Large Language Models for Psychiatric Interview Analysis: An Exploratory Pilot Study

Katarina Lodrant
kl19928@student.uni-lj.si
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria
University of Ljubljana
Ljubljana, Slovenia

Filip Melinščak
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria

Ayse Nur Beris
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Valentin Schneider
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Klara Czernin
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Waltraud Bangerl
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Anselm Bründlmayer
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Frank Scharnowski
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria

Clarissa Laczkovics
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

David Steyrl
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria

Abstract

This exploratory pilot study investigates the use of large language models (LLMs) for automated analysis of psychiatric interviews. Using transcripts from the Structured Interview of Personality Organization (STIPO-R), we tested GPT-4o across three paradigms: direct application of clinical scoring guidelines, emulation of a validated psychometric scale, and exploratory construct elicitation. LLM-derived scores strongly correlated with clinician ratings and captured clinically relevant constructs. Findings highlight opportunities for scalable, theory-driven assessment of patient language, but also underscore challenges including interpretability, reproducibility and data privacy.

Keywords

Large Language Models, Clinical Language Analysis, AI in Mental Health, Sentiment Analysis, Identity Diffusion

1 Introduction

In psychiatry, clinicians are often required to make complex diagnostic judgments without definitive biological markers. Instead, assessments rely on observable behavior, subjective self-report, and, crucially, on language [1]. Patient language provides a uniquely rich source of information: it reflects patterns of thought, emotional states, and interpersonal dynamics, all of which are central to understanding mental functioning [2]. An abundance of naturalistic speech emerges from clinical interviews and therapy sessions, underscoring the need for systematic methods that can both detect subtle psychological cues and handle large volumes efficiently.

Automated methods for language analysis have evolved from early dictionary-based tools such as the Linguistic Enquiry and Word Count (LIWC), which provided interpretable but context-insensitive results [3], to embedding-based models like Word2Vec [4], BERT [5], and RoBERTa [6], which offered greater contextual sensitivity at the cost of interpretability and technical complexity. More recently, large language models (LLMs) such as GPT [7] have emerged as flexible, prompt-driven analyzers.

Researchers have argued that GPT may be a superior tool for automated text analysis, achieving high accuracy on various tasks across languages without training data and with minimal coding demands [8, 9]. Yet others caution that risks of bias, reproducibility, opacity, and overreliance remain. In some contexts, established, validated models still outperform LLMs, and researchers must weigh not only how LLMs can be applied, but whether their use is beneficial given the risks [10].

Analyses of patient language have identified linguistic markers associated with various psychiatric conditions [11, 12, 13, 1, 14]. A 2020 review by Zhang et al. [15] highlighted the growing use of natural language processing (NLP) for mental illness

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.12

detection, noting that social media texts remain the most common data source. In contrast, relatively few studies have examined transcripts of patient speech [16, 17]. This gap likely reflects the scarcity of suitable datasets, as such data is usually not recorded, and when it is, audio recordings and transcripts often contain sensitive personal information and therefore cannot be publicly shared. Moreover, speech data typically require supporting ground-truth measures (e.g., validated questionnaires or clinical assessments) to be useful for research.

Recent advances in automatic transcription, together with the emergence of LLMs, have opened new directions for systematic analysis of patient-generated language. Unlike earlier approaches, LLMs combine ease of use with a seemingly unprecedented sensitivity to linguistic context. In this work, we examine the opportunities and challenges they present through a pilot analysis of transcripts of the Structured Interview of Personality Organization (STIPO-R), a validated psychoanalytic diagnostic instrument.

