all participants. Under this HC reference, the group elevations were essentially undiminished relative to the unadjusted metric and larger than under the full-sample residualization (SSD $\beta = 0.68$ [0.37, 0.99], $P < 0.001$; BD $\beta = 0.46$ [0.06, 0.87], $P = 0.024$; MDD $\beta = 0.014$ [-0.213, 0.240]). Whereas only verbal IQ survived simultaneous adjustment for demographics and diagnosis under the full-sample reference, manic symptoms (YMRS), verbal IQ, lexical and semantic verbal fluency and TLI Impoverished all survived under the HC reference (Supplementary Table 14).

Post-hoc Verbal IQ Analyses

Verbal IQ was available for 355 of 371 participants (170 controls, 185 patients). Within patients, residualized LLM surprisal correlated negatively with verbal IQ ($\rho = -0.26$, $P < 0.001$), whereas within controls the association was weak and non-significant ($\rho = -0.08$, $P = 0.28$); the group-by-IQ interaction was negative at trend level ($\beta = -0.19$, $P = 0.071$). The patient-level association was unchanged after controlling for polypharmacy and chlorpromazine-equivalent antipsychotic dose, and remained significant in the pooled sample after adjusting for patient status ($\rho = -0.18$, $P < 0.001$). A quadratic verbal IQ term improved fit in the pooled sample and within HC alone (Supplementary Table 15), describing a U-shaped relationship with minimum surprisal near average verbal IQ and higher surprisal toward both extremes. The curvature did not differ between groups (group $\times$ quadratic $P = 0.42$); within patients alone it was significant under OLS but did not survive removal of influential observations or trimming of the highest IQ scores (Supplementary Table 16).

Discussion

In this study, we evaluated whether the LLM surprisal of spontaneous narrative speech, quantified as the median cross-entropy loss of a state-of-the-art LLM, is associated with psychiatric diagnosis and with clinical symptom severity in a transdiagnostic sample of patients with SSD, BD, and MDD in comparison with a HC reference. Three principal findings emerged. First, speech from patients with SSD scored consistently higher on LLM surprisal than from HC, with intermediate effects in BD and no detectable elevation in MDD. Second, higher LLM surprisal was associated with lower verbal ability and with greater symptom severity across positive (SAPS total, positive FTD), negative (alogia, abulia/apathy) and manic dimensions; of these, only the association with verbal ability persisted after adjustment for diagnostic group, indicating that the symptom associations were carried largely by between-diagnosis variance. Third, these effects were not reducible to demographic or syntactic confounds: the pattern held after adjustment for demographics, YoE, ASL and for almost all model or prompt configurations. Together, these findings align with earlier research in schizophrenia by demonstrating that LLM-derived predictability—operationalized here as LLM surprisal—captures clinically relevant dimensions of speech disturbance, while extending these insights to the bipolar spectrum and showing no significant deviation in MDD. LLM surprisal overlaps only partially with established semantic and syntactic features.

A central methodological concern was whether higher surprisal in patients simply reflected shorter sentences⁴⁰. ASL was by far the strongest correlate of unadjusted LLM surprisal, substantially larger than any demographic association due to the nature of the measurement (Figure 1). Several convergent analyses argue against this concern. Although ASL was strongly related to unadjusted LLM surprisal, the elevations in SSD and BD persisted after adjustment for sentence length, and in neither group was the effect attributable to patients producing shorter sentences. Likewise, decomposition analyses indicated that most of the SSD effect was attributable to a length-independent component, and LLM surprisal contributed information beyond models that already accounted for ASL and semantic coherence. Across all speakers, longer sentences carried lower surprisal, consistent with

more preceding context making each next word easier to predict. This relationship was the same in SSD patients and HC, but SSD patients sat higher on the surprisal scale at every sentence length. The elevation is therefore a level shift that holds when sentences of equal length are compared, not a by-product of shorter sentences in patients. BD-surprisal showed a more nuanced behavior. The gap from controls was not constant but widened with sentence length: in shorter sentences BD speakers stayed close to controls, in longer ones they departed further. This is not a shortness artifact, since it reflects how surprisal scales with length rather than patients producing shorter sentences, and it sets BD apart from the uniform elevation seen in SSD (Figure 4).

Interpreting LLM cross-entropy loss as a measure of contextual surprise bridges computational psychiatry and clinical theories of FTD. That this signal is reproduced across model architectures, including smaller open-weight models³⁹, suggests the underlying baseline of expected language is a relatively stable property rather than an artifact of any single model. In psychiatric terms, this may reflect reduced contextual constraint during language production, a mechanism closely related to clinical concepts of disorganization, derailment, and impaired discourse coherence. This interpretation extends prior work showing altered perplexity and contextual language processing in psychosis^{31–33}, bipolar disorder³⁴ and dementia^{35,36}. But describing LLM surprisal as a measure of FTD alone is probably too narrow. Because surprisal reflects contextual word predictability, elevated values may arise even in grammatically correct speech when contextual expectations are violated (Figure 1). LLM surprisal may therefore capture a broader dimension of contextual language organization than is captured by conventional clinical ratings of thought disorder. The SSD ablation experiment showed that once surprisal was residualized against confounders, it still improved diagnostic classification on top of clinician-rated positive FTD (Supplementary Table 8), meaning it carries diagnostic information the FTD rating does not. Surprisal did not add value over the TLI, because the TLI scale itself performed no better than chance in our cohort and therefore carried no signal for the distinction between SSD and HC.

