Original Paper

Preprocessing Large-Scale Conversational Datasets: A Framework and Its Application to Behavioral Health Transcripts

Paz Mor Naim[1]; Shiri Sadeh-Sharvit[2,3]; Samuel Jefroykin[2]; Eddie Silber[2,3]; Dennis P Morrison[4]; Ariel Goldstein[1,5,6]

[1]Department of Cognitive and Brain Sciences, Hebrew University of Jerusalem, Jerusalem, Israel
[2]Eleos Health, Waltham, MA, United States
[3]Palo Alto University, Palo Alto, CA, United States
[4]Morrison Consulting, Bloomington, IN, United States
[5]Business School, Hebrew University of Jerusalem, Jerusalem, Israel
[6]Department of Psychology, Azrieli Israel Center for Addiction and Mental Health (Azrieli ICAMH), Hebrew University of Jerusalem, Jerusalem, Israel

Corresponding Author:

Paz Mor Naim
Department of Cognitive and Brain Sciences
Hebrew University of Jerusalem
Mount Scopus
Jerusalem 9190500
Israel
Phone: 972 025882888
Email: paz.naim@mail.huji.ac.il

Abstract

Background: The rise of artificial intelligence and accessible audio equipment has led to a proliferation of recorded conversation transcripts datasets across various fields. However, automatic mass recording and transcription often produce noisy, unstructured data that contain unintended recordings such as hallway conversations, media (eg, TV, radio), or transcription inaccuracies as speaker misattribution or misidentified words. As a result, large conversational transcript datasets require careful preprocessing and filtering to ensure their research utility. This challenge is particularly relevant in behavioral health contexts (eg, therapy, counseling) where deriving meaningful insights, specifically dynamic processes, depends on accurate conversation representation.

Objective: We present a framework for preprocessing large datasets of conversational transcripts and filtering out *non-sessions*—transcripts that do not reflect a behavioral treatment session but instead capture unrelated conversations or background noise. This framework is applied to a large dataset of behavioral health transcripts from community mental health clinics across the United States.

Methods: Our approach integrated basic feature extraction, human annotation, and advanced applications of large language models (LLMs). We began by mapping transcription errors and assessing the number of non-sessions. Next, we extracted statistical and structural features to characterize transcripts and detect outliers. Notably, we used LLM perplexity as a measure of comprehensibility to assess transcript noise levels. Finally, we used zero-shot prompting with an LLM to classify transcripts as sessions or non-sessions, validating its output against expert annotations. Throughout, we prioritized data security by selecting tools that preserve anonymity and minimize the risk of data breaches.

Results: Initial assessment revealed that transcription errors—such as incomprehensible segments, unusually short transcripts, and speaker diarization issues—were present in approximately one-third (n=36 out of 100) of a manually reviewed sample. Statistical outliers revealed that high speaking rate (>3.5 words per second) is associated with short transcripts and answering machine messages, while short conversation duration (<15 min) was an indicator for case management sessions. The 75th percentile of LLM perplexity scores was significantly higher in non-sessions than sessions (permutation test mean difference = −258, P =.02), although this feature alone offered only moderate classification performance (precision =0.63, recall =0.23 after outlier removal). In contrast, zero-shot LLM prompting effectively distinguished sessions from non-sessions with high agreement to expert ratings ($\varkappa$=0.71) while also capturing the nature of the meeting.

Conclusions: This study's hybrid approach effectively characterizes errors, evaluates content, and distinguishes text types within unstructured conversational dataset. It provides a foundation for research on conversational data, key methods, and practical guidelines that serve as crucial first steps in ensuring data quality and usability, particularly in the context of mental health sessions. We highlight the importance of integrating clinical experts with artificial intelligence tools while prioritizing data security throughout the process.

JMIR Form Res2025;9:e78082; doi: 10.2196/78082

Keywords: artificial intelligence; behavioral health; clinical documentation; clinical texts; conversational transcripts; data preprocessing; data quality assessment; health informatics; health information systems; large language models; natural language processing; psychotherapy; text classification

Introduction

As one of our primary communication tools, conversations offer a window into human relationships. Speaking with each other is one of the fundamental aspects of our existence. Consequently, transcripts of conversations have garnered significant interest in research across various disciplines [1]. Analyzing conversational data is a central research focus in fields such as health care [2,3], law [4], customer support [5], negotiations [6], education [7], and behavioral treatment and psychotherapy [8-10].

Among these diverse domains, talk therapy, which relies on a conversation between 2 or more individuals, is a particularly intriguing case for conversation analysis. Analysis of treatment sessions can enhance our understanding of different therapeutic protocols [9,11-14]; improve clinical training [15-17]; and provide insights into the relationships between conversational elements, therapeutic outcomes, and the therapeutic alliances [8,10,18,19]. Conversation analysis research of behavioral health sessions can provide comprehensive insights into the intricate dynamics of therapeutic interactions [2]. By meticulously examining communicative choices, researchers can understand how specific counseling strategies impact client engagement, identify successful intervention techniques, and reveal nonlinear patterns of change within therapy sessions [20,21]. This research methodology allows professionals to study the dyadic or group psychotherapeutic processes, predict potential treatment outcomes, develop more effective interventions, and ultimately enhance the quality of behavioral health care [14,22].

Historically, collecting, transcribing, and analyzing treatment conversations have required significant human effort. This is now changing, thanks to 2 major technological advances. First, recent developments in audio capturing technologies and the ubiquitousness of smart devices simplify the collection of conversation data. Second, the abundance of artificial intelligence (AI)–based speech-to-text tools [23-25] has automated transcription and speaker recognition tasks. These innovations have made it possible to capture high-quality human speech in diverse settings with less manual effort [26]. Consequently, large-scale transcription of audio files is now feasible with unprecedented speed and precision. The automation of these processes introduces new challenges. Transcripts' quality depends on several factors: the recording devices and their placement in the room, background noise, internet connection stability, and accurate speech-to-text and speaker diarization models. Additionally, characteristics of the conversation itself, such as slurred speech, dialects, rare languages, or use of slang, all pose further challenges for automatic speech recognition (ASR) models [27]. Similarly, interrupted, rapid, or overlapping speech, unknown number of speakers as well as same-sex speakers [28,29] each introduce complications for speaker recognition systems.

Moreover, the expected growth in large datasets accompanying this automation amplifies these challenges. Large and diverse datasets might include recordings of different quality levels and therefore are prone to various types of errors. Furthermore, when recordings are made routinely and automatically, unintentional recordings—irrelevant conversations, phone calls, empty moments, or accidental noise—may occur. This illustrates how large volumes of automatically generated transcripts are unstructured and susceptible to errors. Therefore, guidelines and methods for error handling, assessing, and filtering should be established.

These challenges are well-known in the broader field of health care, which has long grappled with extracting meaningful information from unstructured data such as clinical notes, discharge summaries, or patient-generated text [30,31]. As the field has evolved, various preprocessing pipelines and methodological frameworks have been proposed for handling missing metadata, variable data quality, and inconsistent documentation practices [32-34]. However, most existing work focuses on clinical records, and less attention has been given to large-scale, conversational health care data.

Yeomans et al [35] have recently proposed a methodological pipeline for building conversation-based research, starting with the planning and collection of conversations and going through editing and analyzing. However, with the availability of ambient AI and larger datasets, many researchers may be working with secondary data—datasets they have not collected themselves. Therefore, they might not have access to the original recordings, to the speech-to-text models, nor to the transcripts editing process. Lacking a ground truth for verification necessitates expanding these guidelines to address such challenges.

In the case of behavioral treatment sessions, the noisiness of the transcript and its accurate representation of the conversation are key for detecting the interventions applied and extracting treatment insights [25,36]. For instance, misidentifying speakers could compromise insights regarding

the therapeutic alliance [18,21,37]. Understanding which conversations are professional and which are accidental recordings is the first step in this endeavor.

In this paper the abovementioned challenges are addressed, and a methodological approach for preprocessing a large dataset of behavioral treatment transcripts without access to their respective original recordings is presented. First, a systematic approach for characterizing the data is outlined, followed by methods that allow its assessment for future analysis. This methodology is then illustrated by applying it to a large dataset of deidentified behavioral health sessions. Finally, the strengths and limitations of our approach are discussed. We hope to promote the integration of computational tools in traditional talk therapy and to offer relevant methods for any dataset of conversational transcripts.

Methods

Data and Settings

We analyzed 22,337 behavioral treatment sessions from 50 behavioral health programs across the United States collected through the Eleos Health platform between June 2020 and January 2024. Eleos Health's digital platform is designed to promote behavioral treatment quality by providing intervention feedback, supporting clinical decision-making, and enabling progress note automation. Sessions were processed as part of the routine implementation of ambient AI tools within participating behavioral health programs. All sessions underwent transcription, deidentification, and anonymization using Eleos' proprietary models before inclusion in the study. The research team had no access to audio recordings or transcripts prior to this process, and only deidentified data were analyzed. Processed transcripts were stored as comma-separated values files with 3 primary columns: deidentified speaker labels, timestamps, and content. Each content row contained a speech segment from either the therapist or the client—typically up to several sentences—segmented by the Eleos ASR model. Session metadata included random therapist and client ID numbers, organization name, session date, and treatment delivery method (phone, video conferencing, or in-person). This study was conducted in concert with the STROBE Checklist [38].

