Predicting the Effectiveness of Therapy Sessions for Patients Battling Depression

Yuval Arbel
yuval.arbel1@mail.huji.ac.il

Abstract

This study investigates the potential of large language models (LLMs) to automatically assess the effectiveness of psychotherapy sessions for individuals battling depression. Utilizing the "Depressometer" dataset, comprising transcribed therapy session dialogues and patient-reported outcome measures, we finetuned a quantized Llama2 model to predict session impact based on changes in the Outcome Rating Scale (ORS) scores (Miller et al., 2003). Despite challenges related to quantization limitations and GPU resource constraints, this study acted as a pioneer for future investigations in the lab and demonstrated the feasibility for future research in leveraging LLMs for mental health applications. This research contributes to the growing field of AI-assisted mental health care, aiming to develop automated tools that enhance therapy effectiveness and improve patient outcomes.[1]

1 Introduction and Related Work

In recent years, the integration of artificial intelligence (AI) in mental health research has gained significant attention, particularly in the automated analysis of therapeutic processes and the detection of depression and other psychopathologies through text (Delgadillo and Atzil-Slonim, 2022; Le Glaz et al., 2021; Aafjes-van Doorn et al., 2021). Existing studies, such as (Flemotomos et al., 2018), have explored various methodologies for assessing the effectiveness of therapy sessions, often relying on post-session self-reports from patients. However, these approaches face challenges such as patient burden and potential inaccuracies due to the subjective nature of self-reports. Leveraging large language models (LLMs) for automatic analysis offers a promising alternative, potentially reducing the reliance on self-reports while providing continuous monitoring of patient well-being.

This research distinguishes itself from existing studies in several ways. Firstly, it utilizes a longitudinal dataset, allowing us to track changes in patient states throughout the course of their therapy sessions. Secondly, unlike previous studies that used older machine learning methods for detection (Althoff et al., 2016), this research leverages LLMs, known for their ability to process and analyze complex language patterns. Lastly, this study uniquely focuses on predicting the difference in patient well-being between the beginning and end of each session, capturing the immediate impact of therapeutic interventions.

Building on existing work, this study applies a finetuned LLM to predict the effectiveness of therapy sessions for patients battling depression. Specifically, it utilizes the Depressometer dataset, consisting of therapy session manually transcribed from their Hebrew source, and automatically translated into English. The model assesses session impact based on changes in the Outcome Rating Scale (ORS) (Miller et al., 2003) scores provided by patients before and after the session. This approach advances AI in mental health by focusing on session-level impact assessment using natural language processing (NLP) techniques, addressing gaps in current literature.

2 Motivation and Intentions

The primary motivation behind this research is to enhance the automated identification and monitoring of mental health states, particularly in the context of therapy for depression. Traditional methods of assessing the effectiveness of therapy sessions often rely on patients filling out detailed questionnaires, a process that can be burdensome and may not capture real-time fluctuations in mental states. While this self-reporting is applicable in laboratory settings, it is not a feasible long-term solution. By automating the evaluation process using session transcriptions, this research aims to allevi-

[1] Code: https://github.com/schwartz-lab-NLP/psychology_predictions

ate the burden on patients while providing more timely and accurate feedback to therapists. This, in turn, can support early identification of deterioration in mental health, enabling more effective and timely interventions. It is especially crucial given that therapists often overestimate their clients' emotional states (Bar-Kalifa et al., 2016), highlighting the need for objective and data-driven assessment tools.

Moreover, the study seeks to explore the potential for LLMs to identify indicators within therapy session dialogues that correlate with positive or negative outcomes, similar to how topic modeling has been used (Atzil-Slonim et al., 2021; Gaut et al., 2015; Lin et al., 2023a). If successful, this approach could lead to the development of more explainable models, offering insights into what makes a therapy session effective. Such advancements could not only improve patient care but also enhance the ability of therapists to fine-tune their interventions based on data-driven feedback (Lin et al., 2022), ultimately contributing to better long-term outcomes for individuals struggling with depression.

3 Pre-Research Process

3.1 Background Research, and Finding the Research Question

At the outset of my research journey, I was driven by the motivation to create a machine learning tool applicable within the field of psychotherapy, moving toward what could be considered an "artificial psychologist." My goals centered around the themes of affordability, accessibility, and scalability, aiming to broaden the reach of psychological treatments without compromising quality.