A central question is whether LLM surprisal has clinical potential in terms of diagnosis and associations with symptom severity. The dominant pattern in this study was gradual: LLM surprisal followed a diagnostic gradient, clearly elevated in SSD, intermediate in BD, and absent in MDD, and this ordering held across sensitivity analyses. A comparable schizophrenia–bipolar ordering was also reported in a Korean cohort using written language³⁴, suggesting an effect across languages. Symptom associations fell below significance once the diagnostic group was added to the model, indicating they were carried largely by between-diagnosis differences rather than within-patient severity. Just one association behaved dimensionally across the full sample. The correlation with verbal IQ survived simultaneous adjustment for demographics and diagnosis and remained within the MDD subgroup, which showed no group elevation of its own. MDD cohort likewise retained associations with lexical fluency and abulia/apathy despite no group-level elevation. We therefore read LLM surprisal primarily as a graded marker of psychosis-spectrum status, with a narrower dimensional component carried mainly by verbal ability and by residual MDD-subgroup associations. This fits dimensional interpretations only marginally^{49,51}: a continuum account predicts LLM surprisal should track severity across diagnoses, yet most of our signal tracked diagnostic category, with a clearly dimensional component only for

verbal ability and within MDD. Given the strong collinearity between diagnosis and severity, and small sample sizes in SSD and BD we cannot fully separate a categorical from a dimensional interpretation.

The described robust association between surprisal and verbal ability measured as verbal IQ requires careful interpretation. This coupling emerged specifically within patients and was weak and non-significant in HC, so surprisal is not a simple index of verbal ability in unaffected speakers but becomes tied to it in the presence of pathology³⁷. Surprisal also did not collapse into verbal ability: the schizophrenia-spectrum elevation persisted after verbal IQ was added to the deconfounding pipeline, and associations with manic, positive and negative symptoms survived additional adjustment for verbal IQ. Higher surprisal therefore cannot be reduced to lower verbal ability alone; rather, surprisal sits at the intersection of psychopathology and verbal ability. Consistent with surprisal reflecting distance from an inherent norm in LLMs of typical language, its relationship with verbal ability was U-shaped across the full sample and within controls, rising at both extremes. This shows surprisal may reflect more than FTD, being sensitive also to verbal ability.

Sensitivity analyses showed that the shape of the surprisal distribution, not only its median, carried clinical information. Within patients, the upper tail of the sentence-level surprisal correlated with clinician-rated disorganization, and did so after partialling out the median. The association appeared specific to disorganization. This fits the intermittent nature of disorganized speech: derailment and tangentiality surface as occasional sentences that depart sharply from expectation rather than as a uniform elevation. That a text-trained model recovers a disorganization signal specifically in the tail is a small but suggestive pointer to localized breakdowns in language predictability. Results were stable to first word trimming, showing that the noise first tokens with preceding context did not influence the metrics performance.

Three methodological findings emerging from our sensitivity analyses merit further consideration. First, the primary analyses employed full-sample deconfounding, a conservative approach that is likely to attenuate illness-related variance. Consistent with this interpretation, a model estimated exclusively from HC yielded comparable group effects while recovering additional symptom and cognitive associations, suggesting that the reported effects may represent a conservative lower-bound estimate. These findings highlight an unresolved challenge for LLM based language metrics: balancing statistical conservatism against the preservation of clinically meaningful variance. Secondly, surprisal estimates were remarkably stable across prompt configurations. The near-perfect agreement between system-prompt conditions suggests that model-derived estimates of speech surprisal are largely insensitive to simple persona instructions. A plausible explanation is that the statistical regularities acquired during large-scale pretraining dominate next-token predictions, rendering brief prompts insufficient to meaningfully alter the model's expectations. Thirdly, the signal was largely stable across model scales. The smallest multilingual model was, however, the one condition where the weaker BD elevation fell below significance. In contrast to that finding prior work reports that smaller models can be more sensitive³⁹. The optimal scale therefore remains unresolved. One speculative hypothesis could be that smaller models, while adequate for the strong SSD effect, lack the resolution to capture subtler elevations such as of BD patients.

Several limitations temper the conclusions. The design is cross-sectional and observational, so we cannot determine whether elevated surprisal reflects a current clinical state, a stable trait, or both. Although adding surprisal significantly improved model fit for SSD and BD, the absolute cross-validated discrimination was at best moderate for SSD and weak for BD, so the metric does not support standalone diagnostic use. Symptom severity in the patient groups was predominantly mild to moderate, with MDD and BD patients showing low depressive and manic symptom levels, which limits sensitivity and constrains generalization to acutely ill populations. The BD and SSD subgroups were small; the BD incremental results are based on just 26 patients, and the fact that the univariate surprisal model outperformed the fully adjusted model in this subgroup is more plausibly a sign of statistical instability than of superior performance, so it should be regarded as preliminary. Furthermore, while the SSD elevation remained robust after adjusting for verbal IQ, the BD elevation did not survive this control, underscoring that the effect in bipolar disorder is less robust, probably also due to sample size. The data derive from a single site, a single language, and a constrained narrative task (TAT); although cross-model replication with a German and a multilingual model is reassuring, cross-linguistic and cross-task generalizability remains to be established. Finally, cross-entropy loss is mechanistically opaque: it aggregates unpredictability arising from semantic, syntactic, pragmatic, and topical sources and cannot, on its own, identify which of these drives the effect.

