Measuring the quality of therapy sessions against assessment scales using augmented semantic-similarity approaches

Kejian Cui and Simon D'Alfonso and Mike Conway
School of Computing and Information Systems, The University of Melbourne
kejianc@student.unimelb.edu.au
{dalfonso, mike.conway}@unimelb.edu.au

Abstract

Therapist fidelity and competence rating scales provide a way to measure quality assurance and therapist training outcomes. Scores on these scales reflect the extent to which a therapist adheres to specific therapeutic principles during a psychotherapy session. Existing research has employed natural language processing (NLP) techniques to automatically predict scale ratings. However, existing approaches require a model trained on a dataset of therapy sessions annotated with the target rating scale. Recent work has explored directly inferring therapeutic alliance by computing semantic similarity between therapy transcripts and the Working Alliance Inventory, via cosine similarity between sentence embeddings. In this paper, we extend this line of work by computing semantic similarity between therapist talk turns and therapist fidelity scale items to directly infer fidelity to specific therapeutic modalities. We further enhance this method by augmentation with LLM-generated example therapist utterances that instantiate target behaviours (as expressed by scale items) across varied therapeutic contexts. In evaluations on two independent datasets, our example-augmented semantic similarity approach consistently shows effectiveness in discriminating therapeutic modalities and levels of therapist fidelity.

1 Introduction

Psychotherapists use fidelity and competence rating scales for psychotherapy quality assurance and therapist training. Scales such as the Cognitive Therapy Rating Scale (CTRS) (Goldberg et al., 2020) for cognitive behavioural therapy practice and the Motivational Interviewing Skill Code (MISC 2.5) (Houck et al., 2010) for motivational interviewing define good therapy as a set of observable therapist behaviours (e.g., agenda setting, guided discovery, homework). However, applying these scales in practice is expensive and slow: trained raters must review entire therapy sessions and determining scores for each scale item requires focus and effort. This, in turn, creates a bottleneck for research that requires large volumes of fidelity-labelled session transcripts.

A growing body of NLP research aims to automatically predict scale ratings by training supervised machine-learning models on psychotherapy transcripts. For example, Zech et al. (2022) uses a support vector regressor to predict the Facilitative Interpersonal Skills Task for Text (FIS-T) (Zech et al., 2023), and Flemotomos et al. (2021) uses BERT to predict CTRS. Such approaches depend on a sufficiently large corpus annotated with the target metrics. In practice, however, psychotherapy transcripts are often difficult to access due to privacy and governance constraints, and obtaining reliable scale annotations is costly and time-intensive. Moreover, supervised models trained in one setting, or using a specific therapeutic modality do not generally transfer to others, limiting their generalisability.

Recent work has explored using semantic similarity as an alternative approach (Lin et al., 2025): representing therapy utterances and rating scale items in the same embedding space and computing cosine similarity as a proxy for the extent to which therapists' words in a session "match" the behaviours described by the scale. However, many scale items are written as behavioural anchors rather than what therapists actually say in sessions. Moreover, the item descriptions are generally very concise, highly abstract and under-specify the diverse linguistic realisations of therapist behaviour. For instance, the Working Alliance Inventory - Observer Form (WAI-O) (Darchuk et al., 2000), consisting of 36 items, is a standardised assessment tool used by third-party raters to evaluate the quality of the therapeutic relationship between the client and the therapist. One item description of WAI-O is: "There is a mutual liking between

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 271–281

the client and therapist". In a real session, however, such relational qualities are expressed through a wide range of indirect and context-dependent conversational cues rather than explicit statements. Simply embedding these item descriptions may not be enough to cover a wide range of talk turns that match the item.

In this paper, we employ the idea of semantic similarity between therapist utterances and assessment scale items to directly infer therapists' adherence to specific therapeutic techniques and propose an example-augmented semantic similarity framework that reduces the mismatch by enriching scale item representations with LLM-generated exemplar therapist utterances. For each scale item, we extract the manual's description of ideal therapist behaviour and generate LLM-based exemplars that instantiate the behaviour across diverse clinical contexts. They are then merged into an enriched textual representation. We compute turn-level similarity between therapist utterances and augmented items and aggregate these scores into session-level fidelity inference. We evaluate the framework in two experiment settings designed to test both between-modality and within-modality discrimination. Our findings show that the proposed framework consistently discriminates between sessions with different therapeutic modalities or sessions with different levels of adherence to therapeutic techniques. Ablations indicate that exemplar augmentation substantially improves between-modality discrimination, but does not show gains for within-modality quality discrimination.

We thus present the following contributions:

- We introduce a training-free framework that automatically infers therapists' adherence to therapeutic techniques as required by standardised scales.
- We propose two different discrimination tasks for the evaluation of semantic similarity approaches.
- We perform ablation studies to evaluate the impact of important segments within the framework.

