JAMIA Open, 2025, **8(6)**, ooaf127
https://doi.org/10.1093/jamiaopen/ooaf127
Research and Applications

Research and Applications

Automated classification of exposure and encourage events in speech data from pediatric OCD treatment

Juan Antonio Lossio-Ventura, PhD*[,1], Samuel Frank, MS[2], Grace Ringlein, BS[2], Kirsten Bonson, PhD[3], Ardyn Olszko, BS[3], Abbey Knobel, BS[3], Daniel S. Pine, MD, PhD[2], Jennifer B. Freeman, PhD[4], Kristen Benito, PhD[4], David C. Jangraw, PhD[2,3], Francisco Pereira, PhD[1]

[1]Machine Learning Core, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, United States, [2]Emotion and Development Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, United States, [3]Department of Electrical and Biomedical Engineering, University of Vermont, Burlington, VT 05405, United States, [4]Psychiatry and Human Behavior, Warren Alpert Medical School, Brown University, East Providence, RI 02915, United States

*Corresponding author: Juan Antonio Lossio-Ventura, PhD, Machine Learning Core, National Institute of Mental Health, National Institutes of Health, 10 Center Dr, Suite 3D41, Bethesda, MD 20892, United States (juan.lossio@nih.gov)

Abstract

Objective: To develop and evaluate an automated classification system for labeling Exposure Process Coding System (EPCS) quality codes—specifically exposure and encourage events—during in-person exposure therapy sessions using automatic speech recognition (ASR) and natural language processing techniques.

Materials and Methods: The system was trained and tested on 360 manually labeled pediatric Obsessive-Compulsive Disorder (OCD) therapy sessions from 3 clinical trials. Audio recordings were transcribed using ASR tools (OpenAI's Whisper and Google Speech-to-Text). Transcription accuracy was evaluated via word error rate (WER) on manual transcriptions of 2-minute audio segments compared against ASR-generated transcripts. The resulting text was analyzed with transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT), Sentence-BERT, and Meta Llama 3. Models were trained to predict EPCS codes in 2 classification settings: sequence-level classification, where events are labeled in delimited text chunks, and token-level classification, where event boundaries are unknown. Classification was performed either with fine-tuned transformer-based models, or with logistic regression on embeddings produced by each model.

Results: With respect to transcription accuracy, Whisper outperformed Google Speech-to-Text with a lower WER (0.31 vs 0.51). For sequence classification setting, Llama 3 models achieved high performance with area under the ROC curve (AUC) scores of 0.95 for exposures and 0.75 for encourage events, outperforming traditional methods and standard BERT models. In the token-level setting, fine-tuned BERT models performed best, achieving AUC scores of 0.85 for exposures and 0.75 for encourage events.

Discussion and Conclusion: Current ASR and transformer-based models enable automated quality coding of in-person exposure therapy sessions. These findings demonstrate potential for real-time assessment in clinical practice and scalable research on effective therapy methods. Future work should focus on optimization, including improvements in ASR accuracy, expanding training datasets, and multimodal data integration.

Lay Summary

In this study, we developed an automated system to label important therapy events, such as exposures and encourage, during pediatric Obsessive-Compulsive Disorder (OCD) treatment sessions. The system combined automatic speech recognition (ASR) with modern natural language processing tools. We analyzed 360 therapy sessions from 3 clinical trials. Audio recordings were transcribed using ASR software (OpenAI's Whisper and Google Speech-to-Text), and transcription accuracy was measured. OpenAI's Whisper produced more accurate transcripts than Google Speech-to-Text, with fewer errors. The transcriptions were then processed with advanced machine learning models, including BERT, SBERT, and Meta Llama 3, to predict therapy events. Two approaches were tested: one where events were labeled in text chunks, and another where the system had to detect event boundaries on its own. In the chunked setting, Llama 3 achieved the strongest results, especially for identifying exposure events. In the boundary-detection setting, fine-tuned BERT models performed best. Overall, the system showed that current ASR and transformer-based models can reliably code therapy quality. This approach could support real-time feedback for clinicians and help scale research on effective therapy practices. Future improvements may come from better transcription tools, larger training datasets, and combining multiple data sources.

Key words: cognitive behavioral therapy; anxiety; obsessive-compulsive disorder; natural language processing; transformer models; automatic speech recognition; large language models.

Introduction

Exposure therapy, a form of Cognitive Behavioral Therapy (CBT), is a first-line treatment for Anxiety and Related Disorders (ARDs) such as Obsessive-Compulsive Disorder (OCD) and Post-Traumatic Stress Disorder (PTSD),[1–3] yet its use in real-world settings is limited and often inconsistent.[4–6] Scalable validated tools for assessing exposure delivery are lacking, limiting quality monitoring in both clinical care and research.[7–10] To address this gap, we previously developed and validated the Exposure Process Coding System (EPCS),

Received: August 7, 2025; **Revised:** October 1, 2025; **Accepted:** October 2, 2025

an observer-rated framework for identifying therapist and patient behaviors during in-session exposure tasks.[11–13] EPCS demonstrated strong inter-rater reliability and predictive validity, providing a structured alternative to clinician self-report for measuring exposure quality.

Despite the promise of this approach, EPCS is resource-intensive and infeasible for use in practice settings. The presence of exposures itself is often challenging to assess, requiring considerable training for the people coding it. Although EPCS is particularly burdensome by virtue of its micro-analytic approach, this level of detail may be valuable for capturing the complex, dynamic processes unfolding in psychotherapy. As such, detailed but low-burden methods may be particularly important for assessing treatment quality.

Recent advances in Artificial Intelligence (AI), including Natural Language Processing (NLP), are particularly well-suited for handling moment-to-moment data, and have strong potential to reduce the burden of treatment delivery measurement. NLP enables the quantitative analysis at scale of unstructured text, capturing clinically significant linguistic features by transforming words into numerical and graphical representations. NLP capabilities have been increased by advances in deep learning, in particular the emergence of Transformer-based models like BERT,[14] RoBERTa,[15] DistilBERT,[16] as well as sequence-to-sequence models, longer document architectures, and Large Language Models (LLMs) such as Open AI's Generative Pre-trained Transformer (GPT)[17] or Meta's Large Language Model Meta AI (Llama),[18,19] among others. These models can be used as the basis to develop models capable of performing tasks such as intervention revision, session summarization, or data augmentation.

A handful of studies have already used NLP to detect aspects of psychotherapy delivery.[20–23] In the largest study to date, Ewbank et al. used deep learning methods to label broad CBT delivery categories in online, text message-based CBT sessions.[24] Leveraging the increased statistical power of these large-scale automated labels, the quantities of specific techniques deployed were linked directly with outcomes. Results demonstrated that therapist utterances could be classified into delivery categories that predicted clinical results. Although these findings are promising, most psychotherapy sessions do not occur through text messaging, as these did; this is a major challenge for existing approaches. Additionally, most work in this area has focused on validation of NLP methods against globally-rated general CBT delivery factors (ie, using a Likert-type scale at the session level). While these features are valuable, they might miss specific, important process-related events, such as the presence of an exposure.

