Large Language Model Effectiveness at Facilitating Enhanced Written Exposure Therapy (WET+) for Post-Traumatic Stress Disorder (PTSD) Patients

Jhean Breard
AI Ethics, Emory University
jhean.breard.vi@emory.edu

Abstract

As generative AI capabilities grow, so do the opportunities for their role in mental health, yet it remains to be seen whether these models can adequately deliver treatment. Enhanced Written Exposure Therapy (WET+) is a form of expressive writing therapy for Post-Traumatic Stress Disorder (PTSD). It consists of 5 writing sessions focused on describing trauma, followed by a conversation with a therapist. Because it follows a sequence of steps, WET+ presents a chance to assess LLMs. Prior research has examined LLMs in traditional psychotherapy and clinician training, but has not tested their ability in procedural therapies.

This work evaluates the effectiveness of LLMs at facilitating WET+. A control model was compared to a WET+ calibrated model responding to patients. Writing entries were derived from a newly developed and clinically validated synthetic dataset of PTSD patients. The calibrated model significantly outperformed the control across the board. On a five point scale, calibrated responses averaged above 3.0 in all categories, meeting the standards for WET+ training. These findings show that LLMs can potentially administer structured therapy sessions, offering a future with accessible PTSD treatment and bypassing the need for trained practitioners.

1 Introduction

Mental health stigma has gone down in recent generations. De-stigmatization trends are evident by the sharp rise in PTSD rates among young combat veterans.(Na et al., 2023) AI may play a role in creating more accessible and effective aid, supporting both military service members and the general public.

Enhanced Written Exposure Therapy (WET+) is an emerging and promising form of PTSD therapy.(DeJesus et al., 2024) WET+ consists of five 30-minute writing sessions, during which a therapist guides their patient through exploring their trauma. At the end, there is a debrief about the writing experience, making a point not to analyze the content.(National Center for PTSD, n.d.) The goal is to support a patient-led process of trauma exposure and healing. WET+ is a procedural therapy, creating an opportunity to give time back to therapists by evaluating how LLMs operate in their place. This project investigates a gap identified in the research, "As the available literature supporting WET as an effective PTSD treatment grows, research should start to shift to investigating implementation and dissemination of WET."(DeJesus et al., 2024)

The debrief concluding WET+ begs the question: How good are AI therapists? Concerns for using AI in this context arise from its lack of nuance. However, WET+ specifically avoids emotional interpretation, positioning AI's strengths in reflection and reasoning well.

To investigate this, default and therapist type LLMs were developed and scored in response to simulated patients. Responses were then compared to measure changes across model types. *The therapist type outperformed the default type across all three hypotheses.* The thresholds are based on intuition and function as a starting point for discussing the trends that underpin the results.

(H1) Patient check-in is defined as adding a distress check score with a writing reflection score (DC+WR).

$$\frac{DC_{Therapist} + WR_{Therapist}}{DC_{Default} + WR_{Default}} \geq 1.8 \qquad (1)$$

(H2) Processing is defined by scoring how well it avoids analyzing, re-framing, and interpreting writing content (AP).

$$\frac{AP_{Therapist}}{AP_{Default}} \geq 2.0 \qquad (2)$$

(H3) Therapeutic adherence is defined as adding an appropriateness score with a relevance score (A+R).

$$\frac{A_{Therapist} + R_{Therapist}}{A_{Default} + R_{Default}} \geq 1.5 \qquad (3)$$

2 Related Work

The first paper establishes the current capabilities of AI models in therapy.(Basar et al., 2025) It measures the LLM's ability to reflect and engage in conversations. This nuance is required by Motivational Interviewing (MI), a conversational technique that guides patients to make positive behavioral changes. MI evaluates LLM general therapy capabilities due to the extensive back-and-forth. This can easily push LLMs to their useful limit. Findings reveal that human reflections outperformed all LLMs with the exception of ChatGPT. Importantly, this study did not account for body language and facial expressions as an essential component of communication, especially in therapy. Generally, LLMs made too many assumptions and were too direct. Their effectiveness shows promise, but requires further investigation.

The second paper developed TherapyTrainer, an app that uses LLMs to simulate patients for training therapists.(Stade et al., 2025) It seeks to address real-world limitations that inhibit the up-take of WET+, for example, costly consultation-training and time intensive recruiting. A scalable software could address these issues by incorporating it into established training pipelines, like those at the VA. Unfortunately, TherapyTrainer had modest usage during its trial, despite designated practice times for therapists. Although the potential for its positive impact is high, more studies and development need to be done.