As I explored this domain, I began by surveying the existing literature, creating a comprehensive table of 192 articles that spanned various relevant topics, including the intersection of AI & ML with psychology, therapy, emotion, pragmatics/linguistics, and clinical hypnosis. These included key articles from domain experts like Prof. Baihan Lin (Gunal et al., 2024; Lin et al., 2022, 2023b,a) and Prof. Zac Imel (Gaut et al., 2015; Kuo et al., 2024). From this extensive list, I filtered down to 27 of the most promising articles to delve into further.

Simultaneously, I brainstormed potential tools that could assist therapists, such as systems for session transcription, emotion analysis, session summarization, and monitoring of patients' psychological states between sessions. More advanced ideas included predicting future states and therapy outcomes, monitoring emotional regression, and linking emotions to physiological signs.

Throughout this process, several intriguing research questions emerged, including:

- Predicting Attachment Style, rooted in attachment theory (Bowlby, 1969).
- Predicting the Four Temperaments (Merenda, 1987).
- Predicting the Big 5 Personality Traits (Tupes and Christal, 1992; Goldberg, 1993; McCrae and Costa, 1987).
- Monitoring temporal emotional changes throughout therapy sessions and correlating these changes with specific interventions and key moments within the sessions.
- Investigating style transfer from one therapeutic approach to another, such as from Cognitive Behavioral Therapy (CBT) to Psychoanalytic or Internal Family Systems (IFS) (Schwartz and Sweezy, 2019) therapies.
- Predicting emotional regression between therapy sessions.
- Predicting the physiological impact of speech turns (from either the patient or therapist) on the patient.

While these questions provided a broad range of potential research directions, the next phase involved accessing the right data to refine and focus the research question.

3.2 Accessing Data

My focus at this stage was on acquiring a dataset related to psychology treatments that would be suitable for training models, preferably one that was tagged for easier analysis. The first significant dataset I encountered during my literature review was the "Alexander Street Counseling and Psychotherapy Transcripts dataset" (Street, a,b). This resource appeared promising, and I made several attempts to access it online through various channels. However, I was informed that "As is common with aggregated products, we no longer have rights to sell this content." Despite reaching out to Prof. Baihan Lin and Prof. Zac Imel, who

had worked with similar datasets, I was unable to secure access.

Realizing the need for an alternative, I shifted my strategy toward establishing collaborations with psychology labs that might have access to relevant data. I first contacted Dr. Sharon Ziv Beiman at the Mifrasim Institute at the Academic College of Tel Aviv-Yafo (MTA), but unfortunately, they could not assist in this endeavor.

I then reached out to Prof. Anat Brunstein Klomek and Dr. Yael Apter-Levi, head of the Psychology Clinic at Reichman University. They expressed interest in a research collaboration between our institutions. However, the collaboration could not move forward as they were still in the process of obtaining Helsinki approval, which was necessary for accessing their data.

Next, I contacted Prof. Orya Tishby from the Psychology Department at the Hebrew University of Jerusalem (HUJI). Working with her and her PhD students, Shiri Moshe and Gershom Gwer, I was introduced to their dataset, which included short-term dynamic psychotherapy treatments. Notably, their dataset was tagged with each patient's attachment style, defined by a combination of two features: anxiety and avoidance.

I then met with Prof. Shir Atzil, also from HUJI's Psychology Department. After meeting with Monia Masalha, a PhD student in her lab, I was introduced to another intriguing dataset. This dataset involved 5-minute speed dates between student couples and included rich data such as video and audio recordings, valence-arousal tagging (where valence was primarily determined from facial expressions, and arousal from body and facial movements), and physiological measurements like skin electrical conductance (electrodermal activity), heart rate, blood pressure, body temperature, and self-reported tags from participants (including levels of romantic interest, sexual interest, desire to kiss, desire to meet again, etc.).

Despite exploring these various options, the challenge of accessing a suitable dataset for my research persisted until I connected with Prof. Dana Atzil's lab, which provided the comprehensive "Depressometer" dataset that became the foundation for this study.