2 Related Work

2.1 Psychotherapy Rating Scales

Various scales exist to evaluate therapist performance or the quality of the therapeutic relationship during a session. Some scales are designed for specific therapeutic approaches and assess therapists' adherence to the techniques required by a given modality. For example, the Cognitive Behaviour Therapy Rating Scale (CTRS) (Goldberg et al., 2020) is a commonly used scale to assess adherence to Cognitive Behaviour Therapy (CBT) principles. Similarly, the Motivational Interviewing Skill Code (MISC 2.5) (Houck et al., 2010) evaluates the quality of motivational interviewing interactions. While these instruments focus on therapists' technical competence, other scales concern the common factors that cut across therapeutic approaches. For example, the Working Alliance Inventory (WAI) (Horvath and Greenberg, 1989) is a widely used scale to measure the therapeutic alliance between the therapist and the client, with versions from three perspectives: a version to be filled out by clients (WAI-C), a version to be filled out by therapists (WAI-T), and an observer-rated version (WAI-O).

A typical scale contains a set of items, each assessing one aspect of the therapist's behaviour (e.g., empathy, feedback). For Likert-type scales, each item includes a clearly defined set of criteria for every point on the rating scale, specifying what therapist behaviours correspond to lower versus higher scores, which quantifies qualitative judgments into numerical ratings to capture the degree to which a behaviour or skill is demonstrated. In practice, trained raters watch or read session recordings and assign item-level scores based on the scale manual. The item scores are then aggregated (usually by sum) to generate an overall measure. Higher scores generally indicate better performance.

2.2 Supervised NLP Models for Predicting Psychotherapy Ratings

Prior research has widely adopted supervised NLP methods for automated psychotherapy rating. These methods typically require building a transcript dataset manually annotated with scale ratings. Then, statistical and linguistic features are extracted from transcripts and used to train machine learning models to predict session or therapist ratings. Commonly used models include logistic regression and support vector machine/regressor (Flemotomos et al., 2018; Imel et al., 2019).

A recurring pattern in this literature is the conversion of numerical rating scale scores into binary labels, for example, predicting good vs. bad therapist performance (Flemotomos et al., 2021;

Pérez-Rosas et al., 2019), high vs. low engagement quality (Rueda et al., 2025), high vs. low empathy level (Tao et al., 2022; Gibson et al., 2015; Xiao et al., 2016). Fewer studies attempt direct regression on scores (Li et al., 2024; Flemotomos et al., 2021).

2.3 Semantic Similarity Approaches in Psychotherapy Analysis

Psychotherapy transcripts are naturally difficult and time-consuming to annotate according to standardised scales, given that a therapy session often lasts 30-90 minutes. Recently, researchers have used the semantic similarity between the words in transcripts and the descriptions of WAI scale items to directly infer WAI levels. Such an approach starts with embedding talk turns in a transcript and the description of WAI scale items to the same vector space using tools like SentenceBERT (Reimers and Gurevych, 2019), and Doc2Vec (Le and Mikolov, 2014), then computing the cosine similarity between embeddings to obtain a semantic similarity score for each turn, which can then be used for downstream tasks. Conceptually, this method transfers the gap between therapists' behaviour in real sessions and the ideal therapists' behaviour required by the scale into the distance between their contextual embeddings. Lin et al. (2025, 2024) concatenate the similarity scores between talk turns and WAI items with the sentence embeddings of this turn, together serving as the input to a sequence classifier (a Transformer) to classify clinical conditions at the session level. Their results show that adding WAI inference generally improves the classification accuracy, which demonstrates the informativeness of similarity scores. Semantic similarity scores have various usages for different downstream tasks, for example, treating inferred alliance scores as reward signals for reinforcement learning (Lin et al., 2022), evaluating model-generated responses by comparing their similarity to human-written responses (Sedoc et al., 2019; Das et al., 2022).

Thus, instead of manually annotating transcripts with exact scale scores, calculating semantic similarity scores can provide an efficient and alternative way to represent the therapeutic relationship numerically for downstream tasks. In this paper, we take this idea of semantic similarity for scale assessment and construct a broader framework that can be applied to any therapy quality/modality scale.

3 Method

Our proposed method consists of three stages: (1) extract fine-grained natural language descriptions of ideal therapist behaviour for each scale item, (2) enrich these representations using LLM–generated exemple therapist utterances, and (3) compute semantic similarity between real therapist talk turns and these exemplars and summarise to derive session-level scores.

3.1 Extraction of fine-grained behavioural descriptions

In the first step, we extract the most detailed natural language description of ideal therapist behaviour for each scale item. For Likert-type psychotherapist rating scales, the scale manual typically provides not only item labels but also detailed behavioural anchors corresponding to different points on the Likert scale. These behaviour definitions offer richer and more concrete descriptions of therapist behaviour, compared to highly concise and abstract item labels. For example, CTRS contains 11 items scored on a 0-6 Likert scale (item titles are shown in Appendix A). One CTRS item is "Pacing and efficient use of time" (Darchuk et al., 2000). The CTRS manual provides definitions for each even point (0, 2, 4 and 6) for each item. The definition corresponding to the highest score (6) of this item is as follows:

> The therapist used time efficiently by tactfully limiting peripheral and unproductive discussion and by pacing the session as rapidly as was appropriate for the patient. (Darchuk et al., 2000)

In our approach, we adopt the behavioural definition corresponding to the highest Likert score for each item as the canonical description of optimal therapist performance.