2 Methods

2.1 Dataset

We analyzed a subset of data collected by Laczkovics et al. (2025) during the validation of the German STIPO-R for adolescents [18, 19]. The STIPO interview assesses multiple domains of personality functioning. For this study, we focused on the identity domain, which consists of 13 open-ended questions addressing areas such as self-perception, perception of others and engagement in school and recreation. These questions typically elicit rich narrative responses that are well-suited for language analysis. Responses were evaluated by trained clinicians on 15 items, each rated on a three-point scale (0 = no pathology, 1 = moderate pathology, 2 = severe pathology), producing a total identity diffusion score ranging from 0 to 30. This clinician-rated score served as the ground truth for evaluating LLM performance. From the original study sample of 171 participants [18], 70 provided data of sufficient quality for the present analyses: 49 patients with a probable or definite personality disorder (PD) diagnosis and 21 controls without PD. From this set, we derived a subsample of 25 participants (16 patients and 9 controls), aged 14–19 years, using a stratified selection procedure to ensure even coverage of the full spectrum of identity pathology, from consolidated (low diffusion scores) to highly diffused identity.

2.2 LLM Setup

We used GPT-4o [20], accessed via a secure Python API connection under GDPR-compliant data protection. Interview transcripts were in German, while prompts were written in English. Prior work suggests that English prompts improve model performance even when applied to other languages [21, 9]. The model temperature was set to 0, producing consistent outputs for identical prompts.

2.3 Experimental Paradigms

We tested three experimental paradigms that elicited numeric ratings from the LLM, alongside a lexicon-based sentiment analysis baseline for comparison.

First, in a **Direct STIPO Scoring approach**, the official STIPO-R rating guidelines were copied verbatim into prompts, and the model was asked to assign 0–2 scores to individual items, paralleling the procedure used by clinicians in our dataset. Item-level and total scores were compared with clinician ratings.

Second, in a **Scale Emulation paradigm**, we tested whether the LLM could approximate a validated psychometric measure by inferring likely responses to scale items from interview transcripts rather than direct self-report. Specifically, we used the Self-Concept Clarity Scale (SCCS) [22], a 12-item self-report instrument. Each item was presented to the model together with the identity section of the transcript, and the model was instructed to assign a 1–5 Likert score. Item scores were summed to yield a total SCCS score, which we compared with clinician-rated identity diffusion. Conceptually, and as supported by empirical work, Otto Kernberg's notion of identity diffusion assessed in the STIPO is closely related to Campbell's construct of self-concept clarity [23, 24].

Third, we applied an exploratory **Construct Rating** approach, in which we developed rubrics for (a) overall valence (positivity vs. negativity of the response), (b) self-perception (positive vs. negative evaluation of the self), and (c) other-perception (positive vs. negative evaluation of others, including individuals, groups, relationships, or people in general). The construct definitions and prompts were drafted with assistance from ChatGPT-5. Ratings were given on a 1–7 scale, with an NA option if the construct was not referenced. Interviews were split into individual question–answer pairs (24–83 per subject), which served as the unit of analysis. The model was prompted separately for each unit and construct, and subject-level scores were calculated as the mean across units. We compared these mean construct ratings with clinician-rated identity diffusion, hypothesizing that more severe identity diffusion would be associated with more negative language (overall, in descriptions of the self, and in descriptions of others). To evaluate interpretability and reliability, we re-ran analyses where the score was extreme (1 or 7) and asked the model to provide both a score and a brief justification. As an exploratory validity check, a cognitive science master's student (author of this study) reviewed randomly selected transcript excerpts together with LLM ratings and reasonings, assessing whether the assigned scores were plausible and consistent with the intended construct. Additionally, we tested robustness by repeating the analyses with an alternative 0–5 scale.

As a simple and interpretable **Sentiment Analysis** baseline, we used GerVADER [25], a German sentiment lexicon in which each word is assigned a valence score (−4 = very negative, 0 = neutral, +4 = very positive) based on human ratings of perceived positivity or negativity. This choice was motivated by Colibazzi et al. [26], who applied the VADER lexicon [27] to STIPO transcripts. For each question–answer unit, we extracted the patient's response, identified words present in the lexicon, retrieved their valence scores, and calculated three metrics: (a) overall sentiment (mean of all scores), (b) negative sentiment (mean of negative scores only), and (c) positive sentiment (mean of positive scores only). These answer-level values were then averaged across all answers to obtain per-subject scores, which were compared with both the LLM-derived ratings of overall valence and clinician ratings of identity diffusion.