These results point to several next steps. Longitudinal designs that track LLM-derived predictability measures across symptomatic and remitted states would establish whether it indexes state, trait, or both, and whether it is sensitive to clinical change. Larger and acutely symptomatic samples, particularly within the psychosis spectrum, would test clinical utility and reduce the floor effects seen here. Framing the metric explicitly as a normative model⁶⁹, in which expected loss is modelled as a function of demographic and cognitive variables such as age, sex, and ASL, would yield individual deviation scores against a healthy reference and directly absorb these confounders rather than treating them as nuisance covariates, placing loss on a more interpretable, person-level footing. Future work should also explore task-specific fine-tuning by training open-weight LLMs on HC narratives from the exact same TAT task, to enforce better discrimination from HC.

In summary, the median cross-entropy loss, operationalized as LLM surprisal, of a general-purpose LLM applied to spontaneous narrative speech seems to behave as a data-driven model of speech peculiarities: patient speech, most clearly in SSD and partly in BD, deviates from the model's expectation in a way that is not reducible to ASL, demographic confounders or semantic coherence. The effect is graded by diagnosis along the psychosis spectrum, mapping at least partly to FTD. In reference to the growing literature the cross-entropy loss and perplexity of LLMs provides an intriguing hint that these measures can serve as an additional automated marker to characterize language deviations in psychiatric diseases. However, as this field expands, it remains a critical point of discussion as to what the correct deconfounding strategy should be to isolate pure clinical signals from broader demographic, social, or cognitive variance. With longitudinal validation, larger acutely ill samples, and mechanistic dissection, LLM-derived measures may offer a scalable, privacy-preserving, language-level window onto disordered speech across transdiagnostic cohorts.

Methods

Participants

In this cross-sectional, observational subsample, participants were recruited between 2018 and 2021 as part of the FOR2107 cohort⁵⁹ at the Department of Psychiatry, University of Marburg, and affiliated outpatient clinics. This subsample comprised 178 HC and 194 patients, including 48 patients with a SSD (schizophrenia: $n = 11$; schizoaffective disorder: $n = 37$), 119 patients with MDD, and 27 patients with BD. Exclusion criteria included neurological disorders, a history of significant head injury, current substance dependence, other severe medical conditions, current benzodiazepine use, intellectual disability, and insufficient proficiency in the German language. HC participants were additionally screened to exclude any current or past psychiatric diagnosis or treatment. Sample size was determined by data availability. All participants provided written informed consent prior to inclusion. The study protocol was approved by the local ethics committee and conducted in accordance with the Declaration of Helsinki.

Speech Elicitation and Text Processing

Narrative speech was elicited with four pictures from the TAT⁶⁰, with three minutes per picture to tell a story about the depicted scene. Speech was audio-recorded and transcribed verbatim, including hesitations, by trained linguists using f4transkript (www.audiotranskription.de/f4transkript), blind to diagnostic group. These transcripts served as input for both the TLI ratings and all computational analyses.

For the computational analyses, German filler words and verbal hesitations (e.g. "mh", "äh") were removed to prevent tokenization anomalies. Text was segmented into sentences deterministically at terminal punctuation marks, and the average number of words per sentence (ASL) was computed per participant for use as a covariate.

Psychopathological Assessment

Each participant's diagnosis was ascertained through a semi-structured clinical interview conducted in line with the SCID-I, with classification based on DSM-IV-TR criteria⁷⁰. Current psychopathology was characterized using a panel of standardized, clinician-rated instruments. Negative and positive symptom domains were rated with the SANS⁶² and the SAPS⁶¹, respectively. Beyond the total scores, which index global negative and positive symptom severity, FTD was captured through two dedicated measures: the negative FTD score derived from the SANS and the positive FTD score derived from the SAPS. Depressive, anxiety, and manic symptomatology were additionally quantified with the HAM-D-17⁶⁴, the Hamilton Anxiety Rating Scale⁷¹, and the YMRS⁶³.

To characterize disturbances in thought and language more directly, we applied the TLI⁷², which yields impoverishment and disorganization scores from transcribed speech. Trained raters, kept blind to participants' diagnostic group, scored the TLI from verbatim transcripts. Speech elicited by each TAT picture was divided into three successive one-minute segments,

with ratings assigned at the segment level. These segment-level ratings were first summed within each picture and then combined across the four pictures, yielding a single disorganization total and a single impoverishment total per participant. Overall functional status was rated with the GAF⁶⁵. Verbal IQ was measured using the Multiple-Choice Vocabulary Intelligence Test⁷³.

Neurocognitive Assessment

Neurocognitive functioning was characterized using a set of standardized tests spanning verbal memory, executive control, and verbal fluency. Verbal learning and memory were assessed with the Verbal Learning and Memory Test⁷⁴ from which retrieval and recognition indices were obtained. The Trail Making Test⁷⁵ served to index executive functioning and processing speed. Lexical access and semantic retrieval were probed through verbal fluency tasks⁷⁶ administered under semantic (animals), phonemic (p-letters), and alternating conditions (sports/fruits).