3.2 Enhance with LLM-generated exemplars

To capture the diversity of linguistic realisations through which optimal therapist behaviour may be expressed, in the second stage, for each scale item, we prompt an LLM to generate example therapist talk turns that most exemplify that item. The LLM is given the highest-score behavioural definition of a scale item and asked to generate a set of therapist talk turns that could occur in real psychotherapy scenarios, and that strongly reflect the target behaviour. Crucially, no therapy transcript data,

or participant information is sent to a commercial LLM API at any stage. The system prompt used to generate exemplars is included in Appendix B.

The LLM is explicitly required to generate examples that are (i) linguistically distinct from one another, (ii) illustrative of different aspects of the target behaviour definition, and (iii) cover varied therapeutic contexts. The resulting examples are saved specifically for each item. In this step, we have built a corpus of high-fidelity example therapist utterances to compare with real therapist talk turns.

3.3 Semantic similarity computation and aggregation

Given a psychotherapy transcript to be evaluated, we extract all therapist talk turns. Each talk turn and each generated example is embedded independently using the same SentenceBERT (Reimers and Gurevych, 2019) model, each resulting in 384-dimensional vector representations.

The algorithm in Appendix C describes the computation of semantic similarity in our method. To obtain session-level semantic similarity scores based on a scale of N items, we first operate at the talk turn level. For each therapist talk turn in the transcript, calculate the cosine similarity of its embedding with the embedding of each exemplar associated with a given scale item. These similarity scores are averaged across all examples of that item to get an item-specific semantic similarity score for the talk turn. As a result, each therapist's talk turn has N scores, where N is the number of items in the scale. Finally, for each scale item, we aggregate similarity scores across all therapist talk turns within the session by averaging, producing N session-level semantic similarity scores. In the experimental evaluation, we conduct ablation tests to compare alternative calculation and summarisation methods and different numbers of examples used per item.

4 Dataset

Two psychotherapy transcript datasets are used independently in this study. Use of these datasets for secondary data analysis was approved by the Office of Research Ethics and Integrity, University of Melbourne, Australia (Reference Number: 2025-32787-74041-3).

4.1 STEP dataset

The STEP dataset was collected as part of the Staged Treatment in Early Psychosis (STEP) program at Orygen in Melbourne, Australia (Nelson et al., 2018). It comprises 182 English psychotherapy session transcripts involving young individuals identified as being at ultra-high risk (UHR) of psychosis. The sessions include two intervention types: Support and Problem Solving (SPS) and Cognitive Behavioural Case Management (CBCM). CBCM is a specialised manualized psychosocial intervention that delivers CBT within a case management framework, adapted for a young UHR population. Audio recordings of therapy sessions were transcribed using a HIPAA-compliant transcription service, ensuring the security of Protected Health Information (PHI). Speaker roles were annotated, and transcripts are formatted as alternating talk turns between therapist and client. The median number of talk turns per session is 315. Each transcript has metadata denoting the allocated therapeutic modality and the actually delivered type (SPS or CBCM), as annotated by an observer. In total, 134 sessions are labelled as delivered in SPS and 48 sessions as CBCM.

4.2 AnnoMI dataset

The AnnoMI dataset (Wu et al., 2023) is a publicly available transcript dataset that contains 133 transcribed English motivational interviewing (MI) demonstration sessions captured from public video-sharing platforms (YouTube and Vimeo). These sessions are labelled as high-quality or low-quality based on video titles, descriptions, and narrator comments by professional MI practitioners and healthcare organisations. The AnnoMI authors consider this labelling approach sufficiently reliable given uploaders' domain expertise and the well-defined nature of MI quality in the literature. However, these are video-level characterisations rather than independent session-level clinical ratings, and should be interpreted accordingly. This corpus contains 110 high-quality sessions and 23 low-quality sessions. The dataset creator surveyed 6 experienced MI therapists on the quality of the dataset, and 83% reported "agree" or "somewhat agree" that the transcripts overall reflect real-world MI practice. Therapist and client talk turns are explicitly identified in the transcripts, enabling turn-level linguistic analysis.

5 Experiments

We evaluate our proposed method through two different kinds of experiments to check if the proposed method can (i) appropriately distinguish different therapeutic modalities and (ii) distinguish high and low therapist performance. All experiments mentioned below use SentenceBERT as the embedding tool and use the Gemini 3.0 API to generate exemplars. To ensure reproducibility, exemplars were generated once and held fixed throughout all experiments. The same set of exemplars was used across all experimental conditions.

5.1 Discriminating Therapeutic Modalities

In the first experiment, we evaluate whether the proposed method can effectively discriminate between more structured, technique-driven interventions (e.g., CBT, MI) and more general, low-technique approaches such as SPS. In the first part of this experiment, we compute semantic similarity scores between therapy sessions from the STEP dataset and items of the Cognitive Therapy Rating Scale (CTRS). As CTRS specifies CBT-specific behaviour requirements, sessions delivered using CBCM are expected to generally yield higher CTRS-based similarity scores than SPS sessions.

