Reverse Prompting: A Novel Computational Paradigm in Schizophrenia Based on Large Language Models

Ivan Nenchev [1,2], **Christiane Montag** [1], **Sandra Anna Just** [1,3]

[1] Department of Psychiatry and Psychotherapy, Charité Campus Mitte,
Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin,
Humboldt-Universität zu Berlin, and Berlin Institute of Health,
[2] Berlin Institute of Health at Charité – Universitätsmedizin Berlin,
[3] Department of Clinical Medicine, UiT – The Arctic University of Norway,
Tromsø, Norway

ivan.nenchev@charite.de

Abstract

Large language models (LLMs) are increasingly being used to interpret and generate human language, yet their ability to process clinical language remains underexplored. This study examined whether three open-source LLMs can infer interviewer questions from participant responses in a semi-structured psychiatric interview (NET) conducted with individuals diagnosed with schizophrenia (n = 107) and neurotypical controls (n = 66). Using cosine similarity between LLM-generated questions and original questions as a proxy for the precision of the inference, we found that responses from individuals with schizophrenia produced significantly lower similarity scores ($\beta = –0.165$, $p < .001$). Cosine similarity decreased across the nested structure of the interview, with smaller reductions observed in the schizophrenia group. Although all emotions decreased similarity with fear, only sadness showed a significant interaction with diagnosis, suggesting differential processing of emotional discourse. Model type and generation temperature also influenced outcomes, highlighting variability in model performance. Our findings demonstrate that LLMs systematically struggle to reconstruct interviewer intent from responses by individuals with schizophrenia, reflecting known discourse-level disturbances in the disorder.

1 Introduction

Schizophrenia is a serious mental disorder that affects approximately 1% of the global population (Hasan et al., 2020; Finnerty et al., 2024). Although about 20% of the people with schizophrenia experience only one psychotic episode (Alvarez-Jimenez et al., 2011), most follow a recurrent or chronic course. The disorder presents with a broad spectrum of symptoms that impact cognition, perception, affect, and motor functions. Moreover, it also involves disturbances in language, both in comprehension (e.g., concretism Bambini et al., 2020) and production (e.g., incoherence, derailment, tangentiality, word approximations, neologisms, stilted speech, poverty of speech, (Andreasen, 1986)). These linguistic symptoms correlate with symptom severity and are integral to clinical assessment and diagnostic processes, as they provide observable markers of the underlying thought disorder. Interestingly, linguistic symptoms have been conceptualized not only in spontaneous speech (e.g., incoherence) but also at the discourse level across multiple turns of question-and-answer exchanges in clinical interviews. Tangentiality refers to responses that are distant, oblique, or seemingly irrelevant to the question asked.

Currently, symptom severity in schizophrenia is typically assessed using standardized rating scales such as the Thought, Language, and Communication scale (TLC, Andreasen, 1986), the Positive and Negative Syndrome Scale (PANSS, Kay et al., 1987), the Scale for the Assessment of Positive Symptoms (SAPS), and the Scale for the Assessment of Negative Symptoms (SANS, Andreasen, 1989). While these instruments are widely used in both research and clinical practice, their use presents several limitations: clinicians and researchers require extensive training to apply them reliably, the assessments are time-consuming, and

Proceedings of Recent Advances in Natural Language Processing,pages 797–806
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-092

they can impose a significant burden on patients. Recent advances in natural language processing (NLP) offer an alternative approach by enabling objective and replicable analysis of linguistic output in individuals with schizophrenia. Several authors have highlighted the potential of NLP-derived linguistic features as candidate biomarkers of psychosis (Corcoran et al., 2020; de Boer et al., 2020), or even as biosocial markers that integrate clinical, cognitive, and social dimensions of the disorder (Palaniyappan, 2021).

In this study, we propose a novel approach to analyze question-answer turns in schizophrenia using LLMs, which currently represent the state of the art in NLP. Unlike traditional NLP methods that rely on predefined linguistic features or hand-crafted rules, LLMs are capable of capturing complex semantic relationships and contextual dependencies in language. By leveraging their capacity to model coherence and relevance across multiple conversational turns, we aim to operationalize a novel linguistic feature in a manner that is both scalable and sensitive to the subtle discourse patterns observed in clinical interactions.

