Distribution Agreement

Benjamin Dixon April 9th, 2025

Automated Assessment of Concrete Language in Clinical High-Risk for Psychosis: A Novel Large Model Approach

By

Benjamin Dixon

Philip Wolff

Advisor

Department of Linguistics

Philip Wolff

Advisor

Roberto Franzosi

Committee Member

Matthew Segall

Committee Member

2025

Automated Assessment of Concrete Language in Clinical High-Risk for Psychosis: A Novel Large Model Approach

By

Benjamin Dixon

Philip Wolff

Advisor

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Arts with Honors

Department of Linguistics

2025

Abstract

Automated Assessment of Concrete Language in Clinical High-Risk for Psychosis: A Novel Large Model Approach
By: Benjamin Dixon

Language disturbances are key indicators of altered thought processes and serve as reliable markers for emerging psychotic disorders, making them a crucial target for early detection. Current diagnostic methods, relying primarily on behavioral observation and self-reporting, are limited in their ability to predict schizophrenia conversion among clinical high-risk (CHR) populations. Research has shown that individuals with schizophrenia tend to have difficulty processing abstract concepts. Examining concreteness in CHR individuals may reveal deficits in abstract language before psychosis onset. The development of automated tools using large language models offers a novel approach to quantifying these linguistic features objectively and at scale, potentially advancing our ability to detect early warning signs of psychosis. *Participants:* The study includes 225 CHR and 62 matched healthy controls (HC) in a first approach and 385 CHR and 82 HC in a second approach from the Accelerating Medicines Partnership® in Schizophrenia (AMP® SCZ) dataset. All participants underwent an open-ended interview at their baseline visit. *Analysis*: Interviewee speech is extracted. Within each sentence, content words (nouns, verbs, adjectives) are identified and sequentially occluded. For each occluded word, Llama-3 generates a “contrast set” of alternative predictions based on the preceding sentences of context and compares the concreteness of each to the occluded word. A second approach directly prompts the model for a word’s concreteness. Using both approaches, no significant differences between HC and CHR are found at baseline. However, visual examination of CHR pilot data extending past baseline reveals a bimodal distribution, indicating the possibility of a CHR subset with higher levels of concrete speech. The limitations and areas for improvement of the current method are discussed. The novel methodological approach leverages Llama-3 to provide a scalable alternative to manual concreteness ratings. By generating and comparing contextually-appropriate word alternatives, this approach captures subtle linguistic differences that may characterize early psychosis risk. Future research will explore longitudinal changes in concreteness in CHR, as a subset of individuals may exhibit heightened concrete language use that could serve as a predictive marker. Additionally, investigating linguistic concreteness within specific cognitive contexts could help elucidate the heterogeneity of specific deficits in the psychosis spectrum.

Automated Assessment of Concrete Language in Clinical High-Risk for Psychosis: A Novel Large Model Approach

By

Benjamin Dixon

Philip Wolff

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Arts with Honors

Department of Linguistics

2025

Table of Contents

INTRODUCTION

Schizophrenia is a debilitating mental disorder that leads to diminished quality of life in many domains, often resulting in functional disability and increased need for care (Singh et al., 2008; Kadakia et al., 2022). When the development of schizophrenia and other psychotic related disorders is flagged early, patients can receive earlier intervention and prognostic outcomes often significantly improve (Marshall et al., 2005; Larson et al., 2014). Schizophrenia is a disorder of the mind characterized by disturbed thought processes and altered behavior (Schultz et al., 2007). A primary psychological characteristic of schizophrenia is a cognitive deficit in abstract thinking (Pishkin & Bourne, 1981) which results in difficulty processing abstract concepts, but comparable performance to healthy controls in processing concrete concepts which pertain to physical characteristics (Pishkin & Williams, 1976). A sound intuition is that abstractness is compromised as cognitive functions decline. A consistent downstream consequence of altered thought is altered language, an oft-shown phenomenon in the schizophrenia literature (Tan et al., 2021) which is often detectable prior to the development of full-blown psychosis (Hitczenko et al., 2020). Therefore, a lack of abstract language prior to the development of full-blown psychosis may be expected. However, this deficit has never been empirically demonstrated and a literature review does not reveal a clear examination of this topic.

A new set of tools has emerged with natural language processing (NLP) which can help detect subtle disturbances in language prior to full-blown psychosis which may reveal a deficit in abstract language. Recent studies have shown that NLP methods can reliably predict psychosis and crucially provide new insights into the content of the speech produced (see Elvevåg et al., 2010; Mota et al., 2014l; Bedi et al., 2015; Rezaii et al., 2019). The present study also utilizes a novel dataset from the Accelerating Medicines Partnership® in Schizophrenia (AMP® SCZ) program which provides a substantially larger and more diverse sample than previous studies have had access to. I use a state-of-the-art large language model (LLM) and a novel algorithm to detect increased use of concrete language in individuals at clinical high risk (CHR) for psychosis.

If markers for concreteness can be reliably identified in CHR individuals, this may provide clinicians with an objective, non-invasive, and fully automated tool for risk assessment that complements existing methods. Furthermore, this approach can enhance our understanding of the cognitive mechanisms underlying psychosis development, which may lead to improvements in the classification and identification of schizophrenia subtypes.

BACKGROUND

1.1 Linguistic disturbances in schizophrenia

Schizophrenia is a debilitating mental disorder characterized by positive and negative symptoms. Positive characteristic symptoms include delusions, hallucinations, disorganized speech, and negative characteristics symptoms include catatonic behavior, affective flattening, and alogia (American Psychiatric Association, 2013; Schultz et al., 2007). Schizophrenia also manifests with linguistic abnormalities prior to full clinical conversion which have been further characterized in the literature, such as reduced semantic coherence, derailment, tangentiality, and looser associations between topics (Ehlen et al., 2023). Full-blown psychosis, particularly in schizophrenia, is frequently preceded by a period of subclinical symptoms known as the prodromal phase—a period marked by subclinical symptoms in which psychosis is not yet present but manifests in subtler ways (Larson et al., 2014). Both before and during full-blown psychosis, converging evidence suggests that language disturbances remain a stable factor, making them a crucial target for early detection in the prodromal phase (Elvevåg et al., 2010;

Bearden et al., 2011). The nature of these language disturbances is varied, but patients on the psychosis spectrum frequently display negative thought disorder, characterized by disorganized thinking and leading to poverty of speech, poverty of content of speech, and deficits in abstract thinking, which are highly predictive of schizophrenia-related psychosis (Barch & Berenbaum, 1997; Erdeljac et al., 2019; Gooding et al., 2020). Previously, the only available method for evaluation of such linguistic abnormalities was manual transcript review. However, evaluations of speech are subjective and pain-staking, especially during the prodromal phase during which symptoms are subclinical. Such subtle disturbances are only detectable by trained experts and evaluations can suffer from interrater disagreement (Grinker 2010).

