Insight Box: A Multi-agent AI Assistant for Detecting and Addressing Cognitive Distortions

RAM PRIYADHARSHINI RAMACHANDRAN, Independent Researcher, USA

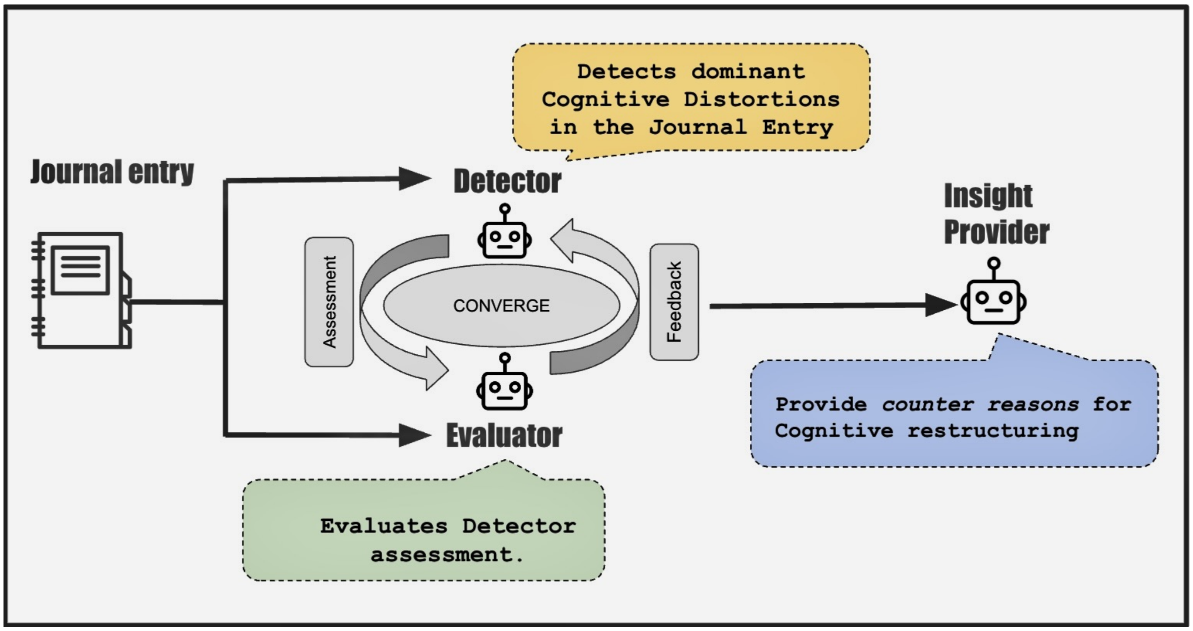
Journal entry
Detector
Detects dominant Cognitive Distortions in the Journal Entry
Assessment
CONVERGE
Feedback
Evaluator
Evaluates Detector assessment.
Insight Provider
Provide *counter reasons* for Cognitive restructuring

Fig. 1. Insight box: An LLM based multi-agent workflow for detecting and addressing Cognitive distortions through counter reasoning

AI-based interventions are well suited for the early detection and management of mental health conditions, serving as a valuable complement to traditional treatments. An early indicator of conditions like depression and anxiety, which predict the severity of its symptoms, is the persistent presence of cognitive distortions - maladaptive negative thought patterns. Previous research has demonstrated the potential of Large Language Models (LLMs) to detect cognitive distortions. This study attempts to develop a conversational multi-agent LLM-based AI system called Insight Box. In addition to the detection of cognitive distortions, Insight Box can also engage in interactions with users in an attempt to augment their reasoning to alleviate detected cognitive distortions. This study preliminarily investigates the ability of Insight Box to detect and address cognitive distortions in simulated journal entries by gently providing counterarguments to challenge negative assumptions and offers insights to help users navigate stressors.

CCS Concepts: • **Computing methodologies** → **Cognitive science**; **Artificial intelligence**.

Additional Key Words and Phrases: cognitive distortion, AI assistant, multi-agent system, augmented reasoning, Generative AI

ACM Reference Format:

Ram Priyadharshini Ramachandran. 2025. Insight Box: A Multi-agent AI Assistant for Detecting and Addressing Cognitive Distortions. 1, 1 (April 2025), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

---

This paper was presented at the 2025 ACM Workshop on Human-AI Interaction for Augmented Reasoning (AIREASONING-2025-01). This is the authors' version for arXiv.