2 Background

2.1 Schizophrenia and NLP

Several prominent NLP approaches currently assess linguistic abnormalities in schizophrenia, targeting vocabulary, syntax, and semantics. Semantic coherence has received the most attention, with generally reduced coherence in patients, though findings vary by methodological choices such as segmentation, embedding type, and illness phase (Corcoran et al., 2020; Parola et al., 2023; Alonso-Sánchez et al., 2022). Lexical diversity findings are mixed, ranging from reduced to increased variability across studies (Voleti et al., 2023; Lundin et al., 2023). Neologisms are rarely explored in NLP-based studies, with only one semi-automated analysis reported (Just et al., 2020). Altered pronoun use—especially increased first-person singular—has emerged as a consistent marker of disrupted self-referential processing (Ziv et al., 2021; Watson et al., 2012; Elleuch et al., 2025). Syntactic complexity appears reduced, particularly in early or high-risk populations, though its predictive value remains unclear (Schneider et al., 2023; Bedi et al., 2015).

In recent years, several approaches have been proposed to quantify semantic divergence between a question and the sentences comprising the response. Elvevåg et al. (2007) introduced a method based on cosine similarity, showing that the slope of a linear regression between sentence position and similarity of consecutive sentences significantly correlated with human ratings of tangentiality in individuals with schizophrenia. Extending this approach, Tang et al. (2021) found that this slope was significantly steeper in individuals with schizophrenia than in neurotypical controls.

Transformer-based language models are currently used primarily to extract contextualized embeddings and assess semantic relatedness between text segments (Tang et al., 2021; Li et al., 2024; Jeong et al., 2023). The application of autoregressive models remains limited. Lawrence et al. (2024) outline three potential domains for large language model (LLM) use in mental health: education, assessment, and intervention. Within assessment, LLMs are proposed to aid in diagnosis and symptom evaluation, including suicide risk and disorganized speech. For example, Pugh et al. (2024) used LLMs to predict clinical ratings from linguistic samples but reported inconsistent results. It is worth noting that such application of LLMs entail unresolved ethical and legal challenges that constrain their clinical integration. A more tractable and ethically acceptable use case may lie in generating synthetic language samples with LLMs, followed by explainable NLP-based analysis. In a notable example, Fradkin et al. (2023) used GPT-2 to simulate language patterns resembling those of individuals with schizophrenia by manipulating generation parameters such as temperature (lexical entropy) and memory span. They showed that higher temperatures or shorter memory spans increased semantic drift between sentences, mimicking aspects of formal thought disorder.

2.2 Model Inversion and Reverse Prompt Engineering

In recent years, several studies have attempted to reconstruct prompts from LLM outputs. Morris et al. (2023) conceptualize this task as language model inversion and show that prompts can be recovered using the next-token probabilities produced by an LLM. Zhang et al. (2024) train a T5-base model with 222 million parameters in an encoder-decoder framework, using a sparse encoder architecture to invert LLM outputs into the prompts that elicited them. Petrov et al. (2024) exploit the low-rank structure of gradients in the self-attention layers and the discrete nature of token embeddings to efficiently verify whether a given token sequence is part of the client data, demonstrating the reconstruction of entire input batches—raising significant privacy concerns. Sha and Zhang (2024) proposed a prompt stealing attack framework that reconstructs original prompts from language model outputs using two components: a parameter extractor and a prompt reconstructor. The parameter extractor first classifies the original prompt into one of three types: direct prompt, role-based prompt, or in-context prompt. Then, the prompt reconstructor uses a language model to infer the prompt, typically by posing a meta-question such as "What question are you asked if you can generate the following answer?"—this works straightforwardly for direct prompts, while role-based and in-context prompts require additional assembly logic. The quality of the reconstructed prompt is evaluated using cosine similarity with the original.

In contrast to this, Reverse Prompt Engineering (RPE) refers to the task of inferring the original prompt or question based on one or more outputs generated by a language model. Li and Klabjan (2025) proposed three strategies to address this task. The One Answer One Shot approach presents the model with a single output and asks it to infer the corresponding prompt; however, this often results in overfitting to specific details of the answer. In contrast, the Five Answers One Shot method improves robustness by extracting a single prompt based on five different outputs. Finally, the Iterative Approach has the model generate several candidate prompts and refine its selection through comparison, gradually converging on the best match.

2.3 Hypothesis

In this paper, we present a novel adaptation of the reverse prompt engineering technique for application in psychiatric research. Traditionally, reverse prompt engineering involves reconstructing an original prompt based solely on the output generated LLM (Li and Klabjan, 2025). In clinical settings, a trained clinician typically poses a question to a patient. We extend the concept of reverse prompt engineering by using LLMs to infer the original interviewer's question from a participant's response in a semi-structured psychiatric interview. Rather than focusing on prompt reconstruction in a generic NLP context, our approach leverages the LLM's reasoning and comprehension abilities to emulate how an AI listener might interpret a response and infer the most probable preceding question.