1.2 The use of natural language processing tools for speech analysis

A new set of tools has emerged with natural language processing (NLP) which can eliminate the subjective and time-intensive nature of speech evaluations. Furthermore, more finely-tuned questions about speech produced during the prodromal phase of psychosis can also be answered through the development of novel algorithms. Recent studies have shown that NLP methods can reliably predict mental illnesses and crucially provide new insights into the content of the speech produced. Elvevåg et al. (2010) demonstrate that computational analysis using semantic, statistical, and surface-level language features could accurately discriminate between patients with schizophrenia, their relatives, and controls with classification accuracy of 77-90% across different comparison groups. Their findings show that semantic measures contribute most significantly to the discriminant models, suggesting that subtle semantic aspects of language can effectively detect psychological differences even between family members, thus providing a foundation for objective computational assessment of speech in psychosis. Bedi et al. (2015) utilizes speech transcripts from a small dataset of CHR participants and approximates semantic coherence by applying latent semantic analysis to speech samples and measuring the average similarity (measured via cosine similarity) between phrases in addition to measuring the frequency of determiners (i.e., ‘what’, ‘that’, ‘which’) and the maximum phrase length. Their classifier predicts psychosis with 100% accuracy on their small dataset. Mota et al. (2014) performs a graph analysis of speech, representing words as nodes and word relationships as edges, and find that patients with schizophrenia produce speech with reduced connectivity. They achieve high classification accuracy between bipolar and control subjects across five languages, supporting the notion that language disturbances are a universal marker for psychosis which can capture fundamental changes in thought processes. Rezaii et al. (2019) perform vector unpacking, a measure of semantic density which transforms sentences into component vectors of meaning and divide by the number of words per sentence. They find that low semantic density and increased content about voices and sounds is highly predictive of future psychosis.

These computational approaches provide objective assessment methods that can detect subtle linguistic markers of psychosis, potentially enabling earlier intervention during the prodromal phase when traditional evaluations may miss subclinical symptoms.

1.3 Concreteness / abstractness in psychosis

The prior studies highlight the need to closely study the semantic quality and content of prodromal speech and address the shortcomings of manual transcript review. Indeed, Hitczenko et al. (2020) emphasize that future computational work should evaluate specific linguistic abnormalities while paying special attention to their cognitive, symptomatic, and clinical correlations. As an unexplored candidate for linguistic analysis in psychosis research, the measurement of concreteness / abstractness in language offers a promising avenue for investigation. Concreteness refers to language that describes tangible, perceptible entities that can be experienced through the senses, while abstractness pertains to concepts, ideas, and qualities that cannot be directly perceived and are often accessible only through language (Brysbaert et al., 2013). This psycholinguistic dimension is relevant to psychosis as those with language processing difficulties, especially those with negative thought disorder, complete linguistic tasks on concrete concepts with greater ease than with abstract concepts (Erdeljac et al., 2019). This phenomenon is potentially explained by the greater ease of processing of concrete words as opposed to abstract words, supporting the idea that as thought becomes disrupted, abstract concepts become more difficult to process for those on the psychosis spectrum (Löhr, 2021; Solovyev, 2021). Abstract and concrete concepts possibly activate separate neural substrates, which further supports the notion that a neurodegenerative / neurodevelopmental disease such as schizophrenia might affect their usage (Solovyev, 2021).

1.4 Importance of context for concreteness / abstractness

Previous methods for assessing concreteness and abstractness have relied on large lexical databases, or dictionaries, constructed by averaging human ratings of individual words presented without context (Löhr, 2021). This dictionary-based approach is limited, as many words are polysemous—their meanings can shift significantly depending on the context in which they appear. For example, the word "table" can signify a piece of furniture with a flat top and legs or a structured arrangement of data in rows and columns. The former is highly concrete as it refers to a perceptible entity, whereas the latter is more abstract because it refers to an idea. To avoid concreteness ambiguity introduced by concreteness dictionaries (see Brysbaert et al., 2013; Scott et al., 2018 for examples), LLMs can be harnessed because of their ability to capture contextual nuance.

Recent advances in pre-trained large language models (LLMs) offer unprecedented capabilities to capture semantic nuances in text by generating representations of meaning which are sensitive to context. These models, trained on vast corpora of human language, are capable of identifying subtle linguistic patterns beyond traditional semantic analysis methods. To the best of the author's knowledge, there has been limited application of LLMs to the study of prodromal speech.

1.5 Using LLMs to make psycholinguistic judgements

To test the reliability of LLMs to make reliable psycholinguistic judgements, Trott (2024) uses GPT-4 to compare its psycholinguistic judgments with human judgements on various psycholinguistic scales. Trott's findings are particularly relevant as they demonstrate that LLMs can reliably generate psycholinguistic judgments like concreteness ratings that closely align with human evaluations. This suggests that LLMs could be efficiently deployed to analyze the semantic content of prodromal speech, potentially enabling the creation of larger, more robust datasets that would not be feasible through manual transcript review alone. Given that patients with negative thought disorder process concrete concepts with greater ease than abstract ones (Pishkin & Bourne, 1981; Erdeljac et al., 2019), Trott's validation of LLM-generated concreteness ratings (which achieved a correlation of 0.81 with human judgments, a higher correlation than the average human achieves) offers a promising methodological approach for capturing this critical psycholinguistic dimension at scale.

1.6 The present study

The present study leverages a state-of-the-art large language model to quantify linguistic concreteness in individuals at clinical high-risk (CHR) for psychosis. By examining concreteness in CHR speech, we aim to detect subtle changes and deficits in abstract thinking that may serve as early warning signs for psychosis onset. Drawing from the Accelerating Medicines Partnership® in Schizophrenia (AMP® SCZ) dataset, which provides a substantially larger and more diverse sample than previous studies, the author develops an automated approach of a novel metric that addresses the limitations of subjective and labor-intensive manual assessments. Transcripts are taken from open-ended interviews at baseline, which is the earlier point at which each participant's speech is recorded in the hope of detecting early language disturbances.

The novel methodological approach employs Llama-3 to generate contextually-appropriate word alternatives for content words in participant speech. The core idea behind this approach is that concreteness is best assessed by evaluating whether, among all the words that could plausibly be used in a given context, people tend to choose more concrete or abstract options. This can be operationalized by prompting an LLM to assess the relative concreteness of a produced word compared to a minimal "contrast set" of alternative words that could have plausibly occurred in the same context. For each occluded word, the model produces the contrast set of ten alternative predictions based on previous context, then compares the concreteness of each suggestion to the original word. This process yields a nuanced, context-sensitive measure of linguistic concreteness that captures the subtle semantic qualities that may characterize prodromal speech. To validate the effectiveness of this approach, the author directly prompts the model for each content word's concreteness as well, given the previous context sentences. If the handcrafted "contrast set" method correlates with the judgements extracted from directly prompting the model, certainty in the contrast set can be more assured from converging validity of methods.

By automating the assessment of linguistic concreteness, this study aims to provide clinicians with an objective, scalable tool for early detection of psychosis risk. Furthermore, this approach opens new avenues for investigating the heterogeneity within the CHR population by characterizing the concreteness of their speech and potentially identifying those at highest risk for conversion to psychosis. Future research will explore longitudinal changes in concreteness patterns and their relationship to clinical outcomes, potentially enhancing our ability to target interventions during the critical prodromal phase when they may be most effective.

METHOD AND MATERIALS

2.1 Dataset

I utilize the Accelerating Medicines Partnership® in Schizophrenia (AMP® SCZ) dataset (https://www.ampscz.org/) of open-ended interviews collected at baseline, which prompt participants to speak about their daily life and prompt natural conversation. These interviews are transcribed and used for subsequent language analysis. Of 532 transcripts, I analyze 62 healthy control and 225 CHR transcripts in the first approach, and 82 healthy control and 358 CHR transcripts in the second approach.

2.2 Model

I run Llama-3-70B, a competitive open-source model which can be feasibly run by clinics and laboratories with access to sufficient computing power. The model is run locally on a GPU cluster to ensure data privacy.

2.3 Interviewee selection

The interviewee remains unlabeled in transcripts. To identify the interviewee, I pass the first 2500 characters of transcript to Llama-3-70b and ask it to identify the interviewee and extract the response (Appendix A).