All comparisons were tested with Pearson correlations (p-values corrected for multiple comparisons; $\alpha = 0.05$).

3 Results

Direct STIPO Scoring. Summed scores produced by GPT-4o strongly correlated with clinician ratings (r = 0.90), as illustrated in Figure 1. Item-level agreement was exact in 66% of cases.

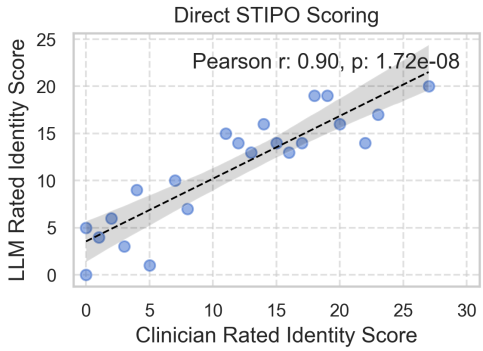

Figure 1: Correlation between clinician-rated and LLM-rated STIPO identity scores.

Scale Emulation. SCCS scores derived from LLM outputs correlated negatively with clinician-rated identity diffusion (r = −0.82; Figure 2). This finding is in line with the conceptual link between identity coherence and self-concept clarity.

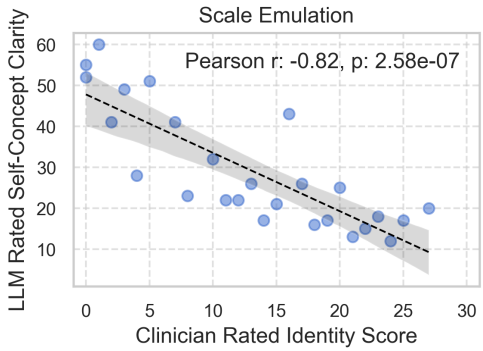

Figure 2: Correlation between clinician-rated STIPO identity scores and LLM-rated Self-Concept Clarity.

Exploratory Construct Ratings. Average overall valence correlated negatively with identity pathology (r = −0.82; Figure 3), suggesting that more severely affected adolescents used more negative language overall.

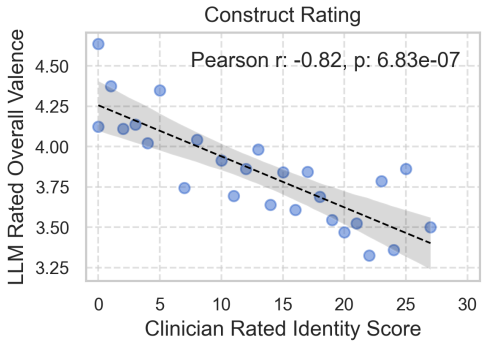

Figure 3: Correlation between clinician-rated STIPO identity scores and LLM-rated overall valence of answers.

Self- and other-perception ratings were also associated with clinician scores (r = −0.81 and −0.57). Manual checks indicated that the model generally distinguished references to the self from references to others. When re-run with prompts requesting both a rating and a brief justification, this distinction improved: all cases lacking a relevant reference were correctly scored as NA. However, providing reasoning noticeably shifted the rating: extreme values were often moderated toward the midrange.

Changing the rating scale from 1–7 to 0–5 did not materially affect the results (r = 0.98). In all cases, the model produced valid outputs in the requested format.

Lexicon-Based Sentiment Analysis. Sentiment analysis with GerVADER revealed a significant correlation between clinician ratings and mean negative sentiment (r = −0.47), but not with mean overall or mean positive sentiment. This correlation was weaker than that between clinician ratings and LLM-derived overall valence, highlighting the limitations of context-insensitive, bag-of-words methods. Manual inspection confirmed that GPT-4o often inferred negativity from conversational context rather than from explicitly negative words. Mean negative sentiment was also significantly correlated with LLM-derived overall valence (r = 0.68).