We hypothesize that responses from individuals with schizophrenia will lead to reconstructed questions that are semantically more divergent from the original interviewer prompts compared to those generated from responses by neurotypical controls. This divergence may reflect key linguistic disruptions such as tangentiality, derailment, or poverty of content of speech. By quantifying this divergence using established semantic similarity measures, our approach offers a novel and scalable method for assessing disorganized speech patterns in clinical populations.

3 Materials and methods

3.1 Reverse prompting

To implement our modified reverse prompt engineering procedure, we used participants' responses from the Narrative of Emotions Task (NET), a semi-structured clinical interview, as input to three open-source large language models (LLMs) hosted on Hugging Face: LLaMA 3 8B Instruct[1] (AI@Meta, 2024), Mistral 7B Instruct-v0.2[2], and Cohere's Aya 8B[3] (Aryabumi et al., 2024). The goal was to evaluate whether LLMs could accurately infer the original interview question solely based on a participant's answer (see Figure 1). Each model was prompted to generate a plausible interview question for a given response. While the core task was consistent across models, each LLM required slightly different prompt formulations to elicit the desired behavior. For reasons of data protection and compliance with ethical guidelines, we deliberately excluded proprietary models and ensured that all processing occurred locally on an NVIDIA A100 GPU at our institution's HPC.

To assess the stability and diversity of generated outputs, we sampled model responses across a range of temperature values (0.001, 0.2, 0.4, 0.6, 0.8, 1.0). This allowed us to examine how sensitive the reconstruction of questions was to stochastic variation during generation. Across all models and

[1] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[3] https://huggingface.co/CohereLabs/aya-23-8B

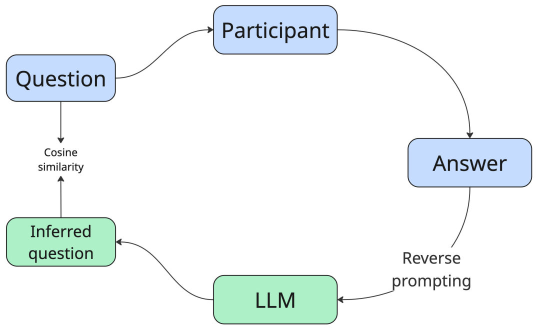

Figure 1: Modification of RPE

temperature settings, we extracted a total of 36,162 reconstructed questions. Both Mistral and LLaMA tended to generate multiple possible inferred questions in response to a single input. For consistency and comparability across models, only the first generated question was selected for further analysis. To evaluate the semantic similarity between the LLM-generated questions and the original NET prompts, we employed Sentence Transformers[4] (Reimers and Gurevych, 2019) to compute cosine similarity between the sentence embeddings of original and reconstructed questions. This similarity metric served as a proxy for measuring the fidelity of question reconstruction.

The following illustration is adapted from a clinical transcript published by Nancy Andreasen in her work on the TLC Scale (Andreasen, 1986). In this example, the participant's response is as follows: "Well, that's a hard question to answer because my parents… I was born in Iowa, but I know that I'm white instead of black so apparently I came from the North somewhere and I don't know—where, you know, I really don't know where my ancestors came from. So I don't know whether I'm Irish or French or Scandinavian or I don't, I don't believe I'm Polish but I think I might be German or Welsh. I'm not but that's all speculation and that, that's one thing that I would like to know and is my ancestors, you know, where did I originate. But I just never took the time to find out the answer to that question." When this response is fed into an LLM with a reverse-prompting instruction, the model infers the question as: "Can you tell me about your family background or ethnic origin?" Although semantically related, the inferred question diverges markedly from the original prompt, which was simply: "What city are you from?" We quantify such semantic drift by computing cosine similarity between the sentence embeddings of original and reconstructed questions, operationalizing tangentiality as the degree of deviation in inferred prompts.

3.2 Participants

A total of 173 German-speaking participants were recruited for the study. Diagnosis was confirmed using the DSM-IV criteria for schizophrenia or schizoaffective disorder. All participants received written and oral information about the project, and written informed consent was obtained prior to inclusion. Descriptive statistics of the sample are presented in Table 1.

3.3 Corpus

We used the Narrative of Emotions Task (NET; (Buck et al., 2014)) to collect speech samples. The NET is a short semi-structured interview with open-ended questions originally developed to assess social cognition. Eliciting speech through (semi-)structured questions is a widely used and cost-effective method in NLP studies (Elvevåg et al., 2007; Just et al., 2020; Iter et al., 2018); it enhances comparability and has been shown to yield more consistent results than analyses of free conversational speech (Morgan et al., 2021). We employed a short version of the NET, translated into German, comprising three questions for each of four basic emotions: sadness, fear, anger, and happiness:

(1) What does [sadness/fear/anger/ happiness] mean to you?

(2) Can you describe a situation where you felt [sadness/fear/anger/happiness]?

(3) Why do you think you felt this emotion in that situation?