2.4 Transcript processing

Transcripts are locally processed with the Stanza NLP Python package. All sentences are individually extracted. All content words (nouns, verbs, adjectives) in all interviewee sentences are extracted for concreteness rating (Appendix B).

2.5 Generating concreteness ratings

In linear fashion, each sentence from the interviewee is given two sentences of prior context from the transcript, which may include interviewer or interviewee speech. This is defined as the target sentence. The concreteness of content word(s) in each sentence are rated. To maximize chances of detecting subtle linguistic differences in concreteness between groups, I utilize two methods. Furthermore, if the results of both methods are similar, it enhances the validity of the results.

The concreteness of nouns, verbs, and adjectives can appear differently. An example of a concreteness / abstract pair of concrete nouns is “rock” and “transcendence.” For verbs, it may be “jog” and “feel”. For adjectives, it may be “sharp” and “honest”.

Method 1. Every content word in each target sentence is occluded and the top 10 full-word token probabilities from Llama-3-70B are taken and filtered until there are ten content words remaining (Appendix C). A token is a word, subword, or punctuation that the LLM takes as input in order to process natural language. For example, the word “Concreteness” is tokenized into “Con”, “cre”, “ten”, “ess” with the Llama-3 tokenizer. The word “Abstract” is simply tokenized as “Abstract”. For tokens which aren’t complete words, I take the next most likely token until a complete word is formed. This is the “contrast set”. The reasoning behind this approach is explained further in the discussion. The occluded word is compared to each word in the contrast set and rated as more or less concrete by the LLM, where a 1 is more concrete and 0 is more abstract (Appendix D). The mean of all judgements determines the concreteness of the word. This is the “Contrast Set Method”.

Method 2. I also directly prompt Llama-3-70B for the concreteness of each contrast word, given the context of previous sentences (Appendix E). The same context used to create alternative words for the contrast set is used to directly prompt the model. This is the “Direct Method”.

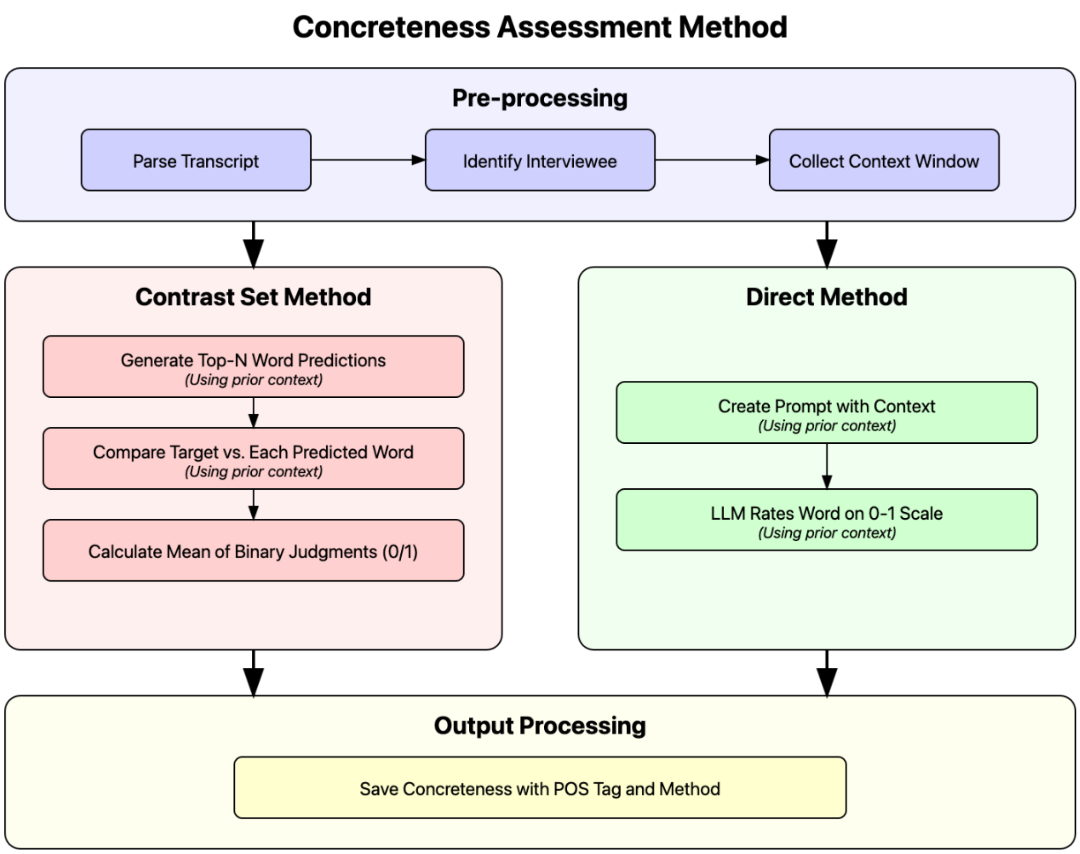

Figure 1. Concreteness assessment method flowchart, showing both the contrast set method and direct method.

2.6 Data analysis plan

The analysis will evaluate differences in language concreteness between Clinical High Risk (CHR) and Healthy Control (HC) groups using both methodological approaches: contrast set comparison and direct prompting. Two approaches will be used: the first quantifies concreteness in all content words per sentence (HC $n = 62$, CHR n = 225) and the second only quantifies concreteness in the final content word per sentence (HC $n = 82$, CHR $n = 358$).

Approach 1. I first analyze the transcripts by finding concreteness of each content word in each sentence and apply both Method 1 and Method 2 (the "Contrast Set Method" and "Direct Method") to calculate concreteness.

Approach 2. However, a possible confound with Approach 1 is that earlier words within each sentence have access to less context than later words, even given two sentences of context prior to the target sentence. It may be possible that earlier words, with a comparative lack of context, may be rated as more concrete. The contrast set produced by the LLM may pertain less to the discourse context and thus the words within the contrast set may be more abstract, and thus after comparison, the occluded word is rated as more concretely. To mitigate this concern, I analyze the transcripts a second time and only find the concreteness of the final content word of each sentence. This ensures that every concreteness rating is done with maximal context and therefore best approximates the overall concreteness of the sentence. Both methodological approaches (Method 1 and Method 2) for calculating concreteness are applied.

Investigation of Part-of-Speech. The investigation will include a stratified analysis of concreteness by grammatical category to evaluate potential differences in linguistic abstraction across parts of speech. This analysis will build upon the preprocessing described in section 2.4, where content words (nouns, verbs, and adjectives) were extracted using the Stanza NLP Python package. For each transcript, concreteness scores will be aggregated separately by part of speech: noun concreteness (typically associated with entities and objects), verb concreteness (typically associated with actions and states), and adjective concreteness (typically associated with qualities and properties). This approach will only be applied to the results of Approach 1, due to the limited variety of POS per-patient in Approach 2, as taking only the final content word significantly reduces the number of nouns, verbs, and adjectives which are analyzed for concreteness per-patient.