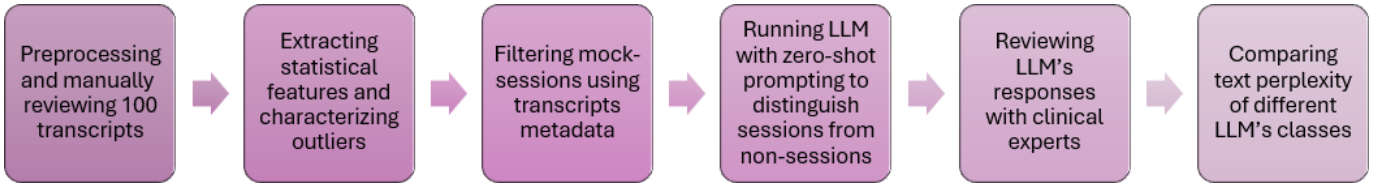
Ethical Considerations

This study was determined exempt from review by the Sterling Institutional Review Board under the Department of Health and Human Services Exemption Category 4. This exemption permits secondary research using identifiable health information when the data are either publicly available or recorded in such a way that subjects cannot be reidentified, and there is no direct interaction with participants. All research procedures adhered to applicable ethical guidelines and regulations. Clients and therapists provided informed consent for the use of anonymized, deidentified session data in secondary research conducted by the company. Both parties retained the option to opt out of having their sessions processed. All data used in this study were deidentified prior to analysis, and no identifiable information was accessed by the research team. No compensation was provided for participation in this secondary analysis, and no images or supplementary materials contain identifiable individuals.

Data Analysis Approach

Our data analysis approach consisted of several stages, each employing different methods for characterizing and filtering the dataset (see Figure 1). The initial stages were exploratory, beginning with a manual review to assess data's content and quality. This process helped us identify categories of files and potential errors in the data, enhancing our understanding of the dataset and the features most suitable for characterization and filtering. Subsequent steps focused on directly classifying transcripts as "sessions" or "non-sessions" (eg, nonprofessional conversations, noise, mock sessions). Using available metadata, we filtered out mock sessions (non-sessions) that had known identifiers. Additionally, we applied zero-shot prompting with a large language model (LLM) to distinguish therapy sessions from irrelevant conversations. Automation played a critical initial role in this phase; however, human expertise - specifically trained psychologists - was incorporated to validate the automated classifications. At this stage, aligned with our primary goal of preparing the dataset for psychological research, various types of interventions were distinguished, and the sessions category was restricted only to formal treatments (excluding peer support, sponsorship, and case management). Notably, we did not prompt the model with these exclusion criteria, testing its ability to make such distinction based solely on therapeutic elements.

Figure 1. Work pipeline—starting with manual review of a dataset subset, followed by statistical feature extraction and outlier detection. Next, we apply zero-shot prompting with a large language model (LLM) to classify transcripts as sessions or non-sessions and validate the LLM's decisions against human annotators. Finally, we compare LLM's perplexity of different classes.



Dataset Preprocessing

Dataset Characteristics

We identified potential duplicates in the dataset by analyzing the similarity between file names. This method is efficient in terms of processing time and computational resources, as it avoids comparing entire text contents. The underlying assumption is that identical files might be mistakenly saved under similar file names. To measure similarity, we used Python's SequenceMatcher function [39]. File names were considered similar if the ratio of their longest matching segment to the total number of characters exceeded 75%. Files meeting this threshold were then manually reviewed to confirm duplication.

Additionally, some transcripts were recorded in languages other than English. To detect the languages, we used the langdetect library in Python [40].

Initial Assessment

A total of 100 randomly selected transcripts were manually reviewed by human raters, with 12 segments analyzed from each transcript to identify common errors and quantify their prevalence. At least 4 categories of errors were defined: (1) non-sessions: transcripts that clearly did not represent a session. At this stage, no distinction was made between types of intervention (eg, peer support, sponsorship, or formal treatment); (2) too short: transcripts with a total duration of less than 2700 seconds (45 min, which approximates the expected session length); (3) unreadable: transcripts with excessive missing words, duplicated segments, or incomprehensible text—readers could not infer the missing content or the meaning of the segment. For example, a transcript with multiple repeated filler phrases (eg, "uh-huh, uh-huh, uh-huh…") was generally deemed unreadable; (4) speaker diarization errors: transcripts with substantial speaker attribution mistakes.

In some cases, annotators indicated that additional context was required to ascertain the presence of an error.

Features Learning

We collected statistics about the whole dataset and each session which included (1) conversation length, (2) speaking rate, (3) frequent words, and (4) segment perplexity.

In addition to characterizing the dataset, these features serve as potential indicators for assessing the data's compatibility for further analysis. For instance, they can highlight recording errors or determine whether the content is session related. In the following section, we elaborate on some of these features and explain how they were calculated for each session.

Conversation Length

Conversation length was calculated by extracting the end timestamp of the final speech segment in the transcript. This method does not account for any time elapsed between when the recording device was activated and the conversation began. However, if the session starts with silence, this approach accurately reflects the length of the conversation.

Speaking Rate

Speaking rate, defined as the average number of words spoken per second, has been employed to predict speaker characteristics [41], detect speech anomalies, and identify changes in conversational dynamics or context [42]. In American conversational speech, the reported average speaking rate ranges from 1.85 to 4.86 words per second (WPS) [41]. Values outside this range may indicate abnormal recordings, such as background noise or transcription errors.

For each transcript, speaking rate was calculated by averaging across segment-level speaking rates. Because silent pauses within segments were not available, we could not adjust for them as is commonly done. However, segments were brief (typically a few seconds) and cut before long silences, making the absence of silent moments negligible and this measure a reasonable approximation. For comparison, we also calculated a global speaking rate by dividing the total word count by the overall conversation length, which incorporates silent periods.

Word Frequency

We explored the content expressed in the transcripts by extracting the most common nouns in the dataset. First, we cleaned each text by removing stop words [43] and technical terms (a list of which can be found in Multimedia Appendix 1). These filtered words were typically used in bureaucratic contexts, such as scheduling meetings or filling out professional forms. Next, we separated each transcript into 2 speaker partitions: therapist and client. Using Python's library, Spacy [44], we collected all the nouns spoken across all transcripts for each speaker. This content analysis revealed common topics discussed during conversations. Additionally, unexpected words that emerged may indicate common errors made by the ASR model.

Perplexity

Perplexity is a measure of a text sequence, indicating its likelihood of being produced by a language model [45]. It is the model's loss when given the words in the sequence and is conceptually similar to the cross-entropy between the model's and the sequence's probability distribution:

$$Perplexity(segment) = exp\left(-\frac{1}{N}\sum_{i=1}^{N} log(P(w_i \mid w_{i-1,\ldots,}w_1))\right)$$

where w_i is a token in the segment's sequence. Due to the short length of segments, the number of tokens is typically smaller than the model's context length. Consequently, the probability of w_i is calculated considering all previous tokens in the sequence. High perplexity suggests that the model is less likely to generate such a segment.

We used OpenAI's *GPT-2* (locally, through the *Hugging Face* platform [46,47]) to compute the perplexity of approximately 5 million segments in 9067 transcripts. Perplexity was calculated at the segment level and then

aggregated within each transcript. We hypothesized that higher perplexity would reflect higher semantic complexity, thereby acting as a marker for actual sessions. However, we also anticipated that extreme values of perplexity could indicate recording errors or corrupted transcripts that do not represent a therapeutic conversation. Thus, we propose that therapeutic conversations will generally have higher average perplexities but will not show the highest maximum values. To test this assumption, we extracted the average perplexity, standard deviation of perplexity, and maximal perplexity for each transcript. Additionally, to get an accurate representation of the upper-bound perplexity values without being over-influenced by outliers, we calculated the 75th percentile of perplexity values for each transcript.

Classification—Distinguishing Sessions From Conversations

Model and Platform

To classify transcripts of behavioral treatment to sessions and non-sessions, we queried an LLM with a zero-shot approach. We used the Amazon Bedrock platform [48] that enables running closed models through a third party, ensuring that the data were not shared or exposed to the model's provider. The model selection out of the available platform's options was based on criteria of cost, context length, and proven reliability. Table S1 in Multimedia Appendix 2 presents the cost of different tokens at the time of writing. Anthropic's Claude [49] was chosen because of its proficiency in human tasks and cost-efficiency [49-51]. Both the LLM and the platform were carefully selected to ensure data security. Amazon's Bedrock platform served as a third party between the model and the user, enabling us to use the model non-locally without exposing our data. Data were not saved on the platform's servers, nor were they used for the benefit of training the model [48,52].

Filtering and Subsetting

Based on the metadata, we excluded files from organizations that are not clinical public entities, such as academic institutes conducting mock sessions and trials (25% filtered). From this subset, we randomly selected 850 transcripts for automatic classification.

Classification With LLM

"Claude-3-Sonnet" was applied through Amazon's Bedrock platform (model ID: "anthropic.claude-3-sonnet-20240229-v1:0") with the following parameters: maximal generated words (max_out): 350; temperature: 0.3; and the remaining parameters were set to default. A detailed prompt with guidelines specifying the criteria for classifying a transcript as a session, and a format for providing a well explained answer were created to ensure the model performed the task as expected. In collaboration with an expert clinical psychologist, we compiled a list of elements that define a behavioral treatment session. The prompt addressed the following 5 core elements of a behavioral treatment session:

Dynamics: a correspondence where one of the sides shares experiences and the other focuses on listening and responding.

Content: the conversation contains personal matters, emotions and experiences, discussions about personal goals, thoughts, or behaviors.

Therapeutic elements: demonstration of active listening and empathy by one side, and use of therapeutic techniques such as reframing, providing coping strategies.

Professional language: therapeutic terminology, references for treatment plans, or previous sessions.

Context clues: mentions of confidentiality, session time limit, or scheduling future appointments.