All interviews were conducted by trained clinicians, recorded and automatically transcribed using OpenAI's Whisper-large-v3 model[5] (Radford et al., 2022). The transcripts were manually preprocessed following established protocols to minimize bias in the subsequent analyzes (Iter et al., 2018; Just et al., 2020). Each transcript was anonymized and segmented by emotion. Verbal fillers (e.g., ”ehm”, ”mhm”) were removed. The

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[5] https://huggingface.co/openai/whisper-large-v3

	Schizophrenia n = 107	Controls n = 66
Age (years)	45.9 (12.6)	41.3 (15.8)
Gender (male)	47 (44%)	26 (39%)
Inpatients	26 (24.3%)	-
Duration of illness (years)	15.5 (7.78)	-
Verbal IQ	106.37 (13.20)	108.4 (12.50)
PANSS	75.8 (34.4)	-

Table 1. Descriptive statistics of the participants.

final data set consisted of 173 recorded and transcribed NET interviews. Table 2 presents the descriptive statistics of the corpus.

Group	n	Sentences Mean(Std)	Tokens Mean(Std)
Schizo-phrenia	107	90.2(69.3)	1093.7(893.8)
Control	66	62.8(35.2)	922.3(549.6)

Table 2. Linguistic corpus

3.4 Statistical analysis

To examine group differences in semantic alignment across emotional conditions and conversational turns, we fitted a linear mixed-effects model using the lmer function from the lme4 package in R (R Core Team, 2024; Bates et al., 2015). The dependent variable was cosine similarity, quantifying the semantic overlap between the original NET interview question and the LLM-generated reconstruction via our reverse prompting procedure. Fixed effects included the interactions between group (schizophrenia vs. control) and emotion (joy, sadness, anger, and fear), group and answer-number (1–3), as well as the interaction between model-type and temperature, standardized response length (length-z), and their main effects. To account for repeated measures and subject-specific variability, we included random intercepts for the participant, and for the nested factors participant:model-type, participant:lenght_z, and participant:emotion. Statistical significance of fixed effects was assessed using Satterthwaite's approximation for degrees of freedom, as implemented in the lmerTest package.

4 Results

To this end, we evaluated the cosine similarity between the original questions from a semi-structured interview and the reconstructed questions across 173 participants. Each participant responded to 12 questions, with each response processed by three different LLMs at multiple temperature settings, resulting in a total of 36,162 question pairs. The cosine similarity values ranged from –0.154 to 1.0 (mean = 0.506, SD = 0.267). We visualized the distribution of cosine similarity values for each interview question in a density plot (Figure 3), where each line represents the similarity distribution across all emotions, models, and temperature settings for a given question.

We analyzed cosine similarity using a linear mixed-effects model with random intercepts for participant and participant-level interactions with emotion, model type, and lenght of the linguistic output. Fixed effects included group, emotion, answer number, model type, temperature, and standardized text length, as well as relevant interactions.

Compared to a null model, which included only a random intercept for participant and yielded an AIC of 4203.8 and a BIC of 4229.3, our full model showed a markedly improved fit, with an AIC of -24986.1 and a BIC of -24782.2. A likelihood ratio test comparing the two models confirmed that the full model significantly outperformed the null model, $\chi^2(20) = 29232$, $p < .001$, indicating that the inclusion of the predictors substantially improved the explanation of variance in cosine similarity.

The schizophrenia group showed significantly lower cosine similarity compared to controls ($\beta =$ –0.165, SE = 0.021, $t(838) = -7.79$, $p < .001$). All emotions were associated with reduced cosine similarity relative to the neutral reference: Freude

Predictor	**Estimate**	**Std. Error**	**df**	**t value**	**Pr($>|t|$)**
(Intercept)	0.8555	0.0168	848.2	51.049	$< .001^{***}$
group: schizophrenia	-0.1653	0.0212	837.8	-7.786	$< .001^{***}$
emotion: joy	-0.0658	0.0194	551.3	-3.398	$< .001^{***}$
emotion: sadness	-0.0572	0.0191	523.1	-2.998	0.0028^{**}
emotion: anger	-0.0580	0.0197	586.2	-2.939	0.0034^{**}
answer number 2	-0.3288	0.0109	3180	-30.131	$< .001^{***}$
answer number 3	-0.4009	0.0101	4243	-39.638	$< .001^{***}$
model: LLaMA	0.0107	0.0050	785.6	2.129	0.0335^{*}
model: Mistral	-0.0543	0.0050	785.6	-10.783	$< .001^{***}$
temperature	-0.0536	0.0042	33660	-12.845	$< .001^{***}$
length (z)	0.0148	0.0045	1186	3.263	0.0011^{**}
schizophrenia × emotion: joy	0.0241	0.0249	582.1	0.968	0.334
schizophrenia × emotion: sadness	0.0622	0.0247	567.2	2.513	0.0122^{*}
schizophrenia × emotion: anger	0.0203	0.0254	621.5	0.797	0.426
schizophrenia × answer number 2	0.0908	0.0142	2848	6.378	$< .001^{***}$
schizophrenia × answer number 3	0.1538	0.0136	3646	11.288	$< .001^{***}$
LLaMA × temperature	0.0217	0.0059	33660	3.682	$< .001^{***}$
Mistral × temperature	0.0521	0.0059	33660	8.826	$< .001^{***}$