The following analytical procedures will be implemented:

2.6.1 Statistical Analysis Framework

Approach 1 & Approach 2. Both Approach 1 and Approach 2 will be analyzed separately using the same statistical pipeline, for each methodical approach (Method 1 and Method 2). For each method:

1. Descriptive Statistics: Means, standard deviations, and coefficients of variation will be calculated for both pooled data (all sentences combined) and patient-level aggregated data (per-patient).
2. Normality and Variance Testing: Shapiro-Wilk tests will assess normality of concreteness distributions within each group, while Levene's test will evaluate homogeneity of variance between groups.
3. Group Comparison Tests: Based on normality and variance assumptions:
   - Parametric analysis: Independent samples t-test if normality assumptions are met
   - Non-parametric analysis: Mann-Whitney U test if data violate normality assumptions
4. Within-Subject Variability: Per-patient standard deviations and coefficients of variation will be calculated to assess within-subject stability of concreteness.
5. Between-Subject Variability: Between-patient variance will be computed to examine heterogeneity within each group.

POS-Specific Concreteness Analysis. For each part of speech category mean concreteness score will be calculated for both HC and CHR groups. To investigate differences between HC and CHR, a comparison is made by applying an independent samples *t*-test.

2.6.2 Visualization Strategy

Visualization will include:

1. Histograms of patient-level mean concreteness scores for CHR and HC groups

All analyses will be conducted using Python 3.8 with pandas, scipy.stats, and matplotlib libraries. Statistical significance will be established at $p < 0.05$.

RESULTS

3.1 Approach 1

Analysis of linguistic concreteness across 36,567 content words (HC: 8,343; CHR: 28,220) revealed no significant group differences in abstractness-concreteness profiles. Both the contrast set method (HC mean = 0.546, SD = 0.063; CHR mean = 0.549, SD = 0.047) and direct prompting method (HC mean = 0.510, SD = 0.082; CHR mean = 0.519, SD = 0.055) produced nearly identical distributions between groups (t-tests: $p = 0.85$ and $p = 0.45$; Mann-Whitney U: $p = 0.85$ and $p = 0.35$). However, between-patient variability was markedly higher in HC than CHR for both methods (contrast set *SD*: HC = 0.0557 vs. CHR = 0.0346; direct method *SD*: HC = 0.0825 vs. CHR = 0.0548), suggesting greater linguistic homogeneity among CHR individuals via the contrast set method.

Notably, while the primary hypothesis of reduced abstractness in CHR was not supported, the high consistency of LLM-derived ratings across methods (mean score difference < 0.015) aligns with prior validations of LLM psycholinguistic judgments (Trott, 2024).

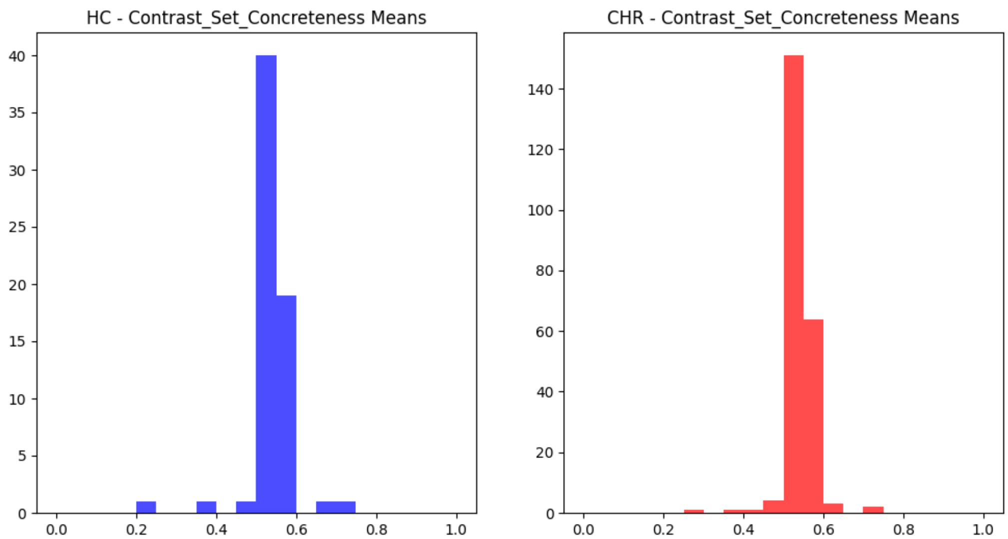

Figure 3. Histograms of concreteness means calculated via contrast set.

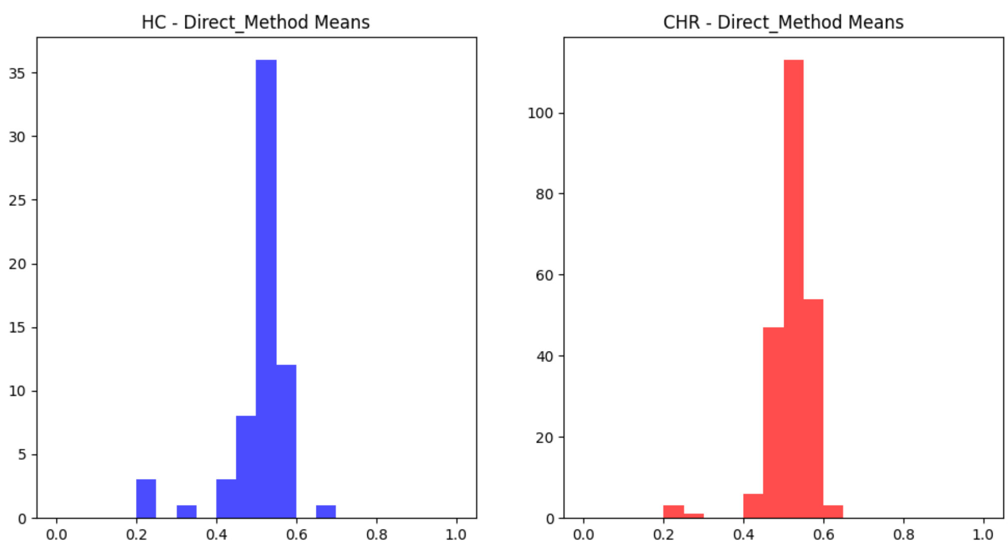

Figure 4. Histograms of concreteness means calculated via direct prompting.

3.2 Approach 2

Analysis of only final content words in each sentence also reveals no significant differences between groups. Examining 32,332 final content words (HC: 10,164; CHR: 22,168) showed highly similar concreteness patterns for both groups. Both the contrast set method (HC mean = 0.540, SD = 0.022; CHR mean = 0.537, SD = 0.022) and direct prompting method (HC mean = 0.582, SD = 0.039; CHR mean = 0.580, SD = 0.041) yielded comparable distributions (t-tests: $p = 0.36$ and $p = 0.35$; Mann-Whitney U: $p = 0.37$ and $p = 0.15$). Unlike Approach 1, between-patient variability was nearly identical across groups for the contrast set method (HC = 0.0005 vs. CHR = 0.0005) and only slightly different for the direct method (HC = 0.0015 vs. CHR = 0.0017). This reinforces the finding that sentence-final content words, which often carry key semantic information, show comparable concreteness profiles across groups.

