Bigger But Not Better: Small Neural Language Models Outperform Large Language Models in Detection of Thought Disorder

Changye Li[1*]   Weizhe Xu[1*]   Serguei Pakhomov[2]

Ellen Bradley[3]   Dror Ben-Zeev[1]   Trevor Cohen[1]

[1]University of Washington
[2]University of Minnesota
[3]University of California, San Francisco
{changyel, xuweizhe}@uw.edu

Abstract

Disorganized thinking is a key diagnostic indicator of schizophrenia-spectrum disorders. Recently, clinical estimates of the severity of disorganized thinking have been shown to correlate with measures of how difficult speech transcripts would be for large language models (LLMs) to predict. However, LLMs' deployment challenges – including privacy concerns, computational and financial costs, and lack of transparency of training data – limit their clinical utility. We investigate whether smaller neural language models can serve as effective alternatives for detecting positive formal thought disorder, using the same sliding window based perplexity measurements that proved effective with larger models. Surprisingly, our results show that smaller models are more sensitive to linguistic differences associated with formal thought disorder than their larger counterparts. Detection capability declines beyond a certain model size and context length, challenging the common assumption of "bigger is better" for LLM-based applications. Our findings generalize across audio diaries and clinical interview speech samples from individuals with psychotic symptoms, suggesting a promising direction for developing efficient, cost-effective, and privacy-preserving screening tools that can be deployed in both clinical and naturalistic settings.

1 Introduction

With an estimated prevalence of 15.2 in 100,000 persons (McGrath et al., 2008), schizophrenia-spectrum disorders (SSDs) are debilitating conditions that can lead to impaired social and occupational functioning, and poor healthcare outcomes including early mortality (Laursen et al., 2014). Formal Thought Disorder (FTD) – a breakdown in the structure of an individual's thinking – is a diagnostic feature of schizophrenia (Kircher et al.,

2018), and is recognized by observing speech that appears incoherent. Traditional evaluation of FTD relies on clinical interviews and standardized rating scales, which require extensive training and can be time-consuming. Natural language processing (NLP) methods have emerged as promising computational tools for automated evaluation of FTD. These data-driven approaches can systematically analyze linguistic patterns and discourse structure in patients' speech, offering objective quantitative measures of semantic coherence that correspond with clinical estimates of FTD severity (Elvevåg et al., 2007; Corcoran et al., 2020; Xu et al., 2021; Sarzynska-Wawer et al., 2021; Xu et al., 2022, *inter alia*).

Among the language impairments detectable by NLP methods, FTD represents a particularly complex set of disruptions in thought and speech organization. Patients with FTD exhibit distinct thinking patterns including tangentiality (gradual topic drift) and derailment (completely/partially unrelated thoughts), which may indicate a relative insensitivity to global discourse context (Kuperberg, 2010a,b). Previous psycholinguistic and neurolinguistic studies in *language comprehension* have shown an impaired ability to use global linguistic context (e.g., information from early in longer sentences) and relatively intact ability for local linguistic context (e.g., information from shorter sentences, local priming) in SSDs (Sitnikova et al., 2002; Swaab et al., 2013). With recent advances in neural language models (LMs) in particular, it is now possible to measure dependence upon global and local context in *language production* in SSDs. Recent work (Sharpe et al., 2024) suggests that disorganized speech can be characterized by comparing the global (probabilities estimated for observed text when including proximal and distal context) and local (probabilities estimated for observed speech when including proximal context only) lexical probabilities retrieved from GPT-3 (Brown

et al., 2020) when applied to speech samples from SSDs patients and neurotypical demographically-matched controls. A key finding from this research is that models that include longer context appear better-equipped to recognize language from participants with SSDs, in accordance with prior work showing SSDs patients are less influenced by overall sentence context during text comprehension (Sitnikova et al., 2002).

Commercially-developed large language models (LLMs) such as GPT-3 are pre-trained on large text corpora to enhance their linguistic capabilities. However, their usage and deployment raise concerns in clinical settings. Healthcare applications require stringent privacy protections, yet commercial LLMs mostly operate through cloud-based APIs, requiring the sharing of patient data with third-party commercial services. Even services compliant with the Health Information Portability and Accountability Act (HIPAA) or European counterpart General Data Protection Regulation (GDPR), LLMs can present certain risks to patient privacy especially when used for health-related purposes outside of a healthcare environment (Marks and Haupt, 2023). In addition to privacy concerns, computational requirements and associated costs of accessing commercial LLMs may also restrict their application in clinical settings. Lastly, the proprietary nature of commercial LLMs and a lack of transparency regarding their training data make it difficult to investigate or mitigate sources of bias.

While these limitations present challenges, there is a pressing need to explore the potential of LLMs in healthcare settings. Healthcare systems continuously face significant workforce shortages, particularly in specialized areas requiring extensive training (Thomas et al., 2009; Butryn et al., 2017). Current diagnostic and monitoring approaches rely heavily on in-person evaluations, creating bottlenecks in patient care and limiting access to specialized services, especially in resource-constrained settings. LLMs, if properly implemented with appropriate privacy safeguards, could help address these challenges by facilitating the development of technology-assisted diagnostic and monitoring tools, potentially improving the efficiency and accessibility of healthcare services.

Beyond these obstacles, we posit that commercial LLMs' extensive exposure to diverse linguistic patterns drawn from the internet and other sources – while beneficial for their remarkable text generation capabilities – may paradoxically reduce their sensitivity to subtle linguistic differences. This hypothesis is supported by recent works suggesting that broad exposure to diverse linguistic data leads LLMs to prioritize general patterns over fine-grained linguistic sensitivity (Lee et al., 2024a; Cong, 2024; Wilson et al., 2023), potentially diminishing their sensitivity to subtle deviations characteristic of FTD. We hypothesize that smaller LMs, such as those in the Pythia suite from EleutherAI (Biderman et al., 2023), may exhibit enhanced sensitivity to these linguistic phenomena. These models, ranging from 70M to 12B parameters, are trained on identical public datasets in the same order. As they differ in size only, it is possible to assess the extent to which they respond to nuanced linguistic patterns differently with the constraints in their capacity. In contrast, larger models' potentially excessive capacity to model complex textual relationships may obscure these subtle linguistic markers beneath layers of broader contextual understanding learned from vast amounts of data.

LMs' sensitivity to linguistic manifestations can be measured with perplexity (PPL). PPL is an intrinsic measure used to evaluate the performance of language models on unseen data. The more different the input is from a LM's training data, the "harder" it is for the model to predict the next word, resulting in higher PPL. Therefore, it is reasonable to hypothesize that PPL may have some degree of diagnostic utility, as has been documented by prior work using PPL to evaluate cognitive impairment in Dementia of the Alzheimer's Type (Orimaye et al., 2018; Fritsch et al., 2019; Cohen and Pakhomov, 2020; Li et al., 2022, 2024, *inter alia*) and psychosis (Colla et al., 2022; He et al., 2024).