4 Discussion

This exploratory study shows that LLMs can approximate expert ratings of psychiatric interviews and apply psychometric constructs to clinical transcripts, while also highlighting barriers that preclude immediate clinical use. In the following, we outline opportunities, risks, and challenges, and suggest pathways for more rigorous validation.

4.1 Opportunities

LLMs perform reliably on structured clinical tasks. Using only verbatim scoring guidelines, GPT-4o approximated expert scoring of the STIPO, a task that typically requires extensive training. While LLMs should not replace clinicians, they could provide secondary checks in research settings or serve as teaching tools to illustrate scoring rules, highlight ambiguities and improve teaching materials.

Applying validated psychometric scales through LLMs anchors automated analyses in established theory. The strong correlation between LLM-rated self-concept clarity and clinician-rated identity diffusion supports the validity of this approach and suggests that LLMs can extend the reach of standardized assessments in scalable ways.

By contrast, defining new constructs ad hoc is more vulnerable to misspecification and requires iterative prompt engineering. Nevertheless, this strategy may capture clinically relevant, context-sensitive phenomena that remain inaccessible to conventional language-processing methods, potentially opening pathways to subtle markers of pathology.

LLMs further offer efficiency in time and cost, scalability to large datasets, cross-linguistic applicability, and the ability to rapidly test new rating schemes or constructs.

4.2 Risks and Challenges

The study also underscores multiple risks.

Interpretability and the black-box problem. LLMs remain opaque, and their internal decision processes are currently inaccessible. Some surface interpretability is possible; for instance, researchers can manually compare scores with text samples, or request rationales from the model. However, such rationales are post hoc, primarily useful for illustrating reasoning, and cannot

be assumed to reflect the actual mechanisms behind the model's ratings.

Reproducibility and test–retest reliability. A key concern is reproducibility across time. Outputs vary not only due to stochasticity but also across different versions of the same model. Because earlier GPT versions are not preserved, analyses cannot be rerun on identical models. Even with temperature fixed at 0 in our study, small prompt variations, such as requesting reasoning, produced measurable differences in outputs, a well-documented phenomenon [28, 21]. Moreover, newer versions are not always improvements: performance can regress on certain tasks [9, 10]. Such variability poses significant challenges for scientific applications, where reproducibility is essential.

Data privacy and ethics. Patient language data are highly sensitive. While GDPR-compliant API contracts ensure encryption and prevent storage or retraining, the ethical stakes remain high. An alternative is to deploy LLMs locally, which enhances data security but requires substantial technical expertise and computing resources. Beyond privacy, there are broader risks of misuse: LLMs could be applied to surveillance or automatic 'flagging' of individuals, raising concerns about autonomy and stigmatization. Awareness of such possibilities is essential to anticipate and counter harmful applications, in line with international guidelines for trustworthy AI [29].

Bias and fairness. Training data for LLMs may embed demographic, cultural, or linguistic biases [10]. In psychiatry, this is particularly dangerous, as dialectical or culturally specific expressions may be misclassified as pathological.

Overreliance and face validity. The fluency and confidence of LLM outputs create risks of undue trust. Clinicians and researchers may treat model scores as authoritative, even when they are unreliable. In healthcare contexts, this raises ethical concerns: automatically generated reports or diagnostic suggestions may be accepted without scrutiny, especially if embedded in clinical workflows.

Prompt engineering. Contrary to claims that LLMs like GPT are easy-to-use, generalist tools that can handle a wide range of text analysis tasks with little coding or training data [9, 8], effective prompting remains challenging and requires significant expertise [21, 28]. A comprehensive 2025 survey of prompting strategies by Schulhoff et al. [21] concluded that robust prompts must balance specificity and flexibility, be iteratively refined, and validated against examples. Well-designed prompts can reduce bias and instability, whereas underspecified prompts yield inconsistent outputs and overly prescriptive prompts risk forcing artificial ratings. Systematic, theory-driven prompt development aligned with established constructs is therefore essential.