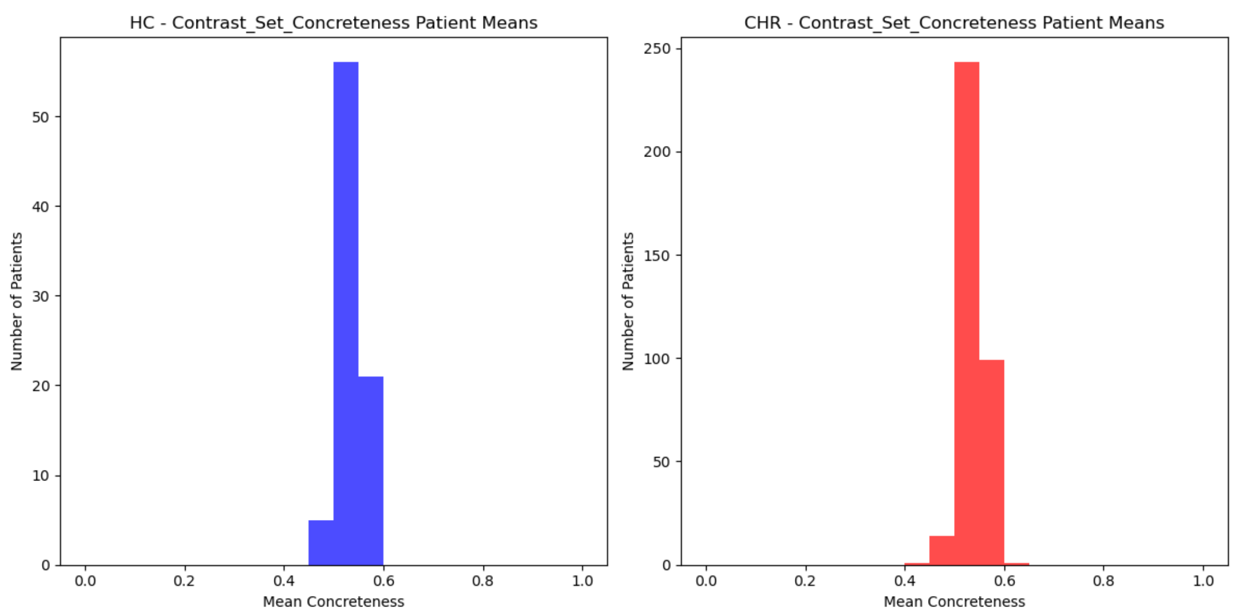

Figure 5. Histograms of concreteness means calculated via contrast set.

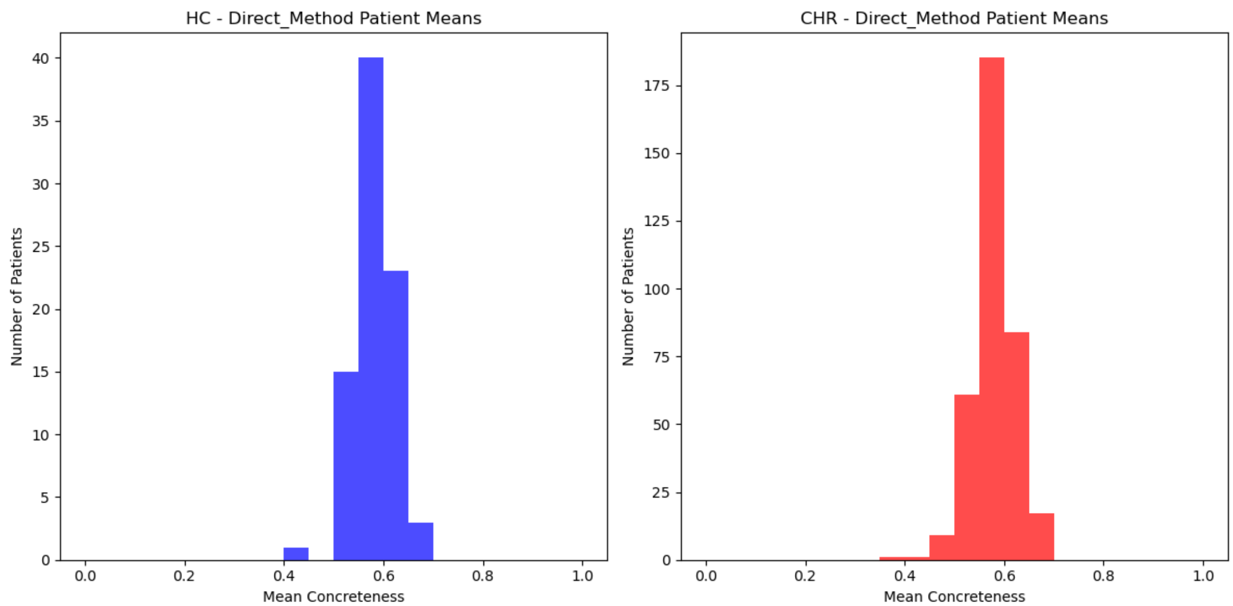

Figure 6. Histograms of concreteness means calculated via direct prompting.

3.3 Analysis by part-of-speech

I average the noun, verb, and adjective concreteness and compare by group. Examining the parts of speech from Approach 1 yielded no statistically significant results. Examining nouns reveals (HC mean: 0.5664, CHR mean: 0.5628) no significant difference, and neither does an analysis of verbs (HC mean: 0.5326, CHR mean: 0.5446) or adjectives (HC mean: 0.5280, CHR mean: 0.5378).

DISCUSSION

4.1 Interpreting absence of concreteness differences in prodromal speech.

In the current study, I analyze the concreteness of speech captured at baseline of the CHR population with two related approaches and by part-of-speech. The absence of significant differences in concrete language between CHR and HC populations at baseline challenges initial expectations about linguistic markers in early prodromal states. While previous research has established cognitive deficits in abstract thinking as a characteristic of schizophrenia (Pishkin & Bourne, 1981), my findings suggest that these deficits may not manifest in speech patterns during the earliest prodromal phases. This temporal dissociation between documented cognitive impairments and their expression in language is noteworthy, as it implies that alterations in abstract thinking may follow a more complex developmental trajectory.

Several interpretations of these null findings merit consideration. First, it's possible that concrete language emerges gradually as the disorder progresses, becoming detectable only as individuals approach conversion to psychosis. The CHR classification encompasses a heterogeneous population, with only a subset eventually developing psychosis. The baseline measurements may have captured a point in the prodromal continuum where linguistic manifestations of concrete thinking have not yet emerged, even though underlying cognitive changes may be initiating. This interpretation aligns with staged models of psychosis development that suggest symptom domains evolve at different rates (Fusar-Poli et al., 2013).

Second, the relationship between cognitive deficits and language production may be mediated by compensatory mechanisms in early stages. Individuals at clinical high risk might employ compensatory linguistic strategies that mask underlying difficulties with abstract conceptualization, particularly in structured interview settings where there is social pressure to communicate effectively.

4.2 Symptom progression may influence concreteness

It is important to note that baseline represents the earliest point at which speech samples are taken from the dataset. As the disorder progresses, subjects may ultimately convert to full-blown psychosis (CHR+), or remit, and see their symptoms attenuate (CHR-). Therefore, we may expect to see any subclinical symptoms (cognitive, linguistic, or otherwise) to become either more or less prevalent, depending on CHR subgroup. There exists the possibility that differences in concreteness may become prevalent as subjects' symptoms become more attenuated or lessened (CHR+ or CHR-). In this case, we would expect to see a bimodal distribution in concreteness where the majority of CHR subjects would have concreteness scores matching those of the HC subjects, in addition to a smaller number of subjects in the CHR population with higher concreteness scores.

4.3 Limitations and possible improvements

The identification of biomarkers, or indicators in the body and brain of the progression of a disease, is an empirical process characterized by extensive trial-and-error. Therefore, a lack of positive results does not immediately signal that a method should be abandoned when improvements can still be made. In a previous pilot-version of the study which included transcripts taken past baseline, a bimodal distribution using the contrast set method was observed, indicating that the current operationalization of concreteness may contain validity but need improvement to successfully distinguish CHR and HC at baseline (see Figure 7). Further work can be done to address limitations and improve the current concreteness analysis pipeline and algorithm in order to enhance detection of concreteness differences at baseline or examine longitudinal changes in concreteness that may emerge as symptoms progress.