After obtaining session-level CTRS scores using Algorithm 1, we compute a total CTRS similarity score per session by adding the semantic similarity scores across all scale items. We then conduct t-tests between the CTRS results for 48 CBT sessions and the CTRS results for 134 SPS sessions, and compute effect sizes (Cohen's d (Cohen, 1977)) on these total CTRS similarity scores to assess the separation between the two therapeutic modalities.

Similarly, in the second part of this experiment, we discriminate SPS from another technique-driven modality, Motivational Interviewing (MI). Motivational Interviewing Skill Code (MISC) 2.5, a widely used scale to evaluate therapists' adherence to MI techniques, is used. MISC contains two distinct parts, global ratings and utterance-level parsing. Here, we operate on six global ratings and sum the similarity scores as the final score. 133 MI transcripts from the AnnoMI dataset and 134 SPS transcripts from the STEP dataset are compared. As MISC specifies MI behaviour requirements, sessions delivered in MI are expected to generally have higher MISC-based similarity scores than SPS sessions.

To further evaluate the effectiveness of the proposed example-augmentation approach, we compare the resulting Cohen's d values against those obtained using a baseline semantic similarity method, in which only the original scale item descriptions are used. A higher Cohen's d (effect size) indicates a stronger ability to discriminate therapeutic modalities. As the details are relevant to some of our subsequent analysis, we here provide the definition of Cohen's d (Cohen, 1977):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p} \tag{1}$$

Where the pooled standard deviation s_p is calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2}$$

5.2 Ablation Study

As part of the first experiment, we conduct ablation studies to evaluate the impact of different summarisation methods and the number of LLM-generated exemplar therapist talk turns.

Specifically, we vary the number of exemplars used per scale item from 0 to 50, where 0 corresponds to the baseline method that uses only the original CTRS behaviour descriptions. This study is conducted for each configuration described below.

We evaluate multiple strategies for computing similarity scores at the talk-turn level. We first compute similarity scores between each exemplar and the therapist talk turn individually, then aggregate these values using different summarisation methods: the mean similarity across all exemplars, the maximum similarity, and the mean of the top-n similarity scores, with three different n values evaluated: 5, 10, 20.

Similarity scores are then summarised at the session level. Multiple summarisation methods are evaluated, including the mean similarity across all talk turns, and the mean of the top-n similarity scores, with four n values evaluated: 5, 10, 20, 40.

For each setting above, t-tests and effect sizes (Cohen's d) are computed on session-level total CTRS similarity scores to quantify the separability between SPS and CBCM sessions. In addition, we conduct a study that evaluates the effectiveness of each CTRS item in distinguishing therapeutic modalities.

5.3 Discriminating Session Quality

In the second experiment, we apply the same framework to the AnnoMI dataset using the Motivational Interviewing Skill Code (MISC). We compute semantic similarity scores between therapy sessions and MISC items to assess therapist fidelity to Motivational Interviewing (MI). Because the MISC specifies behaviour requirements that characterise high-quality MI practice, sessions annotated as high quality are expected to exhibit higher MISC-based similarity scores than those annotated as low quality.

Following the same procedure as in the first experiment, we compute a total MISC similarity score for each session. We then conduct t-tests and compute Cohen's d on the total similarity scores to evaluate the ability of the proposed method to distinguish between high- and low-quality MI sessions, compared with the baseline method.

6 Results and Discussion

6.1 Effectiveness in Discriminating Therapeutic Modalities

Figures 1 and 2 show the Cohen's d value, indicating how effectively the total CTRS similarity score can separate therapeutic modalities across experimental settings. All Cohen's d values are positive, indicating that CBT sessions consistently obtain higher CTRS similarity scores than SPS sessions, as expected given that CTRS measures CBT-specific therapist behaviours. In both figures, the highest convergent value of Cohen's d reaches around 0.85, with LLM-generated examples. Looking at the starting point of each curve (zero examples), the highest Cohen's d value lies around 0.5. Following Cohen (Cohen, 1977) and Lakens (Lakens, 2013), this represents an increase of 0.35 (70%) standard deviation units in group separation, corresponding to a substantially stronger differentiation between the two therapeutic modalities.

Significant improvement is also seen for the MISC-based experiment. As shown in Figure 4, the Cohen's d value significantly improves from zero to one/two examples, then gradually converges to around 0.7. This demonstrates an increase of around 0.55 standard deviation units in the separation between MI and SPS.

We interpret the differences in effect size as evidence that the proposed method captures sufficient fidelity-specific signal (e.g., CBT techniques required by CTRS and MI techniques required by MISC 2.5) to discriminate therapeutic modalities, and the performance improves with LLM-generated exemplars added.

6.2 Ablations

Our ablation studies give insights into different characteristics of our method. Figure 1 and 2 clearly show that the effectiveness of employing LLM-generated exemplars increases with the number of exemplars (N) used to compute semantic similarity. A clear turning point is at N=5, after which the gains gradually plateau. This indicates that only five examples could capture sufficient semantic information about different aspects of scale items. This finding gives us a "sweet spot" for utilising this technique.