The goal of this study is to develop a set of NLP methods designed to automatically capture 2 key exposure delivery techniques, using audio recordings from in-person, exposure-based CBT sessions. Specifically, we use transformer-based models and LLMs to accurately predict moment-to-moment presence of exposure tasks, and therapist use of techniques that encourage patient approach behavior within an exposure task. These can be used to calculate the amount of exposure delivered and therapist time encouraging approach behavior, respectively. Each of these delivery features has been a strong predictor of clinical outcomes in prior studies.[13,25] We hypothesize that NLP methods will reliably detect exposure events and a within-exposure therapist technique (encouraging patients to approach the object of exposure), as compared with EPCS data from human coders. If this automation effort is successful for these 2 EPCS codes, the technique could be generalized to encompass all EPCS codes defined by Benito et al. 2012.[11] We focused on open-source transformer-based models and LLMs deployable on our own servers, which avoids concerns related to protected health information (PHI) or personally identifiable information (PII).

Materials

Dataset

Audio recordings from therapy sessions of 116 youth with OCD were used in this study, including 111 participants with exposure therapy and EPCS coding from 3 multisite Pediatric OCD Treatment Studies (POTS)[26–28] at Brown University, the University of Pennsylvania, and Duke University,[12,13] along with 5 additional POTS participants who did not receive exposure during any treatment sessions, included so as to have some sessions entirely without exposure. On average, participants contributed 4.0 ± 3.1 sessions (mean ± SD), selected to equally represent early (32%), middle (32%), and late (36%) treatment phases. Further details of our dataset are shown in Appendix A Table S1.

Splitting the dataset

We split the dataset into training, validation, and test subsets using a stratified approach based on session counts per site and study (Table 1). Approximately two-thirds of sessions were assigned to training, with the remainder split evenly between validation and test sets. The training set was used for model development, the validation set for model selection, and all results in Experiments and Results Section reflect performance on the held-out test set.

We provide a detailed breakdown of the dataset splits for both sequence and token classification tasks in Table 2. This table presents the total number of sequences and tokens for each event type (Exposure and Encourage), the relative proportions of each class, and how these were allocated across the training, validation, and test sets. Percentages indicate both the class distribution and the proportion assigned to each dataset split. The sequence-level datasets are either balanced or moderately imbalanced across classes (eg, non-exposure vs exposure: 65% vs 35%). In contrast, the

Table 1. Stratified split of EPCS sessions by site (Brown University, University of Pennsylvania, Duke University) and study (POTS I, POTS II, POTS Jr.) into training, validation, and test subsets.

Site	Study	Sessions Training	Validation	Test
Brown University	POTS I	0	0	0
	POTS II	69	19	18
	POTS Jr.	56	13	11
University of Pennsylvania	POTS I	13	3	3
	POTS II	8	2	2
	POTS Jr.	24	5	4
Duke University	POTS I	23	5	5
	POTS II	19	4	5
	POTS Jr.	32	8	9
Total*		**244**	**59**	**57**

* Total number of sessions in the three subsets: training, validation, and test.

Table 2. Distribution of datasets used for sequence classification (fixed text segments) and token classification (dynamic text segments).

For sequence classification (number of segments)						
Events	**Details**	**Total**		**Training**	**Validation**	**Test**
Exposure	All Segments	1006	(100%)	(68%) 683	(16%) 165	(16%) 158
	Class 0 (Non-exposure)	649	(65%)	441	106	102
	Class 1 (Exposure)	357	(35%)	242	59	56
Encourage	All Segments	15 209	(100%)	(68%) 10 274	(14%) 2164	(18%) 2771
	Class 0 (Non-encourage)	7735	(51%)	5225	1104	1406
	Class 1 (Encourage)	7474	(49%)	5049	1060	1365
For Token Classification (number of tokens)						
Events	**Details**	**Total**		**Training**	**Validation**	**Test**
Exposure	All Tokens	2 191 848	(100%)	(67%) 1 478 007	(16%) 351 298	(17%) 362 543
	Class 0 (Non-exposure)	1 749 500	(80%)	1 182 780	287 722	278 998
	Class 1 (Exposure)	442 348	(20%)	295 227	63 576	83 545
Encourage	All Tokens	442 348	(100%)	(67%) 295 227	(14%) 63 576	(19%) 83 545
	Class 0 (Non-encourage)	343 304	(78%)	228 433	48 826	66 045
	Class 1 (Encourage)	99 044	(22%)	66 794	14 750	17 500

For each event type (Exposure and Encourage), the total number of segments/tokens, the relative proportions across classes (Class 0 and Class 1), and the allocation into training, validation, and test splits are provided. Percentages indicate both the relative class distribution and the proportion assigned to each dataset split.

token-level datasets are much more imbalanced, for both sequence and token classification. In particular, the non-exposure tokens make up 80% of the corpus, leaving only 20% exposure tokens. The non-encourage tokens constitute 78% of the data compared to just 22% encourage tokens. Such imbalance is important to consider during training, as it may bias models toward the majority classes if not properly addressed.

Annotation and coding

We conducted 2 types of annotation to support our study. First, manual transcripts were created to evaluate automated speech recognition (ASR) accuracy. Second, EPCS codes were applied to identify therapist behaviors, particularly exposure and encourage events, providing labeled examples for classification models development.

Annotation of manual transcripts for speech recognition

We manually transcribed 2-minute segments from 40 randomly selected sessions to serve as "ground truth" for ASR evaluation in terms of Word Error Rate (WER), a standard performance metric for ASR systems. Lower WER values indicate better ASR performance, with a WER of 0 representing a perfect transcription. Four transcribers participated; one was designated the "criterion transcriber" who undertook the transcription of 10 2-minute segments that each of the other transcribers had already completed. Additionally, the criterion transcriber transcribed 10 segments they had previously transcribed for a second time, to assess potential drift. Inter- and intra-rater reliability were calculated using Cohen's *kappa* coefficient (κ),[29] defined as 1 minus the WER, calculating one value per file and then averaging across files. Each pair of transcribers had a mean κ between 0.6 and 0.8. Figure 1 presents a histogram of the file-wise Cohen's *kappa* values for each transcriber and the criterion transcriber (Transcriber 1*).

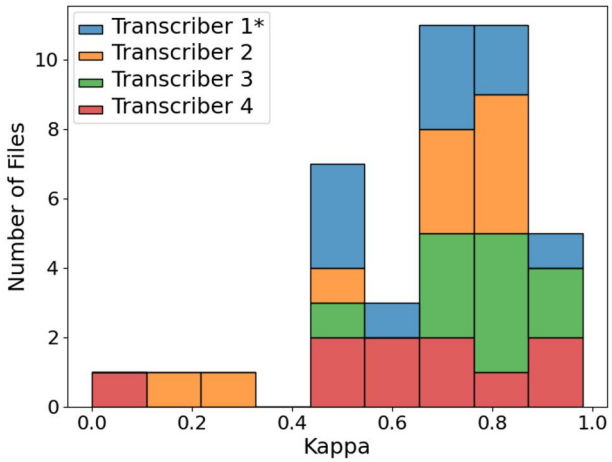

Figure 1. Stacked histogram of file-wise inter-transcriber Cohen's *kappa* values ($n = 40$, 2-minute excerpts). Transcriber 1 is the criterion transcriber (*).

Annotation of therapist behaviors using EPCS codes

A 5-person team at Brown University used the EPCS system to code session recordings. The team consisted of 4 undergraduate-level research assistants and one post-doctoral fellow. EPCS coders underwent comprehensive training to meet established criteria, as described in Benito et al. (2018)[12] and Benito et al. (2021).[13] This training encompassed a range of activities, such as thoroughly reading the EPCS manual, observing seasoned coders in action, coding under the supervision of experienced coders, and independently coding training videos until achieving a set standard (a reliability of κ or intraclass correlation coefficients, ICC, $>70\%$ across all codes). Ongoing training included weekly meetings to discuss EPCS implementation and prevent coder drift, double-coding a minimum of 10% of videos for reliability, and review of an additional 10% of videos by the team's lead therapist.