Sample Characteristics

Group descriptive statistics are reported as mean $\pm$ standard deviation for continuous variables and as counts and percentages for categorical variables. Differences across the four diagnostic groups (HC, BD, MDD, SSD) in continuous demographic and linguistic variables were evaluated using Kruskal–Wallis H-tests. Group differences in categorical variables were assessed using a chi-squared contingency test on the diagnosis $\times$ variable table.

Language Models and Loss Computation

Linguistic feature extraction was executed using three autoregressive LLMs deployed locally on private inference servers via the vLLM engine⁷⁷, guaranteeing data privacy. Qwen2.5-32B-Instruct⁷⁸ was executed across four NVIDIA L40 GPUs, while the smaller model Qwen2.5-7B was executed from two NVIDIA L40 GPUs and LeoMistral-Hessianai-7b⁷⁹ was deployed on a single NVIDIA L40 GPU to serve as an independent architectural validation.

To quantify the structural and semantic predictability of patient speech, we utilized the cross-entropy loss derived from the language models²⁴. While clinical NLP studies frequently report perplexity, we used cross-entropy loss, termed LLM surprisal, because perplexity is the exponential of loss and can disproportionately stretch higher values, increasing skewness, heteroscedasticity, and nonlinear relationships with covariates. Loss therefore provides a more suitable scale for OLS residualization. For every segmented sentence within a participant's transcript, the sequence-level loss was computed as the negative average of the token log-probabilities, conditioned on the prompt text. Formally, for a sequence of N tokens, the cross-entropy loss is defined as:

$$Loss = -\frac{1}{N} \sum_{i=1}^N \log P(t_i | t_{<i})$$

System instructions and prompt tokens provided context for the model but were masked and omitted from the final loss computation. All evaluations were performed using greedy decoding (temperature = 0) to ensure reproducibility. To establish a robust, subject-level metric not too sensitive to outliers, sentence-level losses were aggregated to the participant level using the median across all transcribed sentences.

To systematically evaluate whether model-side contextual expectations introduced systematic bias, this entire pipeline was repeated across four prompt configurations: a healthy persona prompt, a patient persona prompt, a neutral prompt, and a baseline condition devoid of any system prompt (Supplementary Table 9).

The healthy-persona configuration paired with Qwen2.5-32B-Instruct was pre-specified as the primary architecture.

Subject-level agreement between the primary model and the validation models, as well as between the four prompt configurations, was quantified using Spearman rank correlations of the participant-level median LLM surprisal scores.

Confound Control and Residualization

We examined associations between our primary LLM surprisal metric and four potential confounding variables: sex, age, YoE, and ASL, using Spearman rank correlations for all variables. Missing values for YoE were estimated via KNN imputation (k=5)⁸⁰, utilizing the HC distribution as the reference space.

To isolate disease-related semantic and structural predictability from variance driven by baseline demographic and syntactic differences, we employed an OLS regression model⁸¹ predicting median participant surprisal from age, sex, YoE, and log-transformed ASL across the entire participant cohort. The resulting residuals were then extracted. As a sensitivity analysis, we repeated this procedure with the covariate model estimated in HC only and then applied to all participants, so that the residuals express each participant's deviation from the expected surprisal of healthy speakers given the confounders.

Statistical Group Comparisons

To evaluate variation in language predictability across diagnostic groups, we fitted OLS regression models with surprisal as the outcome and diagnostic group as a categorical predictor, using HC as the reference category. This yielded, for each patient group, a standardized beta expressing its difference from controls, with 95% confidence intervals and two-sided p-values, separately for unadjusted and residualized surprisal metrics. The overall patient–control contrast was estimated from the same framework by comparing the combined patient group against controls. Between-patient contrasts (SSD, BD, MDD) were

likewise estimated as OLS group coefficients. Unless stated otherwise, significance was assessed at Benjamini–Hochberg FDR-corrected threshold of $\alpha=0.05$ for this and all following analyses⁸².

Clinical Correlation Analyses

To map how LLM surprisal interacts with distinct clinical and cognitive dimensions, 18 specific variables were selected a priori based on their established relationships with speech production in psychiatric populations⁴³. Exploratory Spearman rank correlation analyses were performed between both unadjusted and residualized LLM surprisal metrics and these 18 clinical variables across patients. To determine whether these clinical associations persisted independently of diagnosis, a secondary OLS regression framework was evaluated. In this model, the clinical variable served as the dependent variable, while unadjusted LLM surprisal, covariates, and diagnosis (dummy-coded using HC as the baseline reference group) were entered simultaneously as predictors.

Finally, to isolate disease-specific effects, Spearman rank correlations were calculated within each individual patient diagnostic subgroup using the residualized LLM surprisal metrics.

In a post hoc exploratory analysis, we used Spearman correlations within patients to test whether distributional features of residualized sentence-level surprisal (standard deviation, 90th and 95th percentiles, and proportion of sentences exceeding the healthy-control 90th percentile) related to thought and language measures (TLI disorganization, impoverishment, and non-specific subscales, and SAPS-FTD).