The prompt specifically instructed the model to look for these elements in its decision-making process, explain its reasoning, and rate its certainty on a scale of 1-5. Additionally, we asked the model to provide a brief summary of the conversation content and to indicate whether it identified nontherapeutic conversational dynamics. These instructions were designed to ensure that the model addressed the entire conversation and understood its content. To maintain consistency, we used XML tags (see Multimedia Appendix 3).

Validation by Clinical Expert and Interrater Reliability

To validate the model's classifications, 2 human raters independently classified 150 randomly selected transcripts. Both raters were graduate students in psychology or cognitive sciences, pursuing either clinical or theoretical training related to psychotherapy. Both were familiar with transcripts of behavioral health sessions prior to this task.

Raters were instructed to read transcript segments (approximately 12 turns) carefully and determine whether the segment belonged to an individual treatment session, excluding family or couple sessions, peer support, and case management calls. If a decision could not be made based on the provided segment, raters were encouraged to consult the full transcript.

These exclusion criteria were not part of the model's automated classification process and may therefore have introduced some discrepancies between human and model decisions. Nevertheless, our aim was to assess whether applying instructions focused on contextual and relational dynamics of the conversation would naturally result in the exclusion of nonprofessional conversations. Interrater agreement was assessed using Cohen kappa and percent agreement. Cohen kappa higher than 0.61 is generally interpreted as a substantial agreement [53]. We then calculated the percent agreement of the human raters with the model for each category (session, non-session).

Perplexity of Sessions Versus Non-Sessions

To measure whether the 2 perplexity distributions can be assigned to 2 different distributions—sessions and

non-sessions—we conducted a permutation test. Out of the transcripts for which we calculated segment perplexity, 335 transcripts were also assigned a class by the LLM: 285 sessions and 50 non-sessions.

We did not control the length of segments and their number, both are expected to affect the perplexity of a file. To ensure that the results are robust under filtering of extremely short transcripts and short segments, we calculated the same results after filtering segments with less than 5 or 10 words, and transcripts for which less than 10 or 20 segments. Finally, based on the results, we assessed the separability of sessions and non-sessions distributions by evaluating a perplexity-based classifier.

Results

Dataset Characteristics

Duplicates Identification and Language Detection

We identified 1 duplicate file and 1 empty file in the dataset. After removing these, the dataset consisted of 22,335 transcripts. The analysis revealed 18 different languages. The most common was English (98%), followed by Hebrew (0.7%).

Initial Assessment

The manual review of 100 transcripts yielded the following results: 46% of the transcripts were comprehensible and had *little to no errors*, 18% required a more thorough review for clear evaluation and 36% contained *clear errors*, which were categorized as follows (some transcripts had multiple error types): speaker identification errors (eg, confusing the therapist with the client, 42%); incomprehensible text (34%); too short to be indicative of the conversation type (22%); either non-session content or group sessions (11%) and *missing words or duplicated segments* (8%).

This preliminary evaluation found that the dataset was highly diverse, comprising the following transcript types:

1. Non-session content: The presence of non-session and group-session transcripts underscores the need for filtering mechanisms to exclude these from analysis.
2. Session quality: Based on this evaluation, we estimate that approximately half of the dataset contains sessions suitable for further analysis.
3. Speaker recognition errors: Given the frequency of speaker diarization errors, features that rely on speaker identification (eg, number of turns, spoken time by speaker) may be unreliable without corrections.
4. Incomprehensible text: This analysis underscores the importance of developing tools to improve text comprehensibility. With appropriate adjustments, more data could be rendered suitable for analysis.

Features Learning

Conversation Length

Conversations length ranged between 0.5 and 18,783 seconds, with an average of 2707 seconds and median of 3062 seconds. Figure 2 shows the full distribution of conversation length (A) and the same distribution after omitting conversations shorter than 15 minutes long (B). A Gaussian mixture model best fitted 3 Gaussian distributions for the full distribution and 2 for the reduced distribution. Figure 3 illustrates a comparison of the model's full distribution fit for different numbers of components, and Table 1 presents the values of the Akaike information criterion and the Bayesian information criterion under each parameter.

Figure 2. Comparison of transcript duration distributions before and after data cleaning. (A) Distribution of durations for all transcripts, including trial sessions and experimental recordings. (B) Distribution after excluding transcripts based on organization category metadata. The comparison highlights that many short-duration transcripts originated from irrelevant or excluded organizations.

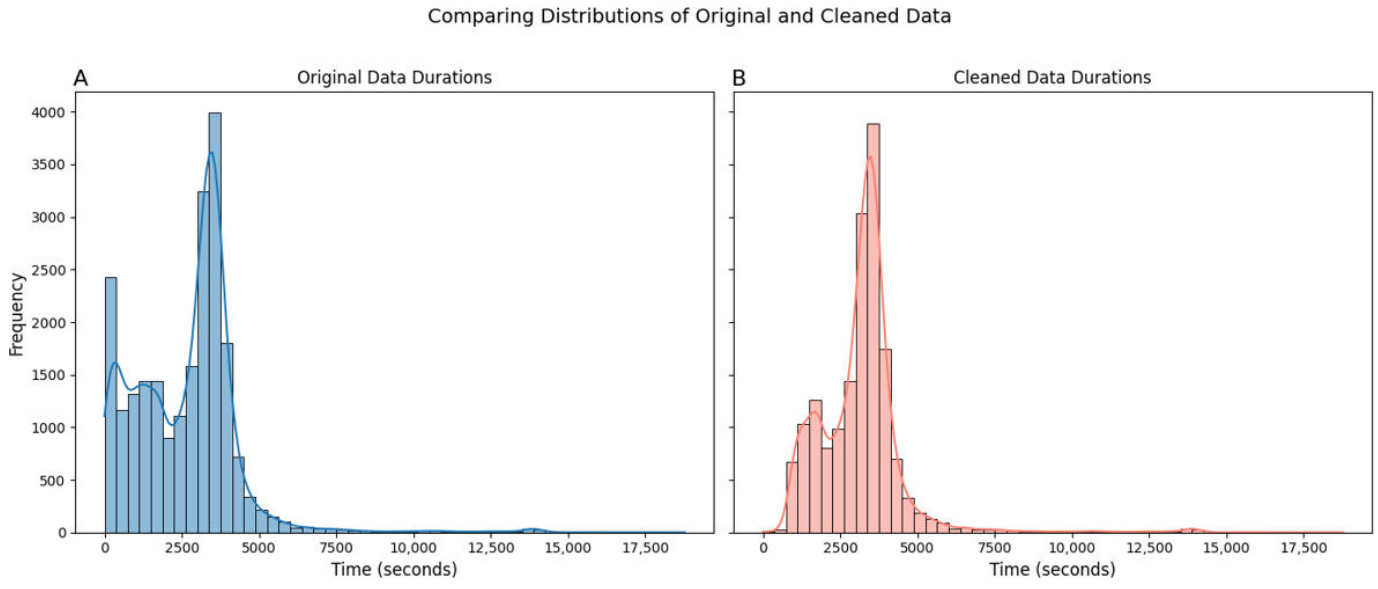


Figure 3. Comparison of Gaussian mixtures models (GMMs) with 2, 3, and 4 components fitted to the distribution of session durations (in seconds) based on transcript timestamps (Figure 2A). The negative log-likelihood is reported for each model, with lower values indicating better fit.

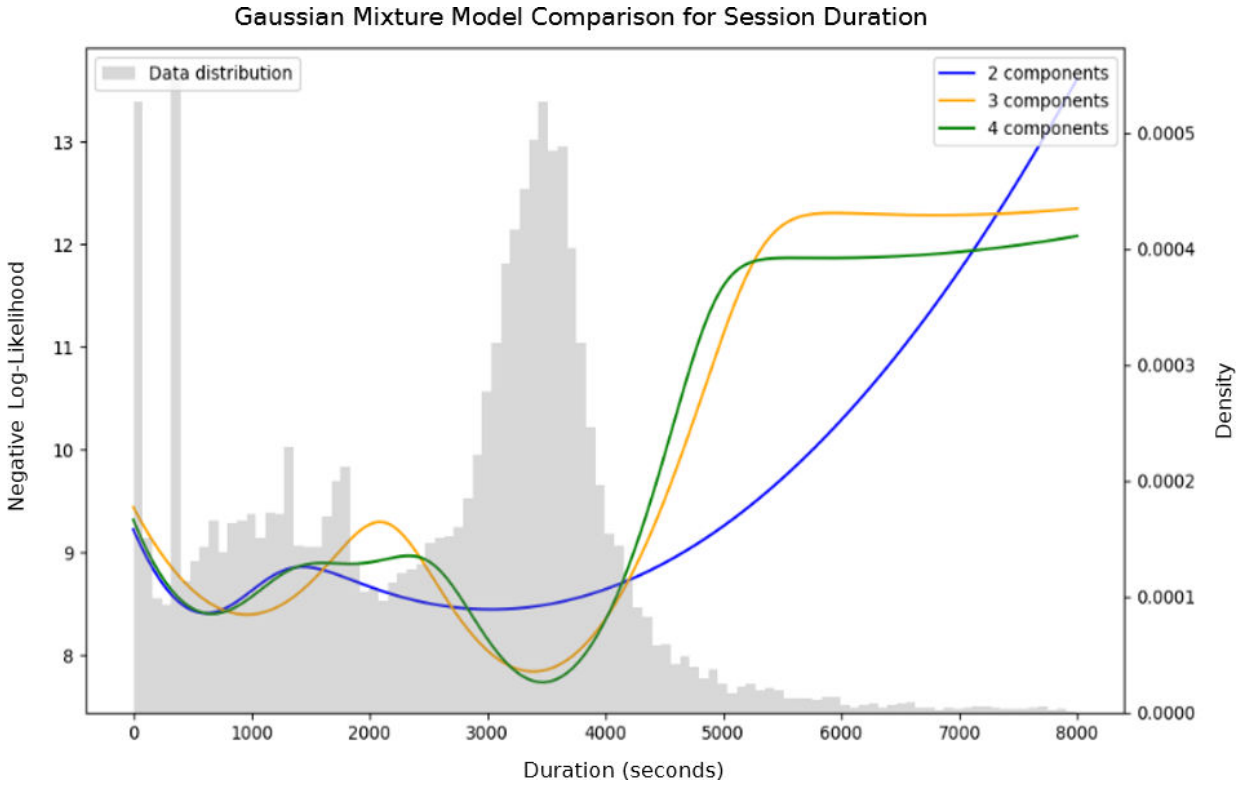


Table 1. Akaike information criterion and Bayesian information criterion for different number of components computed for the Gaussian mixture models presented in Figure 3.