This study focused on 2 EPCS codes: *Exposure Event* and *Encourage Approach*. The start of an *Exposure Event* was defined as the point when at least one of the following occurred: the therapist indicating the start of exposure, a clear presentation of an exposure stimulus, or at least 2 present-focused difficulty ratings (eg, SUDS, a Subjective

Units of Distress Scale, [SUDS]). (A SUDS rating that quantifies a person's distress on a scale from 0 to 10, with 10 being the most distress). The end of an *Exposure Event* was defined as the point when at least one of the following occurred: the therapist indicated the exposure was over, there was a removal of the exposure stimulus, or there was a shift in the session focus to another topic. The code *Encourage Approach* was defined as facilitating physical or mental contact with the exposure stimulus, including redirection (eg, "Try to keep looking at the picture"), discouraging avoidance (eg, "Remember to resist the urge to ask for reassurance"), requests to describe the fear content (eg, "What are the worries saying right now?"), discussion of the stimulus (eg, "Ok, we've got the trash can"), or other actions aimed at keeping the patient engaged.

Methods

Our workflow for detecting *Exposure Event* and *Encourage Approach* behaviors from audio-recorded therapy sessions consists of 2 main steps (Figure 2): (A) Automatic Speech Recognition (ASR), which automatically transcribes session audio, and (B) Automatic Coding of Transcriptions, which identifies target therapist behaviors within the session.

Automatic speech recognition

We used 2 state-of-the-art ASR systems to transcribe session audio: OpenAI's Whisper and Google Speech-to-Text. Whisper is an open-source system which can be deployed locally, whereas Speech-to-Text is a cloud-based commercial system. Below, we describe both systems and their relevant features.

Whisper

Whisper is an ASR system developed by OpenAI.[30,31] It is a multitask transformer model trained on about 680 000 hours of multilingual and multitask supervised data collected from the web. This broad training allows Whisper to perform robustly across a range of acoustic conditions, accents, and speaking styles. In addition to English transcription, the model supports multilingual speech recognition, speech translation, and language identification. Whisper is available as an open-source tool, which makes it accessible for clinical research applications.

Google speech-to-text

Google Speech-to-Text is a cloud-based ASR service and one of the most widely used speech recognition systems, supporting a range of languages and use cases.[32] According to Google Assistant Help, the system incorporates 3 training strategies: (1) conventional learning, which uses audio data collected through Google services, some of which is human-labeled and some used for self-supervised training; (2) federated learning, where models are trained directly on users' devices without uploading audio to Google servers; and (3) ephemeral learning, which temporarily stores and processes audio data for training before deleting it. While Google does not specify how heavily each method is used, the combination suggests a semi-supervised neural architecture.

Automatic coding of transcriptions

The automatic coding task is framed as a binary classification problem: determining whether a segment of text reflects a target therapeutic behavior. This problem can be addressed in 2 different ways: prediction over fixed text segments or over dynamic text segments.

In the fixed segment approach, the dataset is predivided into sequences of text, each labeled with a binary value (0 or 1) indicating the absence or presence of the target behavior (exposure or encourage). Because the segmentation is determined in advance, the model operates at the level of entire sequences (often hundreds or thousands of tokens) and produces one prediction for the entire segment (eg, whether it contains an exposure). While this approach works well for classification at a coarser level, it does not support predictions at a finer temporal resolution, such as for individual tokens or short utterances.

In contrast, the dynamic segment approach treats the task as token-level classification, where the model evaluates each token in its surrounding context to determine whether it is part of the therapeutic behavior. This allows for more flexible and fine-grained identification within text, which is crucial for automatically detecting when a behavioral event occurs. Importantly, classic representations such as BoW and TF-IDF are not well-suited for this setting, as they convert a text sequence into a single global vector and ignore word order and contextual information, which limits their ability to support per-token predictions. For such fine-grained tasks,

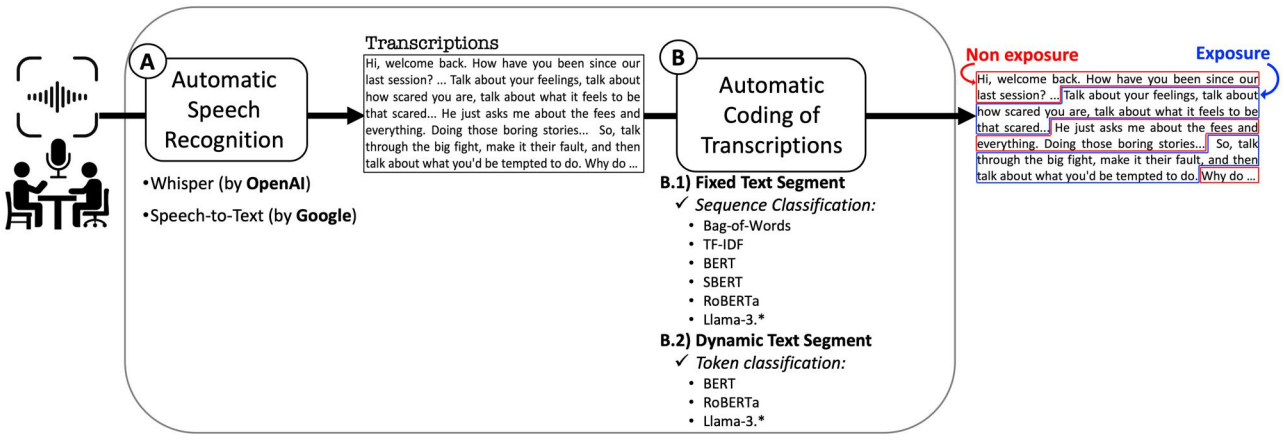

Figure 2. Illustration of the automated therapy coding process, which consists of 2 main steps: (A) ASR and (B) Automatic Coding of Transcriptions. The latter involves 2 types of techniques—sequence and token classification—representing fixed and dynamic text segments analysis, respectively.

models capable of capturing contextual and sequential dependencies remain essential.

Fixed text segments

In this approach, audio transcripts of therapy sessions were segmented into contiguous text sequences that occurred immediately before, during, and after a positive classification event—either a coded *Exposure* or an instance of the *Encourage* approach. All contiguous text sequences were manually defined by time windows within each session. For instance, if a 60-minute session contains a single exposure event from minute 30 to 50, the transcript is segmented into 3 contiguous sequences: ***Seq1.*** non-exposure from 0 to 30 minutes (Class 0); ***Seq2.*** exposure from 30 to 50 minutes (Class 1); and ***Seq3.*** non-exposure from 50 to 60 minutes (Class 0). Each sequence includes all spoken content within the corresponding time interval, regardless of whether it crosses sentence or paragraph boundaries, preserving temporal context and allowing reproducibility. Each sequence was embedded using either classical NLP methods or transformer-based models, then passed to a logistic regression classifier for behavior prediction.

Classical embeddings: Classical embeddings provide simple but effective representations of text by capturing word occurrence patterns within each sequence. Bag-of-Words and TF-IDF embeddings encode each sequence as a fixed-length vector whose dimensionality equals the vocabulary size. In doing so, these embeddings ignore word order and therefore assume that the relevant information is the presence and frequency of individual words. They are computationally efficient and often provide a strong baseline for text classification tasks.

- **Bag-of-Words (BoW)**: Represents text via word frequency counts.
- **Term Frequency-Inverse Document Frequency (TF-IDF)**: Weighs words or phrases based on frequency within a document and rarity across the corpus.