Incremental Predictive Value Analysis

To determine whether LLM surprisal provides unique diagnostic utility beyond basic speech and demographic features, we evaluated a series of nested logistic regression models predicting patient status. In all models, LLM surprisal and the covariates were entered directly as predictors; surprisal was not residualized beforehand, so covariate adjustment was performed within each model rather than as a separate preprocessing step, to avoid data leakage. Model performance was assessed using Leave-One-Out Cross-Validated Area Under the Curve to measure predictive accuracy, and the AIC⁶⁷ to evaluate model fit relative to complexity. Cross-validated predictions were generated with L2-regularized logistic regression (scikit-learn, default: $C = 1.0$) and not tuned, whereas AIC and LRT were based on unregularized maximum-likelihood logistic regression.

To quantify the unique contribution of the LLM, we tested whether adding surprisal improved on a series of increasingly rigorous baseline models along two axes: LRT assessed improvement in model fit, and DeLong tests⁶⁸ for two correlated AUC curves assessed improvement in cross-validated discrimination. The baseline models were: 1) a ASL only model; 2) a full confounder model (age, sex, YoE, and ASL); and 3) a comprehensive semantic model that supplemented the full confounders with a semantic coherence metric.

To calculate semantic coherence, individual utterances were encoded using a multilingual SentenceTransformer^{83,84} (multilingual-e5-large). From the cosine similarities between sentences we derived four coherence features per transcript: mean and standard deviation of first-order (adjacent-sentence) coherence, the minimum first-order value, and mean second-order coherence. These formed the semantic baseline²³. This incremental analysis was performed on the combined, mixed-diagnosis patient cohort, and subsequently stratified for each specific psychiatric subgroup (SSD, BD, and MDD) versus HC.

Separately, in SSD versus HC we tested whether confounder-residualized surprisal improved discrimination (DeLong) over clinician-rated FTD, using SAPS positive FTD and the TLI as baselines.

Additional Sentence-Length Sensitivity Analyses

ASL was the strongest baseline correlate of surprisal, so we ran two analyses to check that group differences in surprisal were not simply driven by differences in ASL. Both used unadjusted surprisal and the full sample.

The first analysis compared patients and controls at the *same* sentence length. Instead of using each participant's median, we modeled every individual sentence, with sentence LLM surprisal as the outcome. The model included diagnostic group, sentence length (log-transformed and centered), and their interaction, plus age, sex, and YoE, and a separate intercept for each participant to account for their multiple sentences. Because length was centered, the group term gives the patient–control difference at average sentence length, and the interaction term tests whether the effect of length on surprisal differs between groups. Models were fitted with restricted maximum likelihood, using HC as the reference, both for each group separately and for all four groups together.

The second analysis asked how much of the group difference in surprisal ran *through* sentence length rather than being independent of it. Using mediation analysis, we set diagnostic group as the predictor, sentence length as the mediator, and median surprisal as the outcome, adjusting for age, sex, and YoE. The part of the effect carried by sentence length was estimated as the product of the relevant coefficients, with 95% confidence intervals from 5,000 bootstrap resamples⁸⁵, using pingouin⁸⁶. This was repeated for each group against controls and for the combined patient group.

Post-hoc verbal IQ analyses

In post-hoc analyses, we computed Spearman correlations between surprisal and verbal IQ separately in controls and patients, a partial correlation in the full sample controlling for patient status, and, within patients, a further partial correlation controlling for the polypharmacy index and chlorpromazine-equivalent dose. To test whether the association differed by clinical status, we fitted an OLS model with verbal IQ, patient status, and their interaction. To test for non-linearity, we compared linear and quadratic models by likelihood-ratio test, pooled and within each group; where a quadratic term was significant, vertices were back-transformed to the raw verbal-IQ scale. The patient-level quadratic term was then

checked for robustness by excluding high-influence observations (Cook's distance $> 4/n$) and trimming the top 5% of verbal IQ. These exploratory analyses were not corrected for multiple comparisons.

Data Availability

The raw data collected in this study is not openly accessible to protect participant consent and confidentiality. Nevertheless, the data used in this study is, in principle, available to qualified scientific researchers affiliated with non-commercial research organizations worldwide. Researchers interested in accessing the data must submit a formal research proposal that clearly defines the specific research questions, methodology, and intended statistical analyses using this form: https://redcap.link/data_sharing.

Code Availability

All code used for the analyses can be found at https://github.com/42elenz/mnm_ppl.

Acknowledgments

We extend our sincere gratitude to all study participants and the research staff for their invaluable contributions. A comprehensive list of acknowledgments is available at www.for2107.de/acknowledgements. The study protocols were approved by the local Ethics Committee of the Medical Faculty at the University of Marburg (AZ 07/14) and adhered to the principles outlined in the Declaration of Helsinki. All participants provided written informed consent prior to participation and received financial compensation for their involvement. The first author acknowledges using ChatGPT (GPT-5.4, OpenAI) for text editing to improve the fluency of the English language in the preparation of this manuscript. The author affirms that the original intent and meaning of the content remained unaltered during editing and that ChatGPT had no involvement in shaping the intellectual content of this work. The authors assume full responsibility for upholding the integrity of the content presented in this manuscript.

This work was supported by the German Research Foundation (DFG) through grant STE3301/1-1 (project number 527712970) and by the Von Behring-Röntgen Society (project number 72_0013), both awarded to Frederike Stein. This study also received support from the DFG Collaborative Research Centre/Transregio 393 (CRC/TRR 393, project number 521379614), with contributions to Tilo Kircher and Frederike Stein.