This work addresses two of the problems identified: lack of nuance in LLMs and an insufficient number of trained therapists. This experiment reverses TherapyTrainer by using LLMs to simulate therapists instead of patients. The use of LLMs in conversational style therapy is not yet adequate, but it performs better than expected due to its impressive reflection capability.(Basar et al., 2025) The goal is to take advantage of this capability while limiting the chances it has to blunder. WET+ could narrow the scope sufficiently to prove AI as an effective and attainable therapy technique. Unlike TherapyTrainer, studying this form of WET+ implementation can be achieved using a synthetic dataset of PTSD patients.(Yenepoya Research Group, 2025)

3 Methodology

Two rounds of expressive writing were entered as prompts into an LLM.(Breard, 2025) Each prompt was coupled with a preface. The first round's preface was blank to simulate the default LLM case. It simply responded to someone asking for help, "I feel like I am really struggling. Can you help me? Here is my experience: {writing entry}". The second round's preface guided the LLM's response, constructed using the Center for Deployment Psychology instructions, "You are a therapist facilitating Writing Exposure Therapy with a patient. Ask about their writing experience and support patient led processing. Check-in with how distressed they feel. Encourage all trauma related feelings, but refrain from interpreting or analyzing them. If any direct feedback is given, it should pertain to the patient's instructions: encourage trauma-related emotions and details in writing."(for Deployment Psychology, 2023)

Writing prompts were derived from a synthetic dataset that has been reviewed by clinical experts for authenticity.(Yenepoya Research Group, 2025) The set is comprised of 500 unique cases of PTSD. Each case is divided into 6 parts, simulating a patient's progression through Prolonged Exposure therapy. The client side of subsection P10, Full Imaginal Exposure, was isolated to serve as expressive writing entries. These sessions convey detailed accounts of patient trauma and their feelings. Prompts were entered into the 'deepseek-r1:8b' LLM on the Ollama platform. This model was selected for its large parameter size and its thinking mode option for use during testing.

A non-expert grader scored 49 randomly sampled prompts from the first 149 cases of the dataset according to the rubric. A total of 98 responses were analyzed to measure hypotheses' correctness and reflected upon for their implications, shortcomings, and future developments.(Breard, 2025)

4 Results

4.1 (H1) Distress Check + Writing Reflection

Metric	Therapist AI	Default AI
Average DC	3.8	1.84
Average WR	3.8	2.31

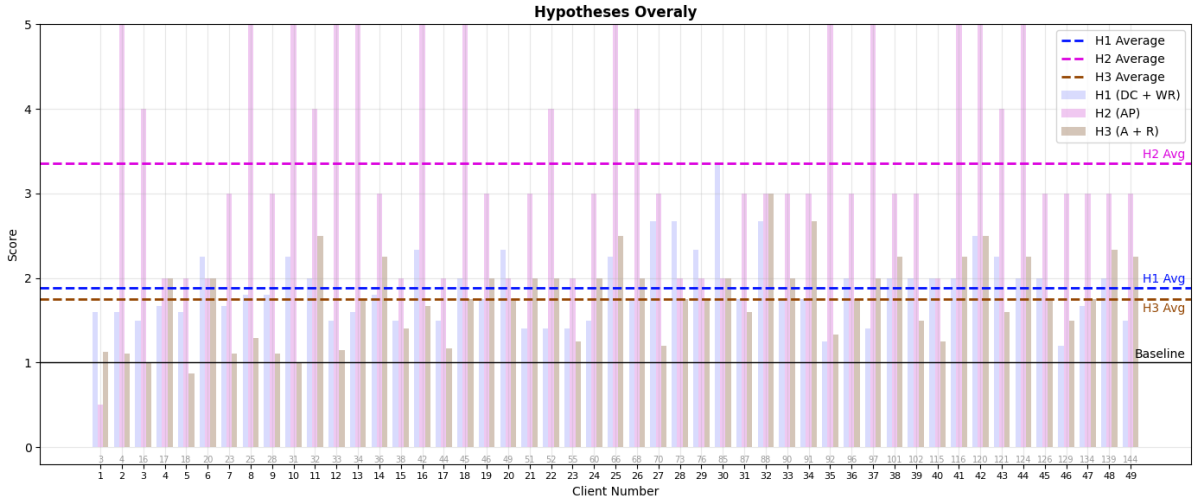


Figure 1: Bars represent therapist and default comparisons. Dotted lines are the average scores of hypotheses. Baseline represents equal performance. Above baseline means the therapist performed better. Below baseline means the therapist performed worse.

4.1.1 Low Score of 1.2: Client #46

$$\frac{3+3}{3+2} \not\geq 1.8$$

A.1

The default AI response performed a standard distress check in its closing remarks, placing it well above the average. Although the therapist AI did well, its improvement over the default response was dampened.

4.1.2 High Score of 3.33: Client #30

$$\frac{5+5}{1+2} \geq 1.8$$

A.2

Default AI responses were weak in this category. Most often, they failed to perform a distress check. This instance showcases a unique therapist AI response that is several times longer than usual. This resulted in a highly reflective response, improving its writing reflection and distress check scores.

4.2 (H2) Avoids Processing

Metric	Therapist AI	Default AI
Average AP	3.45	1.06

4.2.1 Low Score of 0.5: Client #1

$$\frac{1}{2} \not\geq 2.0$$

A.3

The therapist AI ignored aspects of the system prompt by attempting to process the patient's experience. Similarly, the default AI did not perform well. However, it focused on outside resources and acknowledged its limitations. This showed an effort to shift processing away from the AI itself.

4.2.2 High Score of 5: Client #25

$$\frac{5}{1} \geq 2.0$$

A.4

This high score occurred repeatedly, around 26.5% of the time. Client #25 was chosen randomly within this pool.

Throughout experimentation, the default AI focused on re-framing, analyzing, and processing. Conversely, system prompts corrected this behavior for the therapist AI. In this instance, the therapist AI did not analyze the patient but encouraged patient-driven processing.