Figure 1 compares different calculation methods of talk-turn level semantic similarity. Semantic similarity is calculated individually with each example, and then the top n similarity scores are averaged. A clear pattern from the figure is that the performance increases with n, and the best performance is achieved by averaging all scores. However, variation in performance is limited, and all lie around 0.8 after convergence.

Figure 2 compares different summarisation methods. Instead of averaging the top n highest talk-turn level semantic similarity scores, directly averaging the scores of all talk turns achieves the best performance. The intuition of designing this ablation study is to check if the top n similar talk turns are adequate to determine therapists' adherence to CBT techniques, excluding the noise caused by general speeches like greetings. However, the pattern shows that the more talk turns involved in calculating the session-level mean, the better the performance of separation is. This may be due to the fact that the total number of talk turns varies considerably; keeping the same number of talk turns for each session makes longer sessions lose information. On the other hand, averaging all talk turns keeps all information for each session and also normalises by the number of talk turns, resulting in a more representative score on the session level.

Figure 3 shows how the semantic similarity with each CTRS item contributes to the total Cohen's d value. The convergent values of Cohen's d for most items lie between 0.6 and 1.0 except "Agenda", demonstrating medium to high effect size on separating therapeutic modalities according to Cohen (Cohen, 1977). The most effective CTRS item is the 7th: guided discovery, achieving around 0.95

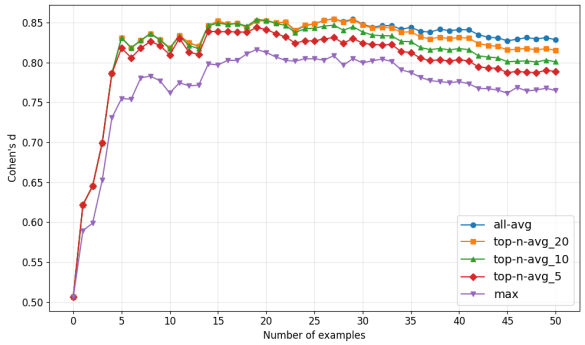

Figure 1: Cohen's d versus exemplar count for CTRS-based modality discrimination (CBCM vs SPS) under different exemplar-aggregation strategies. (Session-level aggregation strategy: average all)

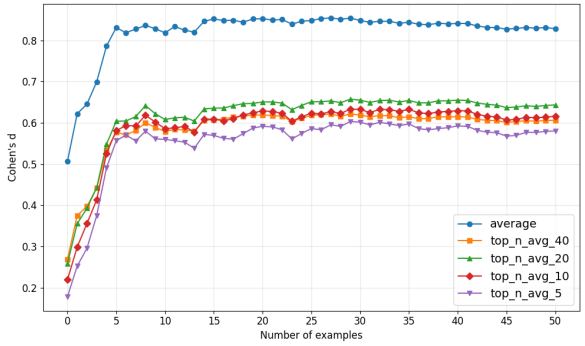

Figure 2: Cohen's d versus exemplar count for CTRS-based modality discrimination (CBCM vs SPS) under different session-level similarity aggregation methods. (Exemplar aggregation strategy: average all)

Cohen's d value. Guided discovery is one of the core techniques of CBT, in which the therapist questions and guides the client to identify automatic thoughts. The use of this technique separates CBT from other modalities like SPS, where guided discovery is not necessarily used. On the contrary, appropriate agenda setting (a more general therapeutic skill), scores the lowest.

6.3 Effectiveness in Discriminating Session Quality

Figure 5 shows how semantic similarity discriminates high- and low-quality MI sessions with respect to the number of exemplars used. The Cohen's d value converges on 0.75, indicating a medium-to-large effect in discriminating between the two groups. We interpret this performance as evidence that example-enhanced semantic similarity scores can discriminate high and low therapists' adherence to a technique-driven modality scale. The pattern in the figure shows that with examples added, the Cohen's d value immediately shrinks,

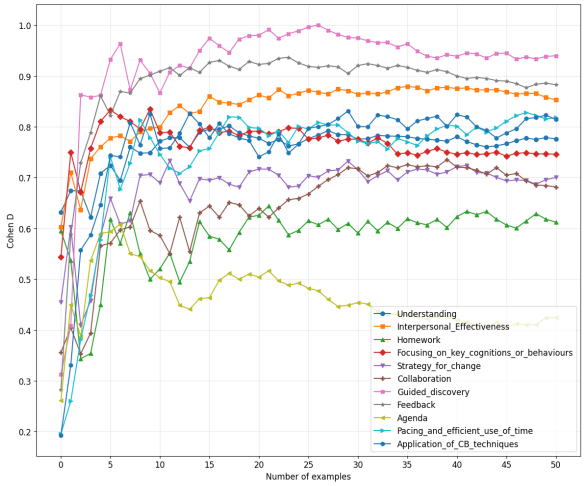

Figure 3: Cohen's d versus exemplar count for individual CTRS items in distinguishing CBCM (CBT) from SPS sessions. (Exemplar aggregation strategy: average all; Session-level aggregation strategy: average all)