Further support was provided by the DFG-funded research unit FOR2107 (project number 240413749), with funding allocated to Tilo Kircher (KI588/14-1, KI588/14-2, KI588/20-1, KI588/22-1). Tilo Kircher, is supported by the DFG Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223), by the DYNAMIC Center (LOEWE program of the Hessian Ministry of Science and the Arts (grant number LOEWE1/16/519/03/09.001(0009)/98), and by the Wellcome Trust funded DIALOG

consortium (Understanding Disorganisation: A Language-Focused Global Initiative in Psychosis; PI: Tilo Kircher; Grant Number 314138/Z/24/Z). This work was supported by the Hector Foundation II (grant number: not applicable). For the publication fee we acknowledge financial support by Heidelberg University. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author Contributions

E.L. and F.S. conceived and designed the study. E.L. designed, conducted and interpreted the statistical analyses. E.L., S.S., J.W., R.R.M., T.K., E.S. and F.S. drafted and critically revised the manuscript. A.M. reviewed the code. F.S. contributed to data curation. T.K., F.S., and E.S. obtained funding. All authors read and approved the final manuscript.

Competing Interests

T.K. has received unrestricted educational grants from Servier, Janssen, Recordati, Aristo, Otsuka, and Neuraxpharm. The remaining authors declare no conflicts of interest and report no financial relationships with biomedical entities. E.S. received speaker fees from bfd buchholz-fachinformationsdienst GmbH and Lundbeckfonden as well as editorial fees from Lundbeckfonden and the Wellcome Trust.

References

1. Gomes, A. I., Jesus, S., Simões, G. & Vicente, S. Symptomatological transversality and the absence of pathognomonic symptoms in psychiatry. *Eur. Psychiatry* **66**, S998–S999 (2023).
2. *Diagnostic and Statistical Manual of Mental Disorders*. (American Psychiatric Publishing, Washington, DC, 2013).
3. Lakhan, S. E., Vieira, K. & Hamlat, E. Biomarkers in psychiatry: drawbacks and potential for misuse. *Int. Arch. Med.* **3**, 1 (2010).
4. Kay, S. R., Fiszbein, A. & Opler, L. A. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* **13**, 261–276 (1987).
5. Bush, G., Fink, M., Petrides, G., Dowling, F. & Francis, A. Catatonia. I. Rating scale and standardized examination. *Acta Psychiatr. Scand.* **93**, 129–136 (1996).
6. Nordgaard, J., Sass, L. A. & Parnas, J. The psychiatric interview: validity, structure, and subjectivity. *Eur. Arch. Psychiatry Clin. Neurosci.* **263**, 353–364 (2013).
7. Andreasen, N. C. Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. *Arch. Gen. Psychiatry* **36**, 1315–1321 (1979).
8. Kircher, T., Bröhl, H., Meier, F. & Engelen, J. Formal thought disorders: from phenomenology to neurobiology. *Lancet Psychiatry* **5**, 515–526 (2018).
9. Roche, E., Creed, L., MacMahon, D., Brennan, D. & Clarke, M. The epidemiology and associated phenomenology of formal thought disorder: A systematic review. *Schizophr. Bull.* **41**, 951–962 (2015).
10. Voppel, A. *et al.* Analysis of conceptual overlap among formal thought disorder rating scales in psychosis: a systematic semantic synthesis. *Schizophrenia (Heidelb.)* **12**, 9

(2025).

11. Kølbaek, P. *et al.* Inter-rater reliability of ratings on the six-item Positive and Negative Syndrome Scale (PANSS-6) obtained using the Simplified Negative and Positive Symptoms Interview (SNAPSI). *Nord. J. Psychiatry* **72**, 431–436 (2018).
12. Kircher, T. *et al.* A rating scale for the assessment of objective and subjective formal Thought and Language Disorder (TALD). *Schizophr. Res.* **160**, 216–221 (2014).
13. Matuszak, J. & Piasecki, M. Inter-Rater Reliability in Psychiatric Diagnosis. *Psychiatr. Times* **29**, (2012)
14. Freedman, R. *et al.* The initial field trials of DSM-5: new blooms and old thorns. *Am. J. Psychiatry* **170**, 1–5 (2013).
15. Regier, D. A. *et al.* DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* **170**, 59–70 (2013).
16. Cohen, A. S. & Elvevåg, B. Automated computerized analysis of speech in psychiatric disorders. *Curr. Opin. Psychiatry* **27**, 203–209 (2014).
17. Hitczenko, K., Mittal, V. A. & Goldrick, M. Understanding Language Abnormalities and Associated Clinical Markers in Psychosis: The Promise of Computational Methods. *Schizophr. Bull.* **47**, 344–362 (2021).
18. Zhang, T., Schoene, A. M., Ji, S. & Ananiadou, S. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digit. Med.* **5**, 46 (2022).
19. Bunnell, B. E. *et al.* Automated detection and prediction of suicidal behavior from clinical notes using deep learning. *PLoS One* **20**, e0331459 (2025).
20. García-Molina, J. T., Downey, M., Méndez, E. & Figueroa-Barra, A. Diagnostic and transition accuracy of natural language processing in high risk for psychosis individuals: A systematic review. *Asian J. Psychiatr.* **112**, 104695 (2025).
21. Bedi, G. *et al.* Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr.* **1**, 15030 (2015).
22. Corcoran, C. M. *et al.* Prediction of psychosis across protocols and risk cohorts using