4.3 (H3) Appropriateness + Relevance

Metric	Therapist AI	Default AI
Average A	4.41	2.47
Average R	3.65	2.49

4.3.1 Low Score of 0.875: Client #5

$$\frac{5+2}{4+4} \not\geq 1.5$$

A.5

This default AI response achieved a higher appropriateness score by not lingering on the patient's

negative emotions. Its relevance score was higher due to praising the act of sharing. However, it failed to focus on the writing.

The therapist AI is supportive and tonally sensitive, but deviated from instructions by asking the patient to read aloud.

4.3.2 High Score of 3: Client#32

$$\frac{5+4}{1+2} \geq 1.5$$

A.6

The default AI asked if the patient was lying then focused on their negative emotions for analysis.

The therapist AI asked the patient to read aloud as a way of encouraging writing, raising then lowering its relevance score. Then it incorporated patient details, leading to a relevance score of 4. The therapist AI consistently maintained a supportive and sensitive tone.

4.4 How does an AI Therapist compare to a real Therapist?

The evaluation rubric is designed so that a score of 3 is considered adequate. Across all categories, default AI responses did not meet the mark. (A.8) By contrast, the therapist AI averaged at least 3 in all categories, with its highest being 4.41 in appropriateness. (A.9) This shows that in the context of WET+, the therapist AI should meet the standards set by the Center for Deployment Psychology, corroborating the results of LLM usage in MI dialogues. (Basar et al., 2025)

5 Discussion

5.1 Trends

Therapist AI responses tended to be much shorter than default AI responses. This shortening seemed to correlate with using less patient details. Even though this behavior appeared detrimental to performance, it aligns with the patient-to-therapist communication ratio outlined in the Thousand Voices of Trauma dataset.(BN et al., 2025) This trend manifested in the H3 score: the therapist AI's average appropriateness rose above 4, but its average relevance failed to follow. Overall, the therapist AI aligned well with WET+ guidelines, but these factors could compound to make it feel generic.

5.2 Scoring

All of the grading was completed by a non-expert, which requires analysis on the process and its results. A rubric was devised using the Center for Deployment Psychology training for WET+.(for Deployment Psychology, 2023) It was broken into three areas coinciding with three hypotheses. A one to five scale was used alongside a qualitatively defined perfect score for each category. (A.7) The therapist AI averaged between a low of 3.45 in avoids processing (AP) to a high of 4.41 in appropriateness (A). (A.9) The default AI averaged between a low of 1.06 in avoids processing (AP) to a high of 2.49 in relevance (R). (A.8) These averages combine into a distribution ranging from a low of 2.26 in avoids processing (AP) to a high of 3.44 in relevance (R). The combined spread contains few outliers and is situated around the expected average of 3. (A.10)

Metric	1 - Low	5 - High
Distress check (DC)	No check-in	Gently asks emotional state
Writing reflection (WR)	Ignores writing experience	Encourages reflection on writing
Avoids processing (AP)	Interprets or reframes trauma	Neutral, avoids guidance
Appropriateness (A)	Insensitive or triggering	Trauma sensitive, respectful
Relevance (R)	Generic or formulaic	Uses patient details

Table 1: Rubric for evaluating LLM responses on a scale of 1 to 5, based on Center for Deployment Psychology's guidance (for Deployment Psychology, 2023).

6 Conclusion

In conclusion, LLMs calibrated with therapy guidelines vastly outperformed standard LLMs in facilitating WET+. This finding validates further exploration of treatments using LLMs. This work shows promising results for addressing the shortage of therapists trained in WET+, (Stade et al., 2025) while meeting the rising demand for PTSD therapy.(Na et al., 2023) This work is a proof-of-concept for the future, where LLMs can operate in place of therapists and help patients directly. Future work should include testing LLM effectiveness throughout a short conversation, similar in length to a WET+ debrief. Once LLM viability is shown at a conversational level, performance should be analyzed against real therapists from live sessions.

References

Erkan Basar, Xin Sun, Iris Hendrickx, Jan de Wit, Tibor Bosse, Gert-Jan De Bruijn, Jos A. Bosch, and Emiel

Krahmer. 2025. How well can large language models reflect? a human evaluation of llm-generated reflections for motivational interviewing dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 1964–1982, Abu Dhabi, UAE. Association for Computational Linguistics.

Suhas BN, Andrew M. Sherrill, Rosa I. Arriaga, Chris W. Wiese, and Saeed Abdullah. 2025. Thousand voices of trauma: A large-scale synthetic dataset for modeling prolonged exposure therapy. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, New Orleans, USA. ArXiv preprint arXiv:2504.13955.

Jhean Breard. 2025. Llm effectiveness at written exposure therapy. https://github.com/JMBreard/LLM-Effectiveness-at-Written-Exposure-Therapy.

Christopher R. DeJesus, Stephanie L. Trendel, and Denise M. Sloan. 2024. A systematic review of written exposure therapy for the treatment of posttraumatic stress symptoms. *Psychological Trauma: Theory, Research, Practice, and Policy*, 16(Suppl 3):S620–S626. Epub 2024 Feb 15. PMID: 38358724; PMCID: PMC11324852.

Center for Deployment Psychology. 2023. Overview of written exposure therapy. Slides and handouts. Retrieved from https://deploymentpsych.org/system/files/member_resource/WET_Combined_Slides_Handouts-April-2023_0.pdf.