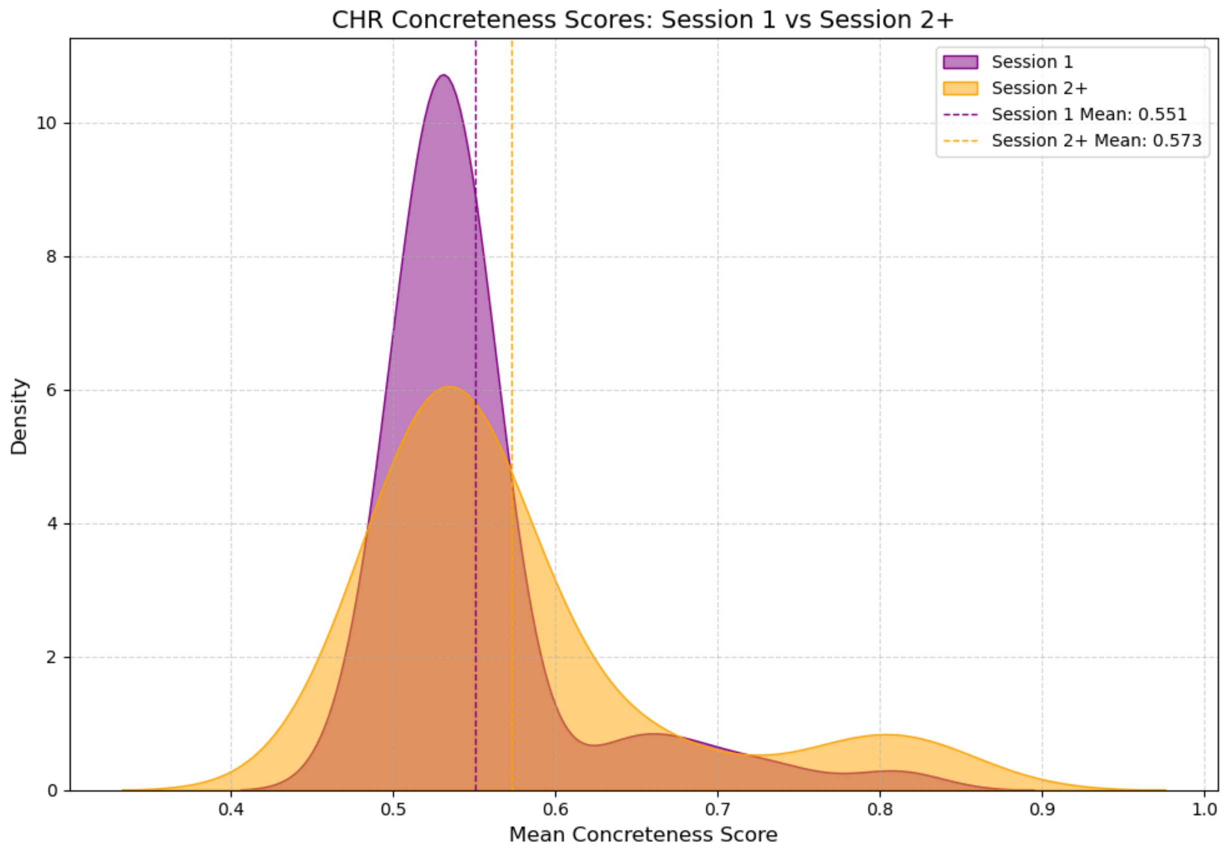

Figure 7. Density distribution chart of concreteness values in CHR population compared between session 1 (baseline) and session 2+.

4.3.1 Employment of compensatory strategies by CHR. Employment of compensatory strategies by CHR subjects may explain the lack of baseline differences. CHR individuals might still have access to abstract language but require more cognitive resources to access it. At baseline, when symptoms are less severe, they may successfully compensate by taking more time to formulate responses or by relying on learned linguistic patterns. However, as the disorder progresses, particularly for those who convert to psychosis, these compensatory mechanisms may become less effective or overwhelmed by increasing cognitive disorganization. This would explain why concreteness differences might only emerge in later sessions or in the CHR+

subgroup, consistent with the bimodal distribution observed in the pilot study that included later sessions. Future work could record the time taken to produce each content word and analyze differences based on the concreteness / abstractness of each.

4.3.2 Insufficient definition of concreteness. Löhr (2021) problematizes the current characterization of concreteness and abstractness. Basing concreteness on the ability to directly sense (through touch, vision, or otherwise) an object is insufficient: consider ATOM, which is too small to perceive yet rated as concrete, or RED, which is perceptually accessible yet not an object. Löhr proposes that concrete concepts are those which apply to "events, actions, properties, relations, or objects whose diagnostic features are perceptually, motorically, or introspectively directly accessible" (p. 559) and for which these features alone can lead to "possession of the concept" (p. 559), and an abstract concept is a concept for which these features are "not sufficient for the possession of the concept". This definition can help reduce theoretical ambiguity about concreteness / abstractness but could also have direct practical benefits for allowing LLMs to make better concreteness judgements about words if placed into the prompt. Improved definitional clarity could help edge cases in which human intuition might differ from the machine's judgement without such instruction.

4.3.3 Contrast set generation. The contrast set method does not prompt the model for possible words in the contrast set. Instead, I take the highest log probabilities of possible tokens and filter for content words until a set of ten exists in order to maximize the naturalistic generalizability of the contrast set. Due to token generation in large language models, not all tokens are complete words. To fix this issue, I take the next-highest token prediction until a complete word is formed. Directly accessing probabilities has the benefit of better assessing the linguistic domain knowledge of the LLM, which has been previously demonstrated to out-perform prompting performance. (Hu & Levy, 2023). However, the contrast set formed by this method may not have close-enough semantic relationships to the target word in order to form meaningful comparisons. A lack of semantic closeness may result in somewhat arbitrary concreteness comparisons that even humans may find difficult to accurately respond to. Consider comparing the concreteness of the bolded word in the following sentences: "The fox jumped over the **fence**" vs. "The fox jumped over the **large** [rest of sentence]". Both bolded words are likely predictions that the model may internally hold. However, the comparison between "fence" and "large" is ambiguous because of the difference in part of speech. Fence is itself an entity, whereas "large" describes a physical characteristic of an entity. To address this limitation, returning to prompt-based methods may be beneficial.

4.4 Possible manifestations of reduced abstract speech in other linguistic domains

While this study focused primarily on direct measures of concreteness, the manifestation of concrete thinking might be better observed in more complex linguistic domains such as creative reasoning, metaphor comprehension, and contextual appropriateness. CHR individuals may maintain superficially normal concrete/abstract word usage while showing subtle deficits in how these words are deployed in creative or ambiguous contexts. This perspective shifts our focus from merely quantifying concreteness to examining how concrete language interacts with broader cognitive processes like semantic association and creative thinking. As explored in the following sections, the relationship between concreteness and looseness of associations may provide a more nuanced understanding of language changes in the prodromal state, particularly in distinguishing between pathological and creative forms of unusual language production.