Author's address: Ram Priyadharshini Ramachandran, Independent Researcher, Tallahasee, Florida, USA, rampriyadarshini@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

1 INTRODUCTION

There is a global need for cheap and novel mental health interventions that will facilitate early detection, continuous monitoring, and immediate assistance for mood disorders. Affordable and accessible AI-based interventions can meet this need by providing ancillary support to traditional assistance, such as the Cognitive behavioural therapy (CBT), a common psycho theraputic treatment for individuals affected with depression and anxiety. One of the main components of CBT is to help individuals identify cognitive distortions/errors in their thinking and restructure them to help them cope with their situations better. This study develops a Multi-agent AI system and explores and evaluates its potential in detecting and addressing distortions based on journal entries by providing insights that can help users cope with their daily struggles and alleviate persistent patterns of cognitive distortions. The cognitive behavioral therapy proposed by Aron Beck [1] states that anxiety and depression emerge from cognitive distortions, maladaptive negative patterns of thinking. A recent study [2] has shown that cognitive distortions predate and predict the severity of symptoms in depression. CBT aims to restructure these thoughts over time by recognizing and challenging negative assumptions to better deal with stressors. There are around 10-20 common types of distortions available in the literature and have been shown to predict severity of depression symptoms. For this study, 10 common types of distortions (Table 1) from the study Shreevastava and Foltz [3] are considered. Recent advances in natural language processing have created new opportunities for AI-based interventions in mental health care. Conversational AI assistants can help bridge gaps in accessibility and affordability within traditional care systems. Interacting with AI assistants may also reduce the fear of judgment, providing users with a safe space to express their thoughts. However AI systems ability to enhance human reasoning through augmented feedback, fostering personal insight and cognitive growth has not been explore much. To address this gap, the present study evaluates a LMM based Multi-agent system named ‘Insight box’ capability for cognitive distortion detection in Journal entries and assist the user through augmented reasoning. It detects the cognitive distortion and interacts with the user by providing counter-reasoning to nudge the person to restructure their distortive thought.

Table 1. Cognitive distortions and counter-reasons used for chain-of-thought prompting

#	Cognitive Distortion Type	Distortion	Counter Reasoning
1	All-or-Nothing Thinking	Viewing situations in absolute terms, as either all good or all bad.	Nudge to see middle grounds.
2	Overgeneralization	Drawing broad conclusions based on limited information.	Allow oneself the benefit of the doubt to remember the evidence to support or refute distress-causing assumption.
3	Mental Filter	Focusing solely on negative details and ignoring positive aspects.	Point out the positive or neutral aspects.
4	Should Statements	Using rigid ’should’ or ’must’ statements about yourself or others that cause distress.	Insist on being kind to oneself and embrace flexibility.
5	Labeling	Reducing yourself or others to a single negative characteristic.	A single action or mistake does not define a person’s life. Instead of labeling, focus on the behavior itself—behaviors can change.
6	Personalization	Blaming yourself for situations outside your control or assigning blame to others unjustly.	People’s actions and emotions are influenced by many factors beyond you.
7	Magnification	Exaggerating the importance of mistakes or negative events.	Reminding mistakes or failures are part of growth and rarely as catastrophic as they feel.
8	Emotional Reasoning	Believing that because you feel a certain way, it must be true.	Feelings aren’t absolute facts.
9	Mind Reading	Assuming you know what others are thinking or why they’re acting a certain way without any concrete evidence.	People’s thoughts aren’t transparent. Give them the benefit of the doubt rather than assuming the worst.
10	Fortune-Telling	Expecting things to go badly without sufficient evidence.	No one can accurately predict the future. Instead of jumping to conclusions, focus on what you can control in the present moment.

2 DATA

Due to lack of open and verified clinical datasets with annotated cognitive distortion labels by clinical provider, a synthetic datset is created to test the prototype AI system. Month-long simulated journal of a college student going through varied emotional states and situations were generated using 'gpt-3.5-turbo' api to assess the feasibility of the protoype. The simulated journal data of an individual contains 30 daily journal entries. The successful operation of the prototype on this simulated dataset illustrates the potential of such an AI system to identify and address condition distortion through augmented reasoning. However, such a system should eventually be tested and validated against real-world data sets with cognitive distortion labels provided by clinicians.