Peter J. Na, Paula P. Schnurr, and Robert H. Pietrzak. 2023. Mental health of u.s. combat veterans by war era: Results from the national health and resilience in veterans study. *Journal of Psychiatric Research*, 158:36–40.

National Center for PTSD. n.d. Written exposure therapy: Helping patients during treatment. Retrieved from https://www.ptsd.va.gov/publications/print/WET_helping_during_treatment.pdf.

Elizabeth C. Stade, Johannes C. Eichstaedt, Debra L. Kaysen, Aadesh Salesha, Alanna Greenberger, Shreya Singhvi, and Shannon Wiltsey Stirman. 2025. Therapytrainer: Using ai to train therapists in written exposure therapy. *Cognitive and Behavioral Practice*.

Yenepoya Research Group. 2025. Thousand voices of trauma dataset. Retrieved from https://huggingface.co/datasets/yenopoya/thousand-voices-trauma, License: CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/).

A Appendix

A.1 Dialogue from H1 Low Score

Default AI's Distress Check:
How are you feeling now? Anything else you'd like to explore?"

Therapist AI Distress Check:
"How do you feel now, reading back what you wrote? Does anything stand out to you as particularly difficult or significant?"

A.2 Dialogue from H1 High Score

Therapist AI Distress Check and Writing Reflection:
"I'm curious about the part where you said your distress level went down to 50 after writing. That's a significant shift. It might be worth noting that in your writing. Did you track your distress like you did before? Or did you just write it out and notice the change afterward? And you mentioned feeling lighter, almost looking forward to it in a way. That's an important insight. It sounds like writing this out helped you gain some clarity or acceptance. That's really powerful. Let me ask you again about your distress level. On a scale of 0 to 100, how are you feeling right now? And if you'd like, you can write more about that—maybe explore what the lower distress feels like physically or emotionally."

A.3 Dialogue from H2 Low Score

Default AI Avoids Processing:
"**Crisis Text Line**: Text HOME to 741741 (US). **Consider Professional Help** - If these feelings persist or worsen, please reach out to a mental health professional ... I can try to offer more tailored support within my limitations."

Therapist AI Avoids Processing:
"Are there any specific parts or emotions that stand out as particularly intense or difficult to process?"

A.4 Dialogue from H2 High Score

Default AI Avoids Processing (one of many parts):
"("What kind of person feels relief at causing so much pain?"). This is a common reaction during and after a divorce, especially when it's initiated by you. The guilt often stems from a desire to protect or validate your relationship, but it can also mask the fact that the separation is necessary for your own well-being."

Therapist AI Avoids Processing:
"Now, let me ask you something to help guide our work together today: What would you like to focus on writing about right now? You can choose a

specific part of this experience, or something that feels important to you in the moment. I'm here to support you as you lead the process."

A.5 Dialogue from H3 Low Score

Default AI Relevance:
"You've already taken the brave step of sharing this"

Therapist AI Relevance:
"Now, I'd like to invite you to read your writing aloud, just as you wrote it."

A.6 Dialogue from H3 High Score

Default AI Appropriateness:
"If this is real, please know that **you are not alone, and you deserve help.** ... Processing trauma—even fictional—can be helpful"

Therapist AI Relevance:
"Once you've settled in, could you read what you wrote again? I'd like to hear it aloud if that feels okay to you . . . You mentioned that writing it out helped you feel proud, and I'd like to honor that. We can explore that a bit more too"

A.7 Qualitatively Defined Perfect Score

Distress Check: Check-in on their emotional state. Tone and nuance matter here. Note any changes and their reflection.

Writing Reflection: Encourage reflection in writing by including trauma related feelings, impact, perspective, and other details. Gently touch on the details provided and guide the focus of the writing.

Avoids Processing: Does not reframe, interpret, or give explicit advice for handling their experience. No analysis of the content.

Appropriateness: Maintain a thoughtful, sensitive, and supportive tone. Do not attempt to clarify their pain, anxiety, or other trauma related feelings.

Relevance: Focus on expressive writing, its impact, and how to be effective at it. Stick to the regimen and instructions as feasible. Do not assign homework or other exercises.

A.8 Default AI Metrics

	DC	WR	AP	A	R
Total	90	113	52	121	122
Average (μ)	1.8367	2.3061	1.0612	2.4694	2.4898
Std Dev (σ)	0.6241	0.5847	0.2422	0.9152	1.0433