4.4.1 Concreteness in creative reasoning. Despite the evidence suggesting a deficit in abstract thinking, laypeople may intuitively hold the belief that schizophrenia actually causes more abstract speech. This is possibly due to "looseness of association", one of the first noted and most primary psychological characteristics of schizophrenia, or the tendency to link distanced semantic concepts closer together (Bleuler, 1911/1950 which results from increased spreading of activation in semantic networks (Mohr et al., 2001). To schizophrenic patients, words with greater semantic distance between them may seem as if they are closer together. In schizotypal speech, this distinction functionally results in speech which may seem "odd" or "unusual", and may even appear to be creative. Mohr et al. (2001) propose that such loosening of associations, even in healthy controls, may be a mechanism of enhanced creativity. Given the similarities between schizophrenic and creative styles of reasoning, it may be productive to find delineating factors between the two.

A plausible candidate for delineation between the two may be the psycholinguistic metric of concreteness. Schizophrenia is associated with deficits in the lexico-semantic system (Erdeljac et al., 2019), which may implicate reduced abstraction as activation spreads in the semantic network. He et al. (2024) further characterizes the loosening of associations theory in patients with schizophrenia and patients at clinical risk and find a tightening between lexical-conceptual associations that occurs simultaneously with a widening of the contextual connections deemed possible between words, challenging the traditional view of semantic deficits in psychosis.

Löhr (2021) notes that abstract words and expressions often have more than one possible interpretation, contributing to the greater difficulty of processing abstract words. The reverse is also true; concrete words oftentimes have a more limited set of interpretations. If the results of

He et al. (2024) generalize to the full CHR population, a tightening lexical-conceptual association may imply a tendency to select concrete words co-occurring with a loosening of associations, a feature of creativity (Mohr et al., 2001). That is, a delineating factor between schizophrenic and creative styles of reasoning may lie in the psycholinguistic characteristics of the words they choose, namely, concreteness. Therefore, detecting deficits in abstract language may be more context-dependent and may first require judgements to be made about the creative intent behind the language production and the appropriateness of concrete word choices in a given communicative context.

4.4.2 Concreteness in metaphor. The “creative” observation in schizophrenia extends to metaphor. Gutierrez et al. (2017) builds on the observation that patients with schizophrenia displays distinctive patterns in language use, particularly in metaphor production, and find that schizophrenia patients use significantly more metaphorical language (6.3%) than healthy controls (5.2%). Based on their findings, they develop a classifier to identify CHR+ individuals with 97.1% accuracy on a small dataset.

This increased metaphor production may seem paradoxical when considered alongside the tendency toward concrete thinking in schizophrenia. However, the cognitive mechanism described by He et al. (2024)—tightened lexical-conceptual associations occurring simultaneously with widened contextual connections—offers a potential explanation. In typical metaphor processing, one maps concrete source domains onto abstract target domains to enhance understanding (Lakoff & Johnson, 1980). The tightened lexical-conceptual associations in schizophrenia might constrain individuals to more concrete source domains while the widened contextual connections allow them to form associations between seemingly unrelated domains.

This cognitive pattern could result in metaphors that are highly concrete in their components but unusual in their combinations, creating speech that appears both concrete and metaphorical simultaneously. An instance of such a paradoxical metaphor was recorded by Andreasen (1986) in patients with thought disorder, who named watches “time vessels” and gloves “hand shoes”. The expressions appear to be odd, creative, and certainly metaphorical, but simultaneously concrete.

While conventional metaphors balance concreteness and abstraction in ways that facilitate shared understanding, the metaphors produced in schizophrenia might fail to bridge this concrete-abstract divide effectively, despite their increased frequency. Future work could further characterize metaphor produced in schizophrenia in regards to concreteness and other psycholinguistic dimensions.

CONCLUSION

In the current study, I introduce the possibility of concreteness / abstractness being a predictive measure for psychosis and utilize a large language model for operationalization, which handles the issue of concreteness ambiguity by understanding context. The application of large language models to psycholinguistic assessment demonstrates promising reliability across measurement approaches, suggesting these tools may be viable for clinical applications when further refined. I examine if prodromal speech collected during open-ended interviews at baseline displays differences from the healthy controls and find no difference, but discuss the possibility of such detection being possible with further improvement. Future research should explore longitudinal changes in concreteness patterns, particularly in CHR individuals who convert to psychosis versus those who remit, as well as investigate how concreteness interacts with other linguistic domains such as metaphor usage and semantic associations. By continuing to develop sophisticated computational approaches to language analysis, we may yet uncover the subtle linguistic markers that signal psychosis risk, potentially enabling earlier and more targeted interventions during this critical window of opportunity.

REFERENCES

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

Andreasen N. C. (1986). Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia bulletin*, *12*(3), 473–482. https://doi.org/10.1093/schbul/12.3.473

Barch, D. M., & Berenbaum, H. (1997). The effect of language production manipulations on negative thought disorder and discourse coherence disturbances in Schizophrenia. *Psychiatry Research*, *71*(2), 115–127. https://doi.org/10.1016/s0165-1781(97)00045-0

Bearden, C. E., Wu, K. N., Caplan, R., & Cannon, T. D. (2011). Thought disorder and communication deviance as predictors of outcome in youth at Clinical High Risk for psychosis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(7), 669–680. https://doi.org/10.1016/j.jaac.2011.03.021

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated Analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia*, *1*(1). https://doi.org/10.1038/npjschz.2015.30

Bleuler, E. (1911). *Dementia praecox; or, the group of Schizophrenias*. International Universities Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Ehlen, F., Montag, C., Leopold, K., & Heinz, A. (2023). Linguistic findings in persons with schizophrenia—a review of the current literature. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1287706

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics*, *23*(3), 270–284. https://doi.org/10.1016/j.jneuroling.2009.05.002

Erdeljac, V., Sekulić Sović, M., Ostojić, D. & Savić, A. (2019). The role of the psycholinguistic characteristics of words in the assessment of language processing of patients with psychosis, ultrahigh risk of psychosis, or schizophrenia. In V. Erdeljac, (Ed.), M. Sekulić Sović, (Ed.), *Interdisciplinary Linguistic and Psychiatric Research on Language Disorders* (pp. 1-24). Zagreb: Filozofski fakultet Sveučilišta u Zagrebu - FF Press. doi: 10.17234/9789531758314.02

Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., Keshavan, M., Wood, S., Ruhrmann, S., Seidman, L. J., Valmaggia, L., Cannon, T., Velthorst, E., De Haan, L., Cornblatt, B., Bonoldi, I., Birchwood, M., McGlashan, T., Carpenter, W., … Yung, A. (2013). The psychosis high-risk state: a comprehensive state-of-the-art review. JAMA Psychiatry, 70(1), 107. https://doi.org/10.1001/jamapsychiatry.2013.269

Gooding, D. C., Ott, S. L., Roberts, S. A., & Erlenmeyer-Kimling, L. (2012). Thought disorder in mid-childhood as a predictor of adulthood diagnostic outcome: Findings from the New York high-risk project. *Psychological Medicine*, *43*(5), 1003–1012. https://doi.org/10.1017/s0033291712001791

Grinker, R. R. (2010). In retrospect: The five lives of the psychiatry manual. *Nature*, *468*(7321), 168–169. https://doi.org/10.1038/468168a

Gutierrez, E. D., Cecchi, G., Corcoran, C., & Corlett, P. (2017). Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2923–2930. https://doi.org/10.18653/v1/d17-1316

He, R., Palominos, C., Zhang, H., Alonso-Sánchez, M. F., Palaniyappan, L., & Hinzen, W. (2024). Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Research*, *333*, 115752. https://doi.org/10.1016/j.psychres.2024.115752

Hitczenko, K., Mittal, V. A., & Goldrick, M. (2020). Understanding language abnormalities and associated clinical markers in psychosis: The promise of computational methods. *Schizophrenia Bulletin*, *47*(2), 344–362. https://doi.org/10.1093/schbul/sbaa141