Building upon Sharpe et al. (2024)'s findings with LLMs, our study seeks to assess smaller LMs' sensitivity to linguistic patterns associated with *positive* FTD by analyzing PPLs derived from Pythia models (exemplifying smaller LMs) and LLaMA (Dubey et al., 2024) (exemplifying a LLM) across both monologue and conversational speech samples from individuals with psychotic symptoms and clinically diagnosed SSDs respectively, and evaluating their correlation with the corresponding clinical ratings.

The contributions of this work can be summarized as follows: a) we provide empirical evidence suggesting that smaller LMs are more sensitive to linguistic patterns associated with FTD; b) we demonstrate that the degree of sensitivity starts to decline after models exceed a threshold of a certain

number of parameters, suggesting a diminishing relationship between model size and detection capability; and c) the sliding window PPL approach generalizes to both monologue and conversational speech samples of individuals with psychotic symptoms and clinically diagnosed SSDs respectively, suggesting its potential utility for screening and monitoring of SSDs in diverse clinical settings.[1]

2 Related Work

2.1 FTD in schizophrenia

Traditionally, FTD is evaluated through clinically administered rating scales such as the thought and language index (Liddle et al., 2002) or thought and language disorder (TALD) scale (Kircher et al., 2014) in research settings, which captures the full variety of FTD phenomenology including subjective experiences. It can also be more conveniently evaluated through self-reporting scales (Barrera et al., 2008). However, there are inherent problems associated with each approach: administering the clinical scales is time-consuming and requires specific training and expertise. Additionally, even when the required expertise is readily available, the clinical assessment only provides intermittent measures during office visits, making it difficult to paint a continuous picture in a more ecological setting. On the other hand, the self-reporting scale lacks objectivity because each patient may have subjective views on the scale's severity settings, and the ability to self-appraise may be impaired in psychosis. These inherent problems and the advancement of computational technologies have inspired the usage of NLP methods to evaluate and quantify the severity of FTD.

2.2 Assessing FTD and SSDs with NLP methods

Advancements in computational systems have introduced innovative methods for automated FTD assessment. A seminal approach by Foltz et al. (1998) utilizes distributional similarity, specifically measuring semantic relatedness between consecutive text segments using latent semantic analysis (LSA) (Landauer and Dumais, 1997) to measure coherence, providing a proxy that was later used to quantify the degree FTD. This method's diagnostic utility was demonstrated by Elvevåg et al. (2007), who found significant differences in automated co-herence metrics when comparing individuals with schizophrenia to healthy controls, as well as among patients with varying levels of thought disorder.

Building on this work, a subsequent study integrated LSA-based coherence metrics into a machine learning classifier that accurately predicted psychosis onset in a small sample of at-risk youth, achieving perfect leave-one-out cross-validation accuracy (Bedi et al., 2015). An adapted version maintained 83% accuracy in predicting psychosis onset in a larger, independent dataset (Corcoran et al., 2018). More recently, neural word embeddings (Mikolov et al., 2013), which represent words as vectors derived from neural networks trained to predict nearby words, have been explored as an alternative to LSA for coherence analysis. Similarity metrics from these embeddings showed promising results in aligning with clinical assessments of thought disorder (Just et al., 2019, 2020).

With advances in NLP methods, recent studies have used sentence embeddings from BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) to identify coherence differences between transcripts from individuals with SSDs and those from healthy controls (Tang et al., 2021). As a transformer-based model, BERT generates context-specific representations of tokens by dynamically incorporating information from surrounding words, unlike LSA or neural word embeddings, which rely on static word representations derived from all of the contexts a word is observed in during training. Prior research also introduced methods for assessing global coherence – estimating the relationship between a unit of text and the overarching theme of a text – using these methods to improve coherence evaluation in automatic speech recognition by extracting time series features for machine learning (Xu et al., 2021, 2022).

With the emergence of autoregressive LMs, some recent studies (Palaniyappan et al., 2023; Fradkin et al., 2023; Sharpe et al., 2024) have examined the assessment of psychosis using such models to demonstrate that LMs can be utilized with in-silico experimental research to gain better understanding of the linguistic manifestation of FTD. In contrast to BERT, which is a bidirectional model that utilizes tokens on both sides of a target token for prediction, autoregressive LMs are designed to predict only the next token in a sequence. While BERT-derived representations are highly effective for estimating semantic relatedness, autoregressive LMs are specifically optimized for generating co-

---

[1]Our code is publicly available on `https://github.com/LinguisticAnomalies/small-lm-sliding-windows`

herent and fluent sequences of text, offering potential for developing alternative approaches to FTD evaluation. However, these approaches have primarily relied on such models without exploring how model size and granular context windows affect sensitivity to linguistic manifestations of FTD.

3 Methods

3.1 Data

AVH Dataset Speech monologue samples from native English speaking participants who experienced auditory verbal hallucinations (AVH) using a smartphone application were collected during the course of a previous study (Ben-Zeev et al., 2020). Participants experiencing AVH were recruited in-person and online, and prompted to describe their experiences of AVH and anything else they would like to share or think would be helpful for the research team to know. Informed consent from participants was obtained through a rigorous procedure involving triple confirmations from a screening questionnaire. The study was approved by the Institutional Review Boards (IRB) of the University of Washington and Dartmouth College. Two annotators labeled the manual transcripts of the audio recordings for their degree of incoherence based on the TALD scale, using the construct of derailment. The TALD score ranges from 0-4 and represents greater incoherence as the score increases. The inter-rater agreement between annotators was 0.71, as measured by weighted Kappa. This set contained samples with a mean TALD score of 1.18 and a standard deviation of 0.83. We select 310 recordings that: a) have manual transcriptions; and b) are annotated with TALD. The transcript-level demographic information for this dataset is summarized in Table A.1 in the Appendix.

Clinical Interview Dataset This set contains semi-structured clinical interviews of San Francisco Bay Area male outpatients diagnosed with SSDs participating in a study of oxytocin conducted independently at University of California, San Francisco (UCSF) (Bradley et al., 2024). All participants are provided with written informed consent and study protocols were approved by the IRB at the UCSF. Following a prior work (Poole et al., 2000), the clinical assessments were conducted by trained raters in the form of a composite score combining the conceptual disorganization item (ranging from 1-7 with increasing severity) (Kay et al., 1987) from the Positive and Neg-

ative Syndrome Scale (PANSS) and the incoherent speech item (ranging from 0-5 with increasing severity) from the Comprehensive Assessment of Symptom and History (CASH) (Andreasen et al., 1992) to supplement the disorganized symptom subscale and the measure of suicidality. To avoid any potential confusion between these terms referring to different types of ratings, in the remainder of this paper we will refer this score as composite PANSS. We use manually transcribed interviews from 39 participants with corresponding composite PANSS between 2 and 8 (in the range of 0-12), with a mean of 3.36 and a standard deviation of 1.80. The transcript-level demographic information for this dataset is summarized in Table A.2 in Appendix.