3 INSIGHT BOX ARCHITECTURE

3.1 Multiagent Workflow

The current 'Insight box' prototype features three distinct LLM-based agents, the Distortion detector, the Evaluator and the Insight provider. The Multi-agent system was implemented using open source python libraries such as *LangGraph* and *Langchain*, and it utilises the *'gpt-3.5-turbo'* model.. The agent workflow is implemented via Langgraph as shown in Figure 1. Each agent has distinct functions in the workflow as follows,

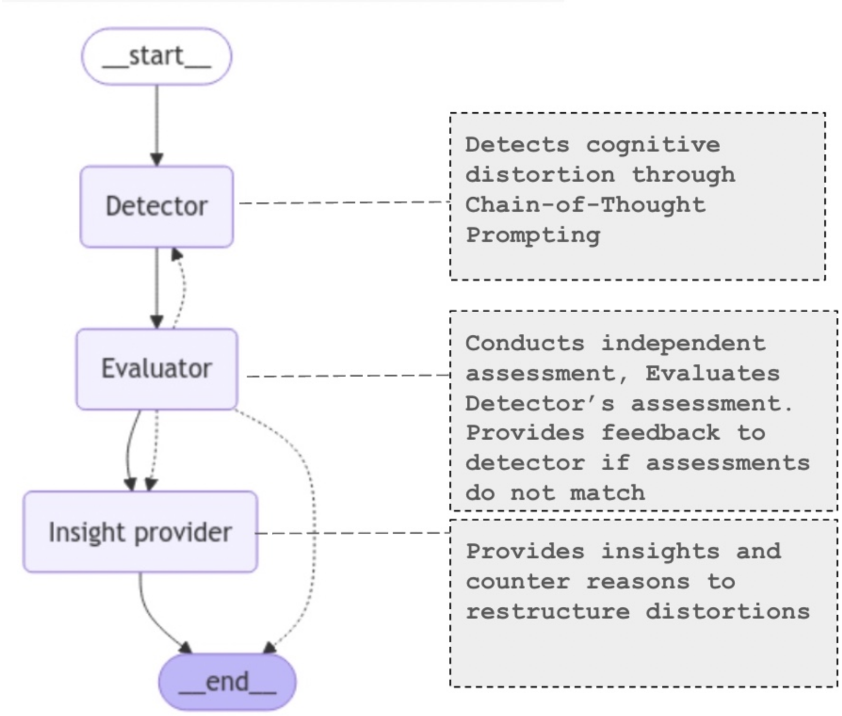
__start__
Detector
Evaluator
Insight provider
__end__
Detects cognitive distortion through Chain-of-Thought Prompting
Conducts independent assessment, Evaluates Detector's assessment. Provides feedback to detector if assessments do not match
Provides insights and counter reasons to restructure distortions

Fig. 2. LangGraph Agentic workflow

3.2. Detector

The detector checks the entry if a cognitive distortion is present or not through the 'chain of thought' prompting, (i) Assessing the situation of the person, and (ii) Question if the person is coping with their situation. (iii) If yes, then respond 'none'. If not, identify the most likely distortion from the provided checklist, which is responsible for not being able to cope. (iii) Justify your choice of distortion.

3.3. Evaluator

The evaluator is a reflective agent whose role is to independently detect cognitive distortions, assess alternate options to the detector and then compare its assessment to detector's assessment. It outputs both the assessment and checks if there is a match or not. If unmatched, it provides feedback to the detector to revise. It further loops for five iterations to converge their findings.

3.4. Insight provider

The insight provider is the end point of this network and the main interactive element with the user. It assimilates the findings of the other two agents to come up with counter reasoning using ’chain of thought’ approach to nudge the user to think contrary to what their distortion might suggest.

4. RESPONSES FROM THE PROTOTYPE SYSTEM

Initial findings indicate that the AI system can effectively detect and address cognitive distortions through augmented reasoning, as demonstrated in Table 2. To illustrate the working through the example below, (i) The entry is first received by the "Detector" and deploys ’chain of thought’ prompting to identify the distortion in the entry. Every step of its decision making such as assessing the situation, judging if the person is coping or not and finally the distortion "All-or-nothing" is chosen and sent to evaluator for further reflection and reevaluation. (ii) The "Evaluator" conducts its own assessment and compares it to "Detectors" assessment. If matched it passess on its assessment to "insight provider". If not it gives feedback to detector to re-evaluate till it matches .(iii) After passing evaluation, insight box compares both assessments, identifies the corresponding "counter reason" for the detected distortion as mentioned in the prompt, and furthermore incorporate it in its reply to the user as shown below. The full results including all 30 synthetic user entries, responses of each agent and the final output is available in the supplementory materials and a sample is also available in the appendix. (https://github.com/ Rampriya92/Insight-Box).