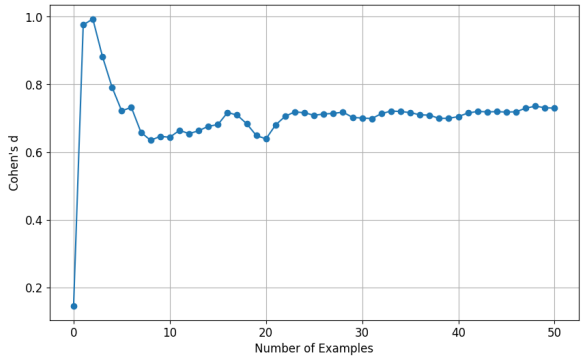

Figure 4: Cohen's d versus exemplar count for MISC-based semantic similarity in distinguishing MI from SPS sessions. (Exemplar aggregation strategy: average all; Session-level aggregation strategy: average all)

then gradually rises and stabilises at a higher value. This suggests that averaging more exemplars stabilises the representation (variance reduces).

However, the high value of Cohen's d at the zero-example point indicates the superiority of directly calculating semantic similarity with scale item descriptions. To dive deeper into why this happens, we look at how statistical values related to Cohen's d (numerator: difference between mean values of two groups; denominator: pooled standard deviation) change with respect to the number of examples used. According to Figure 6, the denominator, the pooled standard deviation, merely changes with n, indicating that adding examples does not add noise to the group deviation. On the other hand, the mean value for both groups shows a notable increase. This suggests that adding examplars in-

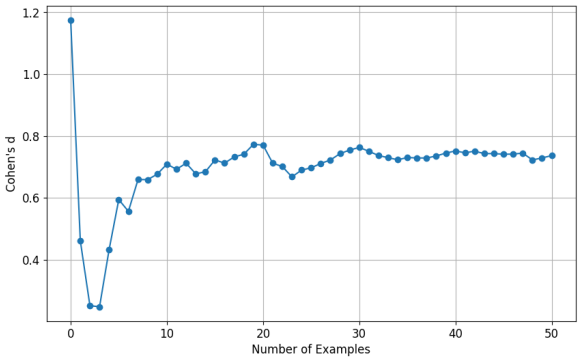

Figure 5: Cohen's d versus exemplar count for MISC-based semantic similarity in distinguishing high- versus low-quality MI sessions.

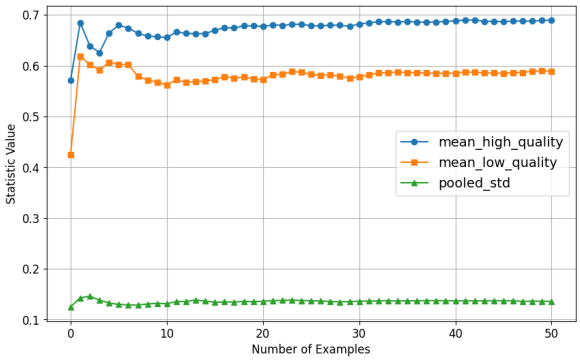

Figure 6: Components of Cohen's d (group means and pooled SD) across exemplar counts for MISC-based MI quality discrimination.

crease the semantic similarity with both high and low quality MI sessions. However, this increment is greater for low-quality sessions, shrinking the high–low mean gap, thus making the Cohen's d value drop. In other words, exemplar augmentation, although it improves the absolute value of semantic similarity, is making high- and low-quality MI sessions look more similar in embedding space.

One reason for the superiority of item-only semantic similarity is that this zero-example baseline actually contains a significant amount of example-related information: MISC items are already exemplar-rich, containing bullet-pointed behavioural indicators. Embedding that text yields a dense representation of the construct, including multiple facets of competence and sometimes a polarity (See Appendix D for an example). That is, the item itself already functions like a "prototype set" covering (i) what good MI looks like for that item, and (ii) what poor/problematic performance looks like, producing a strong high–low separation. On the other hand, LLM-generated exemplars, although closer to the way therapists talk in real psychotherapy sessions, all demonstrate positive therapist behaviours. Adding these exemplars can reduce sensitivity to low-quality sessions in the embedding space.

7 Future Work and Conclusion

In this paper, we introduce an LLM-powered example-augmented semantic similarity framework for directly inferring therapist fidelity signals from transcripts by aligning therapist talk with fidelity/skill scale items in an embedding space. Across two datasets, the approach consistently improves the discrimination of technique-driven therapeutic modality and more general approaches (e.g., MI vs SPS; CBT/CBCM vs SPS). This method offers methodological insights for inferring numerical representations of relative therapists' adherence to specific therapeutic techniques in an annotation-free and training-free approach. These fidelity signals can then be used for downstream computational linguistics tasks.

We also found that augmentation is scale-dependent: for within-modality quality on AnnoMI, the item-only baseline is strong because MISC definitions are already highly detailed and exemplar-like, making additional LLM utterance exemplars partly redundant and sometimes compressing high–low differences.