3.2 Language models

Pythia is the first LLM suite deliberately designed to enable scientific research on LLMs. The Pythia suite offers pre-trained decoder-only autoregressive LMs ranging from 70M to 12B parameters. The Pythia suite is trained on the Pile corpus (Gao et al., 2020), which is a publicly available and curated collection of English language. In particular, we select Pythia checkpoints (70m, 160m, 410m, 1b, 1.4b, 2.8b, 6.9b, and 12b in parameter size) that are pre-trained on a *deduplicated* Pile corpus containing approximately 207B tokens. We select these checkpoints as deduplication has demonstrated its benefits in LLM training process (Lee et al., 2022). The Pythia suite largely follows the architecture and hyperparameters of GPT-3, but differs in several aspects: a) it uses fully dense attention layers; b) it is pre-trained using Flash Attention (Dao et al., 2022) for improved device throughput; and c) it uses rotary positional embeddings (Su et al., 2024) for a flexible mechanism to encode positional information.

We also compare Pythia suite with locally-hosted LLaMA-3.1-405b (Dubey et al., 2024) model, which is quantized with 4-bit precision using ExLlamaV2[2] (prior work indicates that quantization does not significantly degrade model performance (Lee et al., 2024b)). As initial experimentation in previous work showed comparable results to those obtained with the base model (without instruction tuning), we use an instruction-tuned model, hosted locally on a secure server.

---

[2] https://github.com/turboderp-org/exllamav2

93

3.3 Global and sliding window PPLs

We compute PPL for a transcript using two approaches: a) a global PPL that evaluates the full transcript as a single sequence; and b) a local PPL using sliding windows of 8, 16, 32, 64, and 128 for the Pythia suite, and a sliding window of 64 for the LLaMA model, as prior work indicates that restricting to a short input (e.g., context length of 128) can substantially improve the performance of LMs (Press et al., 2021). The sliding window is defined as a window of a corresponding number of tokens moved sequentially through the transcript. PPL is calculated for each window position as it shifts one token at a time until reaching the end of the transcript. If the transcript is shorter than the designated sliding window, then we calculate the global PPL for the transcript instead. As window size increases, the sliding window PPL approach allows the model to have more dynamic context when making each prediction, resulting in a more accurate approximation of the fully-factorized PPL (i.e., the global PPL). This can be particularly useful for evaluating spontaneous speech where the context is more fragmented than with read speech (Auer, 2009; Shriberg, 2001; Agmon et al., 2023; Wang et al., 2010). To generate a transcript-level measure, we use the maximum and the averaged PPL across the estimated sliding window PPLs, in addition to the global PPL for each transcript. For each measure, we compare the Spearman ρ with the TALD and composite PANSS scores for the AVH and clinical interview datasets respectively.

We opt to use maximum sliding window PPL as our primary transcript-level metric for detecting incoherent language. The rationale for this choice is evident in the distinct separation between transcripts that exhibit *mild* derailment (with TALD derailment < 3, labeled as 0) and those that exhibit *severe* derailment (with TALD derailment ≥ 3, labeled as 1) in the AVH dataset (Figure A.1 in Appendix). Transcripts rated with TALD ≥ 3 consistently exhibit higher maximum sliding window PPL spikes across different model sizes (particularly visible with the sliding window length of 64), while transcripts rated below this threshold maintain relatively stable, lower PPL patterns. We observed a similar trend in Figure A.2 in Appendix, the variation of PPL spikes across different severities of composite PANSS, suggesting that maximum sliding window PPL reflects disorganized

speech.[3]

4 Results

4.1 Global PPL as a proxy for FTD-related linguistic patterns

As illustrated in Figure 1, smaller Pythia models (Pythia-70m and Pythia-160m) consistently exhibited higher global PPLs compared to their larger counterparts across both the AVH and clinical interview datasets. Larger models (6.9b and 12b parameters) tended to cluster together at lower PPLs, suggesting diminishing effects on PPL estimation as model size increases. We observed minimal correlation ($\rho < 0.01$) between global PPL and TALD (in the AVH set), and this was statistically insignificant across all model sizes. While correlations between global PPL and composite PANSS (in the clinical interview set) were present (Spearman's ρ between 0.20 and 0.39), statistical significance was achieved only for the larger models, including Pythia-2.8b, Pythia-6.9b, and Pythia-12b, at $\alpha = 0.1$.

## 4.2 Sliding window PPL performance

4.2.1 The AVH dataset

Model	Sliding windows				
	8	16	32	64	128
70m	0.366***	0.375***	0.427***	**0.440***	0.370***
160m	0.369***	0.360***	0.426***	**0.451***	0.378***
410m	0.347***	0.336***	0.430***	**0.458***	0.378***
1b	0.329***	0.328***	0.431***	**0.458***	0.367***
1.4b	0.331***	0.305***	0.423***	**0.486***	0.388***
2.8b	0.315***	0.316***	0.435***	**0.464***	0.365***
6.9b	0.319***	0.310***	0.420***	**0.475***	0.370***
12b	0.317***	0.307***	0.421***	**0.470***	0.372***
LLaMA	–	–	–	0.457***	–

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

Table 1: The AVH dataset Spearman's ρ between the *maximum* sliding window PPL and TALD across model size. **Bold** indicates the highest ρ for a model.

As shown in Table 1, all correlations between *maximum* sliding window PPL and TALD scores were statistically significant (p-value < 0.01) across all model and sliding window sizes. The strongest correlations consistently occurred with a 64-token sliding window, with coefficients peaking at 0.486 for the 1.4b model, and remaining moderately correlated with TALD scores for all

---

[3]Our experiments are conducted on 3 H100 GPUs.
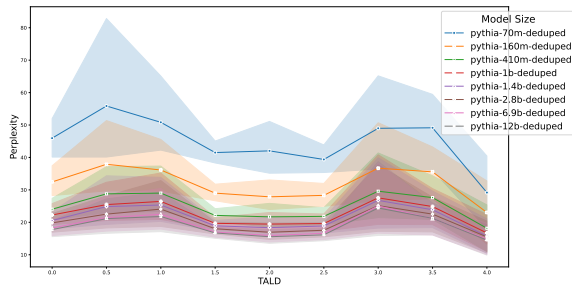
(a) Global PPLs of transcripts in the AVH dataset
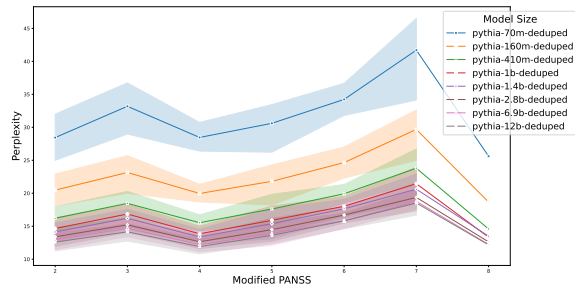
(b) Global PPLs of transcripts in the clinical interview dataset

Figure 1: Global PPLs estimated by the Pythia suite. The shaded area represents 95% confidence intervals of the global PPLs estimated from a pre-trained Pythia model.

model variants. Interestingly, the 4-bit quantized LLaMA 405b model did not outperform the Pythia suite, attaining a lower Spearman ρ of 0.457 on the 64-token sliding window.