Table 3. Fixed Effects from Linear Mixed Effects Model

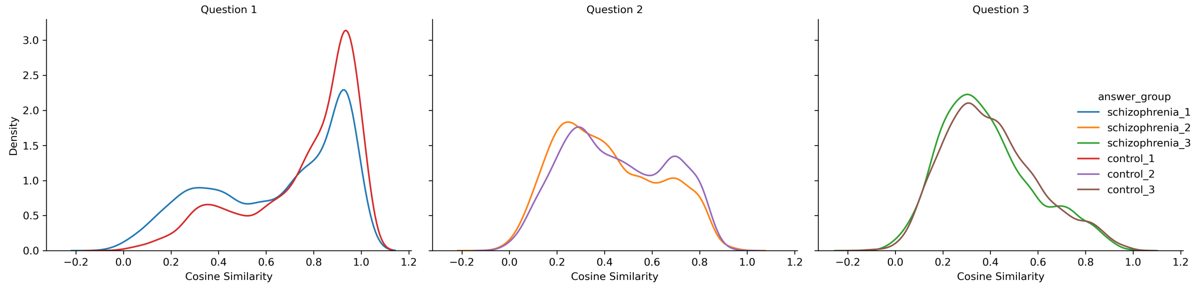

Figure 2: Density plot of cosine similarity values between original and reconstructed questions, grouped by interview question.

($\beta = –0.066$, $p = .001$), Traurigkeit ($\beta = –0.057$, $p = .003$), and Wut ($\beta = –0.058$, $p = .003$). Later answers also showed large decreases in similarity (Answer 2: $\beta = –0.329$, $p < .001$; Answer 3: $\beta = –0.401$, $p < .001$). Figure 3 displays estimated marginal means of cosine similarity across diagnostic groups and interview questions, derived from the fitted linear mixed-effects model.

Model type effects were mixed: Mistral significantly decreased cosine similarity ($\beta = -0.054$, $p < .001$), while LLaMA showed a marginally positive trend ($\beta = 0.011$, $p = .034$). Higher temperature reduced similarity ($\beta = -0.054$, $p < .001$), and longer texts were associated with slightly higher similarity scores ($\beta = 0.015$, $p = .001$).

Significant interactions indicated an increased effect of answer number in the schizophrenia group

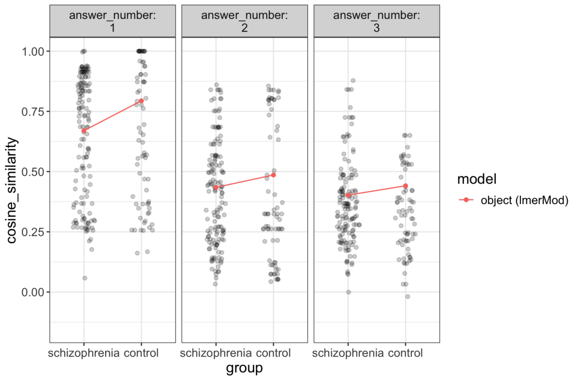

Figure 3: Scatterplot of cosine similarity by diagnostic group and interview question.

(Answer 2: $\beta = 0.091, p < .001$; Answer 3: $\beta = 0.154, p < .001$). Interactions between schizophrenia and emotion were mostly non-significant, except for Traurigkeit ($\beta = 0.062$, $p = .012$). Model-specific temperature effects were positive for both LLaMA ($\beta = 0.022, p < .001$) and Mistral ($\beta = 0.052, p < .001$).

Model diagnostics indicated no serious multicollinearity, with all adjusted GVIF values below 2.5, and residual analyses confirmed the assumptions of linearity, homoscedasticity, and normality were adequately met.

5 Discussion

To this end, we evaluated the capacity of three open-source LLMs to infer interviewer questions based solely on participant responses drawn from a semi-structured psychiatric interview. This was implemented using a modified reverse prompting procedure designed to approximate the interpretive process of a clinical listener. By focusing on the output alone—the participant's answer—we assessed the models' ability to reconstruct the likely original question. The inferred questions were then compared to the actual interviewer questions using cosine similarity between sentence embeddings as a proxy for semantic alignment. To examine potential group-level effects, we analyzed similarity scores using a linear mixed-effects model, comparing responses from individuals with schizophrenia to those from neurotypical controls.