Components	AIC[a]	BIC[b]
2	39,3727	39,3767
3	38,4168	38,4232
4	38,3384	38,3473

[a]AIC: Akaike information criterion.
[b]BIC: Bayesian information criterion.

To support our findings, we examined the metadata to determine whether different conversation types (based on organization) were associated with different transcript length. Among the short transcripts (<15 min) that included conversational content, we found brief case management phone calls. The long transcripts (>1.2 h) included 2 sessions in a row, or a recording extended before or after a session. These patterns suggest the presence of 2 conversation populations, each with a different duration distribution.

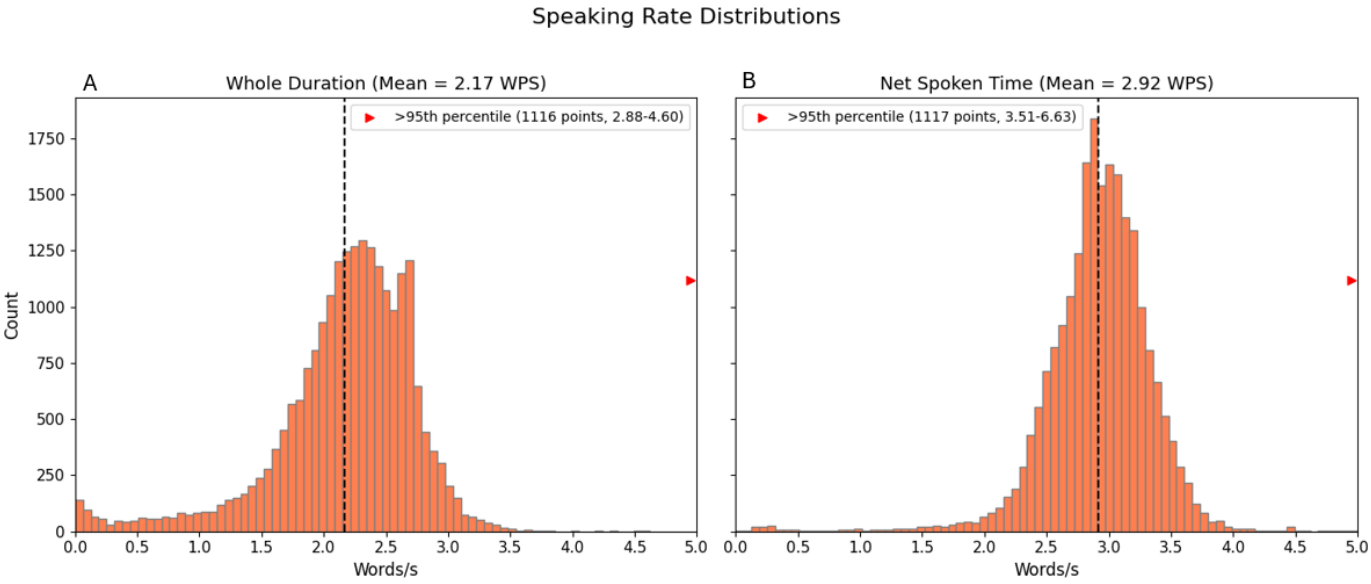
Speaking Rate

Speaking rate ranged from 0 to 4.4 WPS with an average of 2.17. The net speaking rate ranged from 0.04 to 6.63 WPS with an average of 2.9 WPS (Figure 4B). Both averages fall within the reported range of the American average speaking rate in conversation [41] (1.85-4.86 WPS).

The highest speaking rate was observed in a short transcript of an automatic answering machine. Only 30% (n=9) of transcripts with speaking rate above 3.5 WPS (Figure 4A) were longer than 20 minutes, suggesting that high speaking rate often reflected accidental recordings. For instance, of the 9 transcripts with rates greater than 4 WPS, 8 were identified as automatic voice answering machines.

Examining the ratio of overall session time to speech duration, or equivalently the ratio of speaking rate (A) to net speaking rate (B), shows that transcripts with high values (where total conversation time far exceeds spoken time) often reflect timestamp errors or short transcripts with very few words. In either case, such transcripts are not suitable for further analysis.

Figure 4. Distributions of average speaking rates extracted from transcripts with 2 methods: (A) Average number of words per second (WPS) over the whole session (including silences). (B) Average number of words per second for the net spoken time (without silences between segments). The red arrow's height represents the number of values exceeding the 95th percentile.



Frequent Noun Words

After filtering words according to Multimedia Appendix 1 , we identified 8,393,775 nouns in the clients' speech and 32,591 nouns in the therapists' speech. Multimedia Appendix 4 displays the 60 most common words before and after filtering. Therapists' 5 most common words were "thing," "time," "people," "know," and "way," and clients' most common nouns were similar: "thing," "time," "know," "date," and "lot" followed by "people." Among the 15 most common words for clients were also "friend," "mom," "work," "talk," and "life," and therapists shared some of these nouns ("work," "talk," and "date") but also frequently used "help" and "guy."

Classification

Zero-Shot Prompting

Of 850 transcripts that were given to the LLM, it identified 737 as sessions (86.7%, n=737), including 56 transcripts classified as couple's sessions. The remaining 113 transcripts were classified as non-sessions (13.3%, n=113). The model's certainty ratings were skewed toward the higher end of the scale, with 55 transcripts rated at 4 (7.9%) and 645 transcripts rated at 5 (92.1%, highest certainty).

Validation and Interrater Reliability

The raters agreed on 86.3% (44 of 51; see Figure 5A in Multimedia Appendix 5) of the test set transcripts, predominantly classifying transcripts as therapy sessions rather than nontherapy sessions (Rater 1: 61%; Rater 2: 63%). Cohen kappa score was 0.71, indicating substantial agreement [53]. Disagreements primarily revolved around identifying the session type—distinguishing between individual, couple, or family therapy—rather than determining whether a therapy session took place at all. In 43% (3 of 7) of the disagreement cases, both raters classified the transcript as a therapy session but differed on the type. In the remaining 57% (n=4), the disagreement was about whether the conversation was professional or casual. The full distribution of classifications by the raters and the model is shown in Multimedia Appendix 5.

LLM Versus Raters

Unlike the raters, who were instructed to mark non-individual sessions (eg, family or couple therapy) as non-sessions, the LLM was instructed to identify any kind of treatment session, including family and couple therapy. This difference naturally led to disagreements between raters and the model. However, when the model explicitly identified the session type (eg, family or couples) and indicated it clearly in its explanation or summary, we considered the classification an agreement. This verification process can be automated by searching for specific keywords (eg, "couple session") in the summary text or by using another language model to interpret the model's output. Table 2 shows the distribution of rater-model agreement for sessions and non-sessions.

Table 2. Agreement between the model and the raters for transcripts agreed over by raters[a].

	Raters: yes, n	Raters: no, n	Total, n
	26	7	33
	2	9	11
Total, n	28	16	44

[a]The table features the number of transcripts in each category.

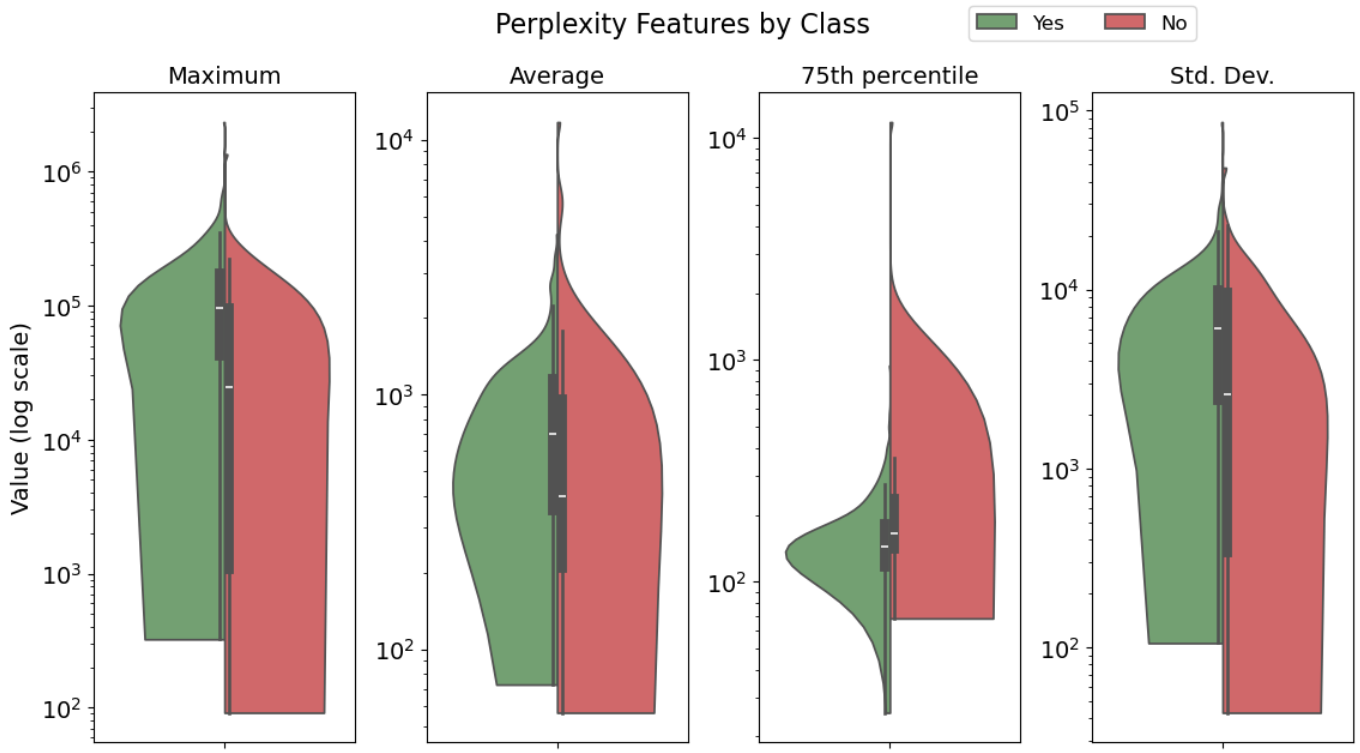
Overall, the model identified 57 sessions as couples or family therapy, demonstrating its ability not only to detect therapy sessions but also to categorize them by type.

