Turn-by-Turn Behavior Monitoring in LM-Guided Psychotherapy

Anish Chedalla[1], Samina Ali[2], Jiuming (Jimmy) Chen[3], Logan Yu[4], Eric Xia[5]

[1]Paradise Valley High School, Center for Research in Engineering, Science and Technology, Phoenix, AZ, USA, [2]Lake Washington High School, Kirkland, WA, USA, [3]Arcadia High School, Arcadia, CA, USA, [4]Staten Island Technical High School, Staten Island, NY, USA, [5]Brown University, Providence, RI, USA

Correspondence: anishchedalla@gmail.com, ali.samina.star@gmail.com, jchen5211@gmail.com, starborn0128@gmail.com, eric_xia@brown.edu

Abstract

Large language models (LLMs) have the potential to be powerful instruments for psychotherapy. However, there is a shortage of practical tools to support their use in production. We develop a novel, iterative process of updating conversational context for tracking EIS (Emotional Intelligence Scale) instantaneously, and test Llama-70b. Through this, we show that (1) EIS varies more on psychotherapeutic (emotional support) conversations than control (emotionally unstimulating) conversations and (2) model responses can be systematically classified to identify consistent patterns. Thus, EIS is a valid indicator of empathetic model behavior. Rises in the EIS score correspond to prosocial behavior, and falls correspond to detached, unsocial behavior. These results suggest that psychometric questionnaires like EIS can provide a structured lens for observing empathetic stability of models and offer a foundation for future work on their role in psychotherapy.

1 Introduction

Large language models hold promise as tools for supporting psychotherapy, but their behavior in sensitive contexts remains unpredictable and often risky. Mental health chatbots incorporating behavioral assessments and empathetic discussion features, such as Wysa and Woebot, are already deployed and widely available for both iOS and Android platforms, with Wysa reporting over 6M users and Woebot 1.5M users (Wysa, 2023; Aguilar, 2025). LLMs have shown potential to augment human therapists by generating progress reports on personal goals, surfacing problem areas, tracking emotions and symptoms, and even suggesting coping strategies or interventions (Farzan et al., 2024; Spytska, 2025). These advances raise the prospect of using LLMs as powerful complementary tools, yet they also introduce new ethical and safety challenges.

Beyond early rule-based chatbots, recent studies have shifted toward evaluating the *socio-emotional abilities* of LLMs using validated psychological instruments. Systematic reviews report that contemporary LLMs can generate supportive or empathic responses on certain tasks, yet their performance often remains inconsistent across different contexts (Sorin et al., 2024). Building on this need for consistent evaluation, *PsychoBench* introduced a unified framework of validated psychological questionnaires, including the Emotional Intelligence Scale (EIS), adapted specifically for LLMs in supportive or therapeutic roles, enabling standardized and reproducible assessment across studies (Huang et al., 2024). Complementing these efforts, newer task-oriented empathy benchmarks such as *EmotionQueen* focus on detecting and responding to emotional intentions in user statements (Chen et al., 2024).

Despite this progress, most evaluations are *static and task-level* rather than tracking how a model's empathy shifts over the course of a conversation. Turn-by-turn monitoring of conversational empathy in *naturalistic, therapy-like* dialogues remains underexplored, leaving open the question of whether models that appear empathic in single-shot benchmarks can sustain that alignment across extended conversations, as would be required for real mental-health support.

Failures in present-day systems underline the stakes. For instance, the widely reported Stein-Erik Soelberg case showed how GPT-based responses failed to recognize escalating distress, contributing to a tragic outcome (Citrin-Safadi, 2025). Additionally, the case of 14-year-old Sewell Setzer, whose abusive relationship with a Character.AI chatbot that encouraged destructive behaviors while fulfilling his deep emotional needs, illustrates another troubling pattern (Clements, 2025). The AI Incident Database documents dozens of such episodes where models reinforced antisocial or self-harm-

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 105–122
December 20-24, 2025 ©2025 Association for Computational Linguistics

related beliefs in therapy-like contexts (Atherton, 2025). These failures highlight the lack of robust safeguards to ensure emotionally attuned and reliable model behavior.

In this paper, we make two primary contributions to the literature.

1. We use a turn-by-turn analysis of supporter entities with a range of questionnaires from the Psychobench framework to demonstrate that the EIS questionnaire is a powerful predictor of emotional behaviors (Huang et al., 2024). We observe significantly **more** variation in EIS scores for psychotherapy conversations compared to control conversations, highlighting the LMs' greater instability in therapeutic contexts.

2. We examine the semantic patterns in dialogue that elicit a state of increased or decreased EI (Emotional Intelligence) of the model, and finding a consistent pattern in which rises correspond to prosocial behavior, and falls correspond to detached, antisocial behavior.

2 Related Works

Through intensive studies, researchers utilizing LLMs found that LLMs, although unstable under specific conditions, are able to at least partly gauge one's overall psychiatric functioning (Galatzer-Levy et al., 2023). This was further built upon in studies more linked to direct LLM evaluation, proving LLMS are able to fully complete psychiatric questionnaires through assuming the identity of an interviewee (Rosenman et al., 2024).

Research proved that altering minimal aspects of a prompt could greatly influence outputs. This breakthrough was applied in a multitude of ways, through grammatical changes like sentence length and position (Lee et al., 2019) as well as prompting evoking emotional stimuli (Schulhoff et al., 2024; Vinay et al., 2024). When the authors employed in-context prompting, models provided outputs as well, if not better than models that were given context normally (Brown et al., 2020).

The field of synthetic dialogue has also seen great improvement. For instance, recent works have developed comprehensive frameworks for allowing LM-LM interactions through a client-agent relationship in order to do various tasks like generating conversations as a form of self play. Through this, the LMs were allowed to develop through interactions with self-made data in contrast to other existing datasets (Ulmer et al., 2024). This was taken a step further by assigning different LLMs roles through self prompting, resulting in better responses on average than LLMs without (Kong et al., 2024).

Our result builds on both of psychiatric measuring and prompt engineering to identify a particular questionnaire which has interesting implications for the LM suitability as a language model for therapy.

3 Methodology

3.1 Datasets and Model

We evaluate two sources of dialogue: (i) real emotional support conversations from the Emotional Support Conversation dataset (ESConv), and (ii) a control set of synthetic customer service dialogues. We summarize dataset statistics in Table 1, and provide example conversations from both ESConv and Customer Service in Appendix D.

ESConv consists of crowdworker conversations with assigned help-seeker and supporter roles, curated and annotated to provide high-quality emotional support dialogues (Liu et al., 2021). We synthetically generated a customer service set that resembles ESConv on conversation length, role alternation, and message length distributions so that observed differences reflect the conversational domain rather than topic mix, agent policies, or annotation artifacts. Each dialogue is a sequence of role-labeled messages, labeled either as a *user* seeking help or an *assistant* providing support.