Table 2. Agents interactions with the Entry

User Entry	Completely bombed the chem quiz today. I knew I wasn’t prepared, but seeing that F staring back at me was still a punch in the gut. It feels like I am failing, but I need to study harder. I’m such a failure. Why can’t I just be smart like everyone else in this class? I skipped tennis practice — what’s the point? If I can’t even manage a simple quiz, how can I hope to be good at anything? Spent the evening scrolling through Instagram, just feeling sorry for myself.
Detected Distortion	All-or-nothing thinking
Detector	**Situation:** Feeling like a failure after performing poorly on a chemistry quiz and skipping tennis practice. **Coping:** No, the person is not coping well with the situation. **Explanation:** Their performance on the quiz is seen as an indication of overall intelligence and worth, using extreme language like "I’m such a failure" and "Why can’t I just be smart like everyone else in this class." This black-and-white thinking doesn’t consider improvement or other factors.
Evaluator	**Evaluator detected distortion:** All-or-nothing thinking **Detector detected distortion:** All-or-nothing thinking **Feedback:** None **Match:** Yes, matched!
Insight Provider	**Acknowledge:** All-or-nothing thinking **Counter reason:** Nudge to see middle grounds **Final Output:** Hey there, I’m sorry to hear that you’re feeling this way after the chemistry quiz. It’s understandable to be disappointed, but remember, one quiz does not define your intelligence or worth. It’s important to not be too hard on yourself. Instead of viewing this as a failure, think of it as a learning opportunity. Maybe you can talk to your teacher or a classmate for study tips or extra help. Also, skipping tennis practice is okay once in a while, everyone needs a break. It’s not a reflection of your abilities in other areas. How about taking a break from social media and doing something that makes you feel good, like going for a walk or reading a book? Remember, you are capable and smart in your own unique way. Let’s work together to come up with a plan to bounce back from this setback. What do you think?

5.LIMITATIONS

One of the main limitations of this study is that it needs further clinical validation with clinically annotated data set to further confirm the efficacy of the system. Each component of the agent workflow needs to be evaluated for efficacy.The lack of an open and verified clinical data set prevented a quantitative evaluation of the performance of this prototype.

6. FUTURE WORK

The multi-agent network needs to be further made more complex to produce more natural conversational responses. It can be extended to be an interactive journal where the user can interact and reason further with the conversational agent. In addition to cognitive distortions, the future development of insight box will be extended to a wide variety of psychological metrics that can be derived from unstructured conversational.The scope for improving the performance of the AI system through user feedback needs to be explored and implemented.

REFERENCES

[1] Adriana Munhoz Carneiro, Danilo Assis Pereira, Fernando Fernandes, Makilim Nunes Baptista, André Russowsky Brunoni, and Ricardo Alberto Moreno. Distorted thoughts as a mediator of depressive symptoms in patients with major depressive disorder: a longitudinal study. *Health and Quality of Life Outcomes*, 21(1):88, August 2023. https://doi.org/10.1186/s12955-023-02178-y

[2] Haaga, D.A., Dyck, M.J., and Ernst, D. Empirical status of cognitive theory of depression. *Psychological Bulletin*, 110(2), September 1991. https://doi.org/10.1037/0033-2909.110.2.215

[3] Sagarika Shreevastava and Peter Foltz. Detecting Cognitive Distortions from Patient-Therapist Interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, pages 151–158. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.clpsych-1.17

Data and supplementary material

Data and code used to analyse the data is openly available here: https://github.com/Rampriya92/InsightBox_CHI-2025_Workshop-submission