Transformer-based embeddings: All transformer-based embeddings were generated by averaging final hidden-layer outputs across tokens. The primary differences between these models included the maximum number of tokens they could process, the size of their internal embedding representations, and the total number of model parameters. We describe the embedding methods used below, with additional details in Table 3.

- **Bidirectional Encoder Representations from Transformers (BERT)**: Contextual word embeddings from bidirectional transformer pretrained on masked language modeling and next sentence prediction.[14]
- **Sentence-BERT (SBERT)**: Fine-tuned BERT in a siamese and triplet network architecture for sentence-level embeddings optimized for similarity tasks.[33]
- **Robustly Optimized BERT Pretraining Approach (RoBERTa)**: An improved BERT variant trained on larger corpora without next sentence prediction.[34]
- **MentalBERT and MentalRoBERTa**: Domain-adapted versions of BERT and RoBERTa, further pretrained on mental health subreddit data (eg, "r/depression," "r/Anxiety," "r/mentalillness," "r/SuicideWatch") to improve performance on psychiatric NLP tasks.
- **Large Language Model Meta AI (Llama 3)**: Meta AI's third-generation decoder-only transformers,[18,19] with 8B and 70B parameter models trained on 15T tokens. Instruction-tuned versions (eg, Llama-3.1-8B-Instruct) are fine-tuned with supervised and reinforcement learning for better alignment.

Logistic regression: We trained logistic regression models to predict labels from text embeddings, using both unweighted and class-weighted loss functions to address label imbalance (errors on positive examples were weighted by the ratio of negative to positive examples for each label). Models were implemented in Python with scikit-learn,[35] using L2 regularization and the LBFGS and LIBLINEAR solvers. Performance was evaluated using 5-fold cross-validation on data stratified by EPCS codes. Hyperparameters were selected via grid search over 7 values of the regularization parameter C $(0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0)$, carried out over the training set in each fold. The scoring metrics for cross-validation were the area under the ROC curve (AUC) and accuracy. Higher values of AUC indicate better discriminative ability, with 1.0 representing perfect performance.

Preprocessing: We considered various data preprocessing strategies for classic embeddings (BoW and TF-IDF), which affect how text is tokenized before vectorization. Tokens can be full words (eg, "funny"), stems (eg, "fun"), or multi-word terms (eg, "funny dog"). Preprocessing has been shown to improve classic NLP performance for various applications.[36–38] Specifically, we tried standard preprocessing techniques, including stop-word removal, *n*-grams (1-3 words), lemmatization, stemming, and restricting vocabulary to specific parts of speech.

These different preprocessing variants reflect assumptions on how varying levels of linguistic abstraction and emphasis on specific parts of speech influence the quality of classic embeddings. Using full words preserves all original lexical information, while lemmatization and stemming reduce sparsity by consolidating word forms. Restricting analysis to particular parts of speech, such as nouns, adjectives, or verbs, highlights content-carrying tokens that are more likely to improve classification, while removing stop words filters out frequent but uninformative terms, and incorporating *n*-grams captures multi-word expressions that may convey richer meaning than individual tokens.

For instance, words such as "embarrassing," "embarrasses," "embarrassed," and "embarrass" would be treated as distinct features without lemmatization or stemming, each reducing frequency and potentially leading to sparsity, poor generalization across training and test sets, and increased risk of overfitting; lemmatization mitigates these issues by consolidating them into the single feature "embarrass." Another example is the phrase "identify negative thoughts"; if only unigrams (identify, negative, thoughts) are used, the connection between them is lost, whereas representing it as a trigram preserves the contextual meaning, which is relevant to CBT. A summary of all preprocessing approaches is presented in Appendix B Table S2. We implemented them using NLTK[39] and spaCy[40]; full details are available in our public GitHub repository.

For BERT-, SBERT-, and Llama-based models, we skipped this preprocessing since these models are trained to handle raw text and are less sensitive to such variations. While tokenization still occurs internally, minor changes (eg, stem vs

Table 3. Context length and hidden dimension sizes for the LLMs used to generate embeddings.

Type	Model	Context Length	Dimension	Parameters
SBERT	*all-mpnet-base-v2*	384	768	109M
	multi-qa-mpnet-base-dot-v1	512	768	109M
	all-distilroberta-v1	512	768	82M
	all-MiniLM-L12-v2	256	384	33M
	multi-qa-distilbert-cos-v1	512	768	66M
	all-MiniLM-L6-v2	256	384	23M
	multi-qa-MiniLM-L6-cos-v1	512	384	23M
	paraphrase-multilingual-mpnet-base-v2	128	768	278M
	paraphrase-albert-small-v2	256	768	12M
	paraphrase-multilingual-MiniLM-L12-v2	128	384	118M
	paraphrase-MiniLM-L3-v2	128	384	17M
	distiluse-base-multilingual-cased-v1	128	512	135M
	distiluse-base-multilingual-cased-v2	128	512	135M
	msmarco-MiniLM-L6-cos-v5	128	512	23M
BERT-based	BERT-base *(bert-base-uncased)*	512	768	110M
	BERT-large *(bert-large-uncased)*	512	1024	340M
	RoBERTa *(roberta-base)*	512	768	125M
	MentalBERT *(mental-bert-base-uncased)*	512	768	110M
	MentalRoBERTa *(mental-roberta-base)*	512	768	125M
Meta-Llama	Llama-3.1-8B	128K	4096	8B
	Llama-3.1-8B-Instruct	128K	4096	8B
	Llama-3.3-70B-Instruct	128K	8192	70B

lemma) have little effect. Instead, we evaluated 5 BERT-based, 3 Llama-based, and 14 Sentence-BERT (SBERT) models.

Finally, note that this approach requires text sequences of a fixed length—often hundreds and thousands of tokens—for label prediction over the entire sequence (eg, whether it contains an exposure or encourage approach behavior). As a result, it does not support generating predictions at a finer temporal resolution, such as for individual tokens (words or expressions) or short sentences.

Dynamic text segments

In dynamic text segment classification, the goal is to classify each token as being part of an EPCS code (either *Exposure* or *Encourage*). We explored 2 approaches: (1) logistic regression applied to token embeddings extracted from transformer-based models and (2) direct token classification by fine-tuning transformer-based models. In this case, the input dataset is token-level, also referred to as word-level, with each word assigned a label—Class 0 for non-exposure or non-encourage, and Class 1 for exposure or encourage. Transformer-based models rely on subword tokenization, such as WordPiece used in BERT[14] or byte-pair encoding[41] used in Llama, which can split a word into one or more subtokens. In these cases, all subtokens inherit the label of the original word they comprise, preserving consistency between annotations and model input. During training, the model predicts labels at the subtoken level, but evaluation is aggregated back to the original word-level labels to maintain alignment with the annotated dataset. For aggregation, we use the first subtoken of each word, which is the most widely adopted approach in token classification tasks, as implemented in Hugging Face.

Note that text sequences were defined by time windows within each session. For instance, a 60-minute session with a single exposure event from minute 30 to 50 was segmented into Seq1. non-exposure (0-30 min, Class 0), Seq2. exposure (30-50 min, Class 1), and Seq3. non-exposure (50-60 min, Class 0). For token classification, each word was assigned the class of the sequence to which it belonged, for example, all words in Seq1. were labeled Class 0, all words in Seq2. were labeled Class 1, and so forth.