3.3 Prof. Dana Atzil's Research Lab

I then met with Prof. Dana Atzil from the Psychology Department at Bar Ilan University. Working with PhD students Amir Eliassaf and Ayal Klein, I was introduced to their comprehensive Depressometer dataset.

The "Depressometer" method is a comprehensive research approach aimed at understanding the mechanisms of change in the treatment of Major Depressive Disorder (MDD) (Otte et al., 2016). This dataset was collected over a period from October 2018 to February 2021 at the Psychotherapy Research Lab, Bar Ilan University, Israel. It involved the collection of multi-modal data—linguistic, vocal, facial, physiological, hormonal, immunological, clinician assessments, and self-reports—repeatedly during short-term psychodynamic psychotherapy and at follow-ups.

Participants

- **Clients:** 62 individuals diagnosed with MDD, aged between 21 and 61, with the majority being women. They were included based on specific diagnostic criteria (e.g., a primary diagnosis of MDD and a Hamilton Rating Scale for Depression (Hamilton, 1960) score of 14 or more) and excluded if they presented conditions like active suicidality, psychosis, or certain physical health issues.

- **Therapists:** Nine therapists (5 men, 4 women) participated, treating 1-4 clients each. They were advanced trainees with 3-7 years of experience, supervised by senior experts.

Procedure

- **Recruitment:** Clients were recruited via social media, with 531 inquiries leading to 128 intake interviews, of which 62 participants were selected.

- **Therapy:** Clients received 16 sessions of supportive-expressive psychodynamic psychotherapy, which included both supportive (e.g., empathic validation) and expressive (e.g., interpretation) techniques. Sessions were recorded using multiple microphones and cameras.

- **Data Collection:** Physiological data, such as ECG and EDA, and saliva samples were collected during five specific sessions (called "Golden Sessions"). Pre, post, and follow-up assessments included clinical interviews and self-report questionnaires.

- **Transcription and Translation:** The therapy sessions were manually transcribed in Hebrew by transcribers who also added timestamp anchors, annotations, and anonymizations to protect client confidentiality. These transcriptions were then divided into individual speech turns, which were subsequently randomized and translated into English using the Google Translate API. This random scrambling of the speech turns effectively removed any coherent thread that could allow an external reader to understand the full context of a highly sensitive therapy session.

Measures

Various standardized instruments were used to assess the clients' psychological and emotional states throughout the treatment, including:

- **Hamilton Rating Scale for Depression (HRS-D):** (Hamilton, 1960) Assesses depression severity.
- **Beck Depression Inventory-II (BDI-II):** (Beck et al., 1961) Measures the severity of depressive symptoms.
- **Outcome Questionnaire 45 (OQ-45):** (Burlingame and Lambert, 2023) Captures functioning in domains like subjective discomfort and social role functioning.
- **Emotion Regulation Questionnaire (ERQ):** (Gross and John, 2003) Assesses cognitive reappraisal and expressive suppression strategies.
- **Difficulties in Emotion Regulation Scale (DERS):** (Gratz and Roemer, 2004) Evaluates problems with emotion regulation.
- **Personality Inventory for DSM-5 Brief Form (PID-5-BF):** (Krueger et al., 2012) Assesses personality traits.
- **Satisfaction With Life Scale (SWLS):** (Diener et al., 1985) Measures global life satisfaction.
- **Profile of Mood States (POMS):** (McNair, 1971) Assesses mood variables.

In addition to these instruments, pre- and post-session self-report questionnaires, such as the Outcome Rating Scale (ORS) (Miller et al., 2003) and the Hopkins Symptom Checklist-short form (HSCL-11) (Parloff et al., 1954), were used to track changes session by session.

My analysis specifically utilized the English translations of the session transcriptions and the self-reported ORS scores collected before and after each therapy session.

This method provides a rich dataset that can help explore the complex interplay of various factors in the treatment of depression, allowing researchers to gain insights into the therapeutic process and outcomes.

4 Methodology

4.1 Preprocessing

The transcription data used in this study was initially presented as tabular data, with each row representing a speech turn by either the patient or the therapist. Additionally, rows marked as TIMESTAMP were used by transcribers to anchor specific moments within the treatment video. The first step in preprocessing involved correcting various human errors in the data, such as removing duplicated sections of rows that occurred across several sessions. I then excluded the first and last sessions of each patient's treatment process, as these sessions tend to be more introductory or summatory and are not representative of the typical therapy sessions analyzed in this study.