Among the 9 disagreements between the model and the human raters, most involved the model misclassifying case management conversations as therapy. Conversely, in cases where the model misidentified conversations tagged as therapy sessions by human raters, its reasoning was because of "lack of therapeutic techniques."

Perplexity of Different Classes

Figure 5 shows the distributions of different statistics of perplexity of transcripts for different classes.

Figure 5. Results of permutation tests between distributions of perplexity metrics of sessions ("Yes") and non-sessions ("No"). Perplexity metrics—maximum, average, 75th percentile, and standard deviation—were calculated over the distribution of segment perplexity for each transcript. The 75th percentile showed a significant result (mean difference = –258, *P*=.01) with sessions having lower values, while the maximum perplexity was higher for sessions (mean difference =73,888 , *P*=.007).

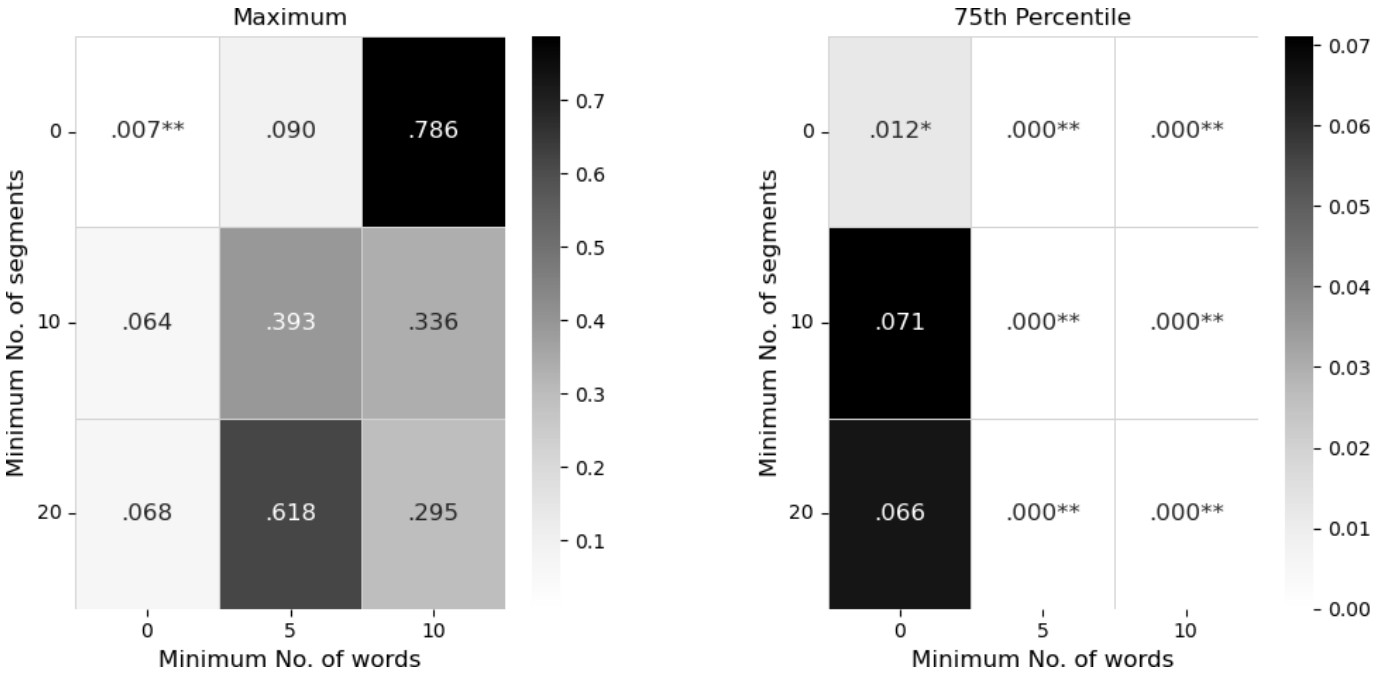


Among the transcripts with high perplexities, 1 transcript had exceptionally high values (mean perplexity of more than 11,000, which was approximately 11 standard deviations more than the average mean perplexity). This 4-segment transcript contained nonwords and backchannels that seemed to be transcribed noises (mostly the sound "Mhm"). Another high-perplexity file was a recorded rap song.

Of the 4 statistics—max, mean, standard deviation, and 75th percentile—a permutation test for means showed a significant result for the 75th percentile (*P*=.01), with sessions having a lower mean 75th percentile than non-sessions (mean difference=–258). In contrast, the permutation test showed that the sessions group had higher maximal values (*P*=.007).

To check the robustness of these results, we repeated the test after limiting the calculation to segments with minimal number of words (minimum words per segment [MWPS]=5, 10) and omitting transcripts with fewer than a minimal number of relevant segments (minimal number of segments [MS]=10, 20). These tests showed that the 75th percentile measure (Figure 6A) remained significant under these parameter changes, whereas the max perplexity measure (Figure 6B) lost significance when limitations were applied (MS>0 or MWPS >0).

Figure 6. *P* values of permutation tests comparing sessions and non-sessions perplexity measures across parameter settings—minimum number of segments and minimal number of words per segment. (A) Maximum perplexity shows no significant values across parameters except when no restrictions are applied and (B) 75th percentile of perplexity is significant for most parameter combinations. Bright colors indicate lower *P* values, **P*<.05, ***P*<.01.



Furthermore, we evaluated the classification performance of a 75th percentile perplexity–based classifier by analyzing receiver operating characteristic (ROC) curves and precision-recall metrics across parameters (MWPS, MS). To align with perplexity values, in this analysis, higher perplexity values correspond to the positive class—non-sessions—while sessions constitute the negative class. The optimal threshold was determined by maximizing the difference between true positive rate and false positive rate, prioritizing accurate classification of non-sessions. The results (Table 3) indicate moderate discriminative ability, with ROC area under the curve values ranging from approximately 0.62 to 0.73 and precision increasing alongside minimal number of segments. Precision tends to increase with larger segment sizes (up to ~0.63), while recall decreases (~0.26).

Table 3. Classification performance of the 75th percentile perplexity–based filter across varying thresholds for minimal words per segment (MWPS) and minimal number of segments (MS).

MWPS	MS	Sessions	Non-sessions	ROC AUC	PR AUC	Precision	Recall
0	0	285	50	0.618	0.240	0.200	0.740
0	5	285	43	0.636	0.198	0.182	0.767
0	10	284	42	0.633	0.196	0.178	0.762
5	0	285	47	0.664	0.354	0.353	0.255
5	5	284	41	0.657	0.331	0.333	0.268
5	10	284	40	0.665	0.335	0.333	0.275
10	0	285	44	0.687	0.378	0.625	0.227
10	5	284	40	0.701	0.380	0.625	0.250
10	10	284	38	0.726	0.392	0.625	0.263

[a]The table reports the number of transcripts remaining in each class after filtering outliers, the receiver operating characteristic area under the curve (ROC AUC), precision-recall AUC (PR AUC), precision, recall, and F1 scores. Increasing the minimal words and segments thresholds improves ROC AUC and precision but reduces recall.

Discussion

Principal Findings

With a dataset of over 22,000 unfiltered transcripts of recordings associated with behavioral treatment sessions, our primary objective in this methodological paper was to illustrate a systematic approach for characterizing the data and implementing a filtering process for subsequent analysis in academic research. Prior research on leveraging machine learning and natural language processing methods for classifying large text datasets, both conversational [54] and

nonconversational [55,56], has focused mostly on content-based classification rather than the contextual framework of conversations. Our proposed methodology integrated human evaluation, statistical analysis, and automated tools, including LLMs, emphasizing the importance of contextual and relational features, which are especially critical for distinguishing therapy sessions from non-sessions.