APPENDIX I

JOURNAL ENTRY 1

Journal Entry	Started the month with a decent swim practice. It felt good to be back in the water, a familiar comfort. Classes are piling up though, and I already feel behind in Organic Chemistry. I hate the idea of a new month, of having to put in the work for another 30 days. Maybe I'm being overly dramatic, but the divorce is still so raw. I miss how things used to be. I talked to Mom tonight, which always helps. I feel calmer and a bit better after catching up with her. I need to find that same peace in my studies.
Detected Distortion	Emotional reasoning
Detector's Note	**Situation:** Feeling overwhelmed by academic workload and emotional stress from a recent divorce, seeking comfort in familiar activities and connections with loved ones. **Coping:** Partially coping—finding comfort in speaking with their mom but still overwhelmed by the workload and emotional turmoil. **Reason:** The person's belief that their negative emotions about the divorce and academic stress dictate their ability to succeed ("I hate the idea of a new month," "Maybe I'm being overly dramatic," "I need to find that same peace in my studies") reflects emotional reasoning, where feelings are seen as evidence for the truth of a situation rather than objectively evaluating their circumstances.
Evaluator's Note	Evaluator detected distortion: Emotional Reasoning Detector detected distortion: Emotional Reasoning Feedback: None Match: Yes, matched!
Insight Provider	**Acknowledge:** Emotional reasoning **Counter reason:** Feelings are valid but not always factual. Your emotions from the divorce and academic stress are real, but they do not define your capabilities or potential for success. **Final output:** It's understandable that you're feeling overwhelmed by everything going on in your life right now. Remember, your emotions are valid, but they don't determine your worth or ability to handle challenges. It's great that talking to your mom helped you feel calmer. Perhaps consider reaching out to a counselor or academic advisor for additional support in managing your workload and emotions. Remember, it's okay to ask for help when you need it. How about scheduling a study session with a classmate or setting aside some time for self-care this week? You've got this! What's one small step you can take today to feel more at peace with your studies?

JOURNAL ENTRY 2

Journal Entry	Completely bombed the chem quiz today. I knew I wasn’t prepared, but seeing that F staring back at me was still a punch in the gut. It feels like I am failing, but I need to study harder. I’m such a failure. Why can’t I just be smart like everyone else in this class? I skipped tennis practice – what’s the point? If I can’t even manage a simple quiz, how can I hope to be good at anything? Spent the evening scrolling through Instagram, just feeling sorry for myself.
Detected Distortion	All-or-nothing thinking
Detector’s Note	**Situation:** Feeling like a failure after performing poorly on a chemistry quiz and skipping tennis practice. **Coping:** No, the person is not coping well with the situation. **Reason:** The person is viewing their performance on the quiz as an indication of their overall intelligence and worth, using extreme language like "I’m such a failure" and "Why can’t I just be smart like everyone else in this class." This black-and-white thinking doesn’t account for the possibility of improvement or other factors contributing to their quiz performance.
Evaluator’s Note	Evaluator detected distortion: All-or-nothing thinking Detector detected distortion: All-or-nothing thinking Feedback: None Match: Yes, matched!
Insight Provider	**Acknowledge:** All-or-nothing thinking **Counter reason:** Nudge to see Middle Grounds **Final output:** Hey there, I’m sorry to hear that you’re feeling this way after the chemistry quiz. It’s understandable to be disappointed, but remember, one quiz does not define your intelligence or worth. It’s important to not be too hard on yourself. Instead of viewing this as a failure, think of it as a learning opportunity. Maybe you can talk to your teacher or a classmate for study tips or extra help. Also, skipping tennis practice is okay once in a while, everyone needs a break. It’s not a reflection of your abilities in other areas. How about taking a break from social media and doing something that makes you feel good, like going for a walk or reading a book? Remember, you are capable and smart in your own unique way. Let’s work together to come up with a plan to bounce back from this setback. What do you think?

JOURNAL ENTRY 3

Journal Entry	Forced myself to go to the library today. Studied for a few hours, which is better than nothing. I'm still feeling down about the quiz, but I am trying to stop thinking about it. I did help this cute freshman with his math homework, which did my self-esteem a little good! Maybe I can use this chem debacle as motivation. My old coach used to say, "fall down seven times, stand up eight." I have to change that negative feeling into a positive and become a better student from this experience. I ordered pizza and watched a movie with my roommate, which was a nice distraction.
Detected Distortion	None
Detector's Note	**Situation:** Feeling down about a quiz but trying to move past it by studying and helping someone with math homework. **Coping:** Yes, the person is making efforts to cope with the negative feelings. **Reason:** The person is acknowledging their feelings, taking positive steps to address them, seeking motivation from past experiences, and engaging in activities to distract themselves. There is no clear indication of a cognitive distortion affecting their ability to cope in this entry.
Evaluator's Note	Evaluator detected distortion: None Detector detected distortion: None Feedback: None Match: Yes, matched!
Insight Provider	**Acknowledge:** None detected **Counter reason:** N/A **Final output:** Great job on recognizing and addressing your feelings about the quiz! It's awesome that you are taking steps to move past the negative emotions by studying and helping others with math homework. Remember, setbacks are a part of the learning process, and your old coach's advice about resilience is spot on. Keep pushing forward and using this experience as motivation to become an even better student. As you mentioned, treating yourself to pizza and a movie with your roommate was a nice distraction. It's important to take breaks and indulge in activities that bring you joy. Remember, self-care is just as important as academic success. Keep up the positive attitude and continue to focus on your growth and improvement. You're doing great! As an interactive element, why not jot down a few things you learned from helping the freshman with math homework? Reflecting on your ability to assist others can boost your self-esteem and reinforce your own understanding of the material. What do you think?