Approach 1: Transformer-based Embedding + Logistic Regression: This method extracts token-level embeddings from transformer-based models and uses logistic regression to classify each token. The process involves:

- **Transformer-based embeddings:** Each chunk of text from a conversation was tokenized and passed through the model, producing a contextualized vector for each token. We used base and large versions of BERT and RoBERTa, as well as domain-specific models such as MentalBERT and MentalRoBERTa. We also included Llama models, including Llama-3.1-8B, Llama-3.1-8B-Instruct, and Llama-3.3-70B-Instruct.
- **Logistic regression:** The procedure was identical to the logistic regression setup described in Fixed Text Segments Section, including 5-fold cross-validation, grid search over C values, and evaluation using AUC and accuracy. Models were trained with unweighted and weighted loss functions to address class imbalance. In addition, since the dataset was highly imbalanced (see Table 2), we applied undersampling of the majority class to match the size of the minority class.

Approach 2: Fine-Tuning Transformer Models: In this approach, transformer-based LLMs were fine-tuned directly for token classification. Instead of relying on extracted embeddings, the entire model was trained end-to-end to predict EPCS labels for each token within a chunk of conversation text. This method leverages the model's full contextual understanding during supervised training to improve token-level predictions. Table 3 provides an overview of the context lengths and hidden dimensions for the models used in our experiments.

We performed full training (fine-tuning) of all model weights. For hyperparameter tuning in our binary token

classification task, we explored several configurations, including a learning rate of [1e−5, 2e−5, 3e−5, 4e−5, 5e−5, 4e−4], batch size of [8, 16, 32, 64], and number of epochs ranging from 1 to 10. We also evaluated both unbalanced and class-balanced settings by employing Cross-Entropy Loss and Weighted Cross-Entropy Loss, respectively.

BoW, TF-IDF, and SBERT were excluded from this approach. These models either lack token-level resolution (BoW/TF-IDF) or produce pooled sentence-level embeddings (SBERT), making them unsuitable for token classification. In contrast, BERT and Llama models preserve contextual information at the token level, making them well-suited for fine-grained prediction. This dynamic approach extends the fixed segment method from Fixed Text Segments Section by supporting per-token predictions, offering finer temporal resolution. Note that, *Exposure* events typically span many tokens, while *encourage* behaviors are often short and sparse (see Figure 3).

Preprocessing: To better capture short and fragmented encourage instances, we merged adjacent encourage segments if separated by fewer than 10 tokens (roughly 4 seconds). This helped preserve continuity across short pauses. Figure 3 illustrates this strategy and highlights how encourage segments can occur within broader exposure events, highlighting the nested nature of these therapeutic techniques.

Experiments and results

Automated speech recognition

To quantify the transcription accuracy of each ASR system, we compared their output transcripts to the corresponding manual reference transcriptions (see Fixed Text Segments Section) and calculated the WER (↓). The comparison revealed that Google Speech-to-Text had a WER of 0.51±0.22 (mean ± SD) whereas OpenAI's Whisper achieved a lower WER of 0.31±0.18, see Figure 4A.

We also tested the baseline Google Speech-to-Text against a variant using vocabulary boosting, which incorporates a list of common exposure-related terms expected in the sessions. Despite this, Whisper outperformed both Speech-to-Text configurations, demonstrating superior transcription quality. Finally, we evaluated how WER directly affects classifier performance. A higher WER—worse transcription quality—resulted in lower classification accuracy, as shown in Figure 4B. Based on these results, we chose to continue using Whisper for the remainder of the work.

Fixed text segments

Table 4 presents the top-performing results for each NLP approach using text sequence classification on Exposure and Encourage Approach events. Llama models (Llama-3.3-70B-Instruct and Llama-3.1-8B) outperformed other models in detecting exposure events, while both MentalBERT (domain-specific BERT variant) and Llama-3.1-8B achieved better performance in identifying encourage-approach events. For BoW and TF-IDF, the best performance was achieved using lemmas for the exposure approach and stems for the encourage events, combined with 1-2 g, and without stop words. Across experiments with both transformer-based and classical embeddings, LIBLINEAR solver typically produced the best results, followed by LBFGS, with $C = 0.01$, for both unweighted and weighted loss functions.

Dynamic text segments

We compared 2 token classification methods: logistic regression on pretrained token embeddings and direct fine-tuning of transformer models. Fine-tuning consistently outperformed logistic regression, as shown by AUC scores in Table 5. Llama models performed well with embeddings but were not fine-tuned due to resource limits, thus, their fine-tuning results are not reported. For logistic regression, the LIBLINEAR solver most often produced the best results with $C = 0.001$ and the weighted loss function, followed by undersampling, where weighting was not required. For fine-tuning, the best results were generally obtained with a learning rate of 4e−4, batch sizes of 8 or 16, a decay rate of 0.05, and 5-7 epochs, using a weighted cross-entropy loss function.

Discussion

Principal findings

We developed a classification framework to automatically label EPCS quality codes during in-person exposure therapy sessions, aiming to enable rapid clinical quality assessment and accelerate research into exposure therapy effectiveness. We evaluated multiple NLP approaches combined with varied preprocessing pipelines, also assessing the impact of audio quality. While this study focused on 2 EPCS codes, the framework is extendable to other codes.

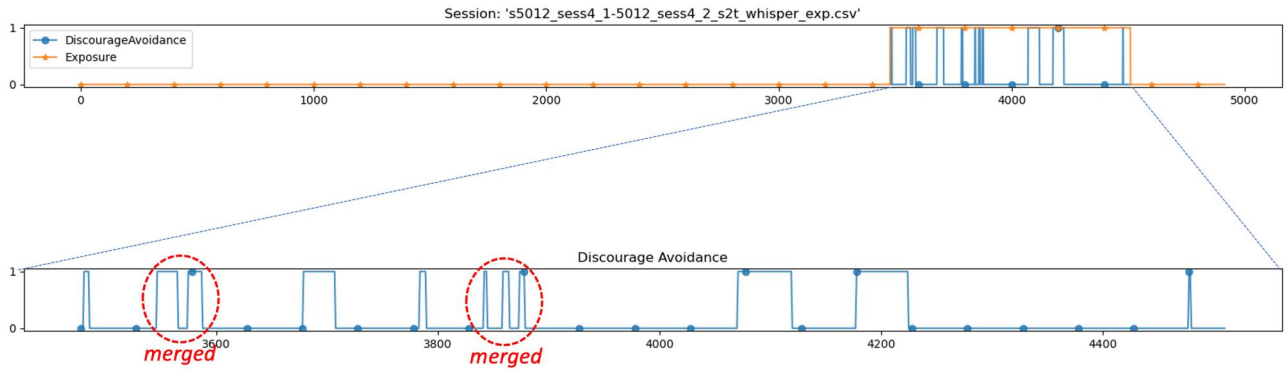

Figure 3. Conversation between a therapist and patients, consisting of ∼5000 tokens. Exposure occurs between tokens 3500 and 4500, along with encouragement in the exposure approach. Processing step for encouragement events: events are merged when the gap between them is shorter than 10 tokens, as illustrated with red circles.

