Mental Health Evaluation through Text Analysis: umbrella project documentation

Gabriel Bonnin

Ruhr University Bochum, Germany

Author Note

Correspondence concerning this article should be addressed to Gabriel Bonnin, Email: gabriel.bonnin@ruhr-uni-bochum.de

Mental Health Evaluation through Text Analysis: umbrella project documentation

Note

RUB
RESEARCH SCHOOL

This project is supported by the Ruhr University Bochum Research School.

Abstract

Psychotherapy is one of the most effective treatments for mental health problems, but its success depends on accurate diagnostic assessments. Current assessment practices largely rely on standardized closed-ended scales that, while reliable, may fail to capture the complexity, context and individuality of patients' mental states. Advances in artificial intelligence (AI) and natural language processing (NLP) enable the measurement of psychological constructs through natural language, offering a promising complement to traditional assessment methods by leveraging patients' own descriptions of their experiences.

While previous NLP-based mental health research has primarily focused on social media language, this project applies state-of-the-art large language models (LLMs) to open-ended intake data from a German outpatient psychotherapy clinic. Before therapy, patients describe the development, context, and perceived causes of their problems, as well as their current difficulties and therapy goals, in their own words. These texts are linked to key clinical measures, including diagnoses, symptom severity, functional impairment, and treatment outcomes, providing an ecologically valid resource for studying language-based assessment in real-world clinical settings.

The project comprises several complementary substudies. First, we examine whether patient language at intake reflects cross-sectional symptom severity and impairment and whether it provides incremental information beyond established self-report questionnaires. Second, we conduct a thematic analysis of patient responses to different open-ended prompts to characterize recurring themes. Third, we evaluate whether pre-therapy language predicts longitudinal treatment outcomes beyond baseline symptom measures.

Together, these studies aim to clarify how patient-generated language can be used for

assessment, interpretation, and prognosis in psychotherapy. The findings are intended to inform the development of clinically meaningful, language-based assessment tools that support personalized care and contribute to the modernization of mental health evaluation.

Introduction

Mental health problems pose a significant global challenge, accounting for a considerable proportion of deaths and disability-adjusted life years (World Health Organization, 2017). Psychotherapy is an effective and sustainable intervention for reducing symptoms and improving quality of life (Chorpita et al., 2011; Wampold & Imel, 2015), but it's success critically depends on accurate assessments (Jensen-Doss & Weisz, 2008; Lutz et al., 2022).

Standardized closed-ended tools such as the Beck Depression Inventory-II (Beck et al., 1996) rely on numerical scales (Likert, 1932) to structure and standardize assessments and are widely used in clinical research and practice. While these methods have advanced replicability and reliability in psychological assessment, they can miss important individual differences by restricting responses to pre-defined categories, limiting the ability to capture the complexity of mental states (Kjell, Kjell, et al., 2024).

Recent advances in AI, particularly transformer-based LLMs (Vaswani et al., 2017), present promising solutions to these limitations (Kjell, Kjell, et al., 2024). LLMs excel in analyzing context-rich natural language with remarkable accuracy across diverse tasks (Devlin et al., 2019). Open-ended response formats, where patients describe their experiences in their own words, provide high-dimensional, context-rich information that remains underutilized in current assessment practices. Empirical studies highlight the potential of NLP-based analysis of open-ended responses, achieving moderate convergence with closed-ended rating scales using traditional NLP methods (Kjell et al., 2019) and nearing theoretical upper limits of accuracy with LLMs (Kjell et al., 2022). Preliminary research also highlights their potential for predicting clinically significant outcomes, including suicide risk (Matero et al., 2019; Mohammadi et al., 2019; Zirikly et al., 2019).

However, much of the existing literature relies on social media or non-clinical text data, raising concerns about ecological validity and clinical relevance. In contrast, open-ended patient responses are routinely collected in clinical settings as part of pre-therapy intake procedures but remain largely underused in empirical research. At the Mental Health Research and Treatment Center at Ruhr University Bochum, patients respond to multiple prompts addressing the development and context of their problems, perceived causes, social reactions, current difficulties, and therapy goals. These narratives are linked to structured diagnostic interviews, repeated symptom assessments, clinician ratings, and longitudinal outcome measures. This unique, large-scale, and longitudinal clinical dataset enables a comprehensive examination of patient language across multiple analytic perspectives.

Accordingly, the present project is organized into three complementary substudies. The first investigates whether pre-therapy language reflects cross-sectional symptom severity and clinician-rated impairment and whether language-based representations provide incremental information beyond standardized self-report questionnaires. The second focuses on the semantic content and structure of patient narratives, using question-specific analyses to identify recurring themes and selective response patterns in how patients conceptualize their mental health problems. The third evaluates the prognostic value of pre-therapy language by testing whether patient narratives predict treatment response and individualized goal attainment over time, beyond baseline symptom severity.

By integrating assessment, interpretive, and prognostic perspectives, this project aims to advance the clinical use of natural language in psychotherapy. Ultimately, the findings seek to support more nuanced, patient-centered assessment practices and contribute to the development of language-based tools that complement existing diagnostic frameworks and inform personalized treatment planning.