Table A.3 in Appendix shows similar patterns of correlation between *averaged* sliding window PPL for a transcript and TALD scores. While all correlations remain significant (p-value < 0.01) across all model sizes and sliding windows, the correlation coefficients are generally lower compared to those for maximum sliding window PPLs. The 64-token sliding window again emerged as the optimal configuration, with correlation coefficients ranging from 0.202 to 0.251. The 1.4b model achieved the strongest correlation ($\rho = 0.251$), followed closely by the 6.9b model ($\rho = 0.249$). The LLaMA model achieved the highest correlation ($\rho = 0.371$) for the sliding window of 64 tokens.

4.2.2 The clinical interview dataset

Model	Sliding windows				
	8	16	32	64	128
70m	0.265	**0.482*****	0.338**	0.356**	0.344**
160m	0.414***	0.433***	0.316*	0.354***	0.325**
410m	0.385*	0.433***	0.316*	0.354**	0.325**
1b	0.415***	0.352**	0.380**	0.410***	0.313*
1.4b	0.458***	0.370***	0.382**	0.418***	0.326**
2.8b	0.425***	0.385**	0.369**	**0.428*****	0.348***
6.9b	**0.478*****	0.352**	0.394**	0.414***	**0.368****
12b	0.441***	0.313*	**0.404****	0.412***	0.357*
LLaMA	–	–	–	0.249	–

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

Table 2: The clinical interview dataset Spearman's ρ between the *maximum* sliding window PPL and modified PANSS across model size. **Bold** indicates the highest ρ for a model.

The patterns observed in Table 2 of the *maxi-*

mum sliding window PPLs on clinical interview dataset vary compared to those from audio diary data shown in Table 1. Pythia-70m with a sliding window of 16 has the highest correlation ($\rho = 0.482$, p-value < 0.01), while the Pythia-6.9b model shows comparable performance with a sliding window of 8 ($\rho = 0.478$, p-value < 0.01). Unlike the patterns with TALD, the sliding window size of 64 tokens was not optimal across all models, though it did yield strong correlations for several model sizes, particularly with the Pythia-2.8b model ($\rho = 0.428$, p-value < 0.01). Similarly, we also observed that LLaMA achieved the lowest and insignificant Spearman ρ with a 64-token sliding window.

Table A.4 in Appendix shows more moderate relationships for *averaged* sliding window PPLs in the clinical interview data, with correlation coefficients ranging from 0.248 to 0.360. The Pythia-1.4b model demonstrated consistently strong performance across all window sizes, achieving the highest correlation coefficients of all models, with its peak at a sliding window of 64 ($\rho = 0.360$, p-value < 0.05). LLaMA achieved the lowest Spearman ρ in the sliding window of 64, despite this being moderately significant.

4.3 The comparison of model sizes and sliding windows

For TALD correlations (Table 1 and Table A.3 in Appendix), larger sliding window sizes (32 and 64) consistently showed stronger correlation as model size increased. However, this trend was less evident with smaller sliding window sizes (8 and 16), where the correlation coefficients remained relatively stable across model sizes. In contrast, the composite PANSS correlations (Table 2 and Table A.4 in Appendix) exhibited a different pattern:

smaller sliding window sizes (8 and 16) showed more variation across model sizes with maximum PPLs, with correlation coefficients fluctuating. For example, with the sliding window of 8, the correlation coefficients ranged from $\rho = 0.265$ to $\rho = 0.478$. The correlation coefficients with averaged PPLs show more consistent behavior, as they gradually increase up to the 1.4b model across all sliding window sizes before they plateau or slightly decline with in larger models.

As can be observed in Table 1, which shows the correlations between maximum PPL for a transcript and the TALD on the AVH dataset, all models exhibit a consistent pattern where correlation coefficients generally increase with a sliding window of 8 and 16, peak at a sliding window of 64, and then decrease with a sliding window of 128. There is a similar trend in Table A.3, but with more moderate increases and decreases. In contrast, there are more variable patterns with the clinical interview results shown in Table 2. Smaller models (70m-410m) tended to achieve peak correlations at smaller window sizes, while larger models show more distributed peaks across different window sizes. The correlation between the averaged sliding window PPL and composite PANSS shows the most consistent pattern across window sizes, as is particularly evident with the 1.4b model, which maintains relatively stable correlations ranging from $\rho = 0.272$ to $\rho = 0.360$ across all window sizes. Notably, a sliding window of 128 consistently produced the weakest correlations in both datasets, suggesting that larger window size may dampen the PPL response to local patterns as compared with medium-sized windows. Interestingly, there is also a general trend of diminishing effects of sliding window size with both datasets, with the correlation coefficients declining with larger models (e.g., billions of parameter size) at the same sliding window size.

5 Discussion

Our key findings are as follows. First, our results suggest LM PPL can potentially serve as an objective computational marker for capturing subtle linguistic patterns associated with FTD. This aligns with previous studies indicating that such abnormal linguistic patterns manifest in ways that can be quantitatively measured (Colla et al., 2022; He et al., 2024; Xu et al., 2021, 2022; Sharpe et al., 2024). Second, our results extend Sharpe et al.

(2024)'s work by examining fine-grained sliding window PPLs to capture semantic variations across longer sequences using models ranging from 70m to 405b parameters for two model families. Our results indicate that model size does not necessarily correlate linearly with its capability for detecting FTD-related linguistic manifestations. Furthermore, our findings suggest that small/medium-sized sliding windows consistently demonstrate optimal performance across different model sizes, indicating an effective balance point between clinical utility and computational efficiency. That performance declines with larger windows may suggest the approximate range within which contextual inconsistencies manifest in FTD. These findings collectively suggest that calibrated smaller LMs can be at least as effective as their larger counterparts, offering practical advantages for real-world deployment while maintaining clinically-validated and robust performance in detecting FTD-related language differences.

Our work also demonstrates a more nuanced relationship between window size and the linguistic manifestations of FTD. While prior work (Sharpe et al., 2024) using GPT-3 indicates that differences in lexical probability (i.e., the intermediate products of PPLs) differ more between cases and controls with larger context windows (i.e., up to 50 tokens) for FTD, our work provides a more granular characterization of optimal sliding window sizes for alignment with human ratings in both monologue and conversational speech samples. This is particularly important for detecting linguistic inconsistencies that span across longer text segments, an aspect of comprehension that prior research suggests may be selectively impaired in people with SSDs (Kuperberg, 2010a,b; Sharpe et al., 2024). In addition, it reveals diminishing correlation coefficients at larger model and sliding window size, suggesting an upper bound to the utility of increasing both context window size and model size. Interestingly, the clinical interview dataset shows more variable optimal sliding window sizes across different model scales, with smaller models performing best at shorter windows and larger models showing distributed peaks across different window sizes. These varied patterns suggest that the manifestation of FTD – particularly in conversational language – may operate at multiple scales, rather than simply becoming more apparent with larger contexts. This in turn suggests that such LM-based methods may require calibrated combinations of both model size

and context window, rather than simply maximizing either dimension.