The preliminary analysis identified a significant proportion of non-session transcripts within the dataset. These non-sessions encompassed brief case management encounters, informal conversations, mock sessions, and incidental recordings captured between sessions. Additionally, some transcripts included errors such as incorrect speaker identification, text duplication, and incomprehensible content [25]. These findings highlight the necessity for systematic filtering mechanisms to exclude low-quality and non-session data. Given that only approximately half of the dataset was found suitable for analysis, leveraging automated classification and quality scoring can significantly enhance the dataset's utility for research purposes.

Recent works have listed features for characterizing texts in corpora [35,57]. In this paper, we have focused on 3 basic ways to extract relevant statistics and examined if outliers could be indicative of problematic transcripts:

Conversation length—Analysis of conversation length distributions revealed that shorter transcripts often represented non-sessions, such as case management encounters, as expected for this type of service, phone calls, or noise. In contrast, unexpectedly long conversations may result from sessions being concatenated or recordings extending beyond the session's actual duration, thus requiring careful preprocessing.

Speaking rate—Speaking rate has been evaluated in many contexts, and it has been shown that its values can predict speakers' features [41], detect speech anomalies, and identify change in conversation dynamics or context [42]. We found that high speaking rate was associated with shorter transcripts, particularly erroneous recordings such as answering machine messages. This finding supports the use of speaking rate as a potential indicator of non-sessions. Additionally, the ratio between speech duration and overall recording duration may serve as a marker for time-labeling errors.

Content analysis—Bag-of-Words is a basic tool for evaluating semantic content. Through word count (mostly nouns), themes and topics can be discovered and ultimately enhance text classification [58-60]. In this paper, we used it as a "reality check" to reveal the contents of the transcripts at hand. Our content analysis provided insights into the vocabulary and topics typical of therapeutic conversations. The most prevalent words were related to everyday life ("relationship," "house," "job," "today," "school," "sleep," and "car"), clients' inner world ("thought," "want," "feel," and "need"), resolving issues ("situation," "problem," "help," and "care"), and relationships ("kid," "love," "date," "sister," "couple," "dad," and "guy"), emphasizing the centrality of these themes in these conversations. Comparing individual transcript themes against the topics that were found in this analysis could be an indicator for being a genuine session. However, it is of note that this analysis ignores the context in which words appear.

Perplexity Analysis

Perplexity is traditionally used to evaluate language models [61,62]. In this study, however, we used it to evaluate the comprehensibility and uniqueness of text segments [63,64] in order to gain insight into potential differences between sessions and non-sessions. Transcripts with higher perplexity scores often contained transcription errors or nonverbal content (eg, noise and backchannels).

Sessions initially exhibited higher maximal transcript perplexity than non-sessions; however, this result was no longer significant when outliers, namely, transcripts with few segments or short segments, were omitted. This suggests that the higher maximal perplexity values in sessions are not necessarily indicative of more verbally complex content but rather stem from transcription errors, backchanneling, discourse markers, or interrupted recordings, which tend to produce short segments and short transcripts. These findings indicate that sessions, in general, do not have higher maximal perplexity than non-sessions. This interpretation is further supported by the finding that the 75th percentile perplexity score for sessions was lower than non-sessions after removing outliers. This may suggest that sessions contain more structured and predictable language, particularly when compared to background noise or accidentally captured conversation fragments. This distinction is particularly useful because it suggests that 75th percentile perplexity could serve as a reliable feature in a non-session filtering model. Unlike maximal perplexity, which was influenced by errors and outliers, the 75th percentile measure remains stable across different parameter settings, reinforcing its robustness as an indicator of session-like structure.

To support this claim and further evaluate its ability to differentiate sessions from non-sessions, we assessed a 75th percentile perplexity–based classifier across varying thresholds for MS and MWPS. The optimal threshold remained consistent across conditions, suggesting its practical applicability as a filtering criterion. However, the results indicate moderate-to-low discriminative performance, with precision improving under stricter outlier exclusion. For instance, removing transcripts with fewer than 10 segments of more than 10 words each yielded the highest precision (~0.63) but substantially reduced recall (~0.26). This pattern likely reflects the dataset imbalance—where non-sessions are the minority—and the observation that outliers were predominantly non-sessions, suggesting they tend to exhibit relatively high 75th percentile perplexity. These findings reinforce the conclusion that this metric is particularly useful when filtering aims to prioritize precision, though caution is needed due to the class imbalance impacting these metrics.

Overall, our results suggest that perplexity can be used to identify flawed, incomprehensible, or highly improbable text segments and may serve as a useful tool for detecting low-quality or non-session transcripts, but not as a standalone

classifier. To validate our results and extend this work, we propose 2 key next steps: (1) to distinguish nonprofessional conversations from flawed transcripts to determine which group is more clearly separable from sessions through perplexity measures. This distinction is critical for understanding the sources of high- and low-perplexity segments. (2) To replicate these analyses with a larger set of labeled transcripts to overcome data imbalance.

Large Language Model Classification

Research on LLMs has explored their use for both text classification and mental health text analysis as separate tasks. Prior work has demonstrated these models' ability to extract key concepts, analyze text dynamics, and identify psychological concepts [14,65-69]. In this study, we integrated both tasks, leveraging zero-shot prompting to analyze mental health conversations with the goal of classification. Our findings indicate that with zero-shot prompting, the model can classify transcripts effectively, showing high agreement with human coders while demonstrating robustness in handling transcript errors. For example, even when speaker identification referenced only a therapist and a patient, the model correctly identified couple or family sessions by interpreting contextual and semantic cues. Additionally, the LLM successfully corrected speaker identification errors, highlighting its potential as an automated error-correction tool.

Instances of disagreement between human coders and the LLM often stemmed from the model's sensitivity to therapeutic techniques. One of the classification guidelines was the presence of such techniques, and the model appeared to weigh them heavily when the conversational framing was ambiguous. Furthermore, in some cases, we found discrepancies between the binary classification and the LLM's explanation. For instance, despite classifying a conversation as a session, the model noted: *The therapist and psychologist discuss updates on multiple client cases, including challenges with clients' behaviors*, recognizing both speakers as therapists. This example underscores the importance of including model explanations as part of the classification process. Future work should investigate these discrepancies and refine classification guidelines accordingly.

Finally, while research on LLMs' ability to analyze conversations through prompting is still evolving, existing studies have yielded inconsistent findings regarding their effectiveness in analyzing conversational contexts [14,65-69]. Our results suggest that when provided with full conversation transcripts, LLMs can capture nuanced textual and relational dynamics, offering valuable insights into participants' interactions. While recent studies have used these capabilities in text-based applications, we demonstrated their applicability to conversations, where relational and dynamic elements are crucial in distinguishing sessions from non-sessions. Thus, although few-shot learning or fine-tuning may further enhance classification accuracy [55,56], our findings suggest that these techniques may not be strictly necessary for effective classification.

Limitations

Defining when a transcript reflects a treatment session presents a challenge: sessions share many aspects with nontherapeutic conversations, and some therapeutic techniques might appear as trivial exchanges. In this paper, we proposed both statistical heuristics and explicit guidelines to evaluate conversations in light of this question.

While we offer a methodological examination of a transcript dataset and highlight considerations for careful filtering, defining a filtering model is beyond the scope of this study. Since ground-truth labels were unavailable for supervised classification, we address this by identifying features that provide valuable insights into the classification process and by demonstrating how LLMs can assist experts when guided by well-designed prompt engineering. Hence, the approach presented here may serve as a foundation for developing a comprehensive classification model.

This study delineates a general framework and provides guidelines for working with conversational datasets in the absence of prior knowledge or structured labels. However, variation exists across datasets in format, content, quality, and the proportion of non-session files and transcription errors. While this paper applies its suggested framework on a large and varied dataset, it still features working with a single data platform. Variation in platforms and datasets calls for researchers to (1) detect any peculiarities relevant to their datasets and acknowledge them before referring to the relevant steps in the framework and (2) cultivate content matter expertise for datasets they analyze or better—conduct studies in multidisciplinary teams and keep a human (expert) in the loop.

Future Research

Our findings suggest that while statistical features can provide broad insights into conversational structure and flag irregularities that may correlate with non-sessions, LLMs add a complementary layer by capturing the essence, dynamics, and nature of conversations. As a proof of concept, this study demonstrates the potential of LLM-based classification; however, further validation, including human annotation on larger samples, is needed. In a treatment setting, we envision a semiautomated pipeline that integrates statistical and semantic features with LLMs. Transcripts would first be screened using metrics such as perplexity, duration, and speaking rate; those within a normal range would then be analyzed by an LLM to classify the conversation type—treatment session or unrelated. Periodic human review of selected transcripts would ensure model validity, with reviewer feedback used to refine prompts or feature thresholds dynamically. This "human-in-the-loop" workflow enables continuous processing while limiting manual review to a small subset of cases. Future work should evaluate different LLM models across diverse behavioral health datasets to determine their ability to capture nuances in treatment styles.

Regarding statistical feature–based outliers filtering, additional research could enhance its efficiency and

interpretability. Incorporating more informative linguistic features—such as the frequency of backchanneling cues, discourse markers, and silent pauses—into interpretable models such as decision trees may improve classification while remaining computationally efficient compared to fine-tuning large-scale models. Moreover, content analysis could reveal how the themes extracted from an individual session differ from those found across the dataset, highlighting deviations that might indicate non-session transcripts. Examining bigrams and multi-word expressions could also refine differentiation between informal conversations and structured therapeutic sessions, while helping to detect common transcription errors, such as misinterpretations caused by background noise.