Shared Methods

All substudies draw on the same clinical cohort and share a common set of intake and outcome variables. The shared dataset comprises (a) pre-therapy intake data, including sociodemographics, standardized psychometric questionnaires, and question-specific open-ended patient narratives, and (b) longitudinal clinical measures collected repeatedly during and after treatment. Textual analyses are based exclusively on pre-therapy narratives, while psychometric and clinician-rated measures are used as cross-sectional outcomes, covariates, or longitudinal endpoints depending on the substudy.

The following sections describe the shared dataset, preprocessing pipeline, and measurement instruments used across all substudies.

Measures

Sociodemographic and context measures

Sociodemographic information included age, sex, marital and relationship status, general educational attainment, vocational qualification, and current work ability. Contextual variables captured prior psychological or psychiatric treatment and the manner in which therapy ended (e.g., regular completion, dropout).

Responses from open-ended questions before therapy

At the start of therapy, patients complete two separate questionnaires designed to assess key aspects of their mental health concerns, functional impairments, and expectations for treatment. Questions 1–9 come from the first questionnaire (*Fragebogen zur Lebensgeschichte*), and questions 10–12 come from the second (*Eingangsfragebogen*). The questions include:

1. **Problem development:** 'Briefly describe how the problems for which you are seeking treatment have developed over time.' (german original question: „Beschreiben Sie kurz, wie sich Ihre Probleme, wegen derer Sie eine Behandlung aufsuchen, im Laufe der Zeit entwickelt haben.”)
2. **Extra stressors:** 'What causes you stress in addition to your everyday problems (e.g. finances, housing situation)?' (german original question: „Was macht Ihnen zusätzlich zu Ihren Problemen im Alltag Stress (z. B. Finanzen, Wohnsituation)¿‘)
3. **Pre-onset changes:** 'Did something special change in your life before the onset of your symptoms? (e.g. death of an important person, divorce or separation, change in work situation or income, addition to the family)' (german original question: „Hat sich vor dem Beginn Ihrer Beschwerden etwas Besonderes in Ihrem Leben verändert? (z. B. Tod einer wichtigen Bezugsperson, Scheidung oder Trennung, Veränderung der Arbeitssituation oder des Einkommens, Familienzuwachs)“)

4. **Event connection:** 'Do you see a connection between the event(s) and the development of your problems?' (german original question: „Sehen Sie einen Zusammenhang zwischen dem Ereignis/den Ereignissen und der Entwicklung Ihrer Probleme¿')
5. **Physical symptoms:** 'Are there any physical side effects when your problems occur?' (german original question: „Gibt es körperliche Begleiterscheinungen, wenn Ihre Probleme auftreten¿')
6. **Problem causes:** 'What do you think are the causes of your problems?' (german original question: „Welche Ursachen sehen Sie für Ihre Probleme¿')
7. **Expected improvements:** 'What would improve in your life if you no longer had your problems?' (german original question: „Was würde sich in Ihrem Leben verbessern, wenn Sie ihre Probleme nicht mehr hätten¿')
8. **Environment response:** 'How does your environment (partner, family, friends, work colleagues) react to your problems?' (german original question: „Wie reagiert Ihre Umwelt (Partner:in, Familie, Freund:innen, Arbeitskolleg:innen) auf die Probleme¿')
9. **No change required:** 'What should not change under any circumstances as a result of the therapy?' (german original question: „Was sollte sich durch die Therapie auf keinen Fall verändern¿')
10. **Problem description:** 'Finally, please describe in your own words the problems for which you would like treatment.' (german original question: „Beschreiben Sie zum Abschluss bitte noch einmal in eigenen Worten Ihre Probleme, deretwegen Sie eine Behandlung wünschen.")
11. **Impacted life areas:** 'In which areas of your life do these problems limit you (e.g. job, relationship)?' (german original question: „In welchen Lebensbereichen schränken Sie diese Probleme ein (z. B. Beruf, Partnerschaft)¿')
12. **Therapy goals:** 'What would you like to achieve for yourself in therapy?' (german original question: „Was möchten Sie in der Therapie für sich erreichen¿')

Psychometric measures

Clinical and psychometric variables were retrieved from the FBZ database and included diagnostic information, self-report symptom measures, therapist- and patient-rated outcome measures, positive mental health indicators, and therapeutic process variables. Diagnoses were coded according to DSM-5 and ICD-10 criteria. Symptom severity and treatment outcomes were assessed using a combination of standardized self-report questionnaires and clinician-rated instruments administered at different points during treatment.

Diagnosis. Diagnosis at the outpatient clinic is conducted using structured clinical interviews. These typically take place before therapy begins, usually at the fourth therapist–patient contact. The interview used is the Diagnostic Interview for Mental Disorders (Margraf et al., 2021), which covers the most frequent DSM-5 disorders encountered in outpatient therapy settings.

Beck-Depression-Inventory II.. Depressive symptoms were assessed using the *Beck Depression Inventory–II* (BDI-II; (Beck et al., 1996)), a widely used self-report questionnaire measuring the severity of depressive symptomatology over the past two weeks.