JOURNAL ENTRY 4

Journal Entry	Woke up feeling more positive. Went for a run, which always clears my head. I am very proud of myself for not skipping tennis practice, even though my ankle is still bothering me. My parents have been bickering non-stop lately. It is my fault that they hate each other now because of the problems that I have been having at school. It's exhausting and makes me feel guilty even though I know it's not my fault. Why can't they just get along for my sake? I need to find a way to block out their drama. I thought about going to the student counseling center, but I'm not sure.
Detected Distortion	Personalization
Detector's Note	**Situation:** Feeling proud of not skipping tennis practice despite an ankle injury, parents bickering, feeling guilty for their relationship issues, considering seeking help at the student counseling center. **Coping:** Partially coping, as the person is trying to find solutions and ways to deal with the situation. **Reason:** The person is taking undue responsibility for their parents' relationship issues, attributing blame to themselves for problems at school and feeling guilty for the parents' bickering. This distortion is evident as they believe their actions are the cause of the parents' problems and feel the need to find a solution to fix their parents' relationship issues.
Evaluator's Note	Evaluator detected distortion: Personalization Detector detected distortion: Personalization Feedback: None Match: Yes, matched!
Insight Provider	**Acknowledge:** Personalization **Counter reason:** You are not responsible for your parents' relationship issues. Their interactions are influenced by many factors beyond your control. **Final output:** It's understandable to feel overwhelmed by your parents' bickering, but it's important to remember that their issues are not your fault. It's not uncommon for children to feel responsible for their parents' conflicts, but it's essential to recognize that their relationship dynamics are independent of your actions. Have you considered talking to a counselor at the student counseling center about how you're feeling? They can provide support and guidance on how to navigate through these challenging emotions. Remember to prioritize your own well-being and self-care during this time. Let's explore some strategies together to help you cope with this situation. How do you feel about reaching out for help at the counseling center?

JOURNAL ENTRY 5

Journal Entry	Had a really good tennis practice. I'm getting my serve back. It felt amazing to just focus on the game and not think about school or my parents. I can't believe I almost quit earlier this week. I am happy that I did not quit playing tennis. Evenings are hard though. I end up just lying in bed, staring at the ceiling and overthinking everything. My mind races with every what-if. I feel so alone, even though I know I have friends and family who care.
Detected Distortion	Fortune-telling
Detector's Note	**Situation:** Feeling happy and accomplished after a good tennis practice, but struggling with overthinking and feelings of loneliness in the evenings. **Coping:** No, the person is struggling with overthinking and feelings of loneliness. **Reason:** The individual is anticipating negative outcomes and assuming that the future will be filled with uncertainties and negative possibilities, as evidenced by the constant "what-if" thoughts and feelings of loneliness despite having supportive friends and family.
Evaluator's Note	Evaluator detected distortion: Catastrophizing Detector detected distortion: Fortune-telling Feedback: The distortion detected by the detector is accurate in this case. The individual is engaging in fortune-telling by predicting negative outcomes and overthinking every possible scenario in the future. It aligns more closely with the fortune-telling distortion than catastrophizing. Match: Yes, matched!
Insight Provider	**Acknowledge:** Fortune-telling **Counter reason:** The future is uncertain. Focus on what you can control in the present. **Final output:** It's great to hear that you had a good tennis practice and are getting your serve back! It's important to remember that the future is uncertain, and overthinking every possible scenario can lead to unnecessary stress and feelings of loneliness. Instead of focusing on the unknown, try to shift your attention to the present moment and what you can control. Have you tried any relaxation techniques or mindfulness exercises to help calm your mind in the evenings? Remember, you're not alone, and reaching out to friends or family for support can make a big difference. Let's work together to find ways to manage your overthinking and feel more connected. How about we explore some strategies to help you unwind and feel more at peace in the evenings?

ADDITIONAL RESULTS

The rest of the results can be accessed in the GitHub repository by the author: https://github.com/Rampriya92/InsightBox_CHI-2025_Workshop-submission