Insight Box: A Multi-agent AI assistant for detecting and addressing cognitive distortions

Ram priyadarshini Ramchandran

Unaffiliated, rampriyadarshini@gmail.com

AI-based interventions are well-suited for early detection and management of mental health conditions, serving as a valuable complement to traditional treatments. One early indicator of conditions like depression and anxiety, which predict 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.

Additional Keywords and Phrases: cognitive distortion, AI assistant, Multi-agent, augmented reasoning.

1. INTRODUCTION

There is a global need for cheap and novel mental health interventions which 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 major components of CBT is to make individuals identify cognitive distortions/errors in their thinking and restructure it 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 behavioural 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 severity of symptoms in depression. CBT aims to restructure these thoughts over time by recognizing and challenging the negative assumptions to better deal with the 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. AI conversational 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, often leading to negative conclusions and misunderstandings in relationships	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 ‘*gemini-1.5-pro-latest’* api to assess the feasibility of the protoype. The simulated journal data of an individual contains 30 daily journal entries. The successful operation on 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 datsets 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 at its core has a reflective agent and was implemented using *LangGraph* and it utilises the *chat-gpt4 model* api. The agent workflow implemented via Langgraph as shown in *Figure 1*. Each agent has distinct functions in the workflow as follows,