Depression Anxiety Stress Scale 42. Depressive symptoms, anxiety symptoms, and general psychological distress were assessed using the *Depression Anxiety Stress Scale–42* (DASS-42(Lovibond & Lovibond, 1995)), which consists of 42 items measuring symptoms of depression, anxiety, and stress on a 4-point likert scale.

Brief Symptom Inventory. Overall psychopathological symptom burden was measured using the *Brief Symptom Inventory* (BSI; (Franke, 2002)), the short form of the Symptom Checklist-90-Revised (SCL-90-R; Derogatis). The BSI consists of 53 items rated on a 5-point Likert scale ranging from 0 (“not at all”) to 4 (“extremely”). Responses to 49 items are assigned to nine primary symptom dimensions, while four items are evaluated separately. These symptom dimensions are summarized into three global indices: the *Global Severity Index* (GSI), reflecting overall psychological distress; the *Positive Symptom Distress Index* (PSDI), indicating symptom intensity; and the *Positive Symptom Total* (PST), representing the number of reported symptoms.

Positive Mental Health Scale. Positive mental health (PMH) was assessed with the nine-item *PMH scale* (Lukat et al., 2016). Responses are given on a 4-point Likert scale from 0 (disagree) to 3 (agree). Item scores are summed to yield a total score ranging from 0 to 27, with higher scores reflecting greater PMH. The scale has been validated as a unidimensional measure with excellent internal consistency (Cronbach's α = .93), good test–retest reliability (Pearson r = .74–.81), and evidence of scalar invariance across samples and over time (Lukat et al., 2016). Furthermore, it shows strong convergent and discriminant validity and is sensitive to therapeutic change across diverse populations (Lukat et al., 2016).

Childhood Trauma Questionnaire. Early adverse experiences were assessed using the *Childhood Trauma Questionnaire* (CTQ) (Bernstein et al., 2003), a widely used self-report instrument for the retrospective assessment of childhood maltreatment. The CTQ measures five domains of adverse experiences: emotional abuse, physical abuse, sexual abuse, emotional neglect, and physical neglect. Items are rated on a 5-point Likert scale ranging from 1 ("not at all") to 5 ("very often"), with higher scores indicating greater exposure to maltreatment. The German version of the CTQ has demonstrated good psychometric properties, including satisfactory reliability and validity in clinical samples (Wingenfeld et al., 2010).

Clinical Global Impression. Clinician-rated symptom severity and improvement were assessed using the *Clinical Global Impression* (CGI) scales.

Severity Scale. The *CGI-Severity* scale evaluates the clinician's global impression of the patient's current level of mental illness, based on their total clinical experience with this population. The item asks: *"Considering your total clinical experience with this particular population, how mentally ill is the patient at this time?"*

Improvement Scale. Treatment-related change was assessed using the *CGI-Improvement* scale. Both patients and therapists rated overall improvement relative to the beginning of therapy, regardless of whether the change was attributed entirely to treatment. Patient and therapist versions differ only in perspective but use equivalent response formats.

Global Improvement

Global therapy outcome was assessed using a six-point global success rating based on two items measuring perceived benefit and satisfaction with therapy (Michalak et al., 2003). These items were completed by both patients and therapists and capture a retrospective evaluation of treatment success. The items assess (1) the extent to which expectations toward therapy have been fulfilled and (2) the overall perceived benefit of therapy. Responses are given on a 6-point Likert scale ranging from 1 (“on the contrary / rather harmful”) to 6 (“completely / very helpful”).

Goal Attainment Scale

Individualized treatment outcomes were assessed using a goal attainment measure inspired by the Goal Attainment Scaling approach (Kiresuk & Sherman, 1968). At the beginning of therapy, patients and therapists collaboratively define individualized treatment goals. During the course and at the end of therapy, patients and therapists retrospectively evaluated the extent to which each of the predefined goals had been achieved.

Goal attainment was rated on a standardized six-point numerical scale ranging from deterioration relative to the initial goal state (-1 = moved away from the goal) to full goal attainment (4 = goal achieved). The scale reflects patients’ subjective assessment of goal progress, with intermediate categories indicating partial progress toward the respective goal.

For each patient, an overall goal attainment score was computed as the mean rating across all individually defined goals, representing the average subjective level of goal progress at the end of therapy.

Measurement time points

Preprocessing

To streamline data collection, an automated transcription pipeline was implemented: The handwritten text data is first recorded by trained employees of the FBZ adult outpatient clinic using a mobile audio recording device. Identifying features (e.g. names, dates of birth, location details) were replaced by placeholders during recording. The transcription was carried out on local hardware using the open source tool Whisper Large v2 (https://github.com/openai/whisper), a state-of-the-art speech-to-text model (Radford et al., 2022). Each recording begins with a structured introduction, including a patient identification code, followed by responses to predefined questions. The transcription pipeline automatically processes all audio recordings, extracts the patient codes, and identifies responses to key questions.