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.306

Kadakia, A., Catillon, M., Fan, Q., Williams, G. R., Marden, J. R., Anderson, A., Kirson, N., & Dembek, C. (2022). The economic burden of schizophrenia in the United States. *The Journal of Clinical Psychiatry*, *83*(6). https://doi.org/10.4088/jcp.22m14458

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science*, *4*(2), 195–208. https://doi.org/10.1207/s15516709cog0402_4

Larson, M. K., Walker, E. F., & Compton, M. T. (2014). Early signs, diagnosis and therapeutics of the prodromal phase of schizophrenia and related psychotic disorders. *Expert Review of Neurotherapeutics*, *10*(8), 1347–1359. https://doi.org/10.1586/ern.10.93

Löhr, G. (2021). What are abstract concepts? on lexical ambiguity and concreteness ratings. *Review of Philosophy and Psychology*, *13*(3), 549–566. https://doi.org/10.1007/s13164-021-00542-9

Marshall, M., Lewis, S., Lockwood, A., Drake, R., Jones, P., & Croudace, T. (2005). Association between duration of untreated psychosis and outcome in cohorts of first-episode patients. *Archives of General Psychiatry*, *62*(9), 975. https://doi.org/10.1001/archpsyc.62.9.975

Mohr, C., Graves, R. E., Gianotti, L. R., Pizzagalli, D., & Brugger, P. (2001). *Journal of Psycholinguistic Research*, *30*(5), 475–483. https://doi.org/10.1023/a:1010461429079

Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., & Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Scientific Reports*, *4*(1). https://doi.org/10.1038/srep03691

Pishkin, V., & Bourne, L. E. (1981). Abstraction and the use of available information by schizophrenic and normal individuals. *Journal of Abnormal Psychology*, *90*(3), 197–203. https://doi.org/10.1037/0021-843x.90.3.197

Pishkin, V., & Williams, W. V. (1976). Schizophrenic hypothesis behavior in concept identification. *Bulletin of the Psychonomic Society*, *8*(5), 385–388. https://doi.org/10.3758/bf03335175

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *Npj Schizophrenia*, *5*(1). https://doi.org/10.1038/s41537-019-0077-9

Schultz, S. H., North, S. W., & Shields, C. G. (2007). Schizophrenia: a review. *American family physician*, *75*(12), 1821–1829

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2018). The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Singh, P., Solanki, R., Midha, A., & Chugh, K. (2008). Schizophrenia: Impact on quality of life. *Indian Journal of Psychiatry*, *50*(3), 181–186. https://doi.org/10.4103/0019-5545.43632

Solovyev, V. (2021). Concreteness/abstractness concept: State of the art. *Advances in Intelligent Systems and Computing*, 275–283. https://doi.org/10.1007/978-3-030-71637-0_33

Tan, E. J., Meyer, D., Neill, E., & Rossell, S. L. (2021). Investigating the diagnostic utility of speech patterns in schizophrenia and their symptom associations. *Schizophrenia Research*, *238*, 91–98. https://doi.org/10.1016/j.schres.2021.10.003

Trott, S. (2024). Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, *56*(6), 6082–6100. https://doi.org/10.3758/s13428-024-02337-z

APPENDIX

Appendix A. *The prompt which performs interviewee selection based on a small portion of the transcript.*

```
{
        "role": "system",
        "content": "Only ever output S1, S2, or S3. Do not say anything else. Just say 'S1', 'S2', or 'S3', and finish."
    },
    {
        "role": "user",
        "content": f"Carefully read the following transcript of an interviewee and interviewer. Your job is to identify the interviewee and output whether they are S1, S2, or S3. Only EVER output one of these three strings, never anything else. Here is the transcript: \n{transcript}."
}
```

Appendix B. *The Python function, using Stanza NLP, filters out all verbs, nouns, and adjectives in a sentence. Further functions in the pipeline reconstruct sentences up until each occluded content word and add its preceding context.*

```
def occlude_content_words(input_string):
    # Process input string with Stanza
    doc = nlp(input_string)
    substrings = []

    for sentence in doc.sentences:
        for word in sentence.words:
            if word.upos in ('VERB', 'NOUN', 'ADJ'):
                substring = (input_string[:word.start_char].rstrip(), " " + word.upos)
                substrings.append(substring)

    return substrings
```

Appendix C. *The Python function is responsible for finding the internal token predictions of the model.*

```
def top_n_tokens(self, sentence: str, occluded_word: str, N: int):
        """
        Each call is passed (substring, N) where each substring should be of type str
        """

        # Tokenize substring
        input_ids = self.tokenizer.encode(sentence, return_tensors='pt').to(self.model.device)

        # Model outputs
        with torch.no_grad():
            outputs_model = self.model(input_ids)

        # Logits for next token
        logits = outputs_model.logits  # Shape: [batch_size, seq_len, vocab_size]
        next_token_logits = logits[:, -1, :]  # Shape: [batch_size, vocab_size]

        # Get top K token IDs
        k = 10 * N  # Generates enough to have :N content words after filtering
        topk = torch.topk(next_token_logits, k=k)
        top_token_ids = topk.indices.squeeze(0).tolist()

        # Decode token IDs to words
        words = []
        for token_id in top_token_ids:
```

```
        # Decode the token ID to get the word
        word = self.tokenizer.decode([token_id]).strip()
        # Skip empty strings and special tokens
        # Special tokens: ['<|begin_of_text|>', '<|eot_id|>']
        if not word or word in self.tokenizer.all_special_tokens or word == occluded_word.strip():
            continue
        words.append(word)

    # POS tagging to filter for content words (verbs, nouns, adjectives)
    content_words = filter_content_words(words)
    content_words = content_words[:N]

    return content_words
```

Appendix D. *The model prompt responsible for the Likert-style ratings between the occluded word and its contrast set. The prompt also adapts the concreteness item from the Glasgow norms (Scott et al., 2018).*

```
{
        "role": "system",
        "content": "Your job is to compare the concreteness of words. Concreteness is a measure of how concrete or abstract something is. A word is CONCRETE if it represents something that exists in a definite physical form in the real world. In contrast, a word is ABSTRACT if it represents more of a concept or idea."
    },
    {
        "role": "user",
        "content":
            f"In the following sentence: '{sentence_context}' + [WORD], compare the concreteness of the occluded word '{occluded_word}' with each word in the following list:\n{indexed_word_list}\nFor each word, output a 1 if the occluded word is more concrete than the word, or a 0 if it is more abstract. Format your answers as follows: 1. 0 2. 1 3. 1 4. 0, and so on. Never output anything else, no matter what."
}
```

Appendix E. *The model prompt used in the "direct" method for assessing concreteness. The prompt is reworded from the concreteness item on the Glasgow norms psycholinguistic rating scale (Scott et al., 2018).*

```
{
        "role": "system",
        "content": "Your job is to compare the concreteness of words. Concreteness is a measure of how concrete or abstract something is. A word is CONCRETE if it represents something that exists in a definite physical form in the real world. In contrast, a word is ABSTRACT if it represents more of a concept or idea. Rate the concreteness on a scale from 0 to 1, where 0 is completely abstract and 1 is completely concrete."
    },
    {
        "role": "user",
        "content": f"{context_text}Please rate the concreteness of the word '{word}' as it appears in the sentence: '{sentence_context}'. Give only a number between 0 and 1, where 0 is completely abstract and 1 is completely concrete. Just respond with the number and nothing else."
}
```