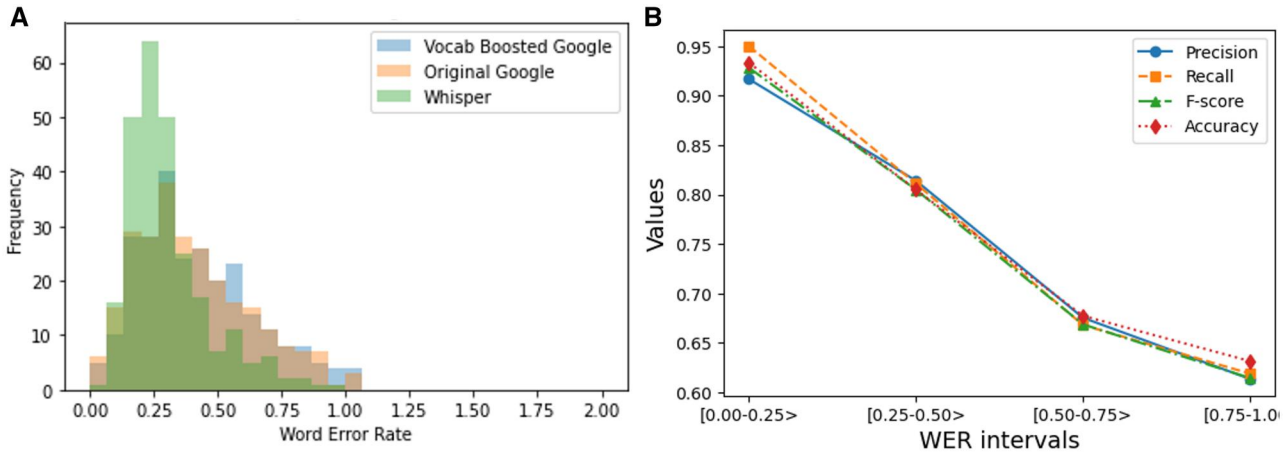

Figure 4. (A) Histograms of the distribution of WER ↓ of Google Speech-to-Text and OpenAI's Whisper, across audio clips, calculated by comparing automatic transcriptions of those clips with manual transcriptions. Speech-to-Text was tested with and without common exposure-related words pre-identified (ie, "vocabulary boosting"). (B) Impact of WER on performance metrics when using BERT to detect exposure events. Average performance metric values for session groups defined by their WER ranges. "[0-0.25 > " indicates $0 \leq \text{WER} < 0.25$.

Table 4. Results of fixed text segment classification (sequence classification) on Exposure and Encourage events: models were selected based on their performance during the validation phase and subsequently evaluated on the held-out test set. The best AUC results are highlighted in bold, and the runner-ups are underlined.

Type	Models	Logistic Regression Exposure	Encourage
Classic	BoW	0.9186	0.6980
	TF-IDF	0.9195	0.6686
SBERT	all-mpnet-base-v2	0.8909	0.6881
	multi-qa-mpnet-base-dot-v1	0.8951	0.7018
	all-distilroberta-v1	0.9028	0.7052
	all-MiniLM-L12-v2	0.8845	0.7036
	multi-qa-distilbert-cos-v1	0.8869	0.6870
	all-MiniLM-L6-v2	0.8908	0.7103
	multi-qa-MiniLM-L6-cos-v1	0.8746	0.7061
	paraphrase-multilingual-mpnet-base-v2	0.9105	0.7001
	paraphrase-albert-small-v2	0.8755	0.6948
	paraphrase-multilingual-MiniLM-L12-v2	0.8915	0.7014
	paraphrase-MiniLM-L3-v2	0.8775	0.6934
	distiluse-base-multilingual-cased-v1	0.8981	0.7211
	distiluse-base-multilingual-cased-v2	0.9028	0.7210
	msmarco-MiniLM-L6-cos-v5	0.8790	0.7222
BERT-based	BERT-base	0.8909	0.7415
	BERT-large	0.8909	0.7372
	RoBERTa	0.9189	0.7397
	MentalBERT	0.9004	**0.7467**
	MentalRoBERTa	0.9111	0.7391
Meta-Llama	Llama-3.1-8B	<u>0.9293</u>	<u>0.7444</u>
	Llama-3.1-8B-Instruct	0.9242	0.7440
	Llama-3.3-70B-Instruct	**0.9534**	0.7398

AUC ↑ scores for logistic regression approaches are reported.

Table 5. Results of dynamic text segment classification (token classification) on Exposure and Encourage events: models were selected based on their performance during the validation phase and subsequently evaluated on the held-out test set. The best AUC results are highlighted in bold, and the runner-ups are underlined.

		Exposure		Encourage	
Type	Models	Log. Reg.	Fine-tuning	Log. Reg.	Fine-tuning
BERT-based	BERT-base	0.7477	0.8396	0.6462	0.7381
	BERT-large	0.7572	**0.8538**	0.6458	**0.7526**
	RoBERTa	0.7562	0.8362	0.6508	<u>0.7504</u>
	MentalBERT	0.7387	<u>0.8470</u>	0.6490	0.7494
	MentalRoBERTa	0.7475	0.8393	0.6383	0.7469
Meta-Llama	Llama-3.1-8B	<u>0.7745</u>	–	<u>0.6609</u>	–
	Llama-3.1-8B-Instruct	0.7735	–	0.6566	–
	Llama-3.3-70B-Instruct	**0.7858**	–	**0.6724**	–

AUC ↑ scores for both logistic regression applied to token embeddings and fine-tuning approaches are reported.

Automatic speech recognition

Whisper outperformed Google Speech-to-Text in transcription accuracy, consistent with prior reports,[30] despite variable audio quality in our in-person therapy recordings. Our dataset's WER (averaging 0.31) was higher than typical ASR benchmarks (averaging 0.13), likely due to older recording equipment, background noise, and the predominance of child speech (most participants were under 14), which is underrepresented in ASR training data[42,43] making the ASR models struggle with the vocal variability. Improved transcription quality (ie, lower WER) correlated with better classification accuracy, as shown in Figure 4B, highlighting the importance of ASR advances and recording quality.

Fixed text segments

For fixed-length text classification, Llama models had the best performance in Exposure detection (AUC = 0.9534), while MentalBERT and Llama-3.1-8B had the best performance for Encourage-approach (AUC = 0.7467 and AUC = 0.7444, respectively).

When classifying Exposures, some of the BoW and TF-IDF classifiers performed as well as the BERT models (see Table 4). One possible reason is that certain words or tokens, such as phobia targets like "spider," are highly informative, allowing simple term-frequency-based features to capture most of the predictive signals without modeling complex dependencies. Moreover, given the relatively small dataset, complex models are more prone to overfitting, while simpler approaches can generalize better. Our feature analysis supports this interpretation, revealing that the most informative unigrams included "rating," "resist," "homework," "anxiety," "temperature," "garbage," "spider," and "germ." These task-specific terms are frequently used in CBT sessions and are strong indicators of exposures, which makes keyword-based methods particularly effective in this setting. These results highlight the importance of strong baselines, showing that simpler approaches can achieve competitive performance at lower computational cost. However, BoW and TF-IDF cannot be directly applied to token-level classification, which is crucial for automatically detecting when a behavioral event occurs. For such fine-grained tasks, models capable of capturing context and sequential dependencies remain necessary.

BERT-based models faced input length limits (512 tokens), requiring chunking and aggregation. This is less effective than using a model capable of handling long sequences—such as Llama, BoW, and TF-IDF—as it results in a fragmented global context, which means there is no way for the transformer attention mechanism to operate between chunks. In contrast, Encourage-approach events tended to be shorter, thus, BERT models could process them in full, explaining their relatively better performance. BoW and TF-IDF classifiers seemed to be particularly sensitive to preprocessing choices when classifying exposures. Users should consider this variability when selecting their preprocessing method. SBERT, by contrast, performed consistently well across the models we tried. Overall, the results show that automating EPCS coding of fixed text sequences using current NLP methods is promising and merits further development.

