From Responses to Trajectories: Modeling the Development of Reflective Listening Skills

Dhruvil Thummar **Verónica Pérez-Rosas**
Texas State University, San Marcos
{kwt32, vperezr}@txstate.edu

Abstract

Reflective listening is a core counseling skill that supports effective communication in mental and behavioral health. Understanding how this skill changes with practice is important for designing scalable training and feedback systems. In this paper, we examine longitudinal patterns in 6,196 trainee responses collected during a three-week counseling training study. We model responses as trajectories in semantic embedding space and use residual embeddings with similarity-based metrics to quantify week-to-week change. Our analyses reveal systematic shifts over time, including increased semantic alignment and reduced variability, consistent with stabilization in trainees' language. We further show that these trajectory patterns are accompanied by modest linguistic shifts relevant to reflective counseling practice. Together, these findings provide a trajectory-based view of how reflective listening practice changes over repeated assignment.

1 Introduction

Reflective listening is a core counseling skill in motivational interviewing (MI) and other patient-centered approaches. It involves generating responses that paraphrase and infer a speaker's underlying meaning or emotional state without introducing advice or judgment (Miller and Rollnick, 2012). For example, given the client statement, "I know I should quit smoking, but I'm overwhelmed right now," a reflective response might be: "You feel torn: you recognize quitting is important, but everything happening in your life makes it feel impossible at the moment." Producing such responses requires nuanced semantic alignment with the speaker's intent and typically develops through repeated practice and feedback.

Its clinical importance is well established, with prior work linking reflective listening to positive therapeutic outcomes (Moyers et al., 2009). However, it remains difficult to teach and assess at scale. Human supervision is time-intensive, qualified supervisors are limited, and direct observation of trainee-client interactions is uncommon in many training programs. As a result, trainees may receive sparse or inconsistent feedback, making it difficult to track how their reflective listening skills develop over time.

Because reflective listening is expressed through language, natural language processing (NLP) offers a promising way to support scalable assessment and feedback. Prior NLP work on reflective listening training has focused on generating alternative phrasings, rewriting trainee responses, and evaluating response quality, enabling scalable feedback on trainee performance (Shen et al., 2022; Min et al., 2022, 2023). Complementing these exchange-level approaches, a longitudinal perspective can help characterize how reflective listening develops across repeated practice.

We present an initial study in this direction by addressing two research questions: (1) Do trainee responses change systematically between practice sessions? (2) Do these changes align with linguistic patterns previously associated with high-quality counseling? We answer these questions by examining longitudinal embedding trajectories and linguistically grounded markers of reflection quality in 6,196 trainee responses collected during a three-week counseling training study. We find systematic shifts over time, including increased semantic alignment and reduced variability, suggesting stabilization in trainees' language. These shifts are accompanied by changes in linguistic features previously associated with high-quality counseling conversations.

2 Related Work

Manual assessment of counseling skills has long provided the clinical foundation for evaluating treatment fidelity and trainee competence. In MI,

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 238–243

coding systems such as the Motivational Interviewing Treatment Integrity (MITI) scale are widely used to assess counselor behaviors, including reflections, questions, MI-adherent behaviors, and relational qualities such as empathy and partnership (Moyers et al., 2016). Similar manual fidelity and competence measures have been developed for other therapeutic approaches, including cognitive behavioral therapy. These instruments provide clinically grounded standards for evaluating counseling quality, but they require trained human coders and are time-intensive to apply. As a result, they are difficult to scale for routine supervision, frequent trainee feedback, or longitudinal monitoring of skill acquisition.

Computational approaches have sought to reduce this assessment burden by automating aspects of counseling skill and fidelity evaluation. Early NLP work on counseling skill assessment focused on classifying discrete counselor behaviors from individual utterances. For example, Pérez-Rosas et al. (2017b,a) used psycholinguistic and n-gram features to predict behaviors such as reflective listening, questions, and empathy, establishing links between language patterns and counseling quality. Building on this work, Pérez-Rosas et al. (2019) analyzed turn-level features such as word exchange, sentiment, and linguistic alignment, showing that higher-quality counselors exhibit more balanced exchange, more positive sentiment, and greater linguistic coordination. These findings motivate our analysis of linguistic patterns associated with reflective practice.

Prior work has also examined automated assessment at the response and session levels. Min et al. (2022) introduced PAIR, a prompt-aware ranking model for scoring reflective responses, and released the dataset used in this work. Other work has modeled broader session-level quality and adherence. For example, Flemotomos et al. (2021) developed a BERT-based model for CBT session evaluation, demonstrating the value of domain-adapted contextual representations, while Ardulov et al. (2022) modeled counseling sessions as dynamical systems, showing that temporal trajectories can be predictive of therapist competence.