Next, I extracted the TIMESTAMP rows into a separate column, ensuring each speech turn was associated with an approximate timestamp by filling the values forward and backward. To standardize the session data, I adjusted all timestamps so that each session began at 00:00. The next step included correcting any timestamp anomalies, such as non-monotonically increasing timestamps due to human error. These corrections enabled accurate segmentation of the sessions based on time. For further analysis, I extracted the "working times" of each session—the first 15 minutes and the last 20 minutes—to shorten the context length for model input, excluding sessions with a total duration of less than 15 minutes.

In the subsequent step, I addressed the issue of non-verbal and redacted speech annotations that were originally included in the Hebrew transcriptions to anonymize personal details and capture non-verbal cues (e.g., '<location A>', '<crying>', '<sighing>'). These annotations were scrambled or altered during the translation process, compli-

Annotation	Replaced String
cry	(crying)
laugh	(laughing)
character	them
figure	them
location	this place
place	this place
fill	hmm
commotion	mm-hmm
hum	mm-hmm
sigh	(sigh)
cluck	(cluck)
shudder	(shudder)
silence	(silence)
q	...

Table 1: Annotated tags and their replacements. All tags enclosed in angle brackets that included the *Annotation* as a substring were replaced with the corresponding *Replaced String*. For example, variations such as <cry>, <crying>, <to cry>, <is crying>, and similar were all standardized to (crying). After these replacements, any remaining tags or annotations were removed from the text.

Speaker	Text Template
Therapist	T: "*{text}*"
Client	C: "*{text}*"
Annotator	Annotation: *{text}*

(a) Templates used for each speech turn line, based on the speaker.

```
**Task: Binary Classification**
**Classes: Yes/No**
**Input Text Description:
  Part of a psychology session transcript,
  with the therapist (T:) and the client (C:)
  speech turns on alternating lines**
**Input Text: {full_session_text}**
**Question:
  Did the session improve the client's wellbeing?**
**Answer (Yes/No):
```

(b) Template used to create the final prompt, including the session speech turns and the LLM objective.

Figure 1: Templates used to format the session text.

	Training	Validation	Test
Clients #	43	7	7
Clients %	75.4	12.3	12.3
Sessions #	251	44	43
Sessions %	74.3	13.0	12.7
Tag Averages	0.1076	0.0909	0.1628

Table 2: Sizes and tag averages of the train, val, and test sets.

cating their accurate preprocessing. Using regular expressions, I standardized these annotations by replacing them with consistent terms (as detailed in Table 1), such as '(crying)' for all variations of crying annotations. Although I considered cross-referencing the original Hebrew transcriptions to restore these annotations accurately, inconsistencies and scrambling led me to rely on regular expression replacement.

Finally, I annotated each session with a tag derived from the SBS (Session-By-Session) questionnaire, representing the difference between the client's ORS (Outcome Rating Scale) score before and after the session. Sessions with an ORS difference greater than 1 were tagged as "1" (indicating a positive improvement), while those with a difference less than 0 were tagged as "-1" (indicating a negative outcome). Sessions with a difference between 0 and 1 were excluded from the analysis due to insufficient impact. The processed speech turns were unified into a single text block per session, using standardized templates for therapist and client speech (Figure 1). The final dataset contained individual session transcriptions, with each row corresponding to one session's working time and tagged with either -1 or 1.

For dataset splitting, the sessions were divided into training (75%), validation (12.5%), and test (12.5%) sets. The splitting was done at the client level, meaning all sessions from a single client were confined to one set, which helped prevent overfitting to specific clients. However, this method did not prevent potential overfitting to specific therapists. After this client-based division, the sets were balanced by session count and by the average tag distribution (positive and negative tags) to ensure similar representation across the training, validation, and test sets. This was achieved by using a constant random seed (chosen through trial and error). The resulting balanced sets are detailed in Table 2.