Future work directions include:

- Validate generalizability across more scales, modalities and clinical settings. In general, for scales that are specific to a particular modality (e.g. CTRS to measure CBT quality as opposed to WAI, which is modality-agnostic), it would be expected that a session employing that modality would score higher for a scale measuring that modality versus a session that is not employing that modality.
- Further refine the method with contrastive augmentation (high-quality vs low-quality exemplars).
- Enhance the aggregation method to emphasise discriminative turns rather than averaging all talk turns.
- Validate the framework against human-rated fidelity scores. For example, correlating framework outputs with human CTRS ratings for CBT sessions.

- Evaluate open-source LLMs as a replacement for commercial APIs in the exemplar generation step. This would improve reproducibility, since model weights are fixed and publicly available.

8 Limitations

While promising, our method and evaluation have some limitations. First is the imbalance and size of the evaluation datasets. Whilst not insignificant datasets in terms of therapy transcripts, both STEP and AnnoMI are relatively small in size (182 and 133). The AnnoMI dataset is highly imbalanced, with only 23 low-quality sessions. The STEP dataset is limited to a certain clinical context (focused on adolescents at ultra-high risk for psychosis in Melbourne, Australia). These factors might affect the generalisability of observed patterns. Similar evaluations should be run on larger and more balanced datasets for more robust results.

A further limitation concerns the cross-dataset confound in the between-modality experiment comparing MI and SPS sessions. MI sessions are drawn from the AnnoMI dataset, which consists of demonstration sessions sourced from public video-sharing platforms, while SPS sessions are drawn from the STEP dataset, which comprises real clinical sessions. These two groups differ not only in therapeutic modality but also in dataset characteristics. Consequently, observed differences in MISC-based similarity scores between MI and SPS sessions may partly reflect these characteristics rather than modality alone. Future work should seek to compare therapeutic modalities using sessions drawn from the same dataset and clinical context in order to isolate the effect of modality.

A critical limitation is the absence of validation against human-rated fidelity scores. The experiments demonstrate that similarity scores can statistically distinguish between therapeutic modalities and between session quality labels, but they do not establish whether these scores correlate with ratings that trained clinicians would assign using the same scales. Establishing this correspondence is a necessary step before clinical applications can be considered.

Another limitation is that the semantic similarity is not a full proxy for competence. It captures conceptual alignment, but many competence signals are interactional (e.g., timing, responsiveness to client change talk) or structural (e.g., reflection question balance, open question rate), which are not well represented by talk-turn level similarity scores. Moreover, aggregating turn-level scores into a single session value can hide short but important events like ruptures and confrontations. A session can score "high" while containing clinically significant low-competence segments.

9 Ethical Considerations

Semantic similarity scores derived from language models have limited interpretability and should not be treated as direct proxies for clinically validated rating scale scores. The mapping between numerical similarity values and clinically meaningful scores is unclear and requires systematic validation. Instead, these scores are more appropriately understood as relative indicators that capture comparative patterns across sessions. Their primary utility lies in supporting computational linguistic analyses or serving as auxiliary features in downstream models.

Semantic similarity scores could inherit bias and fairness issues from the language models they rely on. LLM-generated exemplars reflect LLM's training materials that could inherently contain cultural and linguistic biases. These exemplars may privilege particular styles of therapist talk (e.g. particular clinician norms and dialects), and unfairly lower similarity scores for therapists from underrepresented groups.

References

Jacob Cohen. 1977. The t test for means. In *Statistical Power Analysis for the Behavioral Sciences*, pages 19–74. Elsevier.

Andrew Darchuk, Victor Wang, David Weibel, Jennifer Fende, Timothy Anderson, and Adam Horvath. 2000. Working Alliance Inventory – Observer Form (WAI-O):.

Avisha Das, Salih Selek, Alia R. Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng, and Hua Xu. 2022. Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.

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

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *PLOS ONE*, 16(10):e0258639.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S. Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Interspeech 2015*, pages 1947–1951. ISCA.

Simon B. Goldberg, Scott A. Baldwin, Kritzia Merced, Derek D. Caperton, Zac E. Imel, David C. Atkins, and Torrey Creed. 2020. The structure of competence: Evaluating the factor structure of the cognitive therapy rating scale. *Behavior Therapy*, 51(1):113–122. Publisher: Elsevier BV.

A. O. Horvath and L. S. Greenberg. 1989. Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, 36:223–233.

Jon M Houck, Theresa B Moyers, William R Miller, Lisa H Glynn, and Kevin A Hallgren. 2010. Motivational Interviewing Skill Code (MISC) 2.5.

Zac E. Imel, Brian T. Pace, Christina S. Soma, Michael Tanana, Tad Hirsch, James Gibson, Panayiotis Georgiou, Shrikanth Narayanan, and David C. Atkins. 2019. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy*, 56(2):318–328.

Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint*. ArXiv:1405.4053 [cs].

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Understanding the therapeutic relationship between counselors and clients in online text-based counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.

Baihan Lin, Djallel Bouneffouf, Yulia Landa, Rachel Jespersen, Cheryl Corcoran, and Guillermo Cecchi. 2025. COMPASS: Computational mapping of patient-therapist alliance strategies with language modeling. *Translational Psychiatry*, 15(1):166.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2022. SupervisorBot: NLP-annotated real-time recommendations of psychotherapy treatment strategies with deep reinforcement learning. *arXiv preprint*. ArXiv:2208.13077 [cs].