Responses from individuals with schizophrenia were associated with significantly lower similarity scores compared to neurotypical controls ($\beta = –0.165$, $p < .001$), indicating a main effect of diagnosis on LLM inference performance. This confirms our central finding: LLM-inferred questions based on responses from individuals with schizophrenia were substantially more dissimilar to the original interview questions than those based on responses from neurotypical participants. This suggests that LLMs struggle more to reconstruct the intended question when processing responses from individuals with schizophrenia, likely due to the greater linguistic variability and atypicality characteristic of their speech. These findings are consistent with prior research showing reduced semantic coherence and weakened alignment between questions and answers in schizophrenia (Elvevåg et al., 2007; Iter et al., 2018; Tang et al., 2021). For instance, Tang et al. (2021) demonstrated a significantly steeper decline in semantic similarity across conversational turns in individuals with schizophrenia using BERT-embeddings, reflecting impairments in discourse-level integration.

Although the group effect is statistically significant, its diagnostic utility remains to be determined. The effect may reflect specific symptoms such as tangentiality or incoherence, which are not uniformly present across individuals with schizophrenia. Future work should examine whether the reverse prompting procedure can differentiate individuals with more pronounced positive formal thought disorder and whether it outperforms or complements existing coherence-based approaches. However, the effect appears too weak to be used as a stand-alone classification feature.

In addition to diagnostic group, several other variables significantly influenced the models' ability to infer interview questions from participant responses. First, the semi-structured NET interview includes three nested questions per emotional topic. Our analysis revealed a stepwise reduction in cosine similarity for the second and third questions, reflecting increased difficulty of the LLMs in inferring later questions in a sequence. However, this reduction was not uniform across groups: a significant interaction between question number and group showed that participants with schizophrenia exhibited a notably smaller drop in similarity across the interview turns. This suggests that the consistently lower cosine similarity in the schizophrenia group reflects a general difficulty for the LLM in inferring interviewer prompts, which remains relatively stable across turns, whereas in the neurotypical group, increasing elaboration or digression across nested responses leads to a sharper decline in similarity.

Second, the emotion discussed also modulated similarity. All emotional prompts resulted in reduced cosine similarity compared to the reference emotion fear, indicating that emotional content poses an additional interpretive challenge for LLMs. Yet, a significant interaction was found only for sadness, where individuals with schizophrenia showed slightly higher similarity scores. This suggests that expressions of sadness in this group may follow more predictable patterns or use more canonical language, enabling the model to better reconstruct the original prompt. These findings align with prior research suggesting that emotional processing in schizophrenia can vary by affective domain and may be marked by both flattening and stereotypy.

Third, output length had only a marginal effect on similarity scores. This finding is encouraging, as it indicates that our approach is not overly sensitive to verbosity, a common concern when analyzing spontaneous language in schizophrenia. This robustness increases the feasibility of applying such models in real-world settings with variable-length responses.

Finally, model-specific factors influenced performance. The LLaMA model produced slightly higher cosine similarity scores, while Mistral yielded slightly lower scores relative to our baseline model Aya. As expected, higher decoding temperature significantly reduced similarity, reflecting greater generative randomness and less precise reconstructions. These effects underscore the importance of model selection and decoding parameters in applied LLM workflows for clinical language analysis.

Our modified reverse prompting procedure warrants further investigation as a potential feature in classification tasks and relapse prediction in longitudinal settings. Its associations with symptoms such as tangentiality, incoherence, and poverty of speech should also be systematically examined in future research. Taken together, our findings indicate that smaller open-source LLMs face challenges in processing the language of individuals with schizophrenia. As LLMs gain traction in clinical contexts such as psychotherapy, this raises important questions about the need for domain-specific adaptation and the limitations of general-purpose models in psychiatric applications. This underscores the need for domain-specific models tailored to the linguistic characteristics of psychiatric populations. Future work could also explore whether fine-tuning or prompt engineering might enable models to better process the more disorganized or idiosyncratic responses often found in schizophrenia.