Dynamic text segments

Token-level classification compared logistic regression on fixed embeddings vs fine-tuned transformer models. Fine-tuning improved AUC by over 10%, with BERT-large achieving AUCs of 0.85 (Exposure) and 0.75 (Encourage). These results suggest that BERT-based models are effective in capturing nuanced therapeutic content at the token level. General-purpose BERT outperformed domain-specific variants, challenging assumptions about specialized pretraining.[44]

We also tested newer Llama models using embeddings with logistic regression. Llama-3.3-70B-Instruct showed promising results, but its performance gains over BERT were modest (3%), suggesting that smaller models may suffice in many applications.

Token-level predictions allowed us to estimate "dosage" (ie, total duration of each technique), a useful metric for therapy tracking. The classifiers generalized across multiple sites and held-out participants, indicating robust performance. Token classification is more realistic than fixed-segment classification, as it reflects real-world ambiguity in technique boundaries. It also addresses class imbalance, since exposures and encourages appear less often than other tokens, but loss weighting helped mitigate this issue. We plan to explore detection of even rarer EPCS codes.

This study validates the potential of automated coding and lays the groundwork for broader application. Our work has already begun on extending these methods to other codes and datasets. We also plan to test other LLM families like Mistral to enhance generalizability.

Practical considerations and deployment

Simpler approaches such as BoW and TF-IDF are less computationally demanding than transformer-based models like BERT, which require specialized hardware such as GPUs. This highlights the practical advantage of strong baseline models, particularly in resource-constrained environments such as CPU-only setups. However, training still involves trade-offs, for instance, performing a grid search over multiple parameters can be slow on a CPU, whereas fine-tuning small transformer-based models can be faster on GPUs, since fine-tuning eliminates the need for extensive hyperparameter searches in logistic regression. Transformer-based models and LLMs differ in scale and resource requirements. BERT has hundreds of millions of parameters (BERT-base: 110M, BERT-large: 340M), whereas modern LLMs such as Llama are much larger (8-70 billion parameters) and are trained to generate text in addition to understanding it. Therefore, they require more substantial GPU resources. For example, models such as BERT, RoBERTa, or SBERT can typically be fine-tuned or used for embedding computation on a single NVIDIA A100 (80 GB) GPU. In contrast, Llama-3-8B may require at least 2 A100 (80 GB) GPUs, and Llama-3.3-70B-Instruct may require at least 3 A100 (80 GB) GPUs even for general-purpose inference. Computing embeddings for these LLMs is also time-consuming, so additional GPUs are useful to accelerate the task.

Limitations

Using ASR to extract transcripts from audio has limitations—some addressable with better design and equipment, others inherent to ASR tools. Challenges include background noise, overlapping speech, accents, quiet voices, and mumbling. Though Whisper and Google Speech-to-Text claim strong performance on accents and non-English speech, results should be validated on the target dataset. Both tools

sometimes hallucinated words during silence; but this can be mitigated through certain run settings, although these settings may come with trade-offs. In this paper, we used the default settings for each ASR method. Our transcripts were not diarized, so the classifier lacked information about speaker roles (therapist, patient, or parent), which is relevant for EPCS coding. While single-channel diarization exists, its accuracy varies and was not used here. Future work could incorporate diarization or multi-channel audio to improve speaker identification.

Importantly, this study used a pediatric OCD sample with limited racial and ethnic diversity. Before NLP-based automated exposure coding can widely used, it will be imperative to conduct additional research to test whether findings generalize across the lifespan (ie, with adults and children), across diagnoses (ie, in other disorders treated with exposure therapy such as PTSD), across settings (eg, in research laboratories and in the community), and in historically marginalized groups (for whom optimal treatment and exposure learning may differ[45]). Although the number of individual patients is moderate, the dataset encompasses a large number of therapy sessions collected across diverse clinical sites and treatment settings. This diversity introduces relevant variability in speech patterns and contextual factors, which supports the robustness of our automated classification models for exposure and encouragement events. Moreover, cross-validation across study types and sites was employed to reduce the risk of overfitting and to provide preliminary evidence that the models may generalize to new sessions and settings, though further studies are needed to confirm broader applicability.

There are several limitations and risks associated with using non-open source LLMs, such as ChatGPT, for handling PHI. These models typically require sending data to the cloud or external servers and do not provide transparency regarding how PHI is processed or protected, making it difficult to guarantee compliance with privacy and security standards. They may also have security vulnerabilities that could be exploited to gain unauthorized access to sensitive information. Given the vulnerability of patients with mental health disorders, special care is needed in any future work in this area. We do not recommend using non-open source language models with this or similar datasets unless contractual guarantees are in place to satisfy requirements for handling PII or PHI.

Future work

Multimodal data may improve future classification accuracy. Human EPCS coders had access to session videos to assist coding, whereas our NLP tools did not. Although EPCS coding is content-based, our text-only approach still performed well. However, incorporating prosodic cues (eg, tone of voice) and visual features (eg, body language) may improve accuracy, especially for patient behaviors, where mumbling or avoidance can limit usable text.

Conclusions

By presenting the first complete classification framework capable of automatically labeling EPCS quality codes in in-person exposure therapy sessions, this article lays the groundwork for rapid quality assessment and accelerated psychotherapy research. For these data, the most effective approach for classifying Exposure and Encourage events is found to be the combination of Whisper for ASR and a fine-tuned BERT model for "dynamic text segment classification." As a fallback, estimating the upper bounds of EPCS codes using Llama may be sufficient. The success of our "dynamic text segments classification" method shows that this labeling can remain reliable when the boundaries between techniques are unknown. Our analysis of preprocessing choices and transcription WERs suggests that different approaches are better suited to different classification tasks, and higher-fidelity audio recording improves results. Future work will explore alternative LLMs and extend this feasibility study into the full set of EPCS quality codes.

Acknowledgments

The authors wish to thank Blythe Hattenbach, Katie Paris, and Anjali Poe for their work on manual transcription checks.

Author contributions

Juan Antonio Lossio-Ventura (Conceptualization, Formal analysis, Methodology, Software, Supervision, Validation, Writing—original draft, Writing—review & editing), Samuel Frank (Formal analysis, Software, Writing—review & editing), Grace Ringlein (Formal analysis, Software, Writing—review & editing), Kirsten Bonson (Formal analysis, Software, Writing—original draft, Writing—review & editing), Ardyn Olszko (Data curation, Writing—review & editing), Abbey Knobel (Data curation, Writing—review & editing), Daniel S. Pine (Investigation, Supervision, Writing—review & editing), Jennifer B. Freeman (Supervision, Writing—review & editing), Kristen Benito (Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing—original draft, Writing—review & editing), David C. Jangraw (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing—original draft, Writing—review & editing), and Francisco Pereira (Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

Investigator time was supported in part by the National Institute of Mental Health (NIMH; R01MH135861, K.B.). Collection of coding data was supported by NIMH (R21/R33MH096828, K.B. and J.B.F.). Collection of original trial data was supported by the following NIMH mechanisms: POTS Jr: R01MH79217, J.B.F.; R01MH079154, March; R01MH079377, Franklin. POTS II: R01MH064188, Leonard; R01MH055126, Franklin; R01MH055121, March. POTS I: R01MH055121, March; R01MH055126, Franklin. This research was also supported by the Intramural Research Program of the National Institute of Mental Health (NIMH-IRP), project ZIC-MH002968 (J.A.L.-V., F.P.), and project ZIA-MH002781 (D.S.P., S.F., G.R.).