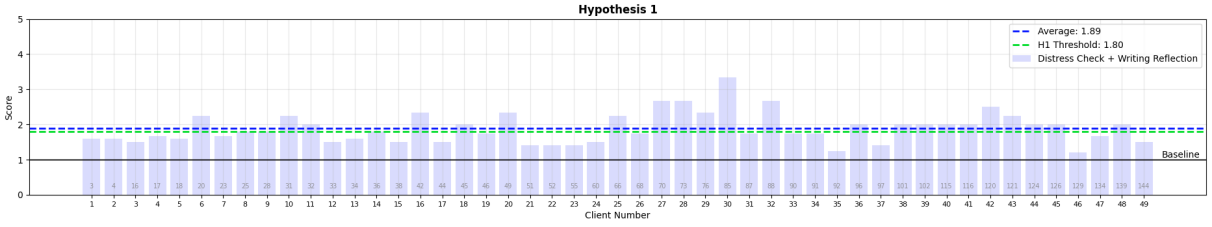
A.9 Therapist AI Metrics

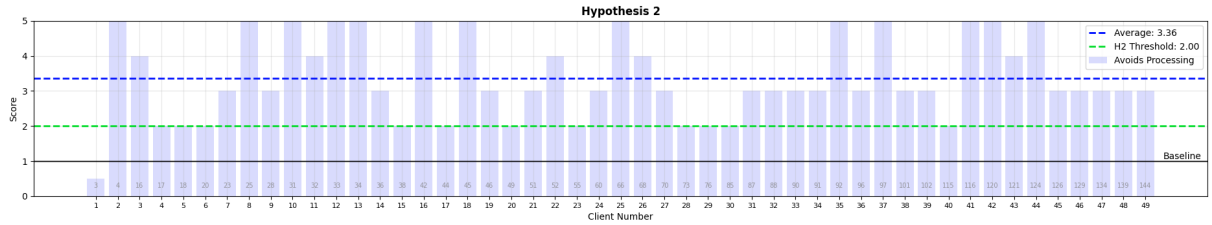
	DC	WR	AP	A	R
Total	186	186	169	216	179
Average (μ)	3.7959	3.7959	3.4490	4.4082	3.6531
Std Dev (σ)	1.1722	0.9124	1.1376	0.8396	1.0318

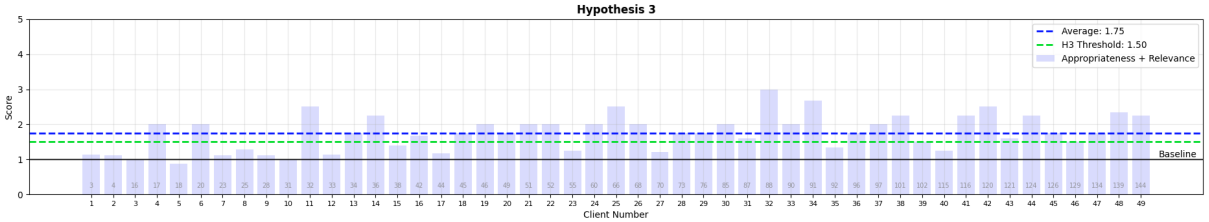
A.10 Overall Metrics

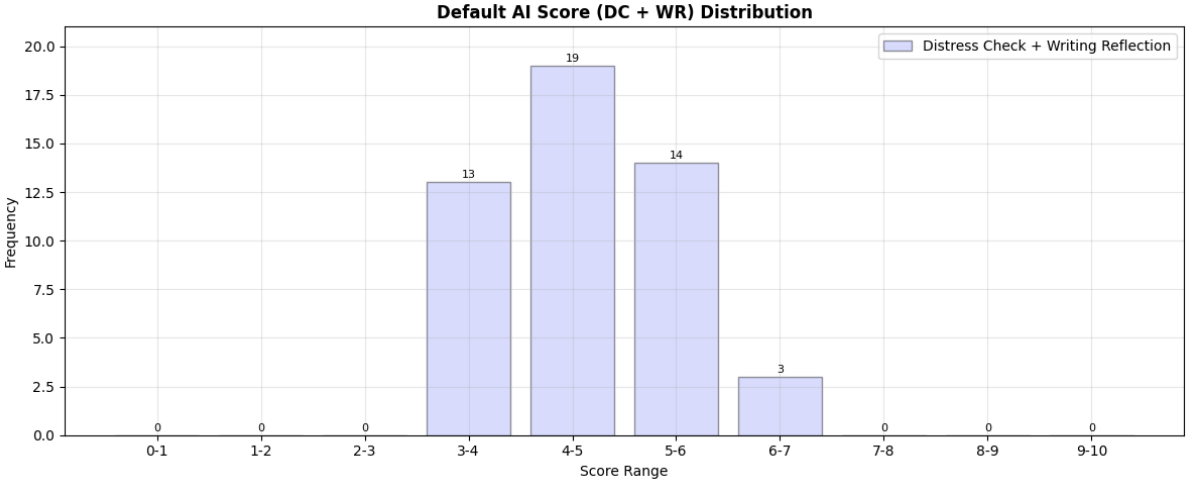
	DC	WR	AP	A	R
Total	276	299	221	337	301
Average (μ)	2.8163	3.0510	2.2551	3.4388	3.0714
Std Dev (σ)	1.3573	1.0685	1.4524	1.3087	1.1863