Together, this literature shows that automated methods can support scalable assessment of counseling language, from utterance-level behavior coding to session-level fidelity evaluation. However, most existing approaches evaluate isolated responses, utterances, or sessions rather than modeling how trainees' reflective listening changes across repeated practice. We build on this work by taking a longitudinal perspective in a structured training setting. By representing trainee responses as residual embeddings and tracking their evolution across weeks, we provide a progression-oriented account of changes in reflective listening practice.

3 Dataset

The dataset used in this study is derived from a user study conducted by Min et al. (2022), in which the authors evaluated an automatic reflection scoring system in a real educational setting. The study was conducted in a graduate-level Motivational Interviewing (MI) training course and was deployed through a web-based learning platform.

As part of three weekly assignments, students were presented with counseling scenarios, or client prompts, designed to elicit reflective responses. For each prompt, students wrote a response and subsequently received automated scoring feedback. Prompt–response interactions generated during the study were securely collected and de-identified, and the original study followed the ethical oversight procedures described by Min et al. (2022). The resulting data consist of client prompts paired with student-generated responses collected over three weeks, capturing longitudinal practice data for each trainee. The full dataset includes 7,264 prompt–response pairs from 90 students.

For longitudinal analyses, we used data from 68 students who participated consistently throughout all three weeks, resulting in 6,196 total responses. The use of these de-identified data was reviewed by the Texas State University Institutional Review Board, which determined it to be exempt. Table 1 shows sample prompts and responses, and Table 2 presents week-level summary statistics.

4 Analyzing Trainee Response Change Over Time

We analyze responses to examine how trainees' language changes over time during reflective listening practice. We first isolate trainee-specific semantic signals via residual embeddings, then quantify week-to-week change using semantic similarity metrics, and finally characterize consistent and variable trajectories to better understand patterns of language adaptation.

Prompt	Response
I have no energy. The thought of working out now is out of the question. I used to bike and play tennis but that seems a lifetime ago. I would love to feel alive again, but I am just so tired all the time.	You are exhausted and thinking about working out seems impossible to you; however, you desire an alternative way to feel revitalized.
Of course, I have to quit at some point, but now is just not the time. I know what it's like to go through withdrawal... I was able to quit before, but with my wife losing her job and working overtime, I don't have the energy right now.	You are dealing with a lot of stress, so quitting smoking is not a priority right now.

Table 1: Sample trainee responses to counseling prompts.

	Word Statistics				
	Avg.	SD	Min	Max	#Responses
Week 1	23.23	14.41	3	165	1586
Week 2	22.90	14.81	3	152	2470
Week 3	25.62	14.49	3	113	2140
All	23.93	14.65	3	165	6196

Table 2: Dataset Statistics

4.1 Capturing Semantic Change via Residual Embeddings

Semantic variation over time can indicate how trainees adapt their responses toward reflective language. We quantify this variation by measuring semantic similarity across weeks. However, because all students respond to the same prompts within each week, similarity may be inflated by shared prompt semantics. To account for this, we compute a *residual embedding* for each response by removing prompt content at the embedding level.

$$\mathbf{r}_{s,w} = \mathbf{e}(\text{response}_{s,w}) - \mathbf{e}(\text{prompt}_{s,w}), \quad (1)$$

where $\mathbf{e}(\cdot)$ is a sentence embedding function, s indexes responses, and w indexes weeks. This formulation relies on approximate linear compositionality, such that subtracting prompt embeddings reduces shared semantic content and highlights response-specific framing. While imperfect, it provides a practical approximation of prompt-conditioned variation.

We use BGE-M3 (Chen et al., 2024) to encode prompts and responses, with L2-normalized embeddings. Residual embeddings are aggregated at the student–week level, and pairwise similarity across weeks is computed using cosine similarity, Euclidean distance, and Manhattan distance. Cosine similarity captures directional alignment (e.g., paraphrasing or stance consistency), while distance metrics reflect the magnitude of semantic change.

Table 4 shows higher cosine similarity and lower distances for Week 2–3 compared to Week 1–2, indicating increasing semantic alignment and reduced variability over time, which may signal the use of consistent linguistic structure while formulating reflections. Compared with the W1-W2 interval, the W2-W3 interval shows higher cosine similarity (0.74 to 0.77) and lower Euclidean and Manhattan distances, reflecting reduced variability in semantic representations. Together, these trends indicate a stabilization of response patterns, with less pronounced semantic shifts in successive weeks.

4.2 Lexical Progression Across Semantic Trajectories

To better understand differences in semantic trajectories, we examine representative students at the extremes of the prompt-adjusted cosine similarity distribution. We label students in the upper and lower quartiles as *consistent* and *variable* trajectory examples, respectively. These labels refer only to movement in prompt-adjusted embedding space, where higher cosine similarity indicates more consistent semantic positioning across weeks.

Table 3 shows sample responses from representative consistent and variable trajectories between Week 1 and Week 3. In these examples, the consistent trajectory remains compact and closely tied to the client's stated perspective, whereas the variable trajectory shows greater variation in structure and elaboration, including multi-clause responses and inferred emotional content.

We further examine lexical anchoring and reformulation using response length and novel-token rate. Novel-token rate is the proportion of response tokens that do not appear in the prompt and captures lexical departure from the prompt's surface wording.

As shown in Figures 1 and 2, response length increased from Week 1 to Week 2 and then stabilized,

Trajectory	Cosine	Week	Response
Consistent	0.79	W1	*"you know cigarettes are bad and you don't see a problem with vaping right now"*
		W2	*"you see more potential in her than just watching TV"*
		W3	*"good and quick are something you look for in eating meals."*
Variable	0.49	W1	*"You want to lose weight, but dieting has not worked for you in the past."*
		W2	*"Despite working hard, you feel frustrated with your experiences. You are motivated, but feel overwhelmed by the lack of results."*
		W3	*"You want to make beneficial dietary changes for your children, but their behavior makes it difficult to follow through."*

Table 3: Illustrative consistent and variable trajectories from Week 1 to Week 3 using residual embeddings. Cosine similarity is computed on aggregated prompt-adjusted residual embeddings; higher values indicate more consistent semantic positioning across weeks.

Interval	Cosine		Euclidean		Manhattan	
	Mean	SD	Mean	SD	Mean	SD
W1–W2	0.74	0.06	0.25	0.03	6.21	0.820
W2–W3	0.77	0.06	0.23	0.04	5.74	0.935
W1–W3	0.68	0.08	0.27	0.04	6.80	1.034

Table 4: Cohort-level residual-embedding similarity across weeks ($n = 68$ students with complete W1–W3 submissions).

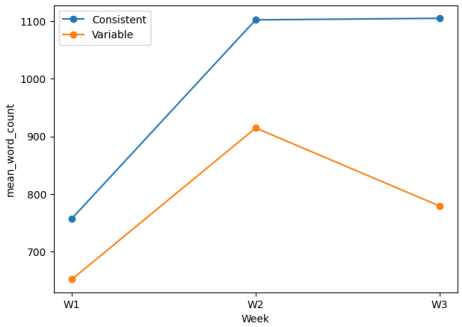

Figure 1: Word count across weeks by trajectory group

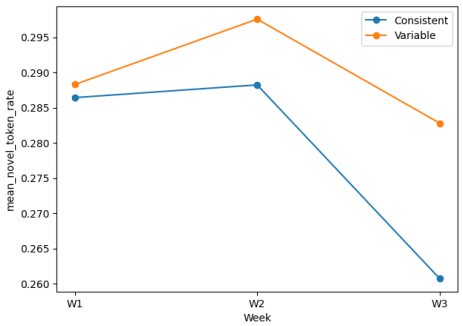

Figure 2: Novel-token rate across weeks by trajectory group

Week	Mean	SD	Min	Max
Week 1	0.146	0.022	0.104	0.200
Week 2	0.166	0.022	0.112	0.242
Week 3	0.147	0.023	0.087	0.227

Table 5: Within-week internal consistency for students ($n = 68$). Higher values indicate greater semantic similarity across a student's responses within the same week.

while novel-token rate also increased from Week 1 to Week 2. This pattern indicates greater lexical departure from the prompt during the middle week. From Week 2 to Week 3, novel-token rate decreased, suggesting a shift back toward greater client-language anchoring. Together, these patterns are consistent with a balance between adding new reflective content and maintaining connection to salient client wording.

4.3 Examining Student-Level Internal Consistency Within Each Week

The previous analyses track semantic change across weeks. Here, we examine within-week variation to assess whether trainees rely on similar response patterns across prompts within the same week. This analysis helps distinguish longitudinal stabilization from repeated use of fixed response templates.

We quantify within-week consistency by computing the average pairwise cosine similarity between each student's responses to different prompts within the same week. Higher values indicate more similar responses across prompts, while lower values indicate greater prompt-specific variation.

As shown in Table 5, within-week similarity is low overall, with means ranging from 0.146 to 0.166. This suggests that trainees maintain substantial variation across prompts rather than relying on highly fixed templates. Similarity increases slightly in Week 2 and decreases again in Week 3, indicating modest fluctuation in within-week consistency across the training period.

Taken together, the longitudinal and within-week analyses suggest that trainees' reflective-listening

responses change over time without becoming rigidly templated. While the longitudinal analyses show increased semantic alignment and changes in lexical anchoring, the low within-week similarity indicates that trainees continue to adapt responses to individual prompts. This pattern is consistent with emerging reflective structure alongside prompt-specific reformulation.

5 Examining Alignment with Reflective Listening Practices

To examine whether the observed trajectory changes align with linguistic patterns relevant to reflective counseling practice (RQ2), we analyze trends in LIWC-based language features. Prior work in counseling dialogue analysis has shown that LIWC categories captured aspects of counseling language, including the use of emotion words (positive and negative emotion), directive markers (auxiliary verbs and certainty), and reflective framing, which includes second-person constructions, cognitive process words, affect, perception, and tentative language (Pérez-Rosas et al., 2019; Tausczik and Pennebaker, 2010). We normalize LIWC counts by response length and average feature values across responses within each week.

Table 6 shows modest and non-uniform changes across linguistic features over time. Most features exhibit relatively small shifts, suggesting that trainees maintain a broadly stable linguistic profile while their responses change in semantic space. Some features decrease in Week 2 and recover in Week 3, including positive emotion and tentative language, echoing the nonlinear pattern observed in the lexical analyses.

A subset of features shows directional changes that are consistent with reflective listening practice. Features that may index directive or closed framing, including auxiliary verbs and certainty terms, decrease from Week 1 to Week 3. In contrast, second-person constructions remain relatively stable, while broader reflective framing features show only modest change. These patterns suggest that trajectory-level changes are not driven by large shifts in surface-feature composition, but by more subtle reorganization of language use.

Overall, the LIWC analysis provides complementary evidence that semantic trajectory changes are accompanied by small but interpretable linguistic shifts. However, these features should be understood as surface indicators associated with reflective practice rather than direct measures of clinical competence.

Feature	W1	W2	W3
Emotion Words	7.74	7.44	7.97
Positive Emotion	3.39	2.84	4.21
Negative Emotion	4.35	4.60	3.76
Directive Markers	12.46	12.53	11.19
Auxiliary Verbs	11.36	11.35	10.39
Certainty	1.10	1.18	0.80
Reflective Framing	47.62	45.25	46.09
You	13.77	14.54	14.22
Cognitive Processes	17.32	16.32	16.78
Affect	8.03	7.75	8.23
Perception	5.63	4.38	4.03
Tentative Language	2.87	2.25	2.82

Table 6: LIWC features across Weeks 1–3.

6 Conclusion

In this work, we model reflective listening practice as a longitudinal process by representing trainee responses as trajectories in semantic embedding space. Using residual embeddings and similarity-based metrics, we identify structured changes in trainee language over a three-week counseling training study. Across analyses, we observe increased semantic alignment and reduced variability, patterns consistent with stabilization in trainees' language. Complementary linguistic analyses show that these trajectory patterns are accompanied by modest shifts in language use rather than large changes in surface-feature composition.

Although our findings do not directly measure clinical competence, they show that embedding-based trajectory representations can capture patterns of change that were interpretable in relation to lexical and LIWC-based analyses in this dataset. More broadly, this work highlights the potential of trajectory-based approaches for studying communication skill development and designing progression-aware feedback systems.

Limitations

This work has several limitations that should be considered while interpreting our findings. First, our dataset is drawn from a single instructional setting over a three-week period, which may limit the generalizability of the observed trajectory patterns. Longer training durations or more diverse educational contexts may exhibit different developmental dynamics. Second, our analyses rely on embedding-based representations of language, which provide an indirect proxy for changes in

language associated with reflective listening behavior. While residual embeddings help isolate response-specific variation, they depend on assumptions about linear compositionality and may not fully capture nuanced aspects of reflective listening. Third, although we interpret trajectory patterns as indicators of changes in reflective listening practice, our study does not directly measure clinical competence. The observed changes reflect structured variation in language use, but should not be taken as definitive evidence of improved counseling performance.

Finally, our linguistic analyses focus on aggregate feature trends, which may overlook more fine-grained discourse phenomena such as turn-taking dynamics or interactional context. Future work could incorporate richer conversational signals and longer-term trajectories to better characterize the development of reflective listening skills.

Acknowledgments

The authors thank the authors of the original dataset for making the study data available for this research. We are also grateful to the anonymous reviewers for their constructive feedback.

References

Victor Ardulov, Victor R. Martinez, Krishna Somandepalli, Allison Lahnala, Jill Burstein, and Shrikanth Narayanan. 2022. Local dynamic mode decomposition for assessing therapist competency in cognitive behavioral therapy sessions. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3018–3022.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2023. VERVE: Template-based ReflectiVE rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10289–10302, Singapore. Association for Computational Linguistics.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. 2016. The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65:36–42.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017a. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017b. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

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.

Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.