As an additional data correction step, the exported transcription table was screened for incomplete entries. Records with missing patient identification codes or without any extracted text were automatically flagged, exported for manual correction, and subsequently re-imported and merged back into the original dataset. The corrected dataset was then used for downstream analyses.

Shared analytic framework

All substudies in this project are based on the same clinical cohort and share a common analytic foundation. Specific operationalizations of text inputs, outcomes, and analytic models differ across substudies and are specified in the corresponding substudy sections below.

Data scope and unit of analysis. Across all substudies, analyses focus on pre-therapy patient narratives collected during intake. Language-based analyses are temporally ordered such that patient narratives precede all clinical outcomes of interest. Outcomes may be assessed cross-sectionally (at intake) or longitudinally (during or after therapy), depending on the substudy. The unit of analysis is the individual patient. Textual data collected during therapy are not used as predictors in any analysis.

Descriptive characterization of open-ended responses. As a shared descriptive foundation, responses to each open-ended prompt are characterized with respect to engagement and heterogeneity. Descriptive statistics include response length and an entropy-based lexical diversity index (Shannon, 1948), computed across pooled responses per question. This characterization provides a common empirical basis for interpreting prompt-specific response patterns across substudies.

Because the open-ended questions were administered in two questionnaire blocks, analyses are restricted to cases in which the respective questionnaire was present. Within these blocks, item-level nonresponse is summarized descriptively and interpreted as potentially informative of selective responding.

Preliminary results

Descriptive statistics

Demographics and context factors

Textual data

Descriptive analyses revealed substantial heterogeneity across questions in response rates, length, and lexical diversity. Holistic questions such as problem description (q10) and therapy goals (q12) showed low missingness, longer median response lengths, and high lexical diversity, indicating that patients readily produced extended and heterogeneous narratives when asked to reflect broadly on their difficulties or desired changes. In contrast, prompts targeting causal connections (q4) or constraints (q11) frequently elicited short or missing responses and exhibited lower diversity, consistent with more constrained or confirmatory response formats. Notably, lexical diversity varied independently of response length, suggesting that some prompts elicited shared narrative scripts despite moderate verbosity (e.g., q7 Expected Improvements).

Diagnosis

Diagnoses were collapsed for display. Only diagnoses with ≥10 occurrences are shown individually; remaining diagnoses are summarized as “Other diagnoses (<10)”.

psychometric measures

Due to substantial dropout and differences in treatment duration, descriptive statistics across therapy phases are based on changing subsamples. In particular, patients continuing into long-term therapy represent a more severe and chronic subgroup. To avoid conflating symptom severity with symptom change, two complementary descriptive tables are reported.

Table 1 presents descriptive statistics for symptom measures at each assessment timepoint. Sample sizes decrease substantially across therapy phases, reflecting treatment completion and dropout. Consequently, means at later timepoints do not represent longitudinal change within individuals and should not be interpreted as symptom worsening.

Table 2 reports descriptive statistics restricted to patients with complete data at both pre-treatment and post-treatment assessments. For complete-case descriptive analyses, the post-treatment endpoint was defined as the assessment at the end of the second short-term therapy phase (KZT2-DUPost), as this represents the last common assessment across the majority of patients. This table provides a descriptive approximation of within-person change, independent of selective dropout.

Substudies

This project comprises multiple complementary substudies that share a common dataset and preprocessing pipeline but address distinct research questions. To avoid analytic flexibility and to maximize interpretability, each substudy pre-specifies its text inputs, outcomes, and evaluation strategy.

Substudy 1: Language as clinical assessment (cross-sectional)

Core research question. Does patients' pre-therapy language provide clinically meaningful information about symptom severity and functioning, and does it add predictive value beyond established self-report questionnaires? Additionally, to what extent is this information captured by large language model (LLM)–based representations compared with simpler linguistic features?

Conceptual contribution. This substudy conceptualizes patient-generated language as an independent assessment modality rather than a proxy for existing symptom scales. While prior work has demonstrated correlations between language and self-reported symptoms, few studies have tested whether language captures incremental clinical information beyond standardized questionnaires, particularly for clinician-rated outcomes.

Text inputs. Q1 Problem developtment, Q10 *Problem description,* Q12 *Therapy goals*

Outcomes.

- Convergent validity outcomes: Cross-sectional self-report symptom (BDI-II, BSI/GSI, DASS-42) and well-being (PMH) measures
- Incremental validity outcomes: Clinician-rated symptom severity (CGI-Severity).

Models & evaluation. Convergent validity with self-report symptom measures (BDI-II, BSI/GSI, DASS-42) will be evaluated by estimating the proportion of symptom variance explained by language-based representations. Incremental validity will be tested by assessing whether language explains additional variance in clinician-rated symptom severity (CGI-Severity) beyond self-report questionnaires.

The Sequential Evaluation with Model Pre-registration (Kjell, Ganesan, et al., 2024) framework will be implemented to ensure robust model development, mitigating overfitting and enabling unbiased performance evaluation. Additionally, evaluating models on prospective data will simulate real-world clinical deployment by assessing performance on new, unseen patient data.