Perplexity analyses could examine the relationship between different segment error types, such as misspelling, and perplexity by categorizing segments accordingly. Additionally, applying a mixed-model analysis to account for the statistical dependencies of perplexities within individual transcripts. Moreover, the interaction between perplexity and factors such as speaker familiarity or amorphous nature of the conversation remains an open question and could help explain cases where high perplexity signals meaningful conversational ambiguity rather than transcription errors. Finally, improving transcript quality remains a critical avenue for future research. Existing methods for correcting speaker diarization errors and ASR mistakes without access to the original audio [70,71] could be integrated into the preprocessing pipeline to enhance transcript reliability before filtering. Implementing these techniques could enhance the preprocessing stage, improve the accuracy of extracted features, and ultimately enhance the filtering process and its success rates.

Conclusion

This study demonstrated the importance of integrating human judgment with automated tools when processing large, unstructured datasets. We assess secondary data—data collected independently of current research—where initial human evaluation is critical for understanding dataset characteristics such as readability, content, and diarization quality. This foundational knowledge can inform the development of effective filtering strategies. While basic statistics, perplexity, and LLM prompting facilitate automated filtering, preliminary human review remains essential for understanding dataset variability and refining classification features. This hybrid approach ensures adaptable and accurate filtering processes, even in the presence of transcription errors.

Disclaimer

This manuscript was partially edited for language and style with the assistance of the AI language models ChatGPT (OpenAI) and Claude (Anthropic). The authors reviewed and edited all AI-generated content and take full responsibility for the final text.

Data Availability

The dataset generated and analyzed during the current study is not publicly available due to privacy and confidentiality reasons.

Conflicts of Interest

SSS, SJ, and ES are employees of Eleos Health whose artificial intelligence platform was used to generate the data analyzed in this study. DPM is the Founder of Morrison Consulting, which provides consulting services to Eleos Health, including his role as Chief Clinical Officer; he is affiliated with Eleos Health in a consulting capacity, but not as an employee. PMN and AG declare no conflicts of interest.

Multimedia Appendix 1

Clean-up words.
[PNG File (Portable Network Graphics File), 122 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Amazon Bedrock models prices.
[PNG File (Portable Network Graphics File), 36 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Prompt.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Most common words: before and after filtering.
[PNG File (Portable Network Graphics File), 403 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Session versus non-session classification results: (A) inter-rater answers distribution and (B) distribution of agreement between the model and the human raters.
[PNG File (Portable Network Graphics File), 102 KB-Multimedia Appendix 5]

References

1. Whalen J, Raymond GT. Conversation analysis. In: Borgatta EF, Montgomery RJV, editors. Encyclopedia of Sociology. 2nd ed. Macmillan Reference USA; 2000. ISBN: 9780028648507
2. Drew P, Chatwin J, Collins S. Conversation analysis: a method for research into interactions between patients and health-care professionals. Health Expect. Mar 2001;4(1):58-70. [doi: 10.1046/j.1369-6513.2001.00125.x] [Medline: 11286600]
3. Park J, Kotzias D, Kuo P, et al. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. J Am Med Inform Assoc. Dec 1, 2019;26(12):1493-1504. [doi: 10.1093/jamia/ocz140] [Medline: 31532490]
4. Fraser H. A framework for deciding how to create and evaluate transcripts for forensic and other purposes. Front Commun. 2022;7:898410. [doi: 10.3389/fcomm.2022.898410]
5. Følstad A, Taylor C. Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. Qual User Exp. Dec 2021;6(1):6. [doi: 10.1007/s41233-021-00046-5]
6. Jeong M, Minson J, Yeomans M, Gino F. Communicating with warmth in distributive negotiations is surprisingly counterproductive. Manage Sci. Dec 2019;65(12):5813-5837. [doi: 10.1287/mnsc.2018.3199]
7. Hennessy S, Calcagni E, Leung A, Mercer N. An analysis of the forms of teacher-student dialogue that are most productive for learning. Language and Education. Mar 4, 2023;37(2):186-211. [doi: 10.1080/09500782.2021.1956943]
8. Goldberg SB, Flemotomos N, Martinez VR, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. J Couns Psychol. Jul 2020;67(4):438-448. [doi: 10.1037/cou0000382] [Medline: 32614225]
9. Spinrad A, Taylor CB, Ruzek JI, et al. Action recommendations review in community-based therapy and depression and anxiety outcomes: a machine learning approach. BMC Psychiatry. Feb 16, 2024;24(1):133. [doi: 10.1186/s12888-024-05570-0] [Medline: 38365635]
10. Atzil-Slonim D, Eliassaf A, Warikoo N, et al. Leveraging natural language processing to study emotional coherence in psychotherapy. Psychotherapy (Chic). Mar 2024;61(1):82-92. [doi: 10.1037/pst0000517] [Medline: 38236227]
11. Sadeh-Sharvit S, Hollon SD. Leveraging the power of nondisruptive technologies to optimize mental health treatment: case study. JMIR Ment Health. Nov 26, 2020;7(11):e20646. [doi: 10.2196/20646] [Medline: 33242025]
12. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res. Jul 10, 2023;25:e46781. [doi: 10.2196/46781] [Medline: 37428547]
13. Imel ZE, Pace BT, Soma CS, et al. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. Psychotherapy (Chic). Jun 2019;56(2):318-328. [doi: 10.1037/pst0000221] [Medline: 30958018]
14. Abdou M, Sahi RS, Hull TD, Nook EC, Daw ND. Leveraging large language models to estimate clinically relevant psychological constructs in psychotherapy transcripts. Preprint posted online on 2025. [doi: 10.1101/2025.03.04.25323338]
15. Ruzek JI, Sadeh-Sharvit S, Bunge EL, et al. Training the psychologist of the future in the use of digital mental health technologies. Prof Psychol: Res Pract. 2024;55(5):395-404. [doi: 10.1037/pro0000567]
16. Flaherty HB. Teaching note—using technology to enhance experiential learning through simulated role plays. J Soc Work Educ. Oct 2, 2023;59(4):1294-1300. [doi: 10.1080/10437797.2022.2050869]
17. Sadeh-Sharvit S, Rego SA, Jefroykin S, Peretz G, Kupershmidt T. A comparison between clinical guidelines and real-world treatment data in examining the use of session summaries: retrospective study. JMIR Form Res. Aug 16, 2022;6(8):e39846. [doi: 10.2196/39846] [Medline: 35972782]
18. Shapira N, Atzil-Slonim D, Tuval Mashiach R, Shapira O. Measuring linguistic synchrony in psychotherapy. Presented at: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology; Jul 14-15, 2022; Seattle, USA. [doi: 10.18653/v1/2022.clpsych-1.14]
19. Atzil-Slonim D, Soma CS, Zhang X, Paz A, Imel ZE. Facilitating dyadic synchrony in psychotherapy sessions: systematic review and meta-analysis. Psychother Res. Sep 2023;33(7):898-917. [doi: 10.1080/10503307.2023.2191803] [Medline: 37001119]
20. Yonatan-Leus R, Gwertzman G, Tishby O. Using machine learning methods to identify trajectories of change and predict responders and non-responders to short-term dynamic therapy. Psychother Res. Sep 2025;35(7):1070-1086. [doi: 10.1080/10503307.2024.2420725] [Medline: 39461002]
21. Atzil-Slonim D, Juravski D, Bar-Kalifa E, et al. Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy. Psychotherapy (Chic). Jun 2021;58(2):324-339. [doi: 10.1037/pst0000362] [Medline: 33734743]