- automated language analysis. *World Psychiatry* **17**, 67–75 (2018).
23. Elvevåg, B., Foltz, P. W., Weinberger, D. R. & Goldberg, T. E. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophr. Res.* **93**, 304–316 (2007).
 24. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
 25. Sarzynska-Wawer, J. *et al.* Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* **304**, 114135 (2021).
 26. Tang, S. X. *et al.* Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *NPJ Schizophr.* **7**, 25 (2021).
 27. He, R. *et al.* Reduced linguistic coherence in psychosis defies semantic similarity accounts and relates to altered large-scale cortical hierarchy. *Sci. Rep.* **16**, 7799 (2026).
 28. Kuperberg, G. R. & Jaeger, T. F. What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* **31**, 32–59 (2016).
 29. Pickering, M. J. & Garrod, S. An integrated theory of language production and comprehension. *Behav. Brain Sci.* **36**, 329–347 (2013).
 30. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving Language Understanding by Generative Pre-Training. Preprint at https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
 31. Ryu, S. *et al.* Perplexity and pseudo-perplexity as objective measures of linguistic deficits in schizophrenia: Evidence from a Korean-Speaking cohort. *Schizophr. Res.* **294**, 147–154 (2026).
 32. Alonso-Sánchez, M. F., Hinzen, W., He, R., Gati, J. & Palaniyappan, L. Perplexity of utterances in untreated first-episode psychosis: an ultra-high field MRI dynamic causal modelling study of the semantic network. *J. Psychiatry Neurosci.* **49**, E252–E262 (2024).
 33. Sharpe, V. *et al.* Selective Insensitivity to Global Versus Local Linguistic Context in

Speech Produced by Patients With Untreated Psychosis and Positive Thought Disorder. *Biol. Psychiatry* **99**, 154–164 (2026).

34. Jo, Y. T. & Joo, Y. H. Semantic abnormalities in schizophrenia and bipolar disorder: A natural language processing approach. *Sci. Prog.* **108**, 368504241308309 (2025).
35. Xiao, Y., Christensen, H. & Goetze, S. Alzheimer's dementia detection using perplexity from paired large language models. in *Interspeech 2025* 1423–1427 (ISCA, ISCA, 2025).
36. Frankenberg, C. *et al.* Perplexity – a new predictor of cognitive changes in spoken language? – results of the Interdisciplinary Longitudinal Study on Adult Development and Aging (ILSE). *Linguist. Vanguard* **5**, 1–10 (2019).
37. Palaniyappan, L. More than a biomarker: could language be a biosocial marker of psychosis? *NPJ Schizophr.* **7**, 42 (2021).
38. Cohen, A. S. *et al.* Natural Language Processing and Psychosis: On the Need for Comprehensive Psychometric Evaluation. *Schizophr. Bull.* **48**, 939–948 (2022).
39. Li, C. *et al.* Bigger But Not Better: Small Neural Language Models Outperform Large Language Models in Detection of Thought Disorder. *Workshop Comput. Linguist. Clin. Psychol.* **2025**, 90–105 (2025).
40. Stein, D., Bobrovskiy, D., Stede, M., Montag, C. & Just, S. A. Cross-linguistic benchmarking of NLP metrics for psychosis research. *NPJ Digit Med* **9**, (2026).
41. Stein, F. & Kircher, T. Transdiagnostic findings across major depressive disorder, bipolar disorder and schizophrenia: A qualitative review. *J. Affect. Disord.* **387**, 119464 (2025).
42. Voleti, R., Liss, J. M. & Berisha, V. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE J. Sel. Top. Signal Process.* **14**, 282–298 (2020).
43. Seuffert, S. *et al.* Linking speech patterns to brain structure in affective and psychotic disorders: an integrative natural language processing approach. *Mol. Psychiatry* **31**,

2057–2068 (2026).

44. Stein, F. *et al.* Dimensions of formal thought disorder and their relation to gray- and white matter brain structure in affective and psychotic disorders. *Schizophr. Bull.* **48**, 902–911 (2022).
45. Stein, F. *et al.* Transdiagnostic types of formal thought disorder and their association with gray matter brain structure: a model-based cluster analytic approach. *Mol. Psychiatry* **30**, 4286–4295 (2025).
46. Stein, F. *et al.* Formal thought disorder in unipolar depression: Prevalence and functional brain correlates from a multimodal MRI study. *J. Affect. Disord.* **407**, 121779 (2026).
47. Stein, F. *et al.* Brain structural network connectivity of formal thought disorder dimensions in affective and psychotic disorders. *Biol. Psychiatry* **95**, 629–638 (2024).
48. Yalincetin, B. *et al.* Formal thought disorder in schizophrenia and bipolar disorder: A systematic review and meta-analysis. *Schizophr. Res.* **185**, 2–8 (2017).
49. Kotov, R. *et al.* The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* **126**, 454–477 (2017).
50. Insel, T. *et al.* Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am. J. Psychiatry* **167**, 748–751 (2010).
51. Hinzen, W. & Palaniyappan, L. The 'L-factor': Language as a transdiagnostic dimension in psychopathology. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **131**, 110952 (2024).
52. Stein, F. *et al.* Factor analyses of multidimensional symptoms in a large group of patients with major depressive disorder, bipolar disorder, schizoaffective disorder and schizophrenia. *Schizophr. Res.* **218**, 38–47 (2020).
53. Kotov, R. *et al.* Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. *World Psychiatry* **19**, 151–172 (2020).
54. Qela, B. *et al.* Predictive coding in neuropsychiatric disorders: A systematic transdiagnostic review. *Neurosci. Biobehav. Rev.* **169**, 106020 (2025).