During the model development phase, preprocessing pipelines will be finalized, and exploratory models developed using advanced cross-validation techniques. Contextual embeddings derived from pretrained LLMs will be linked to clinical outcomes using state-of-the-art prediction models, including ridge regression (Hoerl & Kennard, 1970), lasso regression (Tibshirani, 1996), and random forests (Ho, 1995). The final pipelines will be pre-registered for evaluation (e.g., https://aspredicted.org/). In the evaluation phase, pre-registered models will be tested on held-out datasets, enabling unbiased performance assessments.

Substudy 2: Patient narratives and themes

Core research question. How do patients conceptualize the their mental health problems, as well as anticipated improvements, and how do these narrative patterns vary across individuals and clinical groups?

Conceptual contribution. This substudy emphasizes interpretation and theory generation, using patient language to uncover recurring narrative themes. By analyzing these themes, this substudy provides insight into how narrative content and response patterns relate to diagnostic categories, baseline severity or sociodemographic (e.g. age, gender) and anamnestic (e.g. childhood trauma) characteristics.

Text inputs. Q1 *Problem development*, Q2 *Extra stressors*, Q3 *Pre-onset changes*, Q6 *Problem causes*, Q7 *Expected improvements*, Q8 *Environment response*. Question-wise modeling.

Outcomes. No single predictive outcome is specified, as the primary aim is interpretive and theory-generating rather than predictive.

Models & evaluation. Topic modeling / embedding-based clustering; interpretive labeling; comparison of topic prevalence across patient groups.

Substudy 3: Predicting treatment response and goal attainment (longitudinal prediction)

Core research question. Can patients' pre-therapy language predict clinically meaningful treatment outcomes beyond baseline symptom severity and demographic characteristics?

Conceptual contribution. This substudy extends language-based mental health assessment from cross-sectional validity to prognostic utility. While prior work has primarily examined whether language reflects current symptom severity, far less is known about whether pre-therapy narratives encode information relevant for future treatment response, such as motivation, goal clarity, perceived agency, or narrative coherence. By evaluating the ability of pre-therapy language to predict longitudinal outcomes above and beyond baseline symptom and well-being measures, this substudy tests whether patient-generated text captures clinically actionable signals that are not accessible through standard intake questionnaires alone.

Text inputs. Primary: Q10, Q12, Q7, Q1; Secondary: concatenation models.

Outcomes.

- Clinician-rated improvement (CGI-Improvement) at 6 months-follow-up assessment.
- Global therapy outcome ratings (patient- and therapist-reported) at 6 months-follow-up assessment.
- Goal Attainment Scale at 6 months-follow-up assessment.
- Pre-Post differences in symptom (BDI-II, BSI/GSI, DASS-42) and well-being (PMH) scales .

Models & evaluation.

- Baseline prognostic models: Demographic variables and baseline symptom severity and well-being (e.g., BDI-II, BSI/GSI, PMH, CGI-Severity).
- Language-augmented models: Baseline predictors plus language representations derived from pre-therapy text.

Substudy 4: Prompt-based LLM rubrics for interpretable language assessment

Core research question. Can prompt-based large language models (LLMs) reliably extract interpretable, clinically meaningful rubric scores from pre-therapy narratives, and do these rubric-based language measures add incremental value beyond (a) standardized questionnaires and (b) embedding-based language representations?

References

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II: Beck Depression Inventory manual* (2nd ed.). Psychological Corporation.

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., Stokes, J., Handelsman, L., Medrano, M., Desmond, D., & Zule, W. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse & Neglect*, *27*(2), 169–190. https://doi.org/10.1016/s0145-2134(02)00541-0

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J., Phillips, L., Ward, A., Lynch, R., Trent, L., et al. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice*, *18*(2), 154–172.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Franke, G. (2002). *Franke, g.h. (2000). BSI. Brief symptom inventory - deutsche version. Manual. Göttingen: beltz.*

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67.

https://doi.org/10.1080/00401706.1970.10488634

Jensen-Doss, A., & Weisz, J. R. (2008). Diagnostic agreement predicts treatment process and outcomes in youth mental health clinics. *Journal of Consulting and Clinical Psychology*, *76*(5), 711–722. https://doi.org/10.1037/0022-006X.76.5.711

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal*, *4*(6), 443–453. https://doi.org/10.1007/BF01530764

Kjell, O. N. E., Ganesan, A. V., Boyd, R., Oltmanns, J. R., Rivero, A., Feltman, S., Carr, M. A., Luft, B. J., Kotov, R., & Schwartz, H. A. (2024). *Demonstrating high validity of a new AI-language assessment of PTSD: A sequential evaluation with model pre-registration*. PsyArXiv. https://doi.org/10.31234/osf.io/xw24e

Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, *24*(1), 92–115. https://doi.org/10.1037/met0000191

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Research*, *333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, *12*(1), 3918. https://doi.org/10.1038/s41598-022-07520-w

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.*, *140*(55).