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2024. Working alliance transformer for psychotherapy dialogue classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 64–69, Mexico City, Mexico. Association for Computational Linguistics.

Barnaby Nelson, G. Paul Amminger, Hok Pan Yuen, Nicky Wallis, Melissa J. Kerr, Lisa Dixon, Cameron Carter, Rachel Loewy, Tara A. Niendam, Martha Shumway, Sarah Morris, Julie Blasioli, and Patrick D. McGorry. 2018. Staged treatment in early psychosis: A sequential multiple assignment randomised trial of interventions for ultra high risk of psychosis patients. *Early Intervention in Psychiatry*, 12(3):292–306.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? Learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. *arXiv preprint*. ArXiv:1908.10084 [cs].

Alice Rueda, Argyrios Perivolaris, Niloy Roy, Dylan Weston, Sarmed Shaya, Zachary Cote, Martin Ivanov, Bazen G. Teferra, Yuqi Wu, Sirisha Rambhatla, Divya Sharma, Andrew Greenshaw, Rakesh Jetly, Yanbo Zhang, Bo Cao, Reza Samavi, Sridhar Krishnan, and Venkat Bhat. 2025. Estimating quality in therapeutic conversations: A multi-dimensional natural language processing framework. *arXiv preprint*. ArXiv:2505.06151 [cs].

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Dehua Tao, Tan Lee, Harold Chui, and Sarah Luk. 2022. Hierarchical attention network for evaluating therapist empathy in counseling session. In *Interspeech 2022*, pages 2008–2012. ISCA.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of AnnoMI, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3):110.

Bo Xiao, Chewei Huang, Zac E. Imel, David C. Atkins, Panayiotis Georgiou, and Shrikanth S. Narayanan. 2016. A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ Computer Science*, 2:e59.

James Zech, Victoria Kaitlin Foley, Thomas D. Hull, and Timothy Anderson. 2023. Assessing the quality of digital patient-therapist communication: the development and validation of a text-based facilitative interpersonal skills task. *Psychotherapy Research*, 33(6):743–756.

James M. Zech, Robert Steele, Victoria K. Foley, and Thomas D. Hull. 2022. Automatic rating of therapist facilitative interpersonal skills in text: A natural language processing application. *Frontiers in Digital Health*, 4:917918.

A Cognitive Therapy Rating Scale (CTRS) item titles

Cognitive Therapy Rating Scale (CTRS) item titles (Goldberg et al., 2020)

1. Agenda
2. Feedback
3. Understanding
4. Interpersonal Effectiveness
5. Collaboration
6. Pacing and Efficient Use of Time
7. Guided Discovery
8. Focusing on Key Cognitions or Behaviors
9. Strategy for Change
10. Application of Cognitive-behavioral Techniques
11. Homework

B System prompt for LLM to generate exemplars (CTRS as an example)

"You will be given the description of a Cognitive Therapy Rating Scale (CTRS) item, describing the desired therapist behaviour. Based on this description, generate 10 unique therapist talk turns that could realistically occur in psychotherapy and reflect the specified behaviour. The examples should differ from one another, illustrate different aspects of the item, and span a range of therapy scenarios. Return a valid JSON object with a single key 'examples' whose value is an array of 10 strings. Do not include Markdown or any extra text."

C Algorithm for semantic similarity score computation

Algorithm 1 Computation of Session-Level Semantic Similarity Scores

Require: Session transcript with therapist talk turns $\mathcal{T} = \{t_1, \ldots, t_M\}$
Scale items $\mathcal{I} = \{i_1, \ldots, i_N\}$
Exemplars for each item $\mathcal{E}_n = \{e_{n1}, \ldots, e_{nK_n}\}$
Embedding function $f(\cdot)$
Ensure: Session-level similarity scores $\mathbf{S} = \{S_1, \ldots, S_N\}$
Compute embeddings for all therapist turns: $v_m \leftarrow f(t_m)$ for $m = 1, \ldots, M$
Compute embeddings for all exemplars: $u_{nk} \leftarrow f(e_{nk})$ for all n, k
for $m = 1$ to M **do**
    for $n = 1$ to N **do**
        $s_{mn} \leftarrow \frac{1}{K_n} \sum_{k=1}^{K_n} \cos(v_m, u_{nk})$ ▷ Item-specific similarity for turn t_m
    end for
end for
for $n = 1$ to N **do**
    $S_n \leftarrow \frac{1}{M} \sum_{m=1}^{M} s_{mn}$ ▷ Aggregate across all therapist turns
end for
return S

D MISC 2.5 item definition for the highest point for item "Direction"

Clinician exerts influence on the session and generally does not miss opportunities to direct client toward the target behavior or referral question.

Examples: • Agenda-setting mentions the target behavior • Clinician is transparent in concern about the target behavior • Clinician manages time well and transitions between therapeutic tasks smoothly • Clinician consistently and smoothly directs the client's discourse toward change of a target behavior • Balance of time in the session is spent discussing possible change, rather than the history of the problem Clinician dominates session and does not allow client to wander from target behavior. (Houck et al., 2010)