4.3 Pathways for Validation

As the field is still developing, applications of LLMs for language analysis should be guided by comprehensive validation to help mitigate risks of opacity, instability, and bias. Critical steps include:

- Datasets with multiple ground-truth measures: clinician ratings, validated scales, and demographics to enable triangulation.
- Benchmarking against non-generative approaches: e.g., LIWC, RoBERTa, or traditional machine learning classifiers.
- Cross-model robustness: comparing results across different LLMs (GPT, Claude, Llama).
- Evaluation prompts: asking models to assess their own outputs or, in a multi-agent setup, to evaluate the output of another model (e.g., "Do you agree with this score?").
- Manual inspection: qualitative review of outputs, ideally conducted through interdisciplinary collaboration between domain specialists (e.g., clinicians) and those designing the prompts.
- Perturbation tests: checking stability by slightly altering prompts or text snippets.

5 Conclusion

Language remains psychiatry's most fundamental source of information. Automated analysis of clinical transcripts offers a route toward scalable, theory-driven markers of psychopathology. Our pilot study suggests that LLMs can approximate expert scoring, apply validated psychometric instruments, and flexibly analyze novel constructs with promising validity. Yet these opportunities are tempered by challenges of interpretability, reproducibility, and ethics. We argue that LLMs can serve as valuable research companions and have the potential to benefit clinical diagnostics when integrated cautiously, transparently, and in theory-driven ways.

References

[1] 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*. Biomarkers in the Attenuated Psychosis Syndrome 226, (Dec. 2020), 158–166. doi:10.1016/j.schres.2020.04.032.

[2] Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A. Lindquist. 2022. From Text to Thought: How Analyzing Language Can Advance Psychological Science. EN. *Perspectives on Psychological Science*, 17, 3, (May 2022), 805–826. Publisher: SAGE Publications Inc. doi:10.1177/17456916211004899.

[3] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. en. *Journal of Language and Social Psychology*, 29, 1, (Mar. 2010), 24–54. doi:10.1177/0261927X09351676.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs]. (Sept. 2013). doi:10.48550/arXiv.1301.3781.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Jill Burstein, Christy Doran, and Thamar Solorio, editors. Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. doi:10.18653/v1/N19-1423.

[6] Yinhan Liu et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]. (July 2019). doi:10.48550/arXiv.1907.11692.

[7] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs]. (Mar. 2024). doi:10.48550/arXiv.2303.08774.

[8] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT. arXiv:2303.03186 [cs]. (Mar. 2023). doi:10.48550/arXiv.2303.03186.

[9] Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121, 34, (Aug. 2024), e2308950121. Publisher: Proceedings of the National Academy of Sciences. doi:10.1073/pnas.2308950121.

[10] Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3, 7, (July 2024), pgae245. doi:10.1093/pnasnexus/pgae245.

[11] Robin Quillivic, Yann Auxéméry, Frédérique Gayraud, Jacques Dayan, and Salma Mesmoudi. 2025. Linguistic markers for identifying post-traumatic stress disorder and associated symptoms: a systematic literature review. eng. *Journal of the American Medical Informatics Association: JAMIA*, (May 2025), ocaf075. doi:10.1093/jamia/ocaf075.

[12] Erik C. Nook. 2023. The Promise of Affective Language for Identifying and Intervening on Psychopathology. en. *Affective Science*, 4, 3, (Sept. 2023), 517–521. doi:10.1007/s42761-023-00199-w.

[13] Felipe Argolo et al. 2024. Natural language processing in at-risk mental states: enhancing the assessment of thought disorders and psychotic traits with semantic dynamics and graph theory. *Brazilian Journal of Psychiatry*. doi:10.47626/1516-4446-2023-3419.