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-u

Lukat, J., Margraf, J., Lutz, R., Veld, W. M. van der, & Becker, E. S. (2016). Psychometric properties of the positive mental health scale (PMH-scale). *BMC Psychology*, *4*(1), 8. https://doi.org/10.1186/s40359-016-0111-x

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-Based and Data-Informed Psychological Therapy. *Annual Review of Clinical Psychology*, *18*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Margraf, J., Cwik, J. C., Brachel, R. von, Suppiger, A., & Schneider, S. (2021). *DIPS open access 1.2: Diagnostisches interview bei psychischen störungen.* https://doi.org/10.46586/rub.172.149

Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., & Schwartz, H. A. (2019). Suicide risk assessment with multi-level dual-context language and BERT. In K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, & K. Loveys (Eds.), *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 39–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3005

Michalak, J., Kosfelder, J., Meyer, F., & Schulte, D. (2003). Messung des Therapieerfolgs. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *32*(2), 94–103. https://doi.org/10.1026/0084-5345.32.2.94

Mohammadi, E., Amini, H., & Kosseim, L. (2019). CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, & K. Loveys (Eds.), *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 34–38). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3004

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv Preprint arXiv: 2212.04356*. https://arxiv.org/abs/2212.04356

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379423.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*.

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge.

Wingenfeld, K., Spitzer, C., Mensebach, C., Grabe, H., Hill, A., Gast, U., Schlosser, N., Höpp, H., Beblo, T., & Driessen, M. (2010). Die deutsche Version des Childhood Trauma Questionnaire (CTQ): Erste Befunde zu den psychometrischen Kennwerten. *PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, *60*(08), e13–e13. https://doi.org/10.1055/s-0030-1253494

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*.

Zirikly, A., Resnik, P., Uzuner, Ö., & Hollingshead, K. (2019). CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, & K. Loveys (Eds.), *Proceedings of the sixth workshop on computational linguistics and clinical psychology* (pp. 24–33). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3003

Time-point	DU-DI	DU-Prä	KZT1-DU4	KZT1-DUPost	KZT2-DUPost	LZT1-DUPost	LZT2-DUPost	Kat6
Explanation	Pre-therapy, 4th contact	Pre-therapy, 6th contact	4th therapy session	12th therapy session	24th therapy session	45th therapy session	60th therapy session	6 month after therapy
Diagnosis	X							
Demographics		X						
BSI	X		X	X	X	X	X	X
BDI-II		X	X	X	X	X	X	X
DASS-42		X	X	X	X	X	X	X
PMH		X	X	X	X	X	X	X
CTQ		X						
CGI-S		X						
CGI-I				X	X	X	X	X
Glob-Pt				X	X	X	X	X
GAS				X	X	X	X	X

Table 1

Table X. Descriptive statistics of patient demographics and context factors.

Variable	Valid n	Missing n	Level	Mean ± SD	n (%)
Age at therapy start	586	225		41.25 ± 14.34	
Sex	586	225			
			male		239 (40.8%)
			female		347 (59.2%)
In relationship	318	493			
			yes		189 (59.4%)
			no		129 (40.6%)
Marital status	318	493			
			single		159 (50.0%)
			married		94 (29.6%)
			divorced		42 (13.2%)
			seperated		13 (4.1%)
			widowed		3 (0.9%)
			other		7 (2.2%)

Variable	Valid n	Missing n	Level	Mean ± SD	n (%)
General education	318	493			
			other		3 (0.9%)
			student		4 (1.3%)
			no school-leaving certificate		7 (2.2%)
			lower secondary school certificate		56 (17.6%)
			intermediate secondary school certificate		91 (28.6%)
			higher education entrance qualification		157 (49.4%)
Vocational qualification	318	493			
			Currently in vocational training or studying		34 (10.7%)
			No vocational qualification		37 (11.6%)
			Apprenticeship / vocational training		177 (55.7%)
			University or university of applied sciences degree		46 (14.5%)
			Other		24 (7.5%)

Variable	Valid n	Missing n	Level	Mean ± SD	n (%)
Work ability status	318	493			
			Other		26 (8.2%)
			Able to work		171 (53.8%)
			Unable to work (on sick leave)		100 (31.4%)
			Disability pension		13 (4.1%)
			Old-age pension		8 (2.5%)
Previous psychotherapy	495	316			
			no prior treatment		156 (31.5%)
			outpatient psychotherapy		82 (16.6%)
			inpatient psychotherapy		129 (26.1%)
			both		113 (22.8%)
			exact specification not available		15 (3.0%)
CGI severity	495	316			
			Not assessable		3 (0.6%)

Variable	Valid n	Missing n	Level	Mean ± SD	n (%)
			Normal, not at all ill		2 (0.4%)
			Borderline mentally ill		8 (1.6%)
			Mildly ill		20 (4.0%)
			Moderately ill		137 (27.7%)
			Markedly ill		247 (49.9%)
			Severely ill		76 (15.4%)
			Among the most extremely ill patients		2 (0.4%)

Table 2

Table X. Descriptive statistics of therapist demographics and context factors.