Our findings demonstrate the generalizability of PPL-based computational and automated assessment across both monologue (AVH) and conversational data (clinical interview), suggesting that changes in language associated with FTD can be effectively captured regardless of the communicative setting. Spontaneous speech presents unique complexities due to its impromptu nature, where speakers have minimal time to organize their thoughts. The challenges include a lack of clear syntactic boundaries (Auer, 2009; Shriberg, 2001; Agmon et al., 2023; Wang et al., 2010), complex interaction of linguistic demands due to mental states (Menn and Obler, 1989; Caplan and Hanna, 1998), and context sensitivity. They make it particularly challenging for generalizable computational analysis. However, our results show that PPL-based measures can effectively operate within these complexities, yielding statistically significant correlations with human ratings across both monologue and conversational datasets. This capability to perform consistently across different communication contexts is important for clinical applications, where assessment tools may need to maintain reliability across various real-world scenarios. The consistency of our results across both data sources indicates the potential of sliding-window based LM perplexity as an automated and computational assessment tool.

With respect to model size, 4-bit quantized LLaMA 405b, despite its significantly larger scale and strong performance on open-domain tasks (Lee et al., 2024b), consistently underperformed compared to smaller Pythia models. This finding supports our hypothesis regarding the potential advantages of smaller LMs in detecting subtle linguistic patterns associated with FTD, though it remains to be determined whether this advantage is attributable to model size or a relatively constrained amount of training data. Larger LMs, with their extensive pre-training on vast corpora of text of uncertain provenance, may find the subtle linguistic patterns that characterize FTD more predictable. In contrast, smaller LMs' more limited exposure to coherent language patterns and/or constrained capacity (as proxied by parameter size) may paradoxically enhance their sensitivity to linguistic patterns associated with FTD. This suggests that the relationship between model scale and clinical assessment capability is not strictly linear (i.e., bigger

is not necessarily better), and that optimal performance may be achieved by models that maintain adequate linguistic competence while remaining sensitive to deviations from typical language patterns. A further advantage of the Pythia suite is that their training data are publicly available, and therefore amenable to analyses to identify sources of biased assessment, such as the absence of training data reflecting dialectical variation characteristic of particular population groups. These findings and observations collectively suggest that pursuing ever-larger models may not necessarily yield better clinical assessment capabilities and utilities.

Our analysis across multiple model sizes provides empirical guidance for sliding window size selection in clinical practice. The finding that small- to medium-sized sliding windows (typically 16 to 64 tokens) consistently demonstrate optimal performance across different model sizes suggests an effective range for practical implementation. This observation is consistent with previous studies demonstrating that linguistic inconsistencies manifest as local coherence disruptions (Sitnikova et al., 2002; Swaab et al., 2013; Kuperberg, 2010a,b). The observed performance decline with larger windows (> 128 tokens) further supports this understanding. Notably, the optimal sliding window sizes remain consistent across both shorter monologues (≈ 100 tokens) and longer clinical interviews (> 1000 tokens), suggesting that the linguistic manifestation of FTD operates at a fragmented level independent of overall discourse length or interaction type. This pattern suggests that aspects of FTD may be best characterized in intermittent steps rather than as global narrative incoherence. Additionally, the sliding window size sensitivity remains remarkably consistent across model scales from 70m to 12b, suggesting that PPL, as a computational marker, can effectively capture such linguistic manifestations, providing a compelling evidence for the context-sensitive nature of FTD and its variable expression across different communicative demands.

Our findings suggest that PPL derived from smaller LMs with granular sliding windows offer promising clinical utility in addition to existing assessment methods. Furthermore, The reduced computational requirements of our approach also makes it particularly suitable for resource-constrained settings, potentially enabling automated FTD screening in underserved communities. These models' ability to detect subtle linguistic manifestations of FTD opens several promising application path-

ways in clinical practice. Most notably, the efficiency of smaller LMs (70m-410m parameters) enables privacy-preserving, on-device processing that could streamline the mental health monitoring and early intervention. For example, these lightweight models could be integrated into telehealth platforms to analyze discourse during remote psychiatric consultations in real-time, providing clinicians with immediate linguistic computational markers while ensuring all patient data remains on local devices. In ambient monitoring scenarios, these models could be deployed on smartphones to periodically assess everyday conversation with participants' prior consent, creating longitudinal datasets that track subtle changes in FTD, enabling the comparison of cross-sectional linguistic patterns to identify preliminary warning signs that might surf unnoticed. The resulting data could also used for near-real-time flagging of warning signs, which could be shared with their clinical teams to enable time-sensitive interventions that may help prevent further deterioration (Ben-Zeev et al., 2017).

While our token-level sliding window PPL method demonstrates promising results, we acknowledge that certain sentence-level proximity-based methods (Xu et al., 2022) achieved comparable or higher correlations. However, the sliding window PPL method is responsive to linguistic inconsistencies at varying granularities that complement existing methods, potentially capturing dynamic aspects of FTD that may be missed by static sentence-level measures.

6   Conclusion

We presented experimental findings consistent with prior work suggesting sliding-window PPL as an efficient measurement for linguistic patterns associated with FTD. Surprisingly, our findings suggest that smaller LMs with calibrated sliding window sizes, are more sensitive to such linguistic manifestations. The comparable effectiveness of smaller models opens new possibilities for implementing automated and computational language assessment tools in resource-constrained clinical settings while remaining cost-efficient and privacy-preserving.

Limitations

The work presented here has several limitations. All participants represented in both data sets are English speakers, and it remains unclear the extent to which our findings apply to other languages. Our analysis relied on transcribed speech data, which may not fully capture the nuances of spoken language, including prosody, pauses, and other paralinguistic features that could be clinically relevant (for a related review, see Ehlen et al. (2023)). While our findings demonstrate correlations between PPL and human ratings, these ratings do not constitute clinical diagnoses, which would be needed for case/control comparisons. Furthermore our analysis does not account for potential confounding variables – such as age, gender, origin and socioeconomic status – which may influence language patterns. While smaller LMs show promising results, we have not yet established clear clinical thresholds that would be necessary for diagnosis, or assessed the utility of measurements over time as indicators of change in clinical status. We also note that the severity scores for both datasets are relatively low on average, and that datasets with more representation of severe FTD may be needed to establish optimal parameter settings in this context. Our study focused specifically on positive FTD, a key diagnostic feature for SSDs. Therefore, the extent to which sliding window PPL is responsive to linguistic manifestations of other conditions remains to be established. Future work to address these limitations will be required to reach the potential of these methods for clinical deployment. Finally, while we included a larger language model with a 64-token sliding window, including additional sliding window sizes with LLaMA would make for a more comprehensive analysis that is left for future work.

Acknowledgment

References

Galit Agmon, Manuela Jaeger, Reut Tsarfaty, Martin G Bleichner, and Elana Zion Golumbic. 2023. "um. . . , it's really difficult to. . . um. . . speak fluently": Neural tracking of spontaneous speech. *Neurobiology of Language*, 4(3):435–454.

Nancy C. Andreasen, Michael Flaum, and Stephan Arndt. 1992. The comprehensive assessment of symptoms and history (cash): An instrument for assessing diagnosis and psychopathology. *Archives of General Psychiatry*, 49.

Peter Auer. 2009. On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences*, 31(1):1–13.

Alvaro Barrera, Peter J. McKenna, and German E. Berrios. 2008. Two new scales of formal thought disorder in schizophrenia. *Psychiatry Research*, 157.