22. Althoff T, Clark K, Leskovec J. Large-scale analysis of counseling conversations: an application of natural language processing to mental health. Trans Assoc Comput Linguist. 2016;4:463-476. [Medline: 28344978]
23. Radford A, Kim JW, Xu T, Brockman G, Mcleavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Presented at: Proceedings of the 40th International Conference on Machine Learning; Jul 23-29, 2023:28492-28518.
24. Bain M, Huh J, Han T, Zisserman A. WhisperX: time-accurate speech transcription of long-form audio. Presented at: INTERSPEECH 2023; Aug 20-24, 2023:4489-4493; Dublin, Ireland. [doi: 10.21437/Interspeech.2023-78]
25. Miner AS, Haque A, Fries JA, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med. 2020;3(1):82. [doi: 10.1038/s41746-020-0285-8] [Medline: 32550644]
26. Zolnoori M, Vergez S, Kostic Z, et al. Audio recording patient-nurse verbal communications in home health care settings: pilot feasibility and usability study. JMIR Hum Factors. May 11, 2022;9(2):e35325. [doi: 10.2196/35325] [Medline: 35544296]
27. Graham C, Roll N. Evaluating OpenAI's Whisper ASR: performance analysis across diverse accents and speaker traits. JASA Express Lett. Feb 1, 2024;4(2):025206. [doi: 10.1121/10.0024876] [Medline: 38391582]
28. Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S, Narayanan S. A review of speaker diarization: recent advances with deep learning. Comput Speech Lang. Mar 2022;72:101317. [doi: 10.1016/j.csl.2021.101317]
29. Church K, Zhu W, Vopicka J, Pelecanos J, Dimitriadis D, Fousek P. Speaker diarization: a perspective on challenges and opportunities from theory to practice. Presented at: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Mar 5-9, 2017:4950-4954; New Orleans, LA. [doi: 10.1109/ICASSP.2017.7953098]
30. Hao T, Huang Z, Liang L, Weng H, Tang B. Health natural language processing: methodology development and applications. JMIR Med Inform. Oct 21, 2021;9(10):e23898. [doi: 10.2196/23898] [Medline: 34673533]
31. Elbattah M, Arnaud É, Gignon M, Dequen G. The role of text analytics in healthcare: a review of recent developments and applications. Presented at: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies; Feb 11-13, 2021; Vienna, Austria. [doi: 10.5220/0010414508250832]
32. Bazoge A, Morin E, Daille B, Gourraud PA. Applying natural language processing to textual data from clinical data warehouses: systematic review. JMIR Med Inform. Dec 15, 2023;11:e42477. [doi: 10.2196/42477] [Medline: 38100200]
33. Patel KN, Kiran P. Preprocessing methods for unstructured healthcare text data. IJITEE. 2019;9(2S):715-719. [doi: 10.35940/ijitee.B1024.1292S19]
34. Sedlakova J, Daniore P, Horn Wintsch A, et al. Challenges and best practices for digital unstructured data enrichment in health research: a systematic narrative review. PLOS Digit Health. Oct 2023;2(10):e0000347. [doi: 10.1371/journal.pdig.0000347] [Medline: 37819910]
35. Yeomans M, Boland FK, Collins HK, Abi-Esber N, Brooks AW. A practical guide to conversation research: how to study what people say to each other. Adv Meth Pract Psychol Sci. Oct 2023;6(4). [doi: 10.1177/25152459231183919]
36. Atkins DC, Steyvers M, Imel ZE, Smyth P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implement Sci. Apr 24, 2014;9(1):49. [doi: 10.1186/1748-5908-9-49] [Medline: 24758152]
37. Koole SL, Tschacher W. Synchrony in psychotherapy: a review and an integrative framework for the therapeutic alliance. Front Psychol. 2016;7:862. [doi: 10.3389/fpsyg.2016.00862] [Medline: 27378968]
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. Apr 2008;61(4):344-349. [doi: 10.1016/j.jclinepi.2007.11.008]
39. difflib—helpers for computing deltas. URL: https://docs.python.org/3/library/difflib.html [Accessed 2024-12-25]
40. Langdetect. Dec 2021. URL: https://pypi.org/project/langdetect/ [Accessed 2024-12-25]
41. Yuan J, Liberman M, Cieri C. Towards an integrated understanding of speaking rate in conversation. Presented at: INTERSPEECH 2006; Sep 17-21, 2006; Pittsburgh, PA, USA. URL: https://www.isca-archive.org/interspeech_2006 [doi: 10.21437/Interspeech.2006-204]
42. Wardle M, Cederbaum K, de Wit H. Quantifying talk: developing reliable measures of verbal productivity. Behav Res Methods. Mar 2011;43(1):168-178. [doi: 10.3758/s13428-010-0019-y] [Medline: 21287128]
43. Bird S. NLTK: the natural language toolkit. Presented at: Proceedings of the COLING/ACL on Interactive Presentation Sessions; Jul 17-18, 2006; Sydney, Australia. URL: https://aclanthology.org/P06-4000/ [Accessed 2025-01-05]
44. Honnibal M, Montani I, Landeghem S, Boyd A. SpaCy: industrial-strength natural language processing in python. [doi: 10.5281/zenodo.1212303] [Medline: 7069443]
45. Colla D, Delsanto M, Agosto M, Vitiello B, Radicioni DP. Semantic coherence markers: the contribution of perplexity metrics. Artif Intell Med. Dec 2022;134:102393. [doi: 10.1016/j.artmed.2022.102393]

46. Wolf T, Debut L, Sanh V, et al. HuggingFace's transformers: state-of-the-art natural language processing. Preprint posted online on 2019. URL: https://arxiv.org/abs/1910.03771 [Accessed 2024-12-26]
47. HuggingFace GPT2. URL: https://huggingface.co/gpt2 [Accessed 2024-12-26]
48. Amazon's Bedrock docs. URL: https://docs.aws.amazon.com/bedrock [Accessed 2024-12-26]
49. Anthropic News. Introducing Claude 3.5 Sonnet. Anthropic News. Jun 21, 2024. URL: https://www.anthropic.com/news/claude-3-5-sonnet [Accessed 2025-03-05]
50. Anthropic. claude 3.5 sonnet model card addendum. 2024. URL: https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf [Accessed 2025-03-05]
51. Jin H, Guo J, Lin Q, Wu S, Hu W, Li X. Comparative study of Claude 3.5-Sonnet and human physicians in generating discharge summaries for patients with renal insufficiency: assessment of efficiency, accuracy, and quality. Front Digit Health. 2024;6:1456911. [doi: 10.3389/fdgth.2024.1456911] [Medline: 39703756]
52. Amazon's Bedrock Data Security. URL: https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html [Accessed 2025-02-15]
53. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159. [doi: 10.2307/2529310]
54. Rathor S, Jadon RS. Domain classification of textual conversation using machine learning approach. Presented at: 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT); Jul 10-12, 2018:1-7; Bangalore. [doi: 10.1109/ICCCNT.2018.8494197]
55. Hopkins G, Kalm K. Classifying complex documents: comparing bespoke solutions to large language models. Preprint posted online on Dec 12, 2023. URL: https://arxiv.org/abs/2312.07182 [Accessed 2025-01-07]
56. Edwards A, Camacho-Collados J. Language models for text classification: is in-context learning enough. Mar 26, 2024.URL: https://aclanthology.org/2024.lrec-main.879/ [Accessed 2025-03-04]
57. Guthrie D, Guthrie L, Wilks Y. An unsupervised probabilistic approach for the detection of outliers in corpora. Presented at: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08); Mar 28-30, 2008; Marrakech, Morocco. URL: https://aclanthology.org/L08-1109/ [Accessed 2025-03-04]
58. Yan D, Li K, Gu S, Yang L. Network-based Bag-of-Words model for text classification. IEEE Access. 2020;8:82641-82652. [doi: 10.1109/ACCESS.2020.2991074]
59. Lubis DH, Sawaluddin S, Candra A. Machine learning model for language classification: Bag-of-Words and multilayer perceptron. JITE. 2023;7(1):356-365. [doi: 10.31289/jite.v7i1.10114]
60. Juluru K, Shih HH, Keshava Murthy KN, Elnajjar P. Bag-of-Words technique in natural language processing: a primer for radiologists. Radiographics. 2021;41(5):1420-1426. [doi: 10.1148/rg.2021210025] [Medline: 34388050]
61. Chelba C, Mikolov T, Schuster M, et al. One billion word benchmark for measuring progress in statistical language modeling. Presented at: Interspeech 2014; Sep 14-18, 2014:2635-2639; Singapore. [doi: 10.21437/Interspeech.2014-564]
62. Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. Exploring the limits of language modeling. Preprint posted online on Feb 7, 2016. URL: https://arxiv.org/pdf/1602.02410 [Accessed 2025-01-12]
63. Miaschi A, Alzetta C, Brunato D, Dell'Orletta F, Venturi G. Is neural language model perplexity related to readability. Presented at: Seventh Italian Conference on Computational Linguistics; Mar 1-3, 2021; Milan, Italy. URL: https://ceur-ws.org/Vol-2769/paper_57.pdf [Accessed 2025-01-12]
64. Guo Y, August T, Leroy G, Cohen T. APPLS: evaluating evaluation metrics for plain language summarization. Preprint posted online on May 23, 2023. URL: https://arxiv.org/abs/2305.14341 [Accessed 2025-01-12]
65. Ma J, Na H, Wang Z, et al. Detecting conversational mental manipulation with intent-aware prompting. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Di EB, Schockaert S, editors. Presented at: Proceedings of the 31st International Conference on Computational Linguistics Association for Computational Linguistics; Jan 19-24, 2025:9176-9183; Abu Dhabi, United Arab Emirates. URL: https://aclanthology.org/2025.coling-main.616/ [Accessed 2025-3-5]
66. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. In: Bouamor H, Pino J, Bali K, editors. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6-10, 2023:6056-6077; Singapore. [doi: 10.18653/v1/2023.emnlp-main.370]
67. Zhang X, Yu H, Li Y, Wang M, Chen L, Huang F. The imperative of conversation analysis in the era of llms: a survey of tasks, techniques, and trends. Preprint posted online on Sep 21, 2024. URL: https://arxiv.org/abs/2409.14195 [Accessed 2025-3-5]
68. Lee Y, Goldwasser D, Reese LS. Towards understanding counseling conversations: domain knowledge and large language models. Preprint posted online on Feb 21, 2024. URL: https://arxiv.org/abs/2402.14200 [Accessed 2025-3-5]
69. Long Y, Luo H, Zhang Y. Evaluating large language models in analysing classroom dialogue. NPJ Sci Learn. Oct 3, 2024;9(1):60. [doi: 10.1038/s41539-024-00273-3] [Medline: 39358390]

70. Si M, Cobas O, Fababeir M. Lexical error guard: leveraging large language models for enhanced ASR error correction. MAKE. 2024;6(4):2435-2446. [doi: 10.3390/make6040120]
71. Cheng L, Zheng S, Zhang Q, et al. Improving speaker diarization using semantic information: joint pairwise constraints propagation. Preprint posted online on Sep 19, 2023. URL: https://arxiv.org/abs/2309.10456 [Accessed 2025-1-7]

Abbreviations

AI: artificial intelligence
ASR: automatic speech recognition
LLM: large language model
MS: minimal number of segments
MWPS: minimum words per segment
ROC: receiver operating characteristic
WPS: words per second

Edited by Amaryllis Mavragani; peer-reviewed by Adewumi Adepoju, Mahmoud Elbattah, Sandipan Biswas; submitted 27.May.2025; final revised version received 03.Sep.2025; accepted 09.Sep.2025; published 24.Oct.2025

Please cite as:
Naim PM, Sadeh-Sharvit S, Jefroykin S, Silber E, Morrison DP, Goldstein A
Preprocessing Large-Scale Conversational Datasets: A Framework and Its Application to Behavioral Health Transcripts
JMIR Form Res2025;9:e78082
URL: https://formative.jmir.org/2025/1/e78082
doi: 10.2196/78082