4.2 Baseline Model and Assessment

For the baseline model in this study, I utilized a quantized version of Meta's open-source Llama2 (Touvron et al., 2023) model, accessed via the Hugging Face library, and employing 4-bit Bits-and-Bytes quantization. This allowed for efficient load-

ing and inference of the model on an on-premises GPU hosted by the Data-Science Institution (DSI) at Bar-Ilan University.

Prior to feeding the session data into the model, I performed an additional preprocessing step to accommodate the model's token limit. Whenever the token count for a session prompt exceeded the model's accepted window size, I progressively shortened the prompt by removing entire speech turns, starting from the last one, until the prompt fit within the token limit.

Given that Llama2 is a textual generation model capable of outputting any token sequence, I utilized a method to extract binary Yes/No predictions for the assessment of therapy sessions. Instead of relying on the model to generate free text outputs, I focused on the logits—specifically, the 32,000-dimensional vector representing the output probabilities for each token. To determine the Yes/No response, I isolated the logits corresponding to six specific tokens (three tokens each for "Yes" and "No," accounting for capitalization and word segmentation). The token with the highest weight among these six was selected as the model's binary prediction for the session.

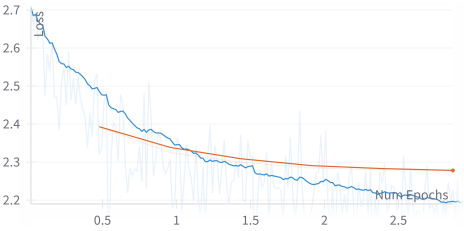
Executing this stage of the study posed certain challenges, particularly in terms of GPU resource management. The GPUs within Bar-Ilan's DSI server group are a limited and highly demanded resource, leading to frequent "GPU Out of Memory" errors during execution. To overcome this issue, I developed code to identify and use the specific GPU with the most available memory space at the time of execution, ensuring that the model could run without interruptions.

Inference on the validation set using this baseline model yielded results in which all predictions were uniformly positive, assigning a "1" to every session. This lack of prediction diversity highlighted the need for further fine-tuning to improve the model's performance and produce more varied and accurate outcomes.

4.3 Training and Testing

To fine-tune the quantized Llama2 (Touvron et al., 2023) model on the training set, I employed QLoRA (Quantized Low Rank Adaptation) (Dettmers et al., 2024) finetuning using Hugging Face's transformers (Wolf et al., 2020), PEFT (Parameter-Efficient Fine-Tuning) (Balne et al., 2024), TRL (Transformer Reinforcement Learning) (von Werra et al., 2020), and accelerate (Gugger et al., 2022) packages. After three epochs, the model began to overfit, as evidenced by the training and validation loss curves shown in Figure 2. Thus, I chose to limit the training to three epochs.

(a) Eval loss demonstrating overfitting as training exceeded three epochs.



(b) Loss curves of the training (*blue*) and validation (*orange*) examples during the three epochs of training.

Figure 2: Loss curves recorded with Weights and Biases during training.

This phase presented several challenges. First, the large language model combined with the extensive context size of the therapy sessions made GPU memory a scarce resource. I was constrained to a maximum batch size of 1, and even a single 32 GB GPU proved insufficient. Additionally, due to the confidential nature of the data, I was limited to using on-premises GPUs hosted by Bar Ilan's DSI, where multiple users shared GPU resources without any resource reservation system in place.

After persistent attempts to secure sufficient GPU resources, I was informed that DSI had implemented RunAI (Geller et al., 2018) to manage and allocate GPU resources among researchers. However, no one in Prof. Atzil's lab had prior experience with RunAI, and it took several weeks of trial and error before I could successfully run my code through the system. This involved creating and uploading a custom Docker image with all necessary requirements to the public Docker Hub, and connecting to volumes mounted on local paths within the DSI network to access the confidential dataset, custom code, and the LLM model being fine-tuned. Even after successfully setting up the

RunAI system, I continued to encounter "GPU Out of Memory" errors. Upon investigation, I discovered that RunAI does not exclusively allocate GPU resources, allowing users physically logged into the servers to use the GPUs despite RunAI's allocations. This made the GPU allocations unreliable and ultimately ineffective for my needs. I eventually managed to train the model by waiting for an opportune moment when three GPUs were simultaneously available and free from interference.