Variable	Valid n	Missing n	Level	Mean ± SD	n (%)
Age at therapy start	586	225		28.56 ± 4.33	
Sex	586	225			
			male		84 (14.3%)
			female		502 (85.7%)

Table 3

Table X. Descriptive statistics of open-ended responses by question.

Question	n_total	missing_pct	very_short_pct	median_words	p10_words	p90_words	diversity_index
Problem development	811	46.97904	2.959309	18	5.9	51.1	710.4708
Extra stressors	811	50.06165	9.494451	10	2.0	36.0	610.5488
Pre-onset changes	811	42.66338	15.782984	9	1.0	35.8	547.6079
Event connection	811	46.48582	24.167694	5	1.0	28.0	468.0259
Physical symptoms	811	45.99260	16.276202	6	2.0	19.0	488.0861
Problem causes	811	48.70530	11.837238	8	2.0	30.0	565.3313
Expected improvements	811	43.52651	5.425401	11	4.0	31.0	374.1694
Environment response	811	42.29346	12.823674	9	2.0	32.3	356.4073
No change required	811	69.42047	8.261406	6	2.0	19.0	273.0941
Problem description	811	11.09741	4.562269	22	6.0	50.0	672.5439
Impacted life areas	811	13.68681	31.196054	6	2.0	31.0	429.0963
Therapy goals	811	10.11097	5.548705	15	5.0	36.0	425.1176

Table 4

DSM-5 diagnoses in the sample (absolute frequencies)

DSM-5 diagnosis	All diagnoses (n)	Primary diagnoses (n)
Major Depression (all subtypes)	287	215
Persistent Depressive Disorder (Dysthymia)	109	80
Social Anxiety Disorder (Social Phobia)	86	45
Posttraumatic Stress Disorder	56	31
Panic Disorder	49	26
Agoraphobia	46	26
Adjustment Disorder with Depressed Mood	34	34
Generalized Anxiety Disorder	32	20
Borderline Personality Disorder	24	13
Obsessive-Compulsive Disorder	24	16
Somatic Symptom Disorder	24	18
Adjustment Disorder with Mixed Anxiety and Depressed Mood	17	17
Insomnia Disorder	13	3
Specific Phobia, Situational Type	13	5
Binge-Eating Disorder	12	3
Separation Anxiety Disorder	12	3
Illness Anxiety Disorder	11	4
Adjustment Disorder, Unspecified	10	8
Other diagnoses (<10)	156	69

Scale	DI	DU-Prä	KZT1-DU4	KZT1-DUPost	KZT2-DUPost
BDI					
Sum		23.48 ± 12.60			
n = 510	19.28 ± 11.31				
n = 244	16.63 ± 12.13				
n = 342	14.97 ± 12.58				
n = 264	15.52 ± 12.49				
n = 110	18.09 ± 12.75				
n = 22	13.14 ± 11.29				
n = 193					
BSI					
GSI	1.24 ± 0.70				
n = 428		1.08 ± 0.66			
n = 245	1.02 ± 0.67				
n = 343	0.85 ± 0.68				
n = 262	1.07 ± 0.74				
n = 109	1.03 ± 0.55				
n = 22	0.85 ± 0.70				
n = 199					
CGI					
improvement patient				2.38 ± 1.14	
n = 297	2.18 ± 1.08				
n = 221	1.94 ± 1.01				
n = 105	1.86 ± 0.56				
n = 22	2.27 ± 1.39				
n = 190					
improvement therapist				2.74 ± 0.79	
n = 219	2.39 ± 0.94				

n = 146	2.07 ± 0.63	
n = 60	1.93 ± 0.47	
n = 14	1.73 ± 1.30	
n = 52		
DASS		
Anxiety		12.03 ± 8.30
n = 341	8.78 ± 6.96	
n = 117	8.30 ± 7.85	
n = 224	7.37 ± 7.32	
n = 186	8.93 ± 7.96	
n = 70	6.38 ± 5.45	
n = 13	6.88 ± 7.42	
n = 193		
Depression		19.86 ± 11.17
n = 341	16.28 ± 10.20	
n = 117	13.31 ± 10.85	
n = 224	12.68 ± 11.35	
n = 186	11.99 ± 10.53	
n = 70	12.08 ± 9.81	
n = 13	11.48 ± 10.79	
n = 193		
Stress		19.37 ± 9.21
n = 341	16.11 ± 8.77	
n = 117	14.86 ± 9.34	
n = 224	12.91 ± 9.39	
n = 186	14.67 ± 10.12	
n = 70	14.15 ± 6.69	
n = 13	13.05 ± 10.04	
n = 193		