Gillinder Bedi, Facundo Carrillo, Guillermo A. Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B. Mota, Sidarta Ribeiro, Daniel C. Javitt, Mauro Copelli, and Cheryl M. Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*.

Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury, Marta Hauser, et al. 2017. Crosscheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric rehabilitation journal*, 40(3):266.

Dror Ben-Zeev, Benjamin Buck, Ayesha Chander, Rachel Brian, Weichen Wang, David Atkins, Carolyn J Brenner, Trevor Cohen, Andrew Campbell, and Jeffrey Munson. 2020. Mobile rdoc: Using smartphones to understand the relationship between auditory verbal hallucinations and need for care. *Schizophrenia Bulletin Open*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Ellen R. Bradley, Jake Portanova, Josh D. Woolley, Benjamin Buck, Ian S. Painter, Michael Hankin, Weizhe Xu, and Trevor Cohen. 2024. Quantifying abnormal emotion processing: A novel computational assessment method and application in schizophrenia. *Psychiatry Research*, 336:115893.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tracy Butryn, Leah Bryant, Christine Marchionni, and Farhad Sholevar. 2017. The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions. *International Journal of Academic Medicine*, 3(1):5–9.

David Caplan and Joy E. Hanna. 1998. Sentence production by aphasic patients in a constrained task. *Brain and Language*, 63(2):184–218.

Trevor Cohen and Serguei Pakhomov. 2020. A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1957, Online. Association for Computational Linguistics.

Davide Colla, Matteo Delsanto, Marco Agosto, Benedetto Vitiello, and Daniele P Radicioni. 2022. Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, 134:102393.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Cheryl M. Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C. Javitt, Carrie E. Bearden, and Guillermo A. Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*.

Cheryl M. Corcoran, Vijay A. Mittal, Carrie E. Bearden, Raquel E. Gur, Kasia Hitczenko, Zarina Bilgrami, Aleksandar Savic, Guillermo A. Cecchi, and Phillip Wolff. 2020. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*, 226:158–166.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Felicitas Ehlen, Christiane Montag, Karolina Leopold, and Andreas Heinz. 2023. Linguistic findings in persons with schizophrenia—a review of the current literature. *Frontiers in Psychology*, 14.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*.

Peter W. Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*.

Isaac Fradkin, Matthew M. Nour, and Raymond J. Dolan. 2023. Theory-driven analysis of natural language processing measures of thought disorder using generative language modeling. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(10):1013–1023. Natural Language Processing in Psychiatry and Clinical Neuroscience Research.

Julian Fritsch, Sebastian Wankerl, and Elmar Nöth. 2019. Automatic diagnosis of alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845. IEEE.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Rui He, Claudio Palominos, Han Zhang, Maria Francisca Alonso-Sánchez, Lena Palaniyappan, and Wolfram Hinzen. 2024. Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Research*, 333:115752.

Sandra Just, Erik Haegert, Nora Kořánová, Anna-Lena Bröcker, Ivan Nenchev, Jakob Funcke, Christiane Montag, and Manfred Stede. 2019. Coherence models in schizophrenia. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136, Minneapolis, Minnesota. Association for Computational Linguistics.

Sandra A. Just, Erik Haegert, Nora Kořánová, Anna Lena Bröcker, Ivan Nenchev, Jakob Funcke, Andreas Heinz, Felix Bermpohl, Manfred Stede, and Christiane Montag. 2020. Modeling incoherent discourse in non-affective psychosis. *Frontiers in Psychiatry*.

S. R. Kay, A. Fiszbein, and L. A. Opler. 1987. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia Bulletin*, 13.

Tilo Kircher, Henrike Bröhl, Felicitas Meier, and Jennifer Engelen. 2018. Formal thought disorders: from phenomenology to neurobiology. *The Lancet Psychiatry*, 5(6):515–526.

Tilo Kircher, Axel Krug, Mirjam Stratmann, Sayed Ghazi, Christian Schales, Michael Frauenheim, Lena Turner, Paul Fährmann, Tobias Hornig, Michael Katzev, Michael Grosvald, Rüdiger Müller-Isberner,

and Arne Nagels. 2014. A rating scale for the assessment of objective and subjective formal thought and language disorder (tald). *Schizophrenia Research*.

Gina R Kuperberg. 2010a. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589.

Gina R Kuperberg. 2010b. Language in schizophrenia part 2: What can psycholinguistics bring to the study of schizophrenia... and vice versa? *Language and linguistics compass*, 4(8):590–604.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Thomas Munk Laursen, Merete Nordentoft, and Preben Bo Mortensen. 2014. Excess early mortality in schizophrenia. *Annual review of clinical psychology*, 10(1):425–448.

Eun-Kyoung Rosa Lee, Sathvik Nair, and Naomi Feldman. 2024a. A psycholinguistic evaluation of language models' sensitivity to argument roles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3262–3274, Miami, Florida, USA. Association for Computational Linguistics.

Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. 2024b. A comprehensive evaluation of quantized instruction-tuned large language models: An experimental analysis up to 405b. *arXiv preprint arXiv:2409.11055*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Changye Li, Zhecheng Sheng, Trevor Cohen, and Serguei Pakhomov. 2024. Too big to fail: Larger language models are disproportionately resilient to induction of dementia-related linguistic anomalies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6363–6377, Bangkok, Thailand. Association for Computational Linguistics.

Peter F Liddle, Elton TC Ngan, Stephanie L Caissie, Cameron M Anderson, Alan T Bates, Digby J Quested, Richard White, and Rowena Weg. 2002.

Thought and language index: an instrument for assessing thought and language in schizophrenia. *The British Journal of Psychiatry*, 181(4):326–330.

Mason Marks and Claudia E Haupt. 2023. Ai chatbots, health privacy, and challenges to hipaa compliance. *Jama*.

John McGrath, Sukanta Saha, David Chant, Joy Welham, et al. 2008. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic reviews*, 30(1):67–76.

Lise Menn and Loraine K Obler. 1989. Cross-language data and theories of agrammatism. In *Agrammatic aphasia*, pages 1369–1389. John Benjamins.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Sylvester Olubolu Orimaye, Jojo Sze-Meng Wong, and Chee Piau Wong. 2018. Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PloS one*, 13(11):e0205636.

Lena Palaniyappan, David Benrimoh, Alban Voppel, and Roberta Rocca. 2023. Studying Psychosis Using Natural Language Generation: A Review of Emerging Opportunities. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(10):994–1004.

John H Poole, Faith C Tobias, and Sophia Vinogradov. 2000. The functional relevance of affect recognition errors in schizophrenia. *Journal of the International Neuropsychological Society*, 6(6):649–658.

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Shortformer: Better language modeling using shorter inputs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5493–5505, Online. Association for Computational Linguistics.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Victoria Sharpe, Michael MacKinley, Samer Nour Eddine, Lin Wang, Lena Palaniyappan, and Gina R Kuperberg. 2024. Gpt-3 reveals selective insensitivity to global vs. local linguistic context in speech produced by treatment-naive patients with positive thought disorder. *bioRxiv*, pages 2024–07.