55. Sterzer, P. *et al.* The Predictive Coding Account of Psychosis. *Biol. Psychiatry* **84**, 634–643 (2018).
56. Corlett, P. R. *et al.* Hallucinations and Strong Priors. *Trends Cogn. Sci.* **23**, 114–127 (2019).
57. Parola, A. *et al.* Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence. *Schizophr. Res.* **259**, 59–70 (2023).
58. Xu, W. *et al.* Perplexity and proximity: Large language model perplexity complements semantic distance metrics for the detection of incoherent speech. *J. Biomed. Inform.* **170**, 104899 (2025).
59. Kircher, T. *et al.* Neurobiology of the major psychoses: a translational perspective on brain structure and function-the FOR2107 consortium. *Eur. Arch. Psychiatry Clin. Neurosci.* **269**, 949–962 (2019).
60. Murray, H. A. *Thematic Apperception Test*. (Harvard University Press, Cambridge, MA, 1943).
61. Andreasen, N. C. *Scale for the Assessment of Positive Symptoms (SAPS)*. (University of Iowa, Iowa City, 1984).
62. Andreasen, N. C. The Scale for the Assessment of Negative Symptoms (SANS): conceptual and theoretical foundations. *Br. J. Psychiatry Suppl.* 49–58 (1989).
63. Young, R. C., Biggs, J. T., Ziegler, V. E. & Meyer, D. A. A rating scale for mania: reliability, validity and sensitivity. *Br. J. Psychiatry* **133**, 429–435 (1978).
64. Hamilton, M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* **23**, 56–62 (1960).
65. Pedersen, G., Urnes, Ø., Hummelen, B., Wilberg, T. & Kvarstein, E. H. Revised manual for the Global Assessment of Functioning scale. *Eur. Psychiatry* **51**, 16–19 (2018).
66. Mülfarth, R. R. *et al.* Linking spontaneous speech, cognition, and psychopathology across affective and psychotic disorders: A network approach. *Eur. Psychiatry* **69**, e21

(2026).

67. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. in *Selected Papers of Hirotugu Akaike* (eds Parzen, E., Tanabe, K. & Kitagawa, G.) 199–213 (Springer New York, New York, NY, 1998).
68. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
69. Marquand, A. F., Rezek, I., Buitelaar, J. & Beckmann, C. F. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol. Psychiatry* **80**, 552–561 (2016).
70. Wittchen, H.-U., Zaudig, M. & Fydrich, T. *Strukturiertes Klinisches Interview Für DSM-IV (SKID)*. (Hogrefe Verlag, Göttingen, Germany, 1997).
71. Hamilton, M. The Assessment of Anxiety States by Rating. *Psychol. Psychother.* **32**, 50–55 (1959).
72. Liddle, P. F. *et al.* Thought and Language Index: an instrument for assessing thought and language in schizophrenia. *Br. J. Psychiatry* **181**, 326–330 (2002).
73. Lehrl, S. *Mehrfachwahl-Wortschatz-Intelligenztest (MWT-B)*. (Spitta, Balingen, Germany, 2005).
74. Helmstaedter, C. & Durwen, H. F. [The Verbal Learning and Retention Test. A useful and differentiated tool in evaluating verbal memory performance]. *Schweiz. Arch. Neurol. Psychiatr.* **141**, 21–30 (1990).
75. Reitan, R. M. *Trail Making Test: Manual for Administration and Scoring*. (Reitan Neuropsychology Laboratory, Tucson, AZ, 1992).
76. Aschenbrenner, S., Tucha, O. & Lange, K. W. *RWT - Regensburger Wortflüssigkeits-Test*. (Hogrefe Verlag, Göttingen, Germany, 2000).
77. Kwon, W. *et al.* Efficient Memory Management for Large Language Model Serving with PagedAttention. Preprint at <https://doi.org/10.48550/arXiv.2309.06180> (2023).

78. Qwen *et al.* Qwen2.5 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2412.15115> (2025).
79. Jiang, A. Q. *et al.* Mistral 7B. Preprint at <https://doi.org/10.48550/arXiv.2310.06825> (2023).
80. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
81. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. *SciPy 2010* <https://doi.org/10.25080/Majora-92bf1922-011> (2010) doi:[10.25080/Majora-92bf1922-011](https://doi.org/10.25080/Majora-92bf1922-011).
82. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
83. Reimers, N. & Gurevych, I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. Preprint at <https://doi.org/10.48550/arXiv.2004.09813> (2020).
84. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv Preprint* at <https://arxiv.org/abs/1908.10084v1> (2019).
85. Preacher, K. J. & Hayes, A. F. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* **40**, 879–891 (2008).
86. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementsRepurposingLargeLanguageModels1.4subnpjdm.docx](#)