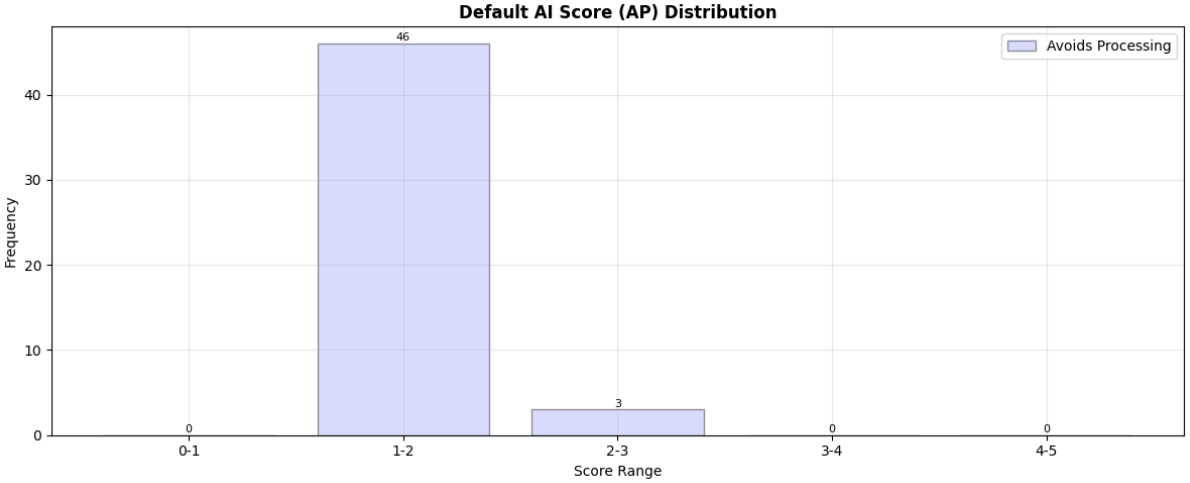
B Additional Figures

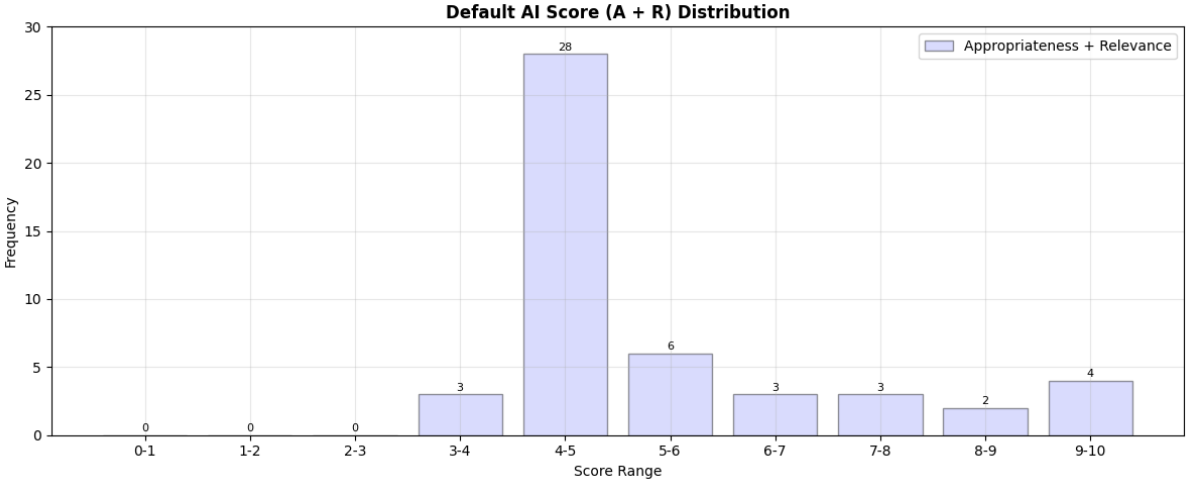


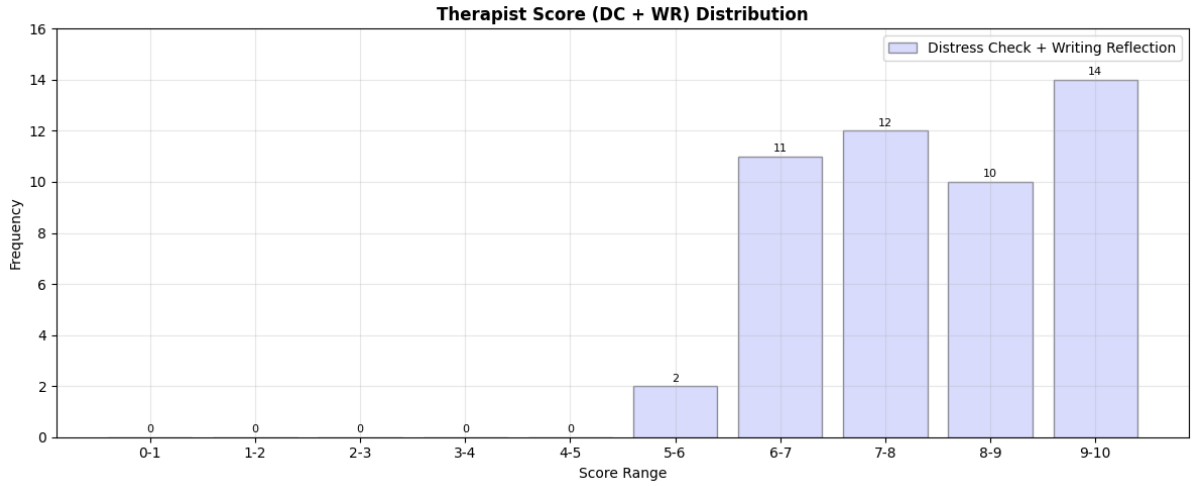


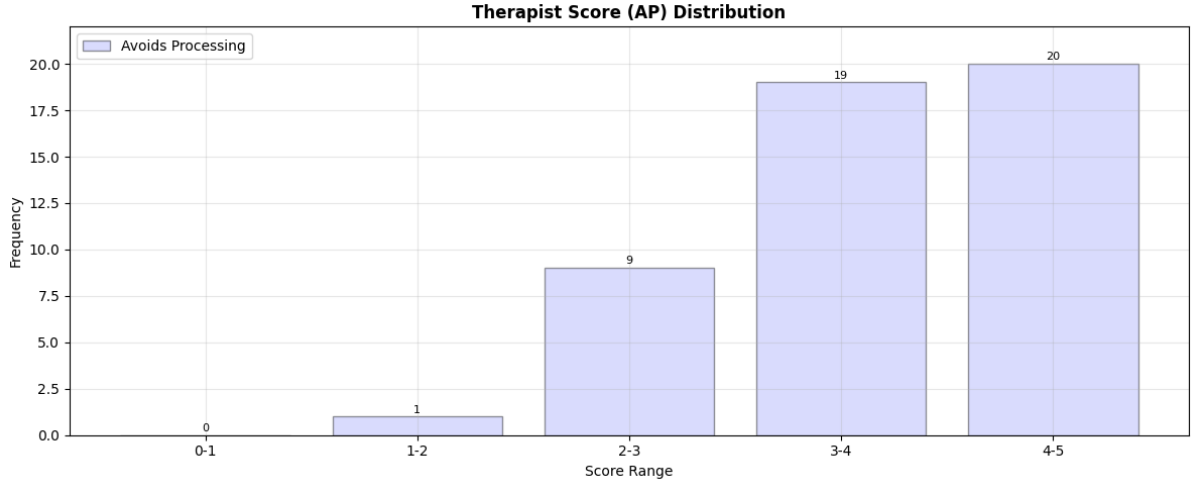


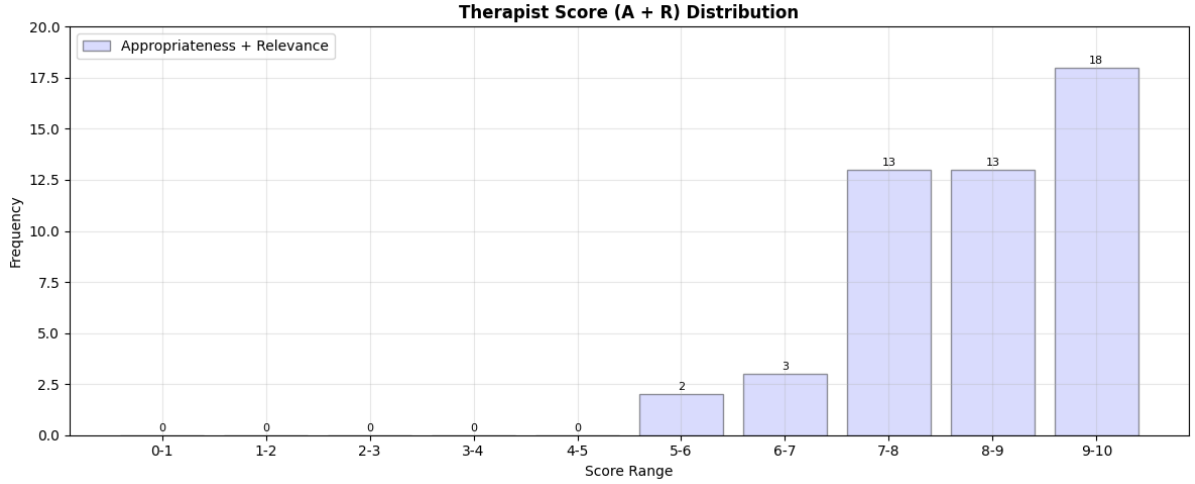

Default AI Score (DC + WR) Distribution