Elizabeth Shriberg. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the international phonetic association*, 31(1):153–169.

Tatiana Sitnikova, Dean F Salisbury, Gina Kuperberg, and Phillip J Holcomb. 2002. Electrophysiological insights into language processing in schizophrenia. *Psychophysiology*, 39(6):851–860.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Tamara Y Swaab, Megan A Boudewyn, Debra L Long, Steve J Luck, Ann M Kring, J Daniel Ragland, Charan Ranganath, Tyler Lesh, Tara Niendam, Marjorie Solomon, et al. 2013. Spared and impaired spoken discourse processing in schizophrenia: effects of local and global language context. *Journal of Neuroscience*, 33(39):15578–15587.

Sunny X. Tang, Reno Kriz, Sunghye Cho, Suh Jung Park, Jenna Harowitz, Raquel E. Gur, Mahendra T. Bhati, Daniel H. Wolf, João Sedoc, and Mark Y. Liberman. 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*.

Kathleen C Thomas, Alan R Ellis, Thomas R Konrad, Charles E Holzer, and Joseph P Morrissey. 2009. County-level estimates of mental health professional shortage in the united states. *Psychiatric services*, 60(10):1323–1328.

Yu-Tsai Wang, Jordan R Green, Ignatius SB Nip, Ray D Kent, and Jane Finley Kent. 2010. Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatrica et Logopaedica*, 62(6):297–302.

Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

Weizhe Xu, Jake Portanova, Ayesha Chander, Dror Ben-Zeev, and Trevor Cohen. 2021. The centroid cannot hold: comparing sequential and global estimates of coherence as indicators of formal thought disorder. In *AMIA Annual Symposium Proceedings*, volume 2020, page 1315.

Weizhe Xu, Weichen Wang, Jake Portanova, Ayesha Chander, Andrew Campbell, Serguei Pakhomov, Dror Ben-Zeev, and Trevor Cohen. 2022. Fully automated detection of formal thought disorder with time-series augmented representations for detection of incoherent speech (TARDIS). *Journal of Biomedical Informatics*, 126:103998.

A Appendix

	Level	Mild	Severe
# of transcripts		292	18
Age (mean (SD))		40.45 (10.71)	38.33 (8.97)
Gender (%)	Female	162 (55.5)	10 (55.6)
	Male	118 (40.4)	8 (44.4)
	Transgendered: FTM	3 (1.0)	0 (0.0)
	Transgendered: MTF	9 (3.1)	0 (0.0)
Education (%)		1 (0.3)	0 (0.0)
	Associates Degree	51 (17.5)	3 (16.7)
	Bachelors Degree	25 (8.6)	0 (0.0)
	Doctorate Degree	3 (1.0)	0 (0.0)
	Grade school	4 (1.4)	4 (22.2)
	High School Diploma /GED	171 (58.6)	9 (50.0)
	Junior High	22 (7.5)	2 (11.1)
	Masters Degree	15 (5.1)	0 (0.0)
Race (%)		1 (0.3)	0 (0.0)
	American Indian or Alaskan Native	6 (2.1)	0 (0.0)
	Asian	6 (2.1)	0 (0.0)
	Black or African American	61 (20.9)	8 (44.4)
	More than one race	35 (12.0)	2 (11.1)
	White	183 (62.7)	8 (44.4)
# of words per transcript (mean (SD))		182.76 (139.17)	300.50 (170.42)
TALD (mean (SD))		1.08 (0.70)	3.33 (0.34)

Table A.1: Basic transcript-level demographic information for the AVH dataset. `Mild` denotes as mild symptoms of positive FTD where TALD score < 3, and `Severe` denotes severe symptoms of positive FTD, where TALD score $\geq$ 3.

	level	2	3	4	5	6	7	8
# of transcripts		20	5	5	3	2	3	1
Gender (%)	Male	20 (100.0)	5 (100.0)	5 (100.0)	3 (100.0)	2 (100.0)	3 (100.0)	1 (100.0)
Age (mean (SD))		32.35 (10.84)	28.20 (6.26)	35.20 (13.77)	41.00 (3.61)	24.50 (3.54)	36.00 (23.39)	58.00 (NA)
Race (%)	African American	4 (20.0)	0 (0.0)	2 (40.0)	0 (0.0)	0 (0.0)	1 (33.3)	0 (0.0)
	Asian	2 (10.0)	1 (20.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (100.0)
	Latino	1 (5.0)	1 (20.0)	0 (0.0)	0 (0.0)	1 (50.0)	0 (0.0)	0 (0.0)
	Other (including mixed)	2 (10.0)	0 (0.0)	1 (20.0)	1 (33.3)	1 (50.0)	1 (33.3)	0 (0.0)
	White	11 (55.0)	3 (60.0)	2 (40.0)	2 (66.7)	0 (0.0)	1 (33.3)	0 (0.0)
Education (%)	Associates Degree	2 (10.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (33.3)	0 (0.0)
	Bachelors Degree	3 (15.0)	2 (40.0)	2 (40.0)	1 (33.3)	0 (0.0)	0 (0.0)	0 (0.0)
	GED	1 (5.0)	1 (20.0)	0 (0.0)	1 (33.3)	0 (0.0)	0 (0.0)	0 (0.0)
	H.S. Diploma	13 (65.0)	1 (20.0)	3 (60.0)	1 (33.3)	2 (100.0)	2 (66.7)	1 (100.0)
	Other	0 (0.0)	1 (20.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
	Vocational Certification	1 (5.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
# of words per transcript (mean (SD))		1494.20 (1053.27)	1178.40 (536.35)	1822.00 (1774.69)	1478.67 (601.23)	2004.50 (340.12)	3023.67 (2197.36)	4339.00 (NA)