Despite successfully completing the training phase, I faced a critical issue during the evaluation and testing of the fine-tuned model. My inference strategy relied on analyzing the logits of the six relevant Yes/No tokens to determine binary predictions. However, the 4-bit quantization process had reduced the logits' precision to just 16 possible values. This caused near-zero logits to collapse to a value of 0, making it impossible to accurately determine the highest value among the six tokens. With the increasing demand for on-premise GPUs, my ability to continue testing through trial and error was severely constrained, forcing me to halt the research prematurely.

5 Challenges and Limitations

While the present study aimed to develop a robust LLM-based system for predicting therapy session effectiveness, several challenges and limitations arose during the research process, ultimately impacting the ability to generate comprehensive results.

5.1 Quantization Limitations and Result Interpretation

The primary challenge encountered involved the limitations imposed by the 4-bit quantization of the Llama2 model. As mentioned above, although quantization allowed for efficient use of limited GPU resources, it significantly reduced the precision of the logits used for binary prediction. This reduction rendered the chosen inference strategy ineffective, as near-zero logits collapsed to a value of 0, making it impossible to accurately distinguish the highest-weighted token among the Yes/No options. This unexpected outcome highlights the critical need to carefully consider the trade-off between computational efficiency and prediction accuracy when working with quantized models, especially when fine-grained analysis of logits is required.

5.2 Resource Constraints and Future Methodological Considerations

Furthermore, access to sufficient and reliable GPU resources proved to be a recurring obstacle throughout the study. The shared nature of the on-premises GPUs, coupled with the absence of a robust resource allocation system, led to frequent interruptions and delays. While the implementation of RunAI offered a potential solution, the system's inability to guarantee exclusive GPU allocation ultimately rendered it ineffective for this research. Future work would greatly benefit from dedicated access to powerful GPU resources or a more reliable allocation system to ensure uninterrupted training and evaluation.

Additionally, the limited batch size imposed by the model's memory requirements may have impacted the fine-tuning process. Exploring alternative model architectures or optimization techniques that allow for larger batch sizes could potentially lead to improved performance.

6 Future Directions

Enhancing Model Accuracy and Explainability

Despite these challenges, this research lays the groundwork for future investigations in applying LLMs to assess therapy session effectiveness. Several promising directions for future work emerge from the limitations encountered:

1. **Exploring Alternative Models and Hyperparameter Tuning:** Future research should explore the performance of different LLM architectures and sizes, including those with larger context windows to accommodate more extensive session data. Rigorous hyperparameter tuning, combined with alternative quantization methods or higher-precision quantization levels, could further enhance model accuracy and address the limitations encountered with logit analysis.

2. **Enhancing Explainability Through Intervention Analysis:** A crucial next step is to incorporate mechanisms for model explainability. This could involve training the model to not only predict session effectiveness but also to identify and highlight specific therapist interventions or dialogue segments that contributed to the positive or negative outcome.

Such insights could prove invaluable for therapist supervision, training, and the development of more targeted interventions.

3. **Incorporating Multimodal Data for Comprehensive Assessment:** The current study focused solely on textual data from session transcripts. Future research should leverage the richness of the multimodal "Depressometer" dataset by incorporating audio and video data. Analyzing vocal tone, facial expressions, and other nonverbal cues alongside the dialogue could provide the model with a more nuanced understanding of the therapeutic interaction, potentially leading to more accurate and insightful predictions.

By addressing these limitations and exploring the proposed future directions, this line of research holds significant promise. It highlights the ability to leverage the power of LLMs to revolutionize mental health care by providing automated, data-driven insights that enhance therapy effectiveness and improve patient outcomes.

References

Katie Aafjes-van Doorn, Céline Kamsteeg, Jordan Bate, and Marc Aafjes. 2021. A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, 31(1):92–116.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Dana Atzil-Slonim, Daniel Juravski, Eran Bar-Kalifa, Eva Gilboa-Schechtman, Rivka Tuval-Mashiach, Natalie Shapira, and Yoav Goldberg. 2021. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. *Psychotherapy*, 58(2):324.

Charith Chandra Sai Balne, Sreyoshi Bhaduri, Tamoghna Roy, Vinija Jain, and Aman Chadha. 2024. Parameter efficient fine tuning: A comprehensive analysis across applications. *arXiv preprint arXiv:2404.13506*.