Limitations

This study has several limitations. First, we focused only on individuals with schizophrenia and neurotypical controls. Other diagnostic groups in both psychiatry (e.g., acute mania, dementia, autism) and neurology (e.g., post-stroke aphasia) may also produce idiosyncratic responses that could challenge LLMs in a reverse-prompting setting. Second, we relied exclusively on smaller, open-source language models due to privacy constraints, which may not match the performance or capabilities of larger proprietary models. Third, our dataset was relatively homogeneous, consisting solely of German-language responses to a brief, semi-structured interview (NET) with twelve fixed questions, potentially limiting linguistic variability and ecological validity. Fourth, our analysis was restricted to one language, which constrains generalizability to other linguistic or cultural contexts. Additionally, model outputs were sensitive to prompt design and required model-specific tuning, which may affect reproducibility. Finally, the constrained and predictable nature of the dataset may have led models to perform disproportionately well, thereby inflating apparent inferential ability.

Ethics statement

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee at [to be filled after review], approval number [to be filled after review]. All participants provided written informed consent prior to participation. All data were anonymized prior to analysis. No proprietary or closed-source models were used; all processing occurred locally to ensure compliance with data protection regulations and institutional guidelines.

Acknowledgements

I.N. is a participant in the BIH Charité Digital Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health at Charité (BIH). The authors acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing computational resources that have contributed to the research results reported in this paper.

References

AI@Meta. 2024. Llama 3 model card.

Maria Francisca Alonso-Sánchez, Sabrina D. Ford, Michael MacKinley, Angélica Silva, Roberto Limongi, and Lena Palaniyappan. 2022. Progressive changes in descriptive discourse in First Episode Schizophrenia: a longitudinal computational semantics study. *Schizophrenia*, 8(1):1–9. Number: 1 Publisher: Nature Publishing Group.

M. Alvarez-Jimenez, J. F. Gleeson, L. P. Henry, S. M. Harrigan, M. G. Harris, G. P. Amminger, E. Killackey, A. R. Yung, H. Herrman, H. J. Jackson, and P. D. McGorry. 2011. Prediction of a single psychotic episode: a 7.5-year, prospective study in first-episode psychosis. *Schizophrenia Research*, 125(2-3):236–246.

Nancy C. Andreasen. 1986. Scale for the assessment of thought, language, and communication (TLC). *Schizophrenia Bulletin*, 12(3):473–482.

Nancy C. Andreasen. 1989. The Scale for the Assessment of Negative Symptoms (SANS): Conceptual and Theoretical Foundations. *The British Journal of Psychiatry*, 155(S7):49–52. Publisher: Cambridge University Press.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Valentina Bambini, Giorgio Arcara, Francesca Bosinelli, Mariachiara Buonocore, Margherita Bechi, Roberto Cavallaro, and Marta Bosia. 2020. A leopard cannot change its spots: A novel pragmatic account of concretism in schizophrenia. *Neuropsychologia*, 139:107332.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

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*, 1:15030.

Janna N. de Boer, Sanne G. Brederoo, Alban E. Voppel, and Iris E. C. Sommer. 2020. Anomalies in language as a biomarker for schizophrenia. *Current Opinion in Psychiatry*, 33(3):212–218.

Benjamin Buck, Kelsey Ludwig, Piper S. Meyer, and David L. Penn. 2014. The use of narrative sampling in the assessment of social cognition: the Narrative of Emotions Task (NET). *Psychiatry Research*, 217(3):233–239.

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.

Dalia Elleuch, Yinhan Chen, Qiang Luo, and Lena Palaniyappan. 2025. Speaking of yourself: A meta-analysis of 80 years of research on pronoun use in schizophrenia. *Schizophrenia Research*, 279:22–30.

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*, 93(1):304–316.

Molly T. Finnerty, Atif Khan, Kai You, Rui Wang, Gyojeong Gu, Deborah Layman, Qingxian Chen, Noémie Elhadad, Shalmali Joshi, Paul S. Appelbaum, Todd Lencz, Sander Markx, Steven A. Kushner, and Andrey Rzhetsky. 2024. Prevalence and incidence measures for schizophrenia among commercial health insurance and medicaid enrollees. *Schizophrenia*, 10(1):1–9. Publisher: Nature Publishing Group.

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.

Alkomiet Hasan, Peter Falkai, Isabell Lehmann, and Wolfgang Gaebel. 2020. Schizophrenia. *Dtsch Arztebl International*, 117(24):412–419.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, New Orleans, LA. Association for Computational Linguistics.

Lydia Jeong, Melissa Lee, Ben Eyre, Aparna Balagopalan, Frank Rudzicz, and Cedric Gabilondo. 2023. Exploring the Use of Natural Language Processing for Objective Assessment of Disorganized Speech in Schizophrenia. *Psychiatric Research and Clinical Practice*, 5(3):84–92. Publisher: American Psychiatric Publishing.

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*, 11. Publisher: Frontiers.

S. R. Kay, A. Fiszbein, and L. A. Opler. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276.

Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Matarić, Daniel J. McDuff, and Megan Jones Bell. 2024. The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health*, 11(1):e59479. Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

Hanqing Li and Diego Klabjan. 2025. Reverse Prompt Engineering. ArXiv:2411.06729 [cs].