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

Due to the private and sensitive nature of the data, it cannot be made available for sharing. The analysis code is publicly available at the GitHub repository https://github.com/juan-lossio/epcs_nlp

References

1. Olatunji BO, Cisler JM, Deacon BJ. Efficacy of cognitive behavioral therapy for anxiety disorders: a review of meta-analytic findings. *Psychiatr Clin North Am*. 2010;33:557-577.
2. Carpenter JK, Andrews LA, Witcraft SM, Powers MB, Smits JA, Hofmann SG. Cognitive behavioral therapy for anxiety and related disorders: a meta-analysis of randomized placebo-controlled trials. *Depress Anxiety*. 2018;35:502-514.
3. Bryant RA, Kenny L, Rawson N, et al. Efficacy of exposure-based cognitive behaviour therapy for post-traumatic stress disorder in emergency service personnel: a randomised clinical trial. *Psychol Med*. 2019;49:1565-1573.
4. Whiteside SPH, Deacon BJ, Benito K, Stewart E. Factors associated with practitioners' use of exposure therapy for childhood anxiety disorders. *J Anxiety Disord*. 2016;40:29-36.
5. Weisz JR, Weiss B, Donenberg GR. The lab versus the clinic. *Am Psychol*. 1992;47:1578-1585.
6. Higa CK, Chorpita BF. *Handbook of Evidence-Based Therapies for Children and Adolescents: Bridging Science and Practice*. Springer; 2008.
7. Perepletchikova F, Treat TA, Kazdin AE. Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *J Consult Clin Psychol*. 2007;75:829-841.
8. Chassin MR, Galvin RW. The urgent need to improve health care quality. Institute of medicine national roundtable on health care quality. *J Am Med Assoc*. 1998;280:1000-1005.
9. The Patient Protection and Affordable Care Act. Pub. L. No. 111–148, 124 Stat. 119, 2010. [Online]. Accessed July 5, 2025. https://www.congress.gov/111/plaws/publ148/PLAW-111publ148.pdf
10. Centers for Medicare & Medicaid Services. *CMS National Quality Strategy*. U.S. Department of Health and Human Services; 2024. [Online]. Accessed July 5, 2025. https://www.cms.gov/medicare/quality/meaningful-measures-initiative/cms-quality-strategy
11. Benito KG, Conelea C, Garcia AM, Freeman JB. CBT specific process in exposure-based treatments: initial examination in a pediatric OCD sample. *J Obsessive Compuls Relat Disord*. 2012;1:77-84.
12. Benito KG, Machan J, Freeman JB, et al. Measuring fear change within exposures: functionally-defined habituation predicts outcome in three randomized controlled trials for pediatric OCD. *J Consult Clin Psychol*. 2018;86:615-630.
13. Benito KG, Machan J, Freeman JB, et al. Therapist behavior during exposure tasks predicts habituation and clinical outcome in three randomized controlled trials for pediatric OCD. *Behav Ther*. 2021;52:523-538.
14. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. Long and Short Papers. Association for Computational Linguistics; 2019:4171-4186. https://aclanthology.org/N19-1423
15. Zhuang L, Wayne L, Ya S, Jun ZA Robustly optimized BERT pre-training approach with post-training. In: Li S, Sun M, Liu Y, Wu H, Liu K, Che W, et al., eds. *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China; 2021: 1218-1227. https://aclanthology.org/2021.ccl-1.108
16. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv: 191001108. 2019, preprint: not peer reviewed.
17. GPT. [Online]. Accessed September 3, 2024. https://openai.com/index/gpt-4/
18. Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models. https://arxiv.org/abs/2407.21783, 2024, preprint: not peer reviewed.
19. Meta A. Llama 3 Model Card. 2024. [Online]. Accessed July 5, 2025. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
20. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. *Transl Psychiatry*. 2023;13:309.
21. Atzil-Slonim D, Eliassaf A, Warikoo N, et al. Leveraging natural language processing to study emotional coherence in psychotherapy. *Psychotherapy*. 2024;61:82-92.
22. Xin AW, Nielson DM, Krause KR, et al. Using large language models to detect outcomes in qualitative studies of adolescent depression. *J Am Med Inform Assoc*. 2024;ocae298.
23. Lin B, Bouneffouf D, Landa Y, Jespersen R, Corcoran C, Cecchi G. COMPASS: computational mapping of patient-therapist alliance strategies with language modeling. *Transl Psychiatry*. 2025;15:166.
24. Ewbank MP, Cummins R, Tablan V, et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*. 2020;77:35-43.
25. Peris TS, Caporino NE, O'Rourke S, et al. Therapist-reported features of exposure tasks that predict differential treatment outcomes for youth with anxiety. *J Am Acad Child Adolesc Psychiatry*. 2017;56:1043-1052.
26. Team P. Pediatric OCD treatment study (POTS) team. Cognitive-behavior therapy, sertraline, and their combination for children and adolescents with obsessive-compulsive disorder: the pediatric OCD treatment study (POTS) randomized controlled trial. *Arch Psychol*. 2004;292:1969-1976.
27. Franklin ME, Sapyta J, Freeman JB, et al. Cognitive behavior therapy augmentation of pharmacotherapy in pediatric obsessive-compulsive disorder: the pediatric OCD treatment study II (POTS II) randomized controlled trial. *Arch Psychol*. 2011;306:1224-1232.
28. Freeman J, Sapyta J, Garcia A, et al. Family-based treatment of early childhood obsessive-compulsive disorder: the pediatric obsessive-compulsive disorder treatment study for young children (POTS Jr)–a randomized clinical trial. *Arch Psychol*. 2014;71:689-698.
29. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22:276-282.
30. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In *ICML'23: Proceedings of the 40th International Conference on Machine Learning*. 1182. Association for Computing Machinery; 2023:28492-28518.
31. Whisper. [Online]. Accessed June 17, 2023. https://openai.com/research/whisper
32. Google Speech-to-Text. [Online]. Accessed June 17, 2023. https://cloud.google.com/speech-to-text/docs/
33. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Inui K, Jiang J, Ng V, Wan X, eds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019:3982-3992. https://aclanthology.org/D19-1410
34. Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv: 190711692. 2019, preprint: not peer reviewed.

35. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learning Res*. 2011;12:2825-2830.
36. Uysal AK, Gunal S. The impact of preprocessing on text classification. *Inf Process Manage*. 2014;50:104-112.
37. Alam S, Yao N. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Comput Math Org Theory*. 2019;25:319-335.
38. Chai CP. Comparison of text preprocessing methods. *Nat Lang Eng*. 2023;29(3):509-553.
39. Natural Language Toolkit. [Online]. Accessed August 15, 2024. https://www.nltk.org/
40. spaCy. [Online]. Accessed August 15, 2024. https://spacy.io/
41. Gage P. A new algorithm for data compression. *C Users J*. 1994;12:23-38.
42. Lee S, Potamianos A, Narayanan S. Acoustics of children's speech: developmental changes of temporal and spectral parameters. *J Acoust Soc Am*. 1999;105:1455-1468.
43. Godwin KE, Almeda MV, Seltman H, et al. Off-task behavior in elementary school children. *Learn Instr*. 2016;44:128-143.
44. Dorfner FJ, Dada A, Busch F, et al. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. *J Am Med Inf Assoc*. 2025;32:1015-1024.
45. Benito K, Pittig A, Abramowitz J, et al. Mechanisms of change in exposure therapy for anxiety and related disorders: a research agenda. *Clin Psychol Sci*. 2025;13:687-719.

JAMIA Open, 2025, 8, 1–12
https://doi.org/10.1093/jamiaopen/ooaf127
Research and Applications