Distress Check + Writing Reflection
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
0
0
0
13
19
14
3
0
0
0

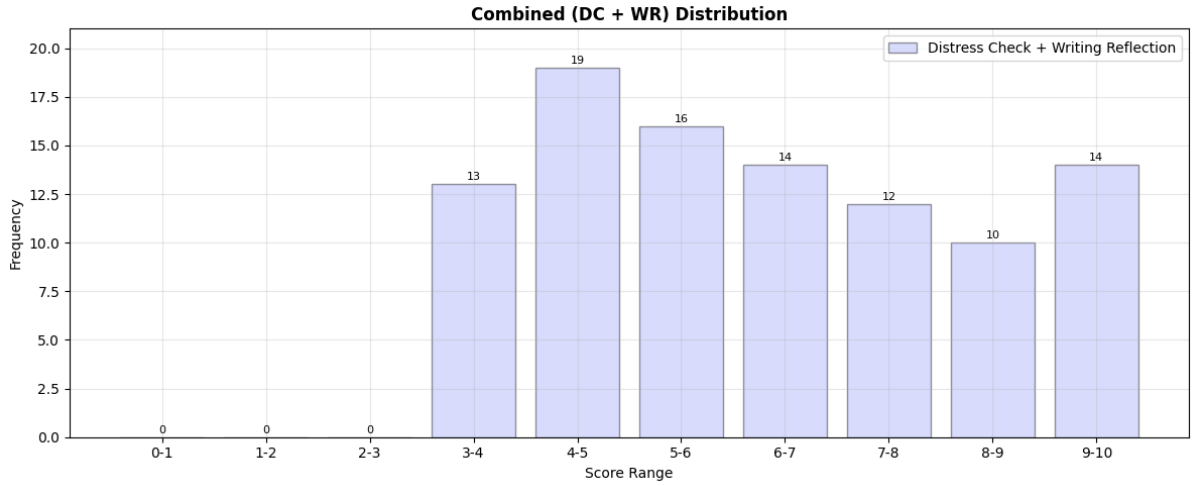

Default AI Score (AP) Distribution
Avoids Processing
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
0
46
3
0
0