Eran Bar-Kalifa, Dana Atzil-Slonim, Eshkol Rafaeli, Tuvia Peri, Julian Rubel, and Wolfgang Lutz. 2016. Therapist–client agreement in assessments of clients' functioning. *Journal of Consulting and Clinical Psychology*, 84(12):1127.

Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. 1961. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

John Bowlby. 1969. *Attachment and loss*. 79. Random House.

Gary M Burlingame and Michael J Lambert. 2023. Outcome questionnaire 45, second version (oq-45.2).

Jaime Delgadillo and Dana Atzil-Slonim. 2022. Artificial intelligence, machine learning and mental health. *eprints.whiterose.ac.uk*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

ED Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. 1985. The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75.

Nikolaos Flemotomos, Victor R Martinez, James Gibson, David C Atkins, Torrey A Creed, and Shrikanth S Narayanan. 2018. Language features for automated evaluation of cognitive behavior psychotherapy sessions. In *Interspeech*, pages 1908–1912.

Garren Gaut, Mark Steyvers, Zac E Imel, David C Atkins, and Padhraic Smyth. 2015. Content coding of psychotherapy transcripts using labeled topic models. *IEEE journal of biomedical and health informatics*, 21(2):476–487.

Omri Geller, Ronen Dar, and Rungtiva Sanprapa. 2018. Run:ai - ai optimization and orchestration.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Kim L Gratz and Lizabeth Roemer. 2004. Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of psychopathology and behavioral assessment*, 26:41–54.

James J Gross and Oliver P John. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Aylin Gunal, Baihan Lin, and Djallel Bouneffouf. 2024. Conversational topic recommendation in counseling and psychotherapy with decision transformer and large language models. *arXiv preprint arXiv:2405.05060*.

Max Hamilton. 1960. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry*, 23(1):56.

Robert F Krueger, Jaime Derringer, Kristian E Markon, David Watson, and Andrew E Skodol. 2012. Initial construction of a maladaptive personality trait model and inventory for dsm-5. *Psychological medicine*, 42(9):1879–1890.

Patty B Kuo, Michael J Tanana, Simon B Goldberg, Derek D Caperton, Shrikanth Narayanan, David C Atkins, and Zac E Imel. 2024. Machine-learning-based prediction of client distress from session recordings. *Clinical Psychological Science*, 12(3):435–446.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Baihan Lin, Djallel Bouneffouf, Guillermo Cecchi, and Ravi Tejwani. 2023a. Neural topic modeling of psychotherapy sessions. In *International workshop on health intelligence*, pages 209–219. Springer.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2022. Supervisorbot: Nlp-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning. *arXiv preprint arXiv:2208.13077*.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2023b. Deep annotation of therapeutic working alliance in psychotherapy. In *International workshop on health intelligence*, pages 193–207. Springer.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

DM McNair. 1971. Manual for the profile of mood states. *Educational and Industrial Testing Services*.

Peter F Merenda. 1987. Toward a four-factor theory of temperament and/or personality. *Journal of personality assessment*, 51(3):367–374.

Scott D Miller, BL Duncan, Jeb Brown, JA Sparks, and DA Claud. 2003. The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of brief Therapy*, 2(2):91–100.

Christian Otte, Stefan M Gold, Brenda W Penninx, Carmine M Pariante, Amit Etkin, Maurizio Fava, David C Mohr, and Alan F Schatzberg. 2016. Major depressive disorder. *Nature reviews Disease primers*, 2(1):1–20.

Morris B Parloff, Herbert C Kelman, and Jerome D Frank. 1954. Comfort, effectiveness, and self-awareness as criteria of improvement in psychotherapy. *American Journal of Psychiatry*, 111(5):343–352.

Richard C Schwartz and Martha Sweezy. 2019. *Internal family systems therapy*. Guilford Publications.

Alexander Street. a. Alexander street counseling and psychotherapy transcripts dataset - legacy url.

Alexander Street. b. Alexander street counseling and psychotherapy transcripts dataset: Volume 1.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ernest C Tupes and Raymond E Christal. 1992. Recurrent personality factors based on trait ratings. *Journal of personality*, 60(2):225–251.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.