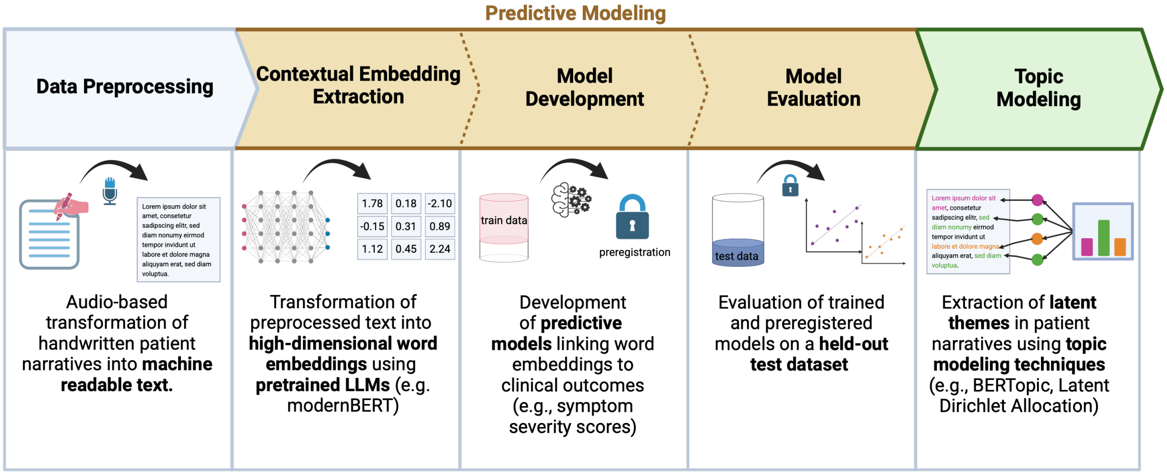
Total		51.25 ± 24.05	
n = 341	41.17 ± 22.03		
n = 117	36.48 ± 24.75		
n = 224	32.95 ± 24.95		
n = 186	35.59 ± 25.92		
n = 70	32.62 ± 19.56		
n = 13	31.40 ± 25.36		
n = 193			
GAS			
pt mean			1.40 ± 1.01
n = 239	1.98 ± 1.12		
n = 194	2.04 ± 1.07		
n = 89	2.42 ± 0.89		
n = 19	2.28 ± 1.14		
n = 82			
th mean			1.39 ± 0.98
n = 179	1.95 ± 1.06		
n = 131	2.36 ± 0.88		
n = 49	2.85 ± 0.60		
n = 11			
Global ratings			
pt benefit			4.54 ± 1.02
n = 344	4.92 ± 0.95		
n = 259	5.01 ± 0.92		
n = 108	5.41 ± 0.67		
n = 22	4.65 ± 1.19		
n = 199			
pt satisfaction			4.07 ± 1.04
n = 344	4.43 ± 1.05		

n = 259	4.42 ± 0.95		
n = 108	4.64 ± 0.73		
n = 22	4.30 ± 1.18		
n = 199			
th benefit			3.85 ± 0.93
n = 261	4.28 ± 1.02		
n = 174	4.66 ± 0.72		
n = 62	5.21 ± 0.43		
n = 14			
th satisfaction			3.62 ± 0.95
n = 261	4.02 ± 1.10		
n = 174	4.31 ± 0.92		
n = 62	4.50 ± 0.76		
n = 14			
PMH			
Sum		10.52 ± 5.89	
n = 504	11.41 ± 6.11		
n = 244	12.72 ± 6.34		
n = 341	13.82 ± 6.72		
n = 263	12.82 ± 6.06		
n = 107	11.64 ± 6.00		
n = 22	15.23 ± 6.87		
n = 199			

Scale	Baseline	Endpoint	Baseline M ± SD	Endpoint M ± SD	ΔM (End – Base)	n (c
BDI						
Sum	DU-Prä	KZT2-DUPost	22.37 ± 12.05	15.11 ± 12.58	-7.26	
BSI						
GSI	DI	KZT2-DUPost	1.19 ± 0.66	0.85 ± 0.68	-0.35	
DASS						
Anxiety	DU-Prä	KZT2-DUPost	11.90 ± 8.16	7.07 ± 7.03	-4.83	
Depression	DU-Prä	KZT2-DUPost	20.30 ± 11.69	13.07 ± 11.92	-7.23	
Stress	DU-Prä	KZT2-DUPost	19.59 ± 9.50	12.65 ± 9.25	-6.94	
Total	DU-Prä	KZT2-DUPost	51.79 ± 24.55	32.79 ± 25.10	-19.00	
PMH						
Sum	DU-Prä	KZT2-DUPost	10.47 ± 5.87	13.68 ± 6.75	3.22	

Figure 1

The step-by-step process of patient narrative analysis, from preprocessing to data evaluation.


Predictive Modeling
Data Preprocessing
Contextual Embedding Extraction
Model Development
Model Evaluation
Topic Modeling
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.
1.78
0.18
-2.10
-0.15
0.31
0.89
1.12
0.45
2.24
train data
preregistration
test data
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.
Audio-based transformation of handwritten patient narratives into machine readable text.
Transformation of preprocessed text into high-dimensional word embeddings using pretrained LLMs (e.g. modernBERT)
Development of predictive models linking word embeddings to clinical outcomes (e.g., symptom severity scores)
Evaluation of trained and preregistered models on a held-out test dataset
Extraction of latent themes in patient narratives using topic modeling techniques (e.g., BERTopic, Latent Dirichlet Allocation)