Renyu Li, Minne Cao, Dawei Fu, Wei Wei, Dequan Wang, Zhaoxia Yuan, Ruofei Hu, and Wei Deng. 2024. Deciphering language disturbances in schizophrenia: A study using fine-tuned language models. *Schizophrenia Research*, 271:120–128.

Nancy B. Lundin, Henry R. Cowan, Divnoor K. Singh, and Aubrey M. Moe. 2023. Lower cohesion and altered first-person pronoun usage in the spoken life narratives of individuals with schizophrenia. *Schizophrenia Research*, 259:140–149.

Sarah E. Morgan, Kelly Diederen, Petra E. Vértes, Samantha H. Y. Ip, Bo Wang, Bethany Thompson, Arsime Demjaha, Andrea De Micheli, Dominic Oliver, Maria Liakata, Paolo Fusar-Poli, Tom J. Spencer, and Philip McGuire. 2021. Natural Language Processing markers in first episode psychosis and people at clinical high-risk. *Translational Psychiatry*, 11(1):1–9. Number: 1 Publisher: Nature Publishing Group.

John X. Morris, Wenting Zhao, Justin T. Chiu, Vitaly Shmatikov, and Alexander M. Rush. 2023. Language Model Inversion. Publisher: arXiv Version Number: 1.

Lena Palaniyappan. 2021. More than a biomarker: could language be a biosocial marker of psychosis? *npj Schizophrenia*, 7(1):1–5. Number: 1 Publisher: Nature Publishing Group.

Alberto Parola, Jessica Mary Lin, Arndis Simonsen, Vibeke Bliksted, Yuan Zhou, Huiling Wang, Lana Inoue, Katja Koelkebeck, and Riccardo Fusaroli. 2023. Speech disturbances in schizophrenia: Assessing cross-linguistic generalizability of NLP automated measures of coherence. *Schizophrenia Research*, 259:59–70.

Ivo Petrov, Dimitar I. Dimitrov, Maximilian Baader, Mark Niklas Müller, and Martin Vechev. 2024. DAGER: Exact Gradient Inversion for Large Language Models. ArXiv:2405.15586 [cs].

Samuel L. Pugh, Chelsea Chandler, Alex S. Cohen, Catherine Diaz-Asper, Brita Elvevåg, and Peter W. Foltz. 2024. Assessing dimensions of thought disorder with large language models: The tradeoff of accuracy and consistency. *Psychiatry Research*, 341:116119.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv:1908.10084 [cs].

Katharina Schneider, Katrin Leinweber, Hamidreza Jamalabadi, Lea Teutenberg, Katharina Brosch, Julia-Katharina Pfarr, Florian Thomas-Odenthal, Paula Usemann, Adrian Wroblewski, Benjamin Straube, Nina Alexander, Igor Nenadić, Andreas Jansen, Axel Krug, Udo Dannlowski, Tilo Kircher, Arne Nagels, and Frederike Stein. 2023. Syntactic complexity and diversity of spontaneous speech production in schizophrenia spectrum and major depressive disorders. *Schizophrenia*, 9(1):1–10. Number: 1 Publisher: Nature Publishing Group.

Zeyang Sha and Yang Zhang. 2024. Prompt Stealing Attacks Against Large Language Models. ArXiv:2402.12959 [cs].

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*, 7(1):25.

Rohit Voleti, Stephanie M Woolridge, Julie M Liss, Melissa Milanovic, Gabriela Stegmann, Shira Hahn, Philip D Harvey, Thomas L Patterson, Christopher R Bowie, and Visar Berisha. 2023. Language Analytics for Assessment of Mental Health Status and Functional Competency. *Schizophrenia Bulletin*, 49(Supplement_2):S183–S195.

Andrew R. Watson, Cağla Defterali, Thomas H. Bak, Antonella Sorace, Andrew M. McIntosh, David G. C. Owens, Eve C. Johnstone, and Stephen M. Lawrie. 2012. Use of second-person pronouns and schizophrenia. *The British Journal of Psychiatry: The Journal of Mental Science*, 200(4):342–343.

Collin Zhang, John Xavier Morris, and Vitaly Shmatikov. 2024. Extracting Prompts by Inverting LLM Outputs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14753–14777, Miami, Florida, USA. Association for Computational Linguistics.

Ido Ziv, Heli Baram, Kfir Bar, Vered Zilberstein, Samuel Itzikowitz, Eran V. Harel, and Nachum Dershowitz. 2021. Morphological characteristics of spoken language in schizophrenia patients - an exploratory study. *Scandinavian Journal of Psychology*.