Table A.2: Basic transcript-level demographic information for the clinical interview dataset. The top row represents the value of the composite PANSS.
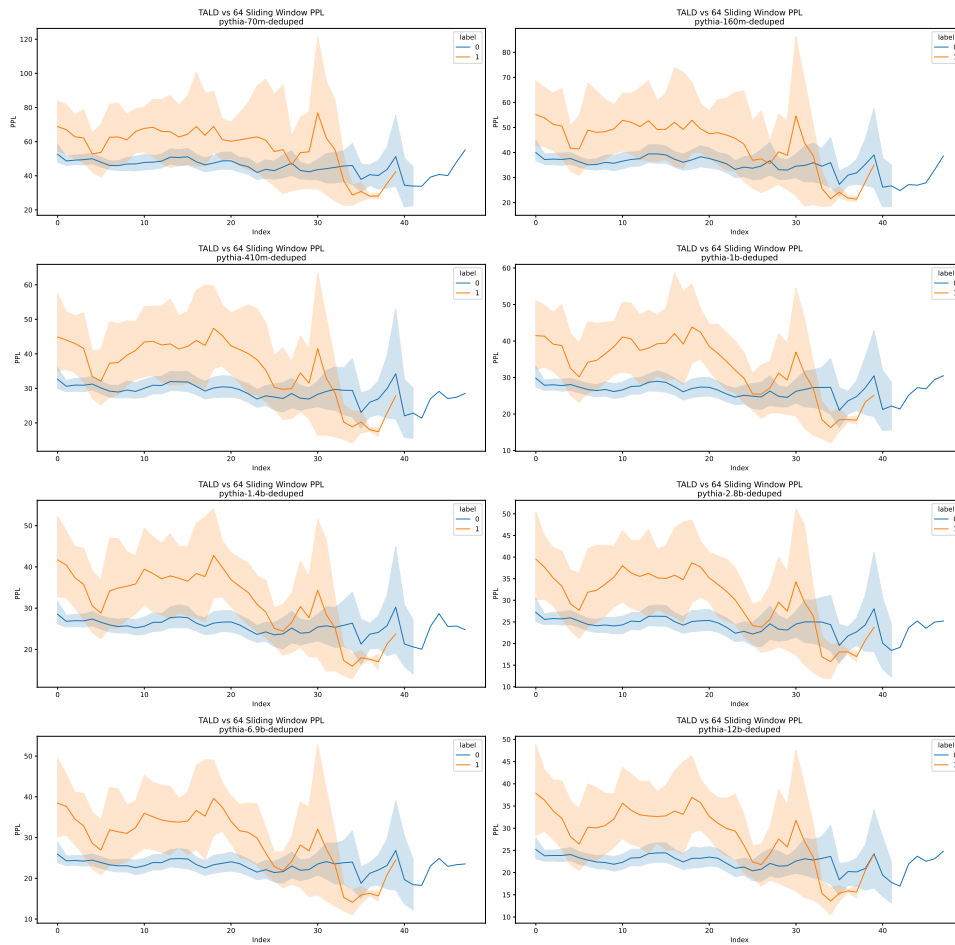
Figure A.1: The distribution of sliding window PPLs using a sliding window of 64 tokens on the AVH dataset, where the x-axis represents the index of sliding window PPLs in transcripts. The shaded area represents the 95% confidence interval of the estimated sliding window PPL on a given index. The label of 0 is defined where TALD < 3, serving as a proxy label for cognitively healthy individuals, whereas the label of 1 serves as the proxy label of FTD individuals.
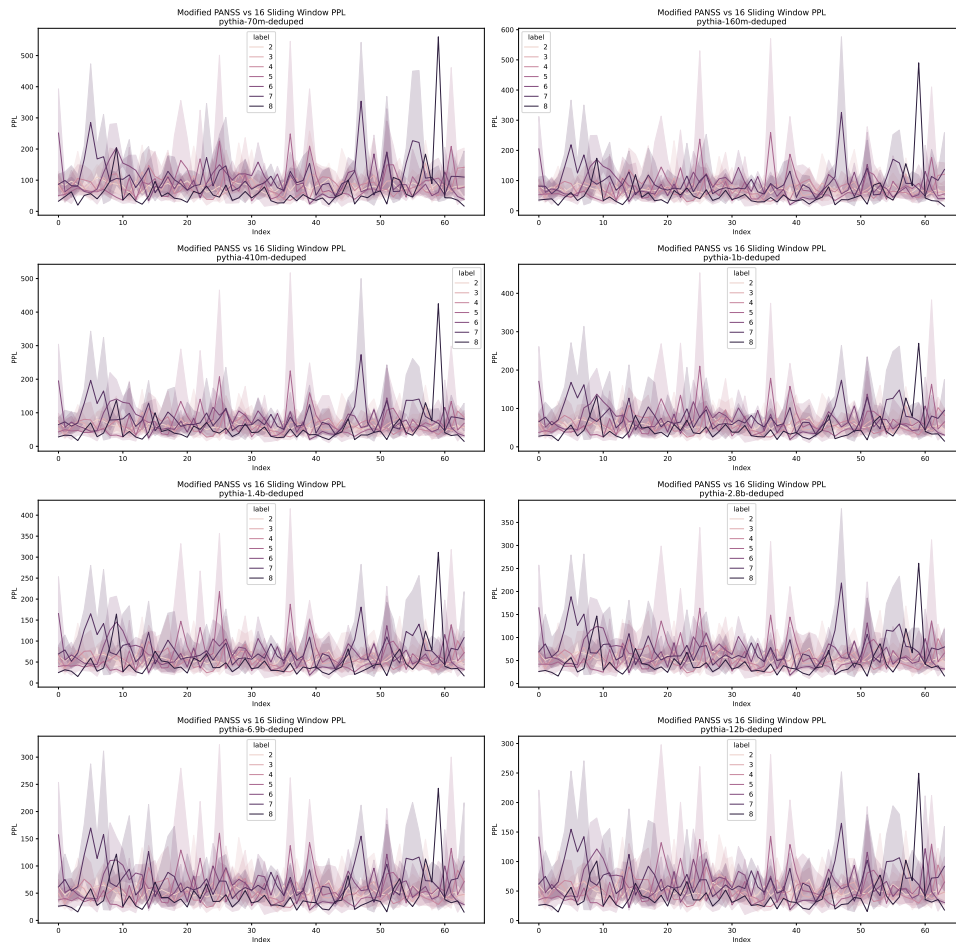
Figure A.2: The distribution of sliding window PPLs using the sliding window of 16 tokens on the clinical interview dataset, where the x-axis represents the index of sliding window PPLs in transcripts. The shaded area represents the 95% confidence interval of the estimated sliding window PPL on a given index.

Model	Sliding windows				
	8	16	32	64	128
70m	0.158***	0.156***	0.176***	**0.202***	0.180***
160m	0.177***	0.170***	0.177***	**0.206***	0.183***
410m	0.175***	0.179***	0.196***	**0.225***	0.189***
1b	0.178***	0.176***	0.205***	**0.230***	0.180***
1.4b	0.181***	0.179***	0.208***	**0.251***	0.201***
2.8b	0.168***	0.165***	0.204***	**0.240***	0.194***
6.9b	0.172***	0.169***	0.206***	**0.249***	0.194***
12b	0.175***	0.174***	0.204***	**0.245***	0.195***
LLaMA	–	–	–	0.371***	–

$^{***}p < 0.01, {}^{**}p < 0.05, {}^{*}p < 0.1$

Table A.3: The AVH dataset Spearman's ρ between the *averaged* sliding window PPL and TALD across model size. **Bold** indicates the highest ρ for a model.

Model	Sliding windows				
	8	16	32	64	128
70m	0.258	0.248	0.274*	0.276*	0.276*
160m	0.264	0.278*	0.296*	0.313*	0.294*
410m	0.263	0.276*	0.324**	0.318**	0.301*
1b	0.266	0.292*	0.318**	0.330***	0.305*
1.4b	**0.272***	**0.334***	**0.355***	**0.360***	**0.342***
2.8b	0.261	0.324**	0.344**	0.343**	0.325**
6.9b	0.269*	0.315*	0.342**	0.315*	0.310*
12b	0.270*	0.302*	0.338**	0.334**	0.326**
LLaMA	–	–	–	0.200*	–

$^{***}p < 0.01, {}^{**}p < 0.05, {}^{*}p < 0.1$

Table A.4: The clinical interview dataset Spearman's ρ between the *averaged* sliding window PPL and composite PANSS across model size. **Bold** indicates the highest ρ for a model.