[14] Cheryl Mary Corcoran and Guillermo A. Cecchi. 2020. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Understanding the Nature and Treatment of Psychopathology: Letting the Data Guide the Way 5, 8, (Aug. 2020), 770–779. doi:10.1016/j.bpsc.2020.06.004.

[15] Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. en. *npj Digital Medicine*, 5, 1, (Apr. 2022), 1–13. Publisher: Nature Publishing Group. doi:10.1038/s41746-022-00589-7.

[16] Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for? - A closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik, and Philip Resnik, editors. Association for Computational Linguistics, New Orleans, LA, (June 2018), 1–12. doi:10.18653/v1/W18-0601.

[17] Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. (Dec. 2016), 136–143. doi:10.1109/SLT.2016.7846256.

[18] C. Laczkovics et al. 2025. Assessment of personality disorders in adolescents – a clinical validity and utility study of the structured interview of personality organization (STIPO). en. *Child and Adolescent Psychiatry and Mental Health*, 19, 1, (May 2025), 49. doi:10.1186/s13034-025-00901-9.

[19] John F Clarkin, Eve Caligor, Barry L Stern, and Otto F Kernberg. 2016. STRUCTURED INTERVIEW OF PERSONALITY ORGANIZATION: STIPO-R. en.

[20] OpenAI et al. 2024. GPT-4o System Card. arXiv:2410.21276 [cs]. (Oct. 2024). doi:10.48550/arXiv.2410.21276.

[21] Sander Schulhoff et al. 2025. The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv:2406.06608 [cs]. (Feb. 2025). doi:10.48550/arXiv.2406.06608.

[22] Jennifer D. Campbell, Paul D. Trapnell, Steven J. Heine, Ilana M. Katz, Loraine F. Lavallee, and Darrin R. Lehman. 1996. Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology*, 70, 1, 141–156. Place: US Publisher: American Psychological Association. doi:10.1037/0022-3514.70.1.141.

[23] Otto F. Kernberg. 1984. *Severe Personality Disorders: Psychotherapeutic Strategies*. en. Google-Books-ID: FIl7opvzgeUC. Yale University Press. ISBN: 978-0-300-03273-4.

[24] J. Wesley Scala, Kenneth N. Levy, Benjamin N. Johnson, Yogev Kivity, William D. Ellison, Aaron L. Pincus, Stephen J. Wilson, and Michelle G. Newman. 2018. The Role of Negative Affect and Self-Concept Clarity in Predicting Self-Injurious Urges in Borderline Personality Disorder Using Ecological Momentary Assessment. *Journal of Personality Disorders*, 32, Supplement, (Jan. 2018), 36–57. Publisher: Guilford Publications Inc. doi:10.1521/pedi.2018.32.supp.36.

[25] Karsten Michael Tymann, Matthias Lutz, Patrick Palsbroker, and Carsten Gips. [n. d.] GerVADER - A German adaptation of the VADER sentiment analysis tool for social media texts. en.

[26] Tiziano Colibazzi, Avner Abrami, Barry Stern, Eve Caligor, Eric A. Fertuck, Michael Lubin, John Clarkin, and Guillermo Cecchi. 2023. Identifying Splitting Through Sentiment Analysis. en. *Journal of Personality Disorders*, 37, 1, (Feb. 2023), 36–48. doi:10.1521/pedi.2023.37.1.36.

[27] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. en. *Proceedings of the International AAAI Conference on Web and Social Media*, 8, 1, (May 2014), 216–225. doi:10.1609/icwsm.v8i1.14550.

[28] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI EA '21). Association for Computing Machinery, New York, NY, USA, (May 2021), 1–7. ISBN: 978-1-4503-8095-9. doi:10.1145/3411763.3451760.

[29] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. en. *Minds and Machines*, 30, 1, (Mar. 2020), 99–120. doi:10.1007/s11023-020-09517-8.