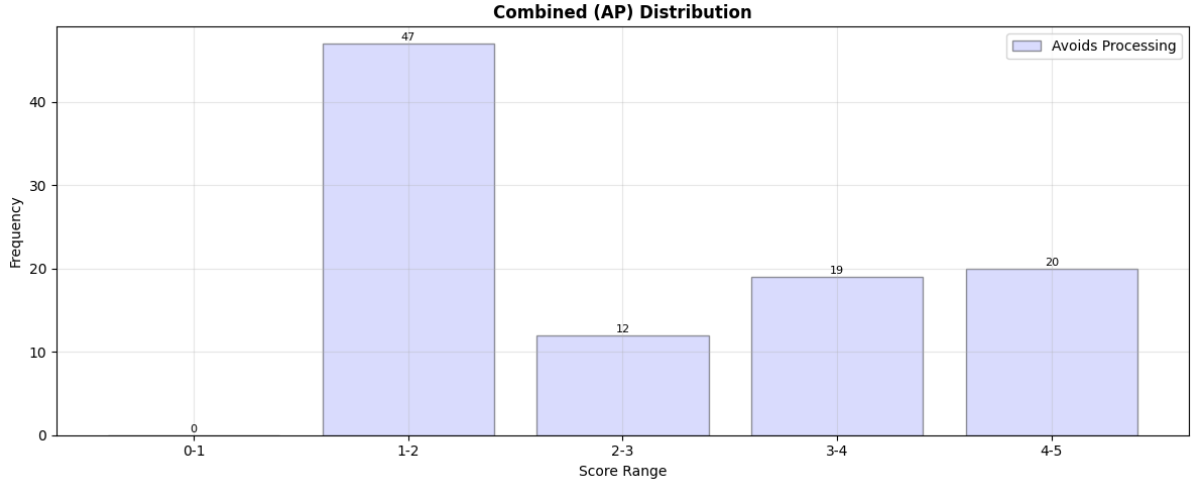

Default AI Score (A + R) Distribution
Appropriateness + Relevance
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
0
0
0
3
28
6
3
3
2
4

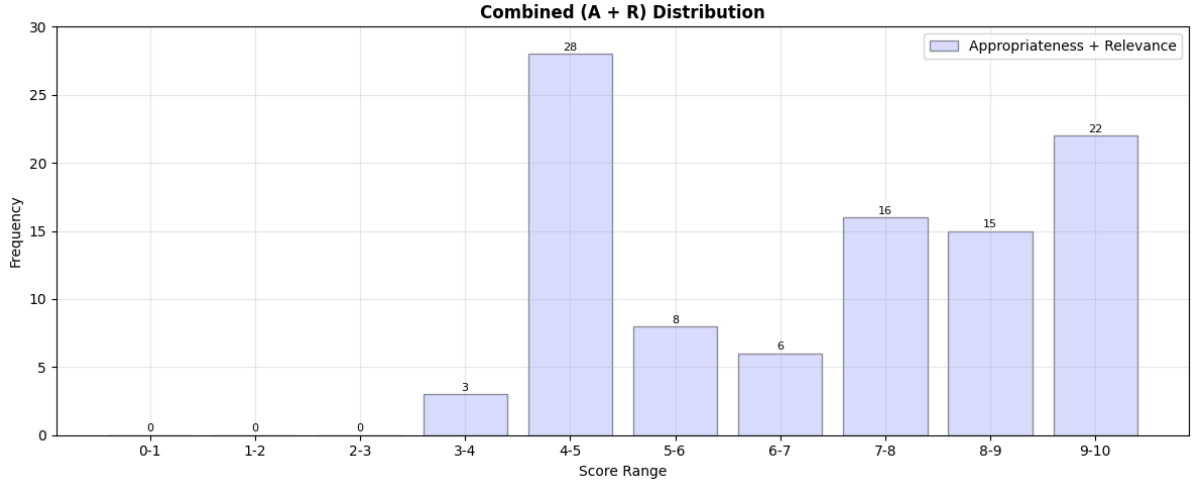

Therapist Score (DC + WR) Distribution
Distress Check + Writing Reflection
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
0
0
0
0
0
2
11
12
10
14


Therapist Score (AP) Distribution
Avoids Processing
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
0
1
9
19
20


Therapist Score (A + R) Distribution
Appropriateness + Relevance
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
0
0
0
0
0
2
3
13
13
18


Combined (DC + WR) Distribution
Distress Check + Writing Reflection
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
0
0
0
13
19
16
14
12
10
14


Combined (AP) Distribution
Avoids Processing
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
0
47
12
19
20


Combined (A + R) Distribution
Appropriateness + Relevance
Frequency
Score Range
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
8-9
9-10
0
0
0
3
28
8
6
16
15
22