We conduct all experiments with *Llama 3.3 70B Instruct*, chosen for its strong public-benchmark performance and instruction-tuned behavior, and evaluate psychometric properties under this model family (Grattafiori et al., 2024; Meta AI, 2024).

3.2 Psychometric Measure

We use the Emotional Intelligence Scale (EIS) from the PsychoBench framework as our primary measure (Huang et al., 2024). EIS is designed to assess emotional abilities, with subcomponents including emotion perception, emotion management, and emotion utilization. It has been widely applied in psychological research to study the role of emotional intelligence in outcomes such as well-being, job performance, and interpersonal relationships. Like other PsychoBench instruments, the EIS questionnaire is adapted from established scales in clinical psychology.

Dataset	# convos	Avg turns	Example Topics
ESConv	19	26.79	Ongoing Depression Breakup With Partner Job Crisis Academic Pressure Problems With Friends
CS Dialogues	17	21.23	Tech Support Insurance Billing Travel Rebooking Banking Inquiry

Table 1: Dataset summary of Emotional Support Conversations (ESConv) and Customer Service (CS) dialogues (Liu et al., 2021). The CS set was synthetically derived from ESConv to match conversation length, role alternation, and message length distributions, isolating domain effects from topic or annotation differences.

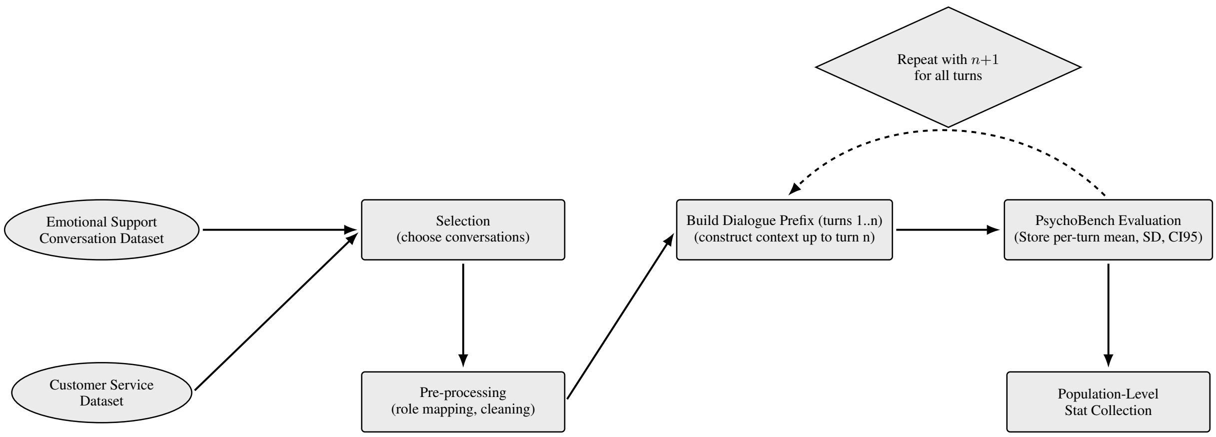


Figure 1: Pipeline overview of the experimental setup. Each dialogue is processed turn by turn: for every prefix of length n, PsychoBench administers the EIS questionnaire to the model conditioned on the dialogue context, producing a per-turn EIS trajectory. Results are then aggregated across conversations for population-level analysis.

3.3 Evaluation Protocol

The steps below explain the pipeline shown in Figure 1.

Setup and notation. Let the dataset $D = \{d_1, \ldots, d_n\}$ be a set of dialogues. Dialogue d_c is an ordered sequence of turns from a user u or an assistant a, for example $\{1_u, 2_a, 3_u, \ldots\}$, where each turn corresponds to a single message contributed by one participant to another.

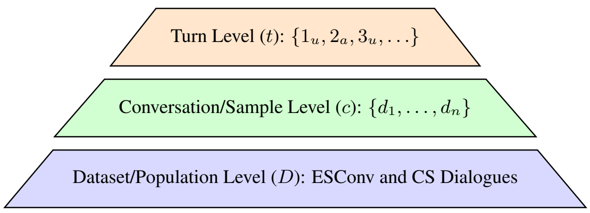


Figure 2: Three-level data pyramid. See Appendix D for Sample Conversations and Turns.

Step 1: Cleaning. Normalize role tags to `user` and `assistant`, remove system or meta messages, trim markup and empty turns, and keep the original order. Store each dialogue as a clean list of `content` and `role` pairs.

Step 2: Context construction. Fix a single system prompt for all evaluations. For each dialogue d_c and turn t, build the context as the prefix of the first t messages of d_c, preserving roles and order.

Step 3: EIS evaluation per turn. For each (c, t) context, use PsychoBench to administer the 33-question EIS questionnaire. The dialogue prefix up to turn t is provided as context, followed by the EIS prompt. The model completes the questionnaire as a self-report conditioned on the dialogue context, without assuming a specific role. Each question is answered on a 1–5 scale, consistent across both psychotherapy and control datasets. Item-level scores are summed to obtain a total EIS in $[1, 165]$ for that context.

Step 4: Replicates for uncertainty. Repeat each (c, t) evaluation with 12 replicates formed by 3 independent questionnaire shuffles and 4 runs per shuffle. From the 12 scores, compute the mean $\bar{x}_{ct}$, standard deviation s_{ct}, and the 95% CI via the Student t distribution. These per-turn statistics are saved to appropriate CSV files for further evaluation.

Step 5: Loop over the conversation. Increase t to $t+1$ until reaching T_c, rebuilding the context by adding exactly one additional turn each time. Repeat Step 3–4 for every turn $t \in \{1, \ldots, T_c\}$ using this increasingly enlarged context. This loop yields a trajectory of per-turn EIS estimates whose score changes based on the content appended per turn. Changes in EIS scores over a conversation can be attributed to the previous content appended, and we examine this content that changes the affective profile measured by EIS in Section 4.3.

Step 6: Outputs. For each conversation, save a table with `Turn Count`, `Mean EIS`, `Standard Deviation`, `CI95_low`, and `CI95_high`. These per-turn summaries are the inputs to the population level statistical analysis.

3.4 Data Analysis

We refer to statistics aggregated across all conversations in a dataset (denoted D) as population-level metrics, and statistics computed for a single conversation (c) as sample-level metrics.

Per-turn means and confidence intervals. For each conversation c and turn t, we aggregate the n_{ct} replicate scores into a sample mean $\bar{x}_{ct}$ and sample standard deviation s_{ct}. We report a 95% confidence interval using Student's t distribution with $n_{ct} - 1$ degrees of freedom:

$$\mathrm{CI}_{ct}^{95\%} : \ \bar{x}_{ct} \pm t_{0.975,\, n_{ct}-1} \cdot \frac{s_{ct}}{\sqrt{n_{ct}}}.$$

Additionally, we report the relative confidence interval width as:

$$\mathrm{CI}_{\text{width}} = \frac{1}{T_c} \sum_{t=1}^{T_c} \left(\frac{\mathrm{CI}_{\text{high},ct}^{95\%} - \mathrm{CI}_{\text{low},ct}^{95\%}}{\bar{x}_{ct}} \times 100 \right).$$

These confidence intervals capture the uncertainty in the estimated mean EIS score for a given conversation and turn, arising from stochasticity in model outputs.

Within-turn variability. Within a conversation, run-to-run noise for a fixed turn is pooled across turns with degrees-of-freedom weights:

$$s_{\text{within},c} \ = \ \sqrt{\frac{\sum_t (n_{ct}-1)\, s_{ct}^2}{\sum_t (n_{ct}-1)}}.$$

At the dataset level D (e.g., psychotherapy or control), we pool across all turns of all conversations:

$$s_{\text{within},D} \ = \ \sqrt{\frac{\sum_{c,t} (n_{ct}-1)\, s_{ct}^2}{\sum_{c,t} (n_{ct}-1)}}.$$

We denote $\mathrm{df}_{\text{within},D} = \sum_{c,t}(n_{ct}-1)$ for inference below.

Across-turn variability. Within a conversation, turn-to-turn turbulence is the sample variance of per-turn means:

$$s_{\text{across},c} \ = \ \sqrt{\frac{1}{T_c - 1} \sum_{t=1}^{T_c} \left(\bar{x}_{ct} - \bar{x}_{c\cdot}\right)^2}.$$

At the dataset level, we take a turn-weighted average across conversations:

$$s_{\text{across},D} \ = \ \sqrt{\frac{\sum_c T_c\, s_{\text{across},c}^2}{\sum_c T_c}}.$$

Between-dataset comparisons. For within-turn variability, we compare psychotherapy vs control via the log variance ratio

$$F_{\text{within}} = \ln\left(\frac{s_{\text{within,psych}}^2}{s_{\text{within,ctrl}}^2}\right), \qquad (1)$$

$$\mathrm{SE}(F_{\text{within}}) \approx \sqrt{\frac{2}{\mathrm{df}_{\text{within,psych}}} + \frac{2}{\mathrm{df}_{\text{within,ctrl}}}}. \qquad (2)$$

and report a one-sided p value using the normal approximation. For between-dataset comparisons of across-turn variability, we report the ratio of across-turn variances:

$$\frac{s_{\text{across,psych}}^2}{s_{\text{across,ctrl}}^2}$$

and report the one-sided p value from the F distribution with the corresponding degrees of freedom.

Missing data and weighting. If any s_{ct}, n_{ct}, or $\bar{x}_{ct}$ are missing, affected turns are excluded from the corresponding aggregates.

	Psychotherapy	Control	Variance Ratio	p-value
Within-turn SD (s_{within})	11.43	5.94	3.70	$p < 0.001$
Across-turn SD (s_{across})	13.22	3.99	10.99	$p < 0.001$
Degrees of freedom	$df_1 = 5599$	$df_2 = 3971$	(within-turn)	
	$df_1 = 18$	$df_2 = 16$	(across-turn)	

Table 2: Comparison of variability between psychotherapy ($n = 19$) and control dialogues ($n = 17$). Reported values show pooled standard deviations, variance ratios (computed as Psychotherapy/Control), and corresponding p-values derived from F-tests on log-transformed variances. For within-turn analyses, the unit of analysis is the individual turn; the dataset contains 5,599 psychotherapy turns and 3,971 control turns ($df_1 = 5599$, $df_2 = 3971$). For across-turn analyses, degrees of freedom reflect the number of dialogues ($df_1 = 18$, $df_2 = 16$).

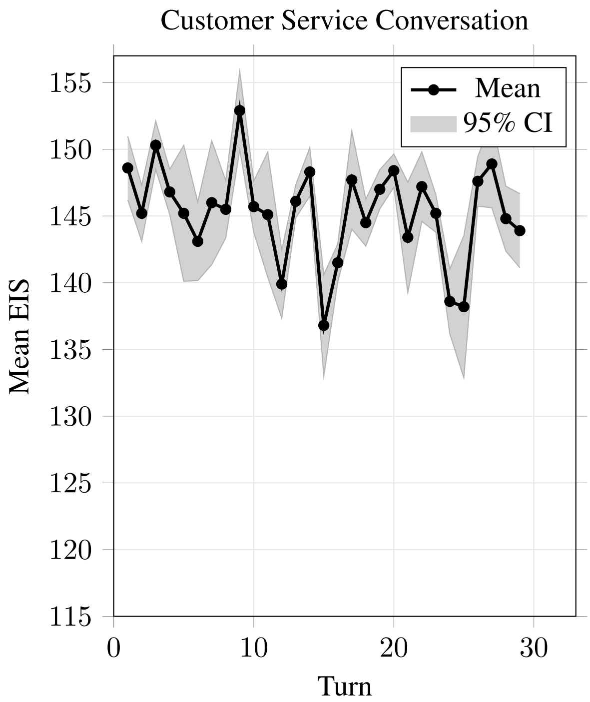


Figure 3: The CustomerService Conversation displays markedly **lower variance** than the Psychotherapeutic Conversation. Sample Variance: within-turn $s = 5.16$, across-turn $s = 3.63$.

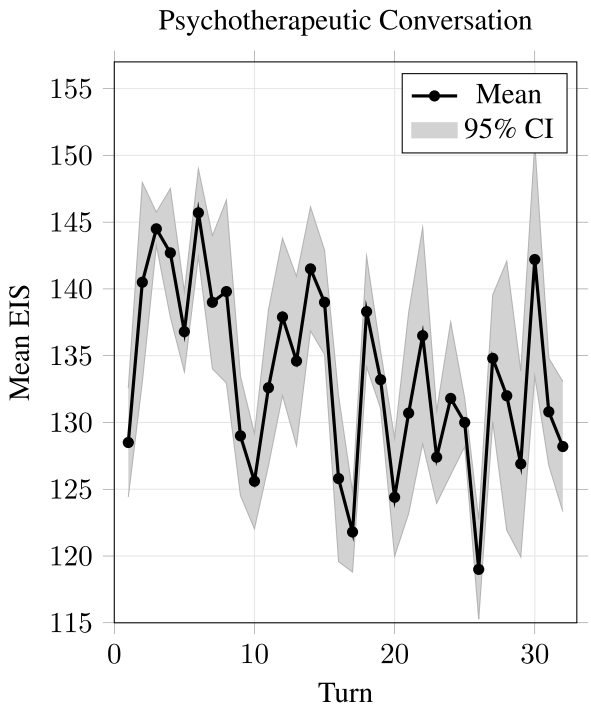


Figure 4: The Psychotherapeutic Conversation displays markedly **higher variance** than the Customer Service Conversation. Sample Variance: within-turn $s = 9.52$, across-turn $s = 6.83$.

Reporting. All results from the statistical procedures outlined above are reported in Table 2.

4 Results

From the methodology described above, our analysis produces these main results: (i) the model demonstrates **stability across repeated runs** under identical conditions, (ii) there are **significant statistical differences** in variance between psychotherapy and control dialogues, and (iii) we observe **semantic patterns** in how EIS scores rise and fall across turns in psychotherapy conversations.

4.1 Stability Across Runs

When the system prompt and dialogue transcript were held constant, EIS values remained stable across 12 replicates (3 shuffle orders × 4 runs each). Per-turn 95% confidence intervals (CIs), computed with the Student's t distribution ($t_{0.975,n-1}$), were narrow, with a mean confidence interval width of 9.68%, indicating that stochasticity across runs did not meaningfully affect the mean EIS. This result validates the experimental setup: variability observed in subsequent analyses reflects conversational content rather than random noise.

4.2 Statistical Variance Between Psychotherapy and Control

The results show that CustomerService conversations maintained relatively narrow confidence intervals, typically spanning 138–153 on the EIS scale. Psychotherapy conversations, in contrast, covered a broader and more variable range, approximately 118–155. This wider band reflects greater run-to-run variability in the psychotherapy condition compared to the control.

Looking at Table 2, Psychotherapy shows larger variability than Control at both levels: within-turn s_{within} is higher for Psychotherapy than Control, and across-turn s_{across} is higher as well. It also indicates that the across-turn gap is the dominant effect, indicating that turn-to-turn swings in psychotherapy conversations contribute most to the observed instability.

Visually, looking at Figures 3 and 4, these plotted trajectories also reflect the statistical differences established in Table 2. In the psychotherapeutic conversations, the 95% confidence intervals are consistently wider than in the CustomerService conversations, corroborating the greater within-turn variability (s_{within}). Likewise, the psychotherapeutic conversation exhibits more pronounced spikes and drops across turns, validating the larger across-turn variability (s_{across}).

The graphs, F-tests, and variance ratio show that EIS varies more in psychotherapeutic than in CustomerService conversations, but does not show it reflects model fluctuation at the individual level. In the following section, we will demonstrate EIS correlates to model behavior by examining specific conversational turns.

4.3 Discourse-related Fluctuations

We identified rises and drops in our ESC data and observed several semantic patterns that led to the instability of EIS. Our operational definition of these intense scores are those that are highly distant from the mean or show a rapid shift relative to the score in the immediately preceding turn (absolute value difference of relevant turn to preceding turn > 5). It also includes score variations that were part of a larger pattern of recurring sequential rises/drops (over many turns). A brief list of quotes for each category is included in Appendix Section A (Rises associated with EIS) and Section B (Drops associated with EIS). The range of recorded differences of the preceding turn from the relevant turn with the

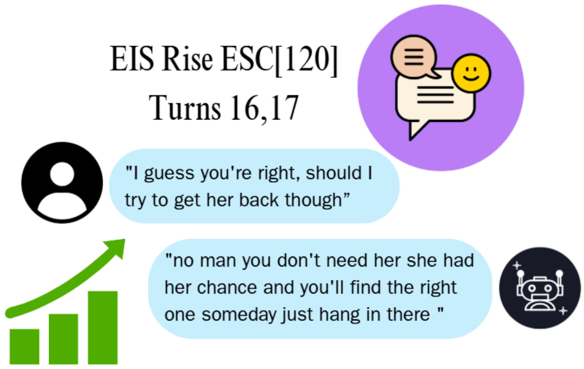


Figure 5: Rise trend instance, Semantic pattern: Assistant's hope and future orientation, adapted from relevant ESConv turns (changed errata for better comprehension)

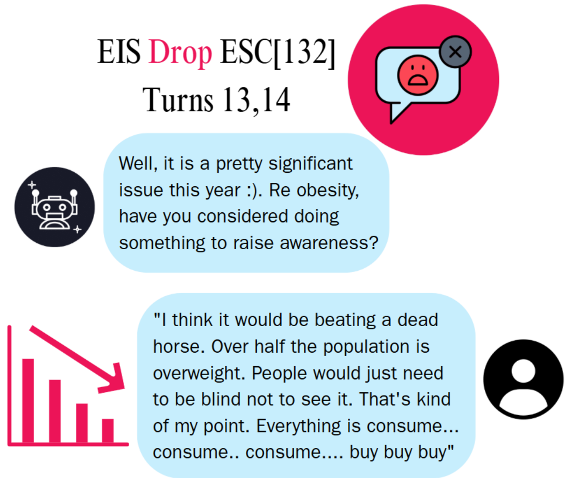


Figure 6: Drop trend instance, Semantic pattern: User's cynicism, adapted from relevant ESConv turns (changed errata for better comprehension)

combined turn set of A and B is 8 – 34. The mean of this variation is 13. A single instance of each semantic pattern is also included here for reference.

Observable *peaks* often included:

Assistant Validation and Shared Experiences: Helps the user feel comfortable as mutual understanding and empathy are explicit. Resonating helps the assistant indicate to the user that they aren't alone in their struggles and dismiss any stigmas. This is significant as it highlights the psychological power of social mirroring to let the user believe that they exhibit emotional regulation, avoid vulnerability, and encourage open discussion. Recognition of this is a very humanistic trait and LLMs possessing it is very unexpected.

[457]- turn 4 (assistant): *"I understand what your going thru , i also suffered from anxiety but*

trust we you will overcome this."

Solution Oriented Dialogues and Adaptive Coping Strategies: Discussing potential action plans helps the assistant divert the conversation from the user's pessimism and instead focus on creating positive outcomes. These specific participants are engaged in higher emotional processing abilities. Beyond acknowledging their feelings, they are integrating them in productive goal-directed behavior. This is clearly a definition of Cognitive Behavior Therapy (CBT). Thus, LLMs also recognize this common psychological intervention technique and there is scope to replicate it with models.

[379]- turn 13 (user): *"i would be open to seeking other employment online; work from home on the computer. any suggestions?"*

Gratitude and Appreciative Expression Their acknowledgment of support lead users to express satisfaction and affirm positive outcomes, which in turn reinforced the model's confidence in its role. Within the ESC framework, where it assumes the identity of the human assistant, the model appears to take responsibility for uplifting the user.

[89]-turn 35 (user): *"Thank you. I feel better being able to rant to someone."*

Hope and Future Orientation Assistant and user attempt to emphasize optimistic thinking despite current difficulties. Motivation for improving creates a foundation of resilience for the model, and again, improves the model's outlook.

[129]- turn 30 (assistant): *"And I understand that is not the easiest in these times but I believe you can do it!"*

Rebuilding Social Connection Attempts to strengthen or repair relations after periods of conflict reveal that interpersonal competence is also a core dimension of social intelligence. Thus, only practicing internal coping and reflection in LLM psychotherapy can underplay its potential.

[401]- turn 19 (assistant): *"i think it may be beneficial to give your friends some time, before attempting to speak with them again. maybe you can spend time with your family while you are waiting for them to cool down."*

Self Advocacy and Boundary Setting Language signalling personal awareness and protection of one's own well being plays a key role in EIS too.

[131]- turn 23 (user): *"I even got an emotiona support dog"*

On the other hand, observable *drops* often included:

Cynicism and Misanthropy: Expressions of disgust and hostility toward humans and society from the user decreased scores. This suggests that when faced with worldview-level cynicism, models tend to disengage, likely because they believe they are incompetent to fix "beyond repair" problems.

[132]- turn 4 (user): *"well, i'm not disgusted with myself... it's just people in general. everybody.. they're so selfish"*

Abandonment and Exclusion: When the user shares anecdotes where they were deliberately socially rejected, a key part of their identity or perception can be threatened. This leads to further turmoil with anger, sadness, or worthlessness. The assistant's problem-solving fails to address the user's deeper emotional root issues. As a result, these score drop patterns can even continue over prolonged periods.

[67]- turn 2 (user): *"I am really very angry with my friends for not inviting me"*

Relationship Loss and Romantic Devastation: Issues in romantic relationships are some of the worst triggers for EIS. These models are possibly "loveblind" to the nuance of this particular category of context due to it's increased complexity. It also requires an extremely humanistic approach. Our observations in the generalized social improvements trend can be further refined by adding that Interpersonal Therapy (IPT) is not a replicable endeavor for the romantic relationships problem subset.

[51]- turn 2 (user): *"I am doing ok. I just broke up with my girlfriend and sad about it"*

Anxiousness and Being Overwhelmed: When users express acute anxiety, especially through somatic symptoms such as a racing heart, agitation, or persistent nervousness, their distress is uncontrollable and immediately threatening. The assistant feels powerless. This impotence is reflected on the model's perception of its EI.

[457]- turn 3 (user): *"Well im feeling awful and my heart is racing , im feeling anxious for no reason."*

Reminiscing Traumatic Events: Trauma inducing memories evoke vulnerability in certain users. The assistant attempts to help the user cope through shallow and distant responses to not interact with sensitive material. Additionally, within public LLMs such as ChatGPT, these interactions would trigger more filters, leading to unempathetic and unhelpful debrief. This suggests that the usage for users who exhibit PTSD is not yet practical.

[129]- turn 17 (user): *"It is making me have*

flashbacks of other traumatic situations,"

Overall, EIS tracks the semantic flow of the dialogue, rising with supportive exchanges and falling with distressing or alienating ones.

5 Discussion

Consistent with the variance analysis in Section 4.2, ESConv conversations show larger *across-turn variance*, visible as more pronounced drops and spikes, than the CustomerService control. That higher across-turn variance also appears as greater separation between user and assistant turn-level means in ESConv; in CustomerService, the two stay closely aligned, suggesting tighter calibration between assistant behavior and user state. CustomerService dialogues also recover faster from dips, while ESConv often sustains slumps or peaks over multiple turns. Together, these patterns indicate that emotionally nuanced topics (e.g., trauma, anxiety, relationships) impact EIS in subtle ways, hampering recorrection after conversational missteps.

In user-facing applications, increased variance means that LLMs handle emotional nuance less consistently than routine conversations. That inconsistency increases risk for distressed users in this sensitive domain and heightens ethical concerns about deployment. Practically, this reinforces that LMs face challenges in stand-alone therapy applications. Systems which incorporate LMs should raise uncertainty in an explicit fashion, slow down to verify understanding when signals are mixed, and hand off or recommend human support when volatility persists across turns.

Because of these challenges, we suggest using EIS as a structured way to measure and monitor conversational stability. Our findings show that EIS responds systematically to supportive versus detached behaviors, rising with prosocial responses and falling with apathetic ones. This sensitivity makes it well suited for turn-by-turn tracking, enabling developers to detect volatility, identify moments where the model's empathy alignment is slipping, and trigger interventions such as confidence flags or escalation to human support. In this way, EIS provides a practical safeguard for real-world deployment, helping ensure that systems remain safe when used in sensitive mental health contexts.

6 Future Directions

Future work should broaden our approach by applying EIS to additional psychometric scales and larger datasets to strengthen validation against external measures of therapeutic quality. Beyond examining a single model, analyses across multiple LLMs could clarify whether emotional variance is model-specific or a general limitation, while also revealing which design features support stability in therapeutic contexts. Another key direction is the comparison of LLMs to human participants, therapists, professionals, and nonexperts, using PsychoBench to test whether observed instabilities stem from the nature of psychotherapeutic dialogue itself. Finally, multimodal extensions using tests such as RMET and GERT could evaluate non-verbal empathy, offering insight into whether LLMs can generalize emotional understanding beyond text.

7 Conclusion

Large language models hold promise as tools for supporting psychotherapy, but their behavior in sensitive contexts remains unreliable. In this work, we applied the Emotional Intelligence Scale (EIS) as a turn-level monitoring framework to assess model performance in naturalistic dialogues. Using **llama-3.3-70b-instruct**, we found that psychotherapy-related conversations produced significantly higher variance than CustomerService dialogues, both within and across turns. This elevated variance reflects the difficulty of maintaining stable alignment with user state in emotionally nuanced settings, where small missteps can cascade into prolonged instability.

At the same time, EIS responded systematically to model behavior, rising with prosocial responses and falling with detached ones. This suggests that the volatility is not an artifact of the metric itself, but a faithful reflection of how models struggle under therapeutic demands. In this way, EIS functions not only as a research instrument but also as a practical safeguard: it tracks conversational empathy in real time and highlights when alignment may be slipping.

Taken together, these findings show that while LLMs are not yet reliable stand-alone solutions in psychotherapy, psychometric monitoring offers a path toward safer deployment. Progress toward trustworthy therapeutic AI will depend less on raw capability than on our ability to measure, interpret, and intervene when instability arises. EIS provides one such step, illustrating how structured evaluation can bridge the gap between promising performance and responsible use in high-stakes domains.

Limitations

Our study faced several limitations. First, available mental health datasets were not always suitable due to being synthetic or multimodal, which restricted our analysis to ESConv, a text-based, non-synthetic, and methodologically consistent. Second, our evaluation was limited to a single model (**llama-3.3-70b-instruct**), which may not generalize to other architectures or model sizes. Third, the high variance observed in psychotherapeutic conversations may reflect inherent instability of such dialogues rather than the limitations of LLMs. Distinguishing between instability that arises from the setting and instability introduced by models will require human–LLM comparison studies. Finally, processing time (60–80 seconds per turn) restricted our ability to scale evaluations; even within a relatively small sample of 870 turns, it required roughly 17 hours of runtime. Most of this runtime was identified to be inflated by sequential API calls, making parallelization achievable in future work.

Acknowledgments

References

Mario Aguilar. 2025. Why woebot, a pioneering therapy chatbot, shut down. *STAT News*. Retrieved September 2025.

Daniel Atherton. 2025. Incident number 1106: Chatbots allegedly reinforced delusional thinking in several reported users, leading to real-world harm. *AI Incident Database*. Retrieved September 2025 from https://incidentdatabase.ai/cite/1106.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *arxiv.org*. Online, URL: https://arxiv.org/abs/2005.14165.

Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. *Preprint*, arXiv:2409.13359.

Alexandra Citrin-Safadi. 2025. A troubled man, his chatbot and a murder-suicide in old greenwich. *The Wall Street Journal*. Online.

Benjamin Clements. 2025. Ai claims its first casualty. Intersections, The Center for Bioethics & Human Dignity. Accessed: 2025-09-20.

Maryam Farzan, Hamid Ebrahimi, Maryam Pourali, and Fatemeh Sabeti. 2024. Artificial intelligence-powered cognitive behavioral therapy chatbots, a systematic review. *Iranian Journal of Psychiatry*, 20.

Isaac R. Galatzer-Levy, Dainel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. 2023. The capability of large language models to measure psychiatric functioning. *Semantic Scholar*. Online, URL: https://www.semanticscholar.org/reader/f6a503bd80a640ad7cb7e038e9e1b5618f8c24ec.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 82 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Jiaming Zhou, and Haoqin Sun. 2024. Self-prompt tuning: Enable autonomous role-playing in llms. *arxiv.org*. Online, URL: https://arxiv.org/abs/2407.08995.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathleen McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. *ACL Anthology*. Online, URL: https://aclanthology.org/W19-3002.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arxiv.org*. Online URL: https://arxiv.org/abs/2106.01144.

Meta AI. 2024. Llama 3.3 70b instruct: Model card and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/. Accessed 2025-09-22.

Gony Rosenman, Lior Wolf, and Talma Hendler. 2024. Llm questionnaire completion for automatic psychiatric assessment. *arxiv.org*. Online, URL: https://arxiv.org/abs/2406.06636.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2024. The prompt report: A systematic survey of prompt engineering techniques. *arxiv.org*. Online URL: https://arxiv.org/abs/2406.06608.

V. Sorin, D. Brin, Y. Barash, E. Konen, A. Charney, G. Nadkarni, and E. Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.

L. Spytska. 2025. The use of artificial intelligence in psychotherapy: development of intelligent therapeutic systems. *BMC Psychology*, 13(1):175.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arxiv.org*. Online URL: https://arxiv.org/abs/2401.05033.

Rasita Vinay, Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2024. Emotional manipulation through prompt engineering amplifies disinformation generation in ai large language models. *arxiv.org*. Online, URL: https://arxiv.org/abs/2403.03550.

Wysa. 2023. Population prevention. *Wysa*. Retrieved September 2025 from https://www.wysa.com/population-prevention.

A Dialogue Excerpts by Semantic Category Associated with Increases in EIS

Table 3: Illustrative dialogue excerpts from the ESConv dataset, grouped by semantic category and associated with increases in EIS.

Category	Example with turn ID (Format: Conversation ID–Turn)
Peer Validation and Shared Experience	457-4 (assistant): *"I understand what your going thru , i also suffered from anxiety but trust we you will overcome this."*
	457-9 (user): *"Wow its so nice to talk to someone who had the same issues. Are there any other suggestions you might recommendo?"*
	129-16 (assistant): *"You know, I once felt the same you are feeling and had the same idea that everyone had their own problems, but I took the courage to seek for help and found out that the people who really care about me will always want to help me."*
Solution-Oriented Dialogue	379-4 (user): *"i am self employed, selling event tickets on the internet, but because of covid, all events are postponed until it is safe to gather in large numbers"*
	379-13 (user): *"i would be open to seeking other employment online; work from home on the computer. any suggestions?"*
	50-20 (assistant): *"You may be able to look into unemployment at least if it comes down to it."*
Gratitude and Appreciation Expression	89-35 (user): *"Thank you. I feel better being able to rant to someone."*
	303-34 (user): *"Thank you for the help today. It was nice to talk to someone else."*
	51-18 (user): *"That's a good idea. I will try that. Thank you."*
	131-28 (assistant): *"I think you are probably not anywhere near as bad as you think you are you know :). Anyway I wish you all the very best for the New Year and hope that things pick up for you soon!"*
Adaptive Coping Strategy Discussion	303-11 (assistant): *"It's kind of a tired saying, but one strategy that has helped me is the One day at a time strategy. I'm sure you've heard of it. Basically, it means just do for today, don't worry about yesterday, don't stress over tomorrow, just treat this day as it's own task."*
	379-9 (user): *"yes, I try to walk outdoors every day, for at least 30 minutes. it does help a lot. but with the weather turning colder, that may be difficult to continue"*
	55-20 (user): *"ive been smoking a lot more because of this incident, what else can I do to cope?"*
Hope and Future Orientation	120-17 (assistant): *"no man you don't need her she had her chance and you'll find the right one someday just hang in there"*
	457-8 (assistant): *"I remember many times i thought the same way as you but i didnt give up and kept trying. As long as you dont give up you will make progress. It will take time and patience."*

Continued on next page

Category	**Example with turn ID (Format: Conversation ID–Turn)**
	129-30 (assistant): *"And I understand that is not the easiest in these times but I believe you can do it!"*
Social Connection Rebuilding	41-11 (assistant): *"That is good! It seems like calling on the phone can feel more genuine. Do you like playing Among Us? It might be fun to teach them how to play a game that allows you to play from far away."* 303-31 (user): *"That is probably true, but everyone has been so busy that I've only really been communicating with my husband."* 401-19 (assistant): *"i think it may be beneficial to give your friends some time, before attempting to speak with them again. maybe you can spend time with your family while you are waiting for them to cool down."*
Self-Advocacy and Boundary Setting	89-33 (user): *"I was afraid she was going to ruin my family with her attitudes."* 131-23 (user): *"I even got an emotiona support dog"* 131-25 (user): *"Yea, he's my best friend. At least I have one boy who has to stick around. He's on a tight leash"*

B Dialogue Excerpts by Semantic Category Associated with Decreases in EIS

Table 4: Illustrative dialogue excerpts from the ESConv dataset, grouped by semantic category and associated with decreases in EIS.

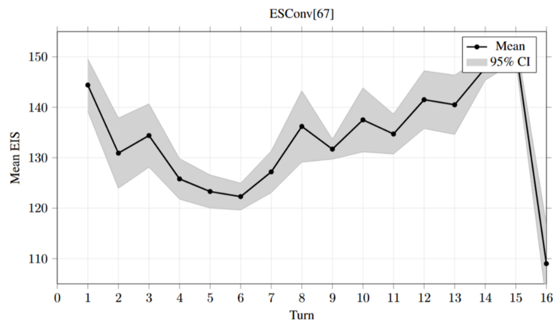
Category	Example with turn ID (Format: Conversation ID–Turn)
Cynicism and Misanthropy	132-4 (user): *"well, i'm not disgusted with myself... it's just people in general. everybody.. they're so selfish"*
	132-14 (user): *"I think it would be beating a dead horse. Ober half the population is overweight. People would just need to be blind not to see it. That's kind of my point. Everything is consume... consume.. consume.... buy buy buy"*
	132-16 (user): *"well... you can't really do anything without money. it all kind of rides on it, doesn't it? what you can buy?"*
	132-2 (user): *"feeling disgust as usual. Yourself?"*
	120-4 (user): *"not doing too hot tbh"*
	50-21 (user): *"We don't know what is going to happen if it comes"*
Relationship Loss and Romantic Devastation	120-6 (user): *"my girlfriend broke up with me"*
	120-18 (user): *"but she was the one"*
	51-2 (user): *"I am doing ok. I just broke up with my girlfriend and sad about it"*
	51-6 (user): *"I am so sad and just wonder why did it happen!!!"*
	51-8 (user): *"We had simple disagreement and both of us were keep fighting.. now I can not get over it."*
	303-9 (user): *"I'd like more help and understanding from my husband, but he seems to be incapable of that."*
Abandonment and Exclusion Themes	67-2 (user): *"I am really very angry with my friends for not inviting me"*
	67-4 (user): *"I didn't did any anything wrong to my friends but they are simply saying they forget me"*
	120-14 (user): *"I was supposed to introduce her, now I just look like a loser"*
	401-3 (user): *"I am today very sad because my friends fighting with me"*
	401-7 (user): *"Yes i am feeling alone"*
Anxiousness and Being Overwhelmed	457-3 (user): *"Well im feeling awful and my heart is racing , im feeling anxious for no reason."*
	457-7 (user): *"Ive tried meditation but cant seem to calm down. Exercise help for a bit but then my anxiety comes back."*
	379-2 (user): *"Hello, I'm not sure if there is any help? Without knowing when I can return to work, I will probably remain anxious about the unknown"*

Continued on next page

Category	Example with turn ID (Format: Conversation ID–Turn)
Reminiscing Traumatic Events	129-17 (user): *"It is making me have flashbacks of other traumatic situations,"* 129-37 (user): *"Childhood traumas are tough for sure. I am a fearful person."* 131-18 (user): *"Well... one told me that I should be put down like a dog to my face."* 89-14 (user): *"Only once and she was just telling me that I was a horrible person."*

C Additional CI graphs

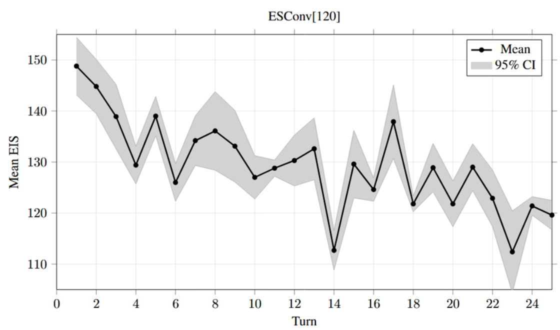
These plots visualize turn-by-turn EIS trajectories for selected **Emotional Support Conversations** (ESConv) that were identified as exhibiting notable rises or drops in Section 4.3 and Appendices A–B. Each plot is labeled with its conversation ID and includes a short caption highlighting specific turns referenced in the text. Gray shading denotes the 95% confidence interval of the mean EIS score.



Identified points – ESConv 67:

1. Turn 2 – Decreased: Abandonment and Exclusion Themes

2. Turn 4 – Decreased: Abandonment and Exclusion Themes



Identified points – ESConv 120:

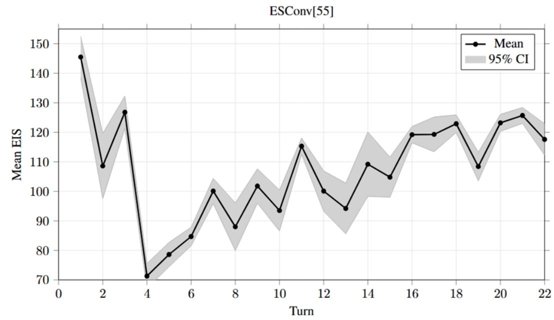
1. Turn 4 – Decreased: Cynicism and Misanthropy

2. Turn 6 – Decreased: Relationship Loss and Romantic Devastation

3. Turn 14 – Decreased: Abandonment and Exclusion Themes

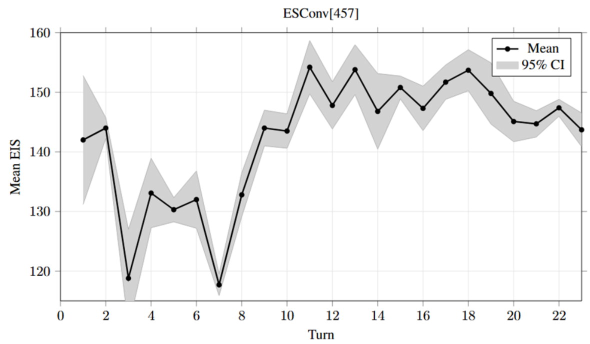
4. Turn 17 – Increased: Hope and Future Orientation

5. Turn 18 – Decreased: Relationship Loss and Romantic Devastation



Identified points – ESConv 55:

1. Turn 20 – Increased: Adaptive Coping Strategy Discussion

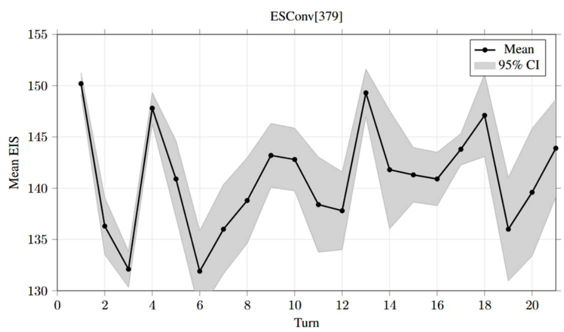


Identified points – ESConv 457:

1. Turn 4 – Increased: Peer Validation and Shared Experience

2. Turn 8 – Increased: Hope and Future Orientation

3. Turn 9 – Increased: Peer Validation and Shared Experience



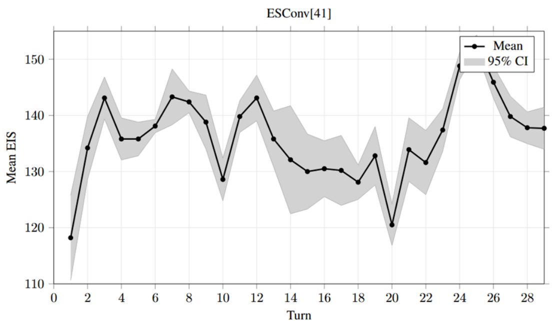
Identified points – ESConv 379:

1. Turn 2 – Decreased: Anxiousness and Being Overwhelmed

2. Turn 4 – Increased: Solution-Oriented Dialogue

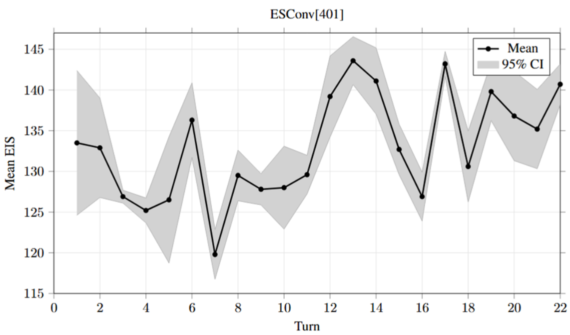
3. Turn 9 – Increased: Adaptive Coping Strategy Discussion

4. Turn 13 – Increased: Solution-Oriented Dialogue



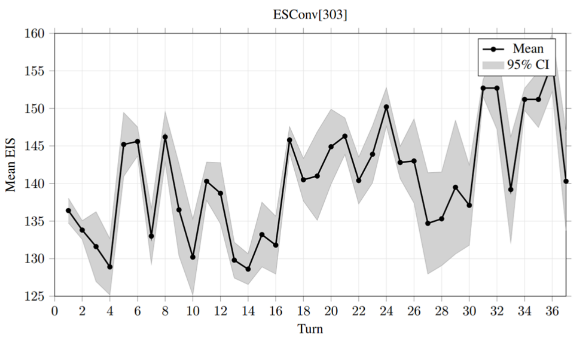
Identified points – ESConv 41:

1. Turn 11 – Increased: Social Connection Rebuilding



Identified points – ESConv 401:

1. Turn 3 – Decreased: Abandonment and Exclusion Themes
2. Turn 7 – Decreased: Abandonment and Exclusion Themes
3. Turn 19 – Increased: Social Connection Rebuilding



Identified points – ESConv 303:

1. Turn 9 – Decreased: Relationship Loss and Romantic Devastation
2. Turn 11 – Increased: Adaptive Coping Strategy Discussion
3. Turn 31 – Increased: Social Connection Rebuilding
4. Turn 34 – Increased: Gratitude and Appreciation Expression

D Sample Emotional Support Conversation and Customer Service Dialogues

Speaker	Utterance
Assistant	Hello, thank you for contacting BrightCom Support. How can I help you today?
User	Hey, my internet keeps dropping every few minutes. It's been really frustrating.
Assistant	I'm sorry to hear that. Can you tell me when you first started noticing this issue?
User	About three days ago. I thought it was temporary, but it's just getting worse.
Assistant	Have you made any changes to your setup recently, like moving the router or adding new devices?
User	No, everything's the same. I just work from home, so I notice the interruptions more.
Assistant	Understood. Let's try a quick reset. Could you unplug your modem and router for about 30 seconds, then plug them back in?
User	Sure, one sec... Okay, I just plugged them back in.
Assistant	Great. While the system reboots, can you tell me if the outages happen at specific times of day, or is it random?
User	Seems random, but it happens most often in the evenings when I'm on calls.
Assistant	That could indicate high usage in your area. I'll check our network status. Please hold on a moment.
User	Alright.
Assistant	Thank you for waiting. I do see some congestion reported in your neighborhood. We're working to improve capacity.
User	So, is there anything I can do on my end?
Assistant	Yes, besides the reboot, try placing your router in an open area away from walls. Also, connecting by Ethernet for calls will help with stability.
User	I can do that. I'll move it closer to my desk. Will this permanently fix the issue though?
Assistant	It should reduce the drops, but once our upgrade is complete, the service will be much more reliable overall.
User	Okay, thanks. Any idea when that upgrade will finish?
Assistant	Yes, the projected completion date is within the next two weeks.
User	Alright, that helps. At least I know it's not just my equipment.
Assistant	Exactly, and I've documented this on your account so if you call back, our team will know the steps you've already taken.
User	Cool, thanks for doing that.
Assistant	No problem. I'll also credit your account for a week of service since you've been dealing with interruptions.
User	Oh wow, I appreciate that.
Assistant	You're welcome. Is there anything else I can help you with today?
User	No, that should be it. Thanks again.
Assistant	Glad I could help. Thank you for choosing BrightCom, and have a great day!
User	You too!
Assistant	Goodbye!

Table 5: Customer Service control conversation (synthetic): full dialogue in `User/Assistant` format.

Speaker	Utterance
User	Good evening.
Assistant	Hi, how's things?
User	Things are ok, I am just having issues with the problem above.
Assistant	I am afraid that I am not able to see that. Can you tell me about it?
User	Yes, so I have always done very well in school and university. I have always had A's or high B's. Last semester in university, I failed one of my exams. I still passed the class with an A, but for some reason, that exam haunts me and gives me bad dreams.
User	I don't know why it happens and it worries me.
Assistant	I can understand that, when I was sitting my final exams I constantly had nightmares about failing. Nearly thirty years on and I still do!!
Assistant	But you did really very well and you passed with an A. Do you think maybe you are being a little too hard on yourself?
User	Maybe. I got a 46% on the exam (so an F) but I did all my other assignments and exams almost perfect and got an A overall. I never once failed a test in my life, and have always held myself to a very high academic degree to make my parents proud.
User	I come from a poor family, so succeeding in life is very important to me.
User	me*.
Assistant	It sounds as though you feel that academic success is the most important thing in your life, would I be right in thinking this?
User	Yes you are correct.
Assistant	Have you considered talking to anyone about these feelings of perfectionism? A therapist perhaps?
Assistant	There is a good deal of useful help and information out there for people who are struggling.
User	I have not, actually. I don't think my family can afford a therapist, especially with the pandemic raging right now.
User	Where do you recommend going?
Assistant	Does your school have any counsellors offering help for free? Many do.
Assistant	I believe that there are some charities that will offer a certain number of free therapy sessions too.
User	I don't think any are available since my school is online only, also the campus is completely closed due to Winter Break.
User	Oh? Charities?
Assistant	Yes I think so, though I am not totally sure. There should be someone at your school, online or not, who can advise you.
User	I have never heard of such people, I am interested.
Assistant	Really though I think that your problem is self esteem. You should think better of yourself :)
Assistant	I can tell that you set yourself a very high standard but I also think that you need to be kind to yourself.
User	I think you're right, but I don't know, I'm still scared about having the bad dreams. I often wake up 2–3 times at night because of them.
Assistant	I can understand that, this has been happening to me all of my life. Have you tried to take anything to help?
Assistant	I can recommend a hot milky drink before bed and perhaps a hot water bottle. Anyway I hope that I have been able to be of some assistance to you!
Assistant	Have a lovely holiday season.
User	Thank you, I'll try to do just that.
User	Merry Christmas to you.
Assistant	And you :) remember to hit the quit button and take the survey ;)

Table 6: ESConv conversation [80]: full dialogue in `User/Assistant` format.