Automated scoring of the Thought and Language Disorder Scale in schizophrenia using a large language model: reliability and comparison with clinicians*

Iannotta F.[1], Ceparano C.[1], Ricci C.[1], Iasevoli F.[1*]

[1] Department of Neuroscience, Reproductive Science and Odontostomatology, University *School of Medicine "Federico II", Via Sergio Pansini, 5, Naples 80131, Italy*

Abstract

Background: Formal thought disorders (FTDs) are a characteristic of schizophrenia symptomatology, but are difficult to score objectively. The Thought and Language Disorder (TALD) scale has been validated but not yet fully incorporated with NLP tools. We tested whether a large language model (LLM) could score the TALD reliably against clinicians.

Methods: Thirty-three individuals with schizophrenia (SCZ, n = 19) or treatment-resistant schizophrenia (TRS, n = 14) were evaluated and scored on the TALD by experienced clinicians. Recordings were also evaluated with an LLM trained on predetermined language measures. Intraclass correlation coefficients (ICCs) for total scores and weighted Cohen's kappa for items were used to determine reliability. A mixed-design ANOVA was used to test group effects.

Results: The LLM consistently provided higher TALD total scores than clinicians ($p = 0.001$), but replicated the absence of differences between SCZ and TRS patients. ICCs showed good overall agreement, and most items reached moderate-to-near-perfect concordance, but more atypical features (e.g., logorrhea, dissociation of thinking) showed smaller kappa values.

Conclusions: Automated TALD scoring approximates clinician ratings with good reliability, and therefore has potential as an objective and scalable assessment of thought disorder in schizophrenia.

Keywords

schizophrenia, large language model, formal thought disorders.

EMPATH-IA: Workshop on EMpowering PAtients THrough AI, Multimedia, and Explainable HCI: Innovations in Personalized Healthcare, in conjunction with the 16th Biannual Conference of the Italian SIGCHI Chapter (CHITALY 2025), 6–10 October 2025, Salerno, Italy

[1*] Corresponding author.

federica.iannotta@unina,it (F. Iannotta); caterina.ceparano3@gmail.com (C. Ceparano); claric86@gmail.com (C. Ricci), felice.iasevoli@unina.it (F. Iasevoli).

0000-0002-4938-7627 (F. Iannotta); 0009-0005-7975-5744 (C. Ceparano); 0009-0004-0997-6576 (C. Ricci); 0000-0002-7051-5013 (F. Iasevoli).

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4040/paper5.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Introduction

Approximately 1% of the global population is affected by schizophrenia, a highly heterogeneous disorder with pleiotropic manifestations, including cognitive deficits, disturbances in thought form and content, perceptual abnormalities, and impairments in social cognition and emotional regulation. This phenotypic heterogeneity likely reflects underlying neurobiological diversity, with multiple independent causal mechanisms converging on various pathophysiological pathways that give rise to broadly similar behavioral outcomes.

Consequently, efforts to identify a single, unifying etiology or pathophysiological mechanism for schizophrenia have been repeatedly frustrated by inconclusive results, and even advances in molecular, genomic, and neuroimaging research have been hampered by inconsistent findings. These challenges have hindered the development of reliable disease biomarkers, including emerging digital biomarkers derived from the rapidly growing field of medical artificial intelligence.

However, in recent years, formal thought disorder (FTD) has emerged as a core, relatively homogeneous behavioral phenotype of schizophrenia, with clear genetic underpinnings [1]. Individuals with schizophrenia often demonstrate reduced verbal productivity and verbal fluency, frequently producing disjointed and fragmented speech in which discourse lacks logical organization and coherence. Such semantic incoherence not only hampers effective communication but also reflects the underlying cognitive disorganization characteristic of schizophrenia [2-4] and is likely closely linked to the neurobiological substrates of this phenotype. Notably, in our previous work, the disorganization dimension—largely considered to capture formal thought disorder—was found to be the most predictive psychopathological domain for non-response to pharmacological treatments in patients with schizophrenia [5]. Furthermore, disorganization has been identified as the only psychotic dimension to correlate with impaired metabolic patterns of the prefrontal cortex in an FDG-PET study of patients with schizophrenia [6]. Together, these findings support the hypothesis that FTD may represent a promising candidate for research into schizophrenia-related biomarkers, including digital biomarkers.

Historically, Andreasen argued that in thought disorder the speaker “violates the syntactical and semantic conventions which govern language usage” [7]. She developed and validated the Scale for the Assessment of Thought, Language and Communication (TLC) [8] which distinguishes “positive” and “negative” thought disorder [9]. Positive thought disorder includes reductions in semantic or discourse coherence (e.g., tangentiality, derailment, circumstantiality), whereas negative thought disorder includes poverty of speech and poverty of content; while positive thought disorder ratings predicted psychosis, negative thought disorder was specifically predictive of schizophrenia [10].

Against this clinical backdrop, computational approaches have demonstrated that discourse abnormalities can be quantified and localized reproducibly. Foundational work using Latent Semantic Analysis (LSA) showed that automated measures of semantic coherence discriminate patients with schizophrenia from healthy controls and align with clinical ratings [11]. LSA is both a cognitive model of knowledge acquisition and a practical tool for concept-based text analysis: it learns semantic structure from large corpora by factorizing the term-by-context matrix via Singular Value Decomposition (SVD) to obtain a low-dimensional "semantic space" (ca. 100–500 dimensions) [12]. Within this tradition, LSA has localized where incoherence emerges during sentence production and predicted the degree of disorganization as well as class membership (patient vs. control) with accuracies around 80–82% [11]. Extending beyond patients, automated analyses have detected subtle deviations in relatives of individuals with schizophrenia, consistent with intermediate phenotypes and familial liability [13]. These and subsequent reviews established that linguistic biomarkers—coherence, referential cohesion, syntactic complexity, semantic density—can index psychosis risk, correlate with clinician ratings, and support prediction tasks [14-15].

Modern NLP has expanded this toolkit. Contextual embedding models (e.g., BERT/RoBERTa) capture bidirectional context and enable sentence-level embeddings sensitive to subtle disruptions in flow and meaning. A common approach measures semantic dissimilarity across adjacent utterances via distances in embedding space; additional features include next-sentence probability and surprisal, which operationalize, respectively, contextual fit and unexpectedness of an utterance [15-18]. These transformer-based measures detect subclinical language disturbance even when conventional clinical scores show no group differences, capturing increased tangentiality (larger embedding distances) and shifts in function-word usage in schizophrenia-spectrum disorders [15]. Complementary morpho-syntactic features, via part-of-speech (POS) tagging, quantify complexity (e.g., clause usage, sentence length), while Coh-Metrix-style indices such as type–token ratio (TTR) are often reduced and correlate with thought-disorder ratings [14; 18]. On the semantic content side, vector unpacking estimates semantic density—how many distinct meaning vectors are needed to reconstruct a sentence's meaning from distributional embeddings—thus indexing poverty of content; critically, low semantic density has been linked to increased risk of conversion from CHR to psychosis [19].

Beyond their predictive capability, NLP-derived metrics have proven sensitive to subclinical differences and generalize across tasks and settings: they discriminate SSD from controls even when clinician-rated scores do not [15], and they can be computed from open-ended verbalizations and short free-speech samples collected online, enabling scalable remote assessment [20-22]. Cross-linguistic studies indicate both shared and language-specific patterns in coherence and syntax, underscoring the need for language-aware tokenization/normalization and domain adaptation [23-25]. State-

of-the-art approaches integrate automated speech recognition (ASR) with semantic NLP to improve scalability in naturalistic settings [26]. Longitudinal work shows that composite NLP markers track within-person change in disorganization and negative symptom burden, supporting measurement-based care [15;27]; cluster analyses before/during/after onset reveal divergent trajectories in discourse features, with implications for early warning and personalized intervention [28]. Real-world deployments in Electronic Health Records demonstrate population-scale phenotyping (e.g., negative/cognitive symptoms extraction, duration of untreated psychosis timelines), while observational studies on social media reveal reduced coherence in naturalistic posts [29-33].

Despite decades of "proof-of-concept," clinical translation has lagged due to psychometric blind spots and fairness concerns. Criterion/content validity is often shown, but test–retest reliability, divergent validity (to address generalized-deficit concerns), and bias from demographics/context are under-evaluated. A comprehensive psychometric agenda—explicitly argued and exemplified in recent work—shows that model performance depends on contextual moderators (e.g., at home vs. away, alone vs. around strangers) and that systematic racial/sex biases can emerge if covariates are not modeled [34]. Accordingly, next-generation frameworks should adopt psychometrics-by-design (reliability, validity, measurement invariance), harmonized data-collection and preprocessing, and transparent reporting, with human-in-the-loop safeguards for high-stakes outputs [10; 34].

In sum, converging evidence from psycholinguistics, computational semantics, and deep NLP supports language as a quantitative phenotype for schizophrenia. Classical constructs (TLC positive vs. negative thought disorder) map onto measurable features (coherence, complexity, density). Foundational LSA-based studies established feasibility and validity [10;11;13;14] transformer-based models extend sensitivity to context-dependent disruptions [15-17], and longitudinal/cross-linguistic/real-world analyses demonstrate scalability and clinical relevance [20-25;27-33]. Nonetheless, artificial intelligence methods have been widely applied to support differential diagnosis across heterogeneous mental disorders and to predict disease trajectories. However, despite the substantial potential clinical impact of these applications, the reliability of such digital biomarkers remains uncertain, owing to the marked phenotypic and neurobiological heterogeneity of the underlying constructs they seek to measure. Some authors have leveraged artificial intelligence methods to predict symptom severity by correlating speech alterations, as assessed through NLP techniques, with scores on established clinical rating scales. However, the direct computation of psychopathology scores—through algorithm-based detection and quantification of speech and thought disturbances—has received little attention, likely due to the challenges of capturing the complexity and contextual nuances of clinical rating criteria via automated approaches.

Taken these elements together with our previous findings on the clinical and neurobiological salience of the disorganization/formal thought disorder (FTD) dimension [5; 6], these considerations strongly motivate the search for objective, scalable language markers as potential schizophrenia-related biomarkers—including digital biomarkers—within a harmonized, psychometrically rigorous framework. Here, we present a fully automated ASR + NLP pipeline specifically designed to directly quantify the level of formal thought disorder in patients by generating both item-level and total scores on the Thought and Language Disorder (TALD) Scale [35], achieving high consistency and reliability when compared with ratings from trained human evaluators. We propose that this pipeline could serve as a foundation for developing more refined systems aimed at enhancing model performance while preventing or mitigating inherent biases. We further anticipate that this pipeline could yield a more robust digital biomarker by anchoring measurement to a well-defined, homogeneous clinical phenotype—such as formal thought disorder—tightly linked to specific genomic and neurobiological substrates.

1. Methods

2.1 Populations.

We included 33 patients: 19 with a diagnosis of Schizophrenia (SCZ), and 14 with Treatment-resistant schizophrenia (TRS), recruited from July 2025 until August 2025.

All subjects were enrolled at the Outpatient Unit for Neurodevelopmental Disorders and treatment-resistant psychoses, Department of Neuroscience, Reproductive sciences and Dentistry, of the University of Naples Federico II. This study was part of the Supporting schizophrenia PatiEnts Care wiTh aRtificiAl intelligence (SPECTRA) project, a Research Projects of Significant National Interest (PRIN) 2022 PNRR, which the local Ethics Committee approved with protocol number 146/2025. All patients provided written informed consent before the study. All the study procedures were conducted following the principles of the 1975 Declaration of Helsinki, revised in 2008. The inclusion criteria were: i) age > 18; ii) capacity of giving written informed consent; iii) diagnosis of schizophrenia according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [36]. Exclusion criteria were: i) a diagnosis of intellectual disability or other neurodevelopmental conditions; ii) neurological disorders or cognitive decline; iii) current substance abuse. The schizophrenia diagnosis was performed by two trained clinicians, under the DSM-5 edition criteria. The definition of drug resistance was based on the indications of the American Psychiatric Association, subsequently redefined by the Treatment Response and Resistance in Psychosis (TRRIP) guidelines [37].

All the participants underwent a clinical interview conducted by two trained psychiatrists. The interview was recorded with the Tascam DR-05X digital audio recorder. After the interview, the two clinicians completed the TALD (Thought and Language Disorder Scale) scoring. The recordings were also rated by a Large Language Model (LLM), which was trained for TALD scoring according to predefined metrics, and produced independent TALD ratings for each participants.

2.3 Statistical analysis.

Statistical analyses were performed using R Studio (2.4.2024 version). Descriptive statistics were used to examine the TALD mean scores from clinicians and LLM ratings. For the descriptive statistics, the TALD total scores assigned by clinicians and by LLM for each group (all patients together, SCZ group, and TRS group) were reported as the mean and standard deviation (SD). To provide additional information on data distribution, the median and interquartile range are presented in boxplots. A mixed-design repeated measures ANOVA was performed to evaluate the differences within clinicians and LLM scoring and between SCZ and TRS groups, with Rater (clinicians vs. LLM) as the within-subjects factor and Group (SCZ vs. TRS) as the between-subjects factor. Partial eta squared (η^2p) was reported as a measure of effect size. Subsequently, a concordance analysis was performed. For each item of the TALD, the weighted Cohen's kappa (quadratic weights), with 95% confidence intervals and raw agreement percentages. For total TALD scores, intraclass correlation coefficients (ICC, model A,1) were computed. A p-value <0.05 was considered statistically significant, and the False Discovery Rate (FDR) adjustment was applied to account for multiple comparisons.

2. Results

3.1 Descriptive statistics

Clinicians and LLM produced similar TALD scores across groups (Table 1). For all patients, the TALD mean total score was 26.42 ± 10.67 for clinicians and 28.03 ± 7.96 for LLM. For the SCZ group, mean scores were 24.73±10.75 (clinicians' scores) vs. 28.21 ± 8.65 (LLM'scores), while in the TRS group, scores were 24.00 ±10.96 (clinicians' scores) vs. 27.78 ± 7.38 (LLM'scores).

Table 1. Descriptive statistics. Mean TALD scores for each patient group. SCZ = schizophrenia; TRS = Treatment-resistant schizophrenia; LLM = Large Language Model.

	Clinicians	**LLM**
All patients	26.42 +- 10.67	28.03 +- 7.96
SCZ group	24.73 +- 10.75	28.21+- 8.56
TRS group	24.00 +- 10.96	27.78+-7.38

3.2 Mixed design ANOVA

We used a mixed-design repeated measures ANOVA to examine differences between raters (clinicians and LLM, within-subjects factor) and between diagnostic groups (SCZ and TRS groups, between-subjects factor) in TALD total scores. This allowed us to examine both whether the LLM systematically differed from clinicians across score attribution, and whether the effect varied between patient groups.

The mixed-design repeated measures ANOVA revealed a significant Rater main effect ($F(1,31) = 12.67$, $p = 0.001$, $\eta^2p = 0.29$), indicating that the LLM provided higher TALD scores compared to clinicians. No significant Group main effect ($F(1,31) = 0.03$, $p = 0.861$, $\eta^2p = 0.001$) or Group × Rater interaction ($F(1,31) = 0.04$, $p = 0.836$, $\eta^2p = 0.001$) was found (Table 2). Boxplots of TALD total scores (median and IQR) are presented in Figure 1 and Figure 2 to display SCZ and TRS group distribution of the scores, separately for clinician and LLM ratings.

Table 2. Results of the mixed-design repeated measures ANOVA comparing TALD total scores between raters (clinicians VS LLM) and groups (SCZ vs TRS). SCZ = schizophrenia; TRS = Treatment-resistant schizophrenia; LLM = Large Language Model; η^2p = partial eta squared.

	F(1,31)	**p-value**	**η^2p**
Status (SCZ vs TRS)	0.03	0.861	0.001
Rater	12.67	0.001	0.290
Interaction (Status:Rater)	0.04	0.836	0.001

Figure 1. Boxplot displayed median and IQR for TALD scores across each group of patients (SCZ vs TRS) by the clinician's ratings. No significant differences were found between the groups.

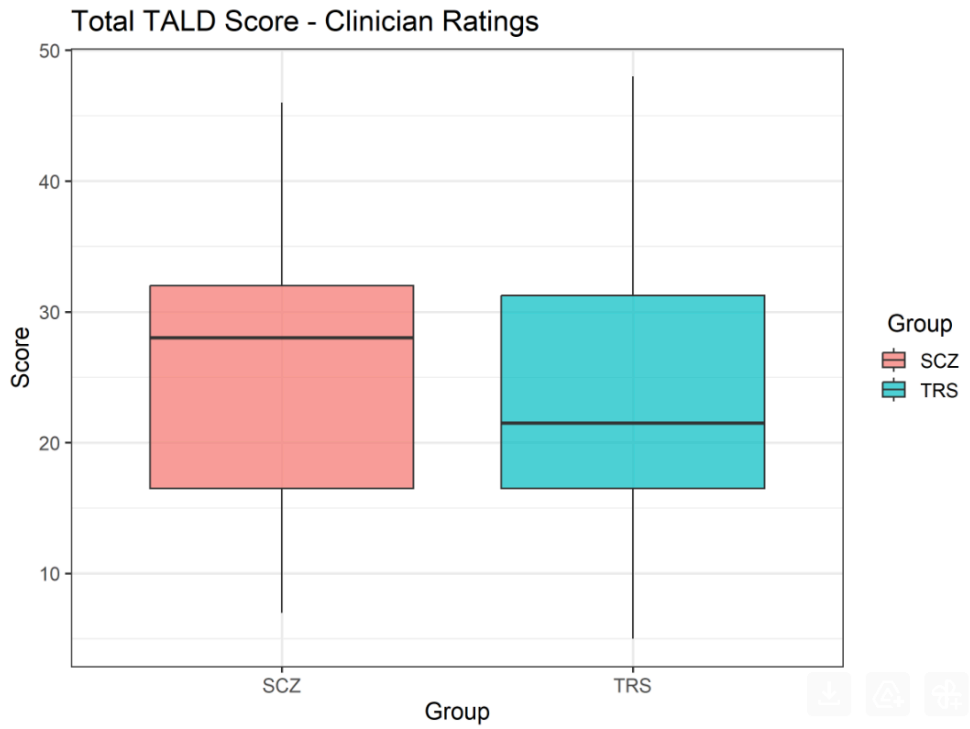
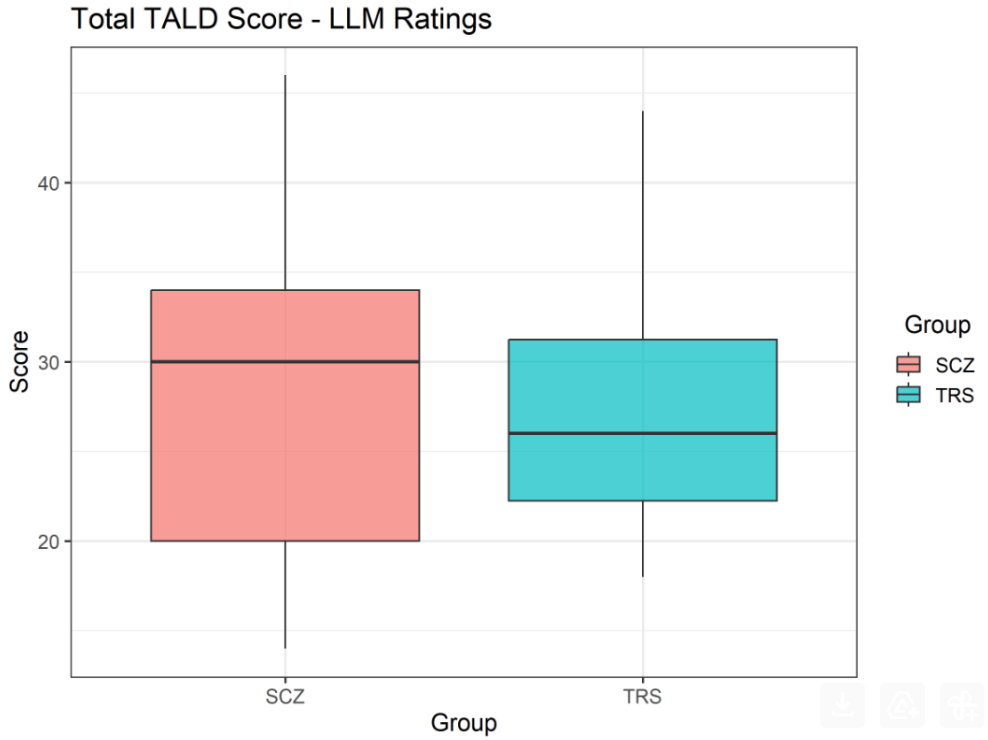


Figure 2. Boxplot displayed median and IQR for TALD scores across each group of patients (SCZ vs TRS) by LLM ratings. No significant differences were found between the groups.



3.3 Concordance on total TALD scores

Agreement on total TALD scores was evaluated using the ICC with a two-way mixed effects model, absolute agreement, single measures (ICC, 1, A), comparing the LLM scoring with clinician'scoring. The level of agreement was assessed for all patients and for each specific group (SCZ and TRS) to evaluate any differences in scoring that could depend on the group to which the patients belonged. The ICC indicated a good overall agreement between clinician ratings and LLM ratings (ICC = 0.84, 95% CI 0.42–0.94, p = 0.001). Good concordance was also observed separately within the SCZ group (ICC = 0.86, 95% CI 0.44–0.95, p = 0.001) and the TRS group (ICC = 0.83, 95% CI 0.33–0.95, p = 0.002) (Table 3).

Table 3. Intraclass correlation coefficient (ICC) for total TALD scores between clinicians and LLM for all patients, SCZ group, and TRS group. (SCZ schizophrenia; TRS = Treatment-resistant schizophrenia, LLM = Large Language Model).

Measure	ICC (A,1)	95% CI	p-value	Interpretation
Total TALD scores- all patients	0.842	0.424-0.941	0.001	Good agreement
Total TALD scores- SCZ group	0.858	0.444-0.954	0.001	Good agreement
Total TALD scores- TRS group	0.828	0.327-0.950	0.002	Good agreement

3.4 Item-level agreement between clinicians and LLM

Since the individual items of the TALD scale are ordinal variables, Cohen's weighted kappa was used to assess the agreement between clinicians and LLMs on each item. Weighted Cohen's kappa coefficients (quadratic weights) indicated variable agreement across items of the TALD. Blockage, interference of thought, and receptive speech dysfunction exhibited almost-perfect agreement ($\kappa \approx 0.94$), while numerous others showed substantial to moderate agreement ($\kappa = 0.60–0.80$). Scores were lower for logorrhea ($\kappa \approx 0.25$) and dissociation of thinking ($\kappa \approx 0.33$) (Table 4).

Table 4. Item-level agreement between clinician and AI (LLM) ratings on the TALD, reported as weighted Cohen's kappa coefficients (quadratic weights), 95% confidence intervals, p-values, and raw agreement percentages.

TALD item	Observations	Weight kappa (quadratic)	95% CI	p-value	p-value adjusted (FDR)	%Raw Agreement
Blockage	33	0.942	0.881-0.991	<0.001	0.001	87.87%
Circumstantiality	33	0.822	0.662-0.972	<0.001	0.001	75.76%
Perseverance	33	0.658	0.378-0.938	<0.001	0.001	78.81%
Concretism	33	0.903	0.800-1.00	<0.001	0.001	87.88%
Derailment	33	0.864	0.733-0.993	<0.001	0.001	84.85%
Crosstalk	33	0.592	0.201-0.984	<0.001	0.001	84.85%

Manneristic speech	33	0.716	0.368-1.062	<0.001	0.001	87.88%
Pressured speech	33	0.585	0.222-0.946	<0.001	0.001	81.82%
Dysfunction of Thought Initiative and Intentionality	33	0.895	0.809-0.981	<0.001	0.001	81.82%
Expressive speech dysfunction	33	0.893	0.845-0.939	<0.001	0.001	66.66%
Receptive speech dysfunction	33	0.940	0.894-0.987	<0.001	0.001	84.85%
Dissociation of thinking	33	0.326	-0.066-0.719	0.042	0.041	54.55%
Echolalia	33	0.449	-0.020-0.918	0.003	0.003	72.73%
Thought interference	33	0.942	0.898-0.985	<0.001	0.001	84.85%
Logorrhea	33	0.251	-0.164-0.666	0.148	0.148	75.76%
Neologism	33	0.539	0.040-1.030	0.001	0.001	81.82%
Fonemic paraphasia	33	0.547	-0.004-1.140	<0.001	0.001	90.91%
Semantic paraphasia	33	0.521	-0.044-1.087	<0.001	0.001	81.82%
Inhibited thinking	33	0.895	0.817-0.917	<0.001	0.001	78.79%
Slowed thinking	33	0.852	0.752-0.951	<0.001	0.001	72.73%
Restricted thinking	33	0.893	0.761-1.024	<0.001	0.001	87.88%
Perseverance	33	0.658	0.378-0.937	<0.001	0.001	72.73%
Poverty of content of speech	33	0.905	0.802-0.988	<0.001	0.001	81.82%
Poverty of speech	33	0.892	0.788-0.996	<0.001	0.001	81.82%
Poverty of thinking	33	0.536	0.237-0.833	0.001	0.001	66.67%
Pressure/rush of thoughts	33	0.870	0.755-0.985	<0.001	0.001	78.79%
Rupture of thought	33	0.441	0.147-0.734	0.009	0.009	69.70%
Rumination	33	0.871	0.765-0.986	<0.001	0.001	78.79%
Tangentiality	33	0.628	0.353-0.902	<0.001	0.001	66.67%

Verbigeration	33	0.593	0.0453-1.191	<0.001	0.001	90.91%

4. Discussion

Our study compared clinician and LLM ratings on the Thought and Language Disorder (TALD) scale in patients with SCZ and TRS. Although there have been previous attempts in the literature to identify FTDs in SCZ using LLM [13-22; 38], to our knowledge, this is the first study to propose a trained model for scoring an entire scale routinely applied in clinical practice.

Three main findings emerged. First, the mixed-design ANOVA showed a significant main effect of Rater, indicating that the LLM systematically assigned higher TALD ratings than human raters. Importantly, this difference was consistent across both diagnostic groups, as neither a Group effect nor a Group × Rater interaction was detected. This suggests that the LLM tends to overestimate (or, alternatively, that clinicians tend to underestimate) TALD severity in both diagnostic subgroups.

Second, despite this systematic shift in absolute values, agreement between clinicians and the LLM was good. ICCs for total TALD scores indicated good concordance for the overall sample as well as for the SCZ and TRS groups, confirming the stability of the automated scoring system. At the item level, weighted Cohen's kappa values ranged from moderate to almost perfect for most TALD items, with the highest concordance observed for blockage, thought interference, and receptive speech dysfunction. However, for some items (logorrhea, dissociation of thinking, echolalia, rupture of thought, paraphasias, verbigeration, pressured speech, crosstalk) the kappa values ranged from low to modest agreement, and broad confidence intervals. This trend likely reflects the rarity of these phenomena in our data: when things are uncommon, kappa estimates are volatile and even slight disagreements between raters inordinately lower the coefficient. That is, these findings may not reflect a systematic problem of the model, but rather the need for larger datasets with sufficient instances of rare phenomena to allow for more stable training and testing.

Third, at the group level, the machine-scored TALD mirrored the pattern of clinician ratings. Both clinicians and the LLM failed to detect differences in total TALD scores between SCZ and TRS groups, indicating that the automated system reproduced human judgment in relative terms. While TRS patients are generally considered more severe and higher scores would have been expected, this absence of difference can be explained in two ways: first, all patients included had been clinically stable for at least six months and individuals with acute exacerbations were excluded; second, SCZ and TRS may not differ in total TALD scores but only in qualitative aspects of formal thought disorder assessed by the scale. These aspects were not further investigated as

they were beyond the scope of this study. Importantly, however, the model did not introduce bias related to subgroup status, supporting its reliability.

Overall, these findings suggest that an NLP-based TALD rating can approximate experienced clinicians' ratings with high reliability. The consistent elevation of LLM scores may be due to the absence of an "emotional calibration" that clinicians implicitly apply when rating disorganized speech. Psychometrically, this upward shift could represent a strength, making the system more sensitive to subtle disturbances and less prone to underestimation due to clinical habituation or subjective thresholds. Conversely, calibration may be required if the tool is to be adopted in clinical decision-making based on predefined cutoffs.

From a clinical perspective, having an objective tool to assess thought disorder in psychosis would represent an important support for clinical practice. An LLM-based system can capture subtle and nuanced psychopathological alterations, providing a standardized assessment that is less influenced by emotional factors or socio-cultural knowledge of the patient. Future developments of similar tools could allow trained models to detect early exacerbations or subtle psychopathological changes, enable remote monitoring, and facilitate the identification of individuals at risk of developing psychosis.

Nevertheless, our study has limitations. The most relevant is the small sample size, which calls for cautious interpretation of the results. A modest clinical sample may lead to overestimation, as suggested by the wide confidence intervals despite strong statistical significance. Furthermore, the use of a single language and clinical context raises the risk of cultural bias. Moreover, the model itself could overestimate the results. Therefore, methodological transparency, interpretability of results, and clinical supervision of the outputs are still indispensable.

Acknowledgments

The authors thank the anonymous participants.
This research has been financially supported by the European Union NEXTGenerationEU project and by the Italian Ministry of University and Research (MUR), through a Research Project of Significant National Interest (PRIN) 2022 PNRR, project no. D53D23017290001 entitled *"Supporting schizophrenia Patients' Care with Artificial Intelligence (SPECTRA)"*, Principal Investigator: Rita Francese.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to perform grammar and spelling checks.

References

[1] Legge SE, Cardno AG, Allardyce J, Dennison C, Hubbard L, Pardiñas AF, Richards A, Rees E, Di Florio A, Escott-Price V, Zammit S, Holmans P, Owen MJ, O'Donovan MC, Walters JTR. Associations Between Schizophrenia Polygenic Liability, Symptom Dimensions, and Cognitive Ability in Schizophrenia. JAMA Psychiatry. 2021 Oct 1;78(10):1143-1151. doi: 10.1001/jamapsychiatry.2021.1961. PMID: 34347035; PMCID: PMC8340009.

[2] Meyer L, Lakatos P, He Y. Language Dysfunction in Schizophrenia: Assessing Neural Tracking to Characterize the Underlying Disorder(s)? Front Neurosci. 2021 Feb 22;15:640502. doi: 10.3389/fnins.2021.640502. PMID: 33692672; PMCID: PMC7937925.

[3] Compton MT, Ku BS, Covington MA, Metzger C, Hogoboom A. Lexical Diversity and Other Linguistic Measures in Schizophrenia: Associations With Negative Symptoms and Neurocognitive Performance. J Nerv Ment Dis. 2023 Aug 1;211(8):613-620. doi: 10.1097/NMD.0000000000001672. Epub 2023 May 31. PMID: 37256631; PMCID: PMC11140903.

[4] Hinzen, W., & Rosselló, J. (2015). The linguistics of schizophrenia: Thought disturbance as language pathology across positive symptoms. *Frontiers in Psychology, 6,* Article 971.

[5] Barone A, De Prisco M, et al. Disorganization domain as a putative predictor of Treatment Resistant Schizophrenia (TRS) diagnosis: A machine learning approach. J Psychiatr Res. 2022 Nov;155:572-578. doi: 10.1016/j.jpsychires.2022.09.044.

[6] Iasevoli F, D'Ambrosio L, et al. Altered Patterns of Brain Glucose Metabolism Involve More Extensive and Discrete Cortical Areas in Treatment-resistant Schizophrenia Patients Compared to Responder Patients and Controls: Results From a Head-to-Head 2-[18F]-FDG-PET Study. Schizophr Bull. 2023 Mar 15;49(2):474-485. doi: 10.1093/schbul/sbac1

[7] Andreasen NC, Grove WM. Thought, language, and communication in schizophrenia: diagnosis and prognosis. Schizophr Bull. 1986;12(3):348-59. doi: 10.1093/schbul/12.3.348. PMID: 3764356.

[8] Andreasen NC (1979a) Thought, language, and communication disorders. I. Clinical assessment, definition of terms, and evaluation of their reliability. Arch Gen Psychiatry 36:1315–1321

[9] Andreasen NC (1979b) Thought, language, and communication disorders. II. Diagnostic significance. Arch Gen Psychiatry 36:1325–1330

[10] Corcoran CM, Cecchi GA. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biological Psychiatry: Cognitive Neuroscience*

and Neuroimaging. 2020;5(8):770-779. doi:10.1016/j.bpsc.2020.06.004.

[11] Elvevåg B, Foltz PW, Weinberger DR, Goldberg TE. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*. 2007;93(1-3):304-316. doi:10.1016/j.schres.2007.03.001.

[12] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284. https://doi.org/10.1080/01638539809545028

[13] Elvevåg B, Foltz PW, Rosenstein M, DeLisi LE. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*. 2010;23(3):270-284. doi:10.1016/j.jneuroling.2009.05.002.

[14] Corcoran CM, Mittal VA, Bearden CE, et al. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*. 2020;226:158-166. doi:10.1016/j.schres.2020.04.032.

[15] Tang SX, Kriz R, Cho S, et al. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*. 2021;7(25). doi:10.1038/s41537-021-00154-3.

[16] Iter D, Yoon J, Jurafsky D. Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL). 2018; lines 146–152. doi:10.18653/v1/N18-2023.

[17] Hoffman RE, Hampson M, et al. Theory-driven language analysis in schizophrenia. *Schizophrenia Bulletin*. 2011;37(3):431-439. doi:10.1093/schbul/sbq094.

[18] Tanaka H, Chen H, Nagai T, et al. Automatic detection of disorganized speech in schizophrenia: Development and evaluation of a web-based system. *JMIR Mental Health*. 2020;7(8):e16829. doi:10.2196/16829.

[19] Rezaii N, Walker E, Wolff P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia*. 2019;5(9). doi:10.1038/s41537-019-0077-9.

[20] Jeong L, Lee M, Eyre B, et al. Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in Schizophrenia. *Psychiatry Research: Clinical Practice*. 2023;5(2):84-92. doi:10.1176/appi.prcp.20230003.

[21] Valverde-Albacete FJ, Peláez-Moreno C, et al. Detecting mental illness from short utterances on the web: Feasibility study. *JMIR Formative Research*. 2021;5(3):e24179. doi:10.2196/24179.

[22] Min MJ, Park S, Kim Y, et al. Automated detection of psychosis and at-risk mental state from short free-speech audio using deep learning. *Frontiers in Psychiatry*. 2022;13:818808. doi:10.3389/fpsyt.2022.818808.

[23] Parola A, Simonsen A, Bliksted V, Fusaroli R. Voice patterns in schizophrenia: A cross-linguistic systematic review and meta-analysis. *Schizophrenia Research*. 2020;216:17-24. doi:10.1016/j.schres.2019.11.031.

[24] Çabuk M, Altintas E, et al. Speech graph measures in Turkish patients with schizophrenia. *Clinical Linguistics & Phonetics*. 2021;35(10):915-930. doi:10.1080/02699206.2021.1884835.

[25] Arslan B, Öztürk A, et al. Coherence measures in Turkish schizophrenia speech: An NLP approach. *BMC Psychiatry*. 2022;22:447. doi:10.1186/s12888-022-04148-4.

[26] Tanaka S, Maezawa Y, Kirino E. Classification of schizophrenia patients and healthy controls using p100 event-related potentials for visual processing. Neuropsychobiology. 2013;68(2):71-8. doi: 10.1159/000350962. Epub 2013 Jul 19. PMID: 23881066.

[27] Carrillo F, Cecchi GA, Sigman M, et al. Language patterns before and after onset of psychosis: A longitudinal case study. *Schizophrenia Research*. 2018;197:627-633. doi:10.1016/j.schres.2018.01.003.

[28] López-Jaramillo C, Vargas C, et al. Longitudinal analysis of speech disorganization in first-episode psychosis. *Schizophrenia Research*. 2020;215:211-218. doi:10.1016/j.schres.2019.11.038

[29] Viani N, Pellizzer G, et al. Automatic detection of negative symptoms in schizophrenia from clinical records. *Journal of Biomedical Informatics*. 2020;110:103531. doi:10.1016/j.jbi.2020.103531.

[30] Chandran D, Radhakrishnan B, et al. Using NLP to identify obsessive-compulsive symptoms in serious mental illness. *BMC Psychiatry*. 2019;19:372. doi:10.1186/s12888-019-2365-6.

[31]Low DM, Rumker L, et al. Natural language processing reveals vulnerable mental health support groups and heightened COVID-19 anxiety on Reddit. *Internet Interventions*. 2020;20:100315. doi:10.1016/j.invent.2020.100315

[32] Reilly S, Planner C, et al. Negative symptoms and healthcare use in schizophrenia: Using NLP to investigate service use. *BMC Psychiatry*. 2021;21:119. doi:10.1186/s12888-021-03128-8.

[33] Fernandes AC, Dutta R, et al. Identifying first-episode psychosis in clinical records: An NLP approach. *BMJ Open*. 2013;3:e002764. doi:10.1136/bmjopen-2013-002764.

[34] Cohen AS, Rodriguez Z, Warren KK, et al. Natural Language Processing and Psychosis: On the Need for Comprehensive Psychometric Evaluation. *Schizophrenia Bulletin*. 2022;48(5):939-948. doi:10.1093/schbul/sbac051.

[35] Kircher T, Krug A, Stratmann M, Ghazi S, Schales C, Frauenheim M, Turner L, Fährmann P, Hornig T, Katzev M, Grosvald M, Müller-Isberner R, Nagels A. A rating scale for the assessment of objective and subjective formal Thought and Language Disorder (TALD). Schizophr Res. 2014 Dec;160(1-3):216-21. doi: 10.1016/j.schres.2014.10.024. Epub 2014 Nov 17. PMID: 25458572.

[36] American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

[37] Howes OD, McCutcheon R, Agid O, de Bartolomeis A, van Beveren NJ, Birnbaum ML, Bloomfield MA, Bressan RA, Buchanan RW, Carpenter WT, Castle DJ, Citrome L, Daskalakis ZJ, Davidson M, Drake RJ, Dursun S, Ebdrup BH, Elkis H, Falkai P, Fleischacker WW, Gadelha A, Gaughran F, Glenthøj BY, Graff-Guerrero A, Hallak JE, Honer WG, Kennedy J, Kinon BJ, Lawrie SM, Lee J, Leweke FM, MacCabe JH, McNabb CB, Meltzer H, Möller HJ, Nakajima S, Pantelis C, Reis Marques T, Remington G, Rossell SL, Russell BR, Siu CO, Suzuki T, Sommer IE, Taylor D, Thomas N, Üçok A, Umbricht D, Walters JT, Kane J, Correll CU. Treatment-Resistant Schizophrenia: Treatment Response and Resistance in Psychosis (TRRIP) Working Group Consensus Guidelines on Diagnosis and Terminology. Am J Psychiatry. 2017 Mar 1;174(3):216-229. doi: 10.1176/appi.ajp.2016.16050503. Epub 2016 Dec 6. PMID: 27919182; PMCID: PMC6231547.

[38] Pugh SL, Chandler C, Cohen AS, Diaz-Asper C, Elvevåg B, Foltz PW. Assessing dimensions of thought disorder with large language models: The tradeoff of accuracy and consistency. Psychiatry Res. 2024 Nov;341:116119. doi: 10.1016/j.psychres.2024.116119. Epub 2024 Aug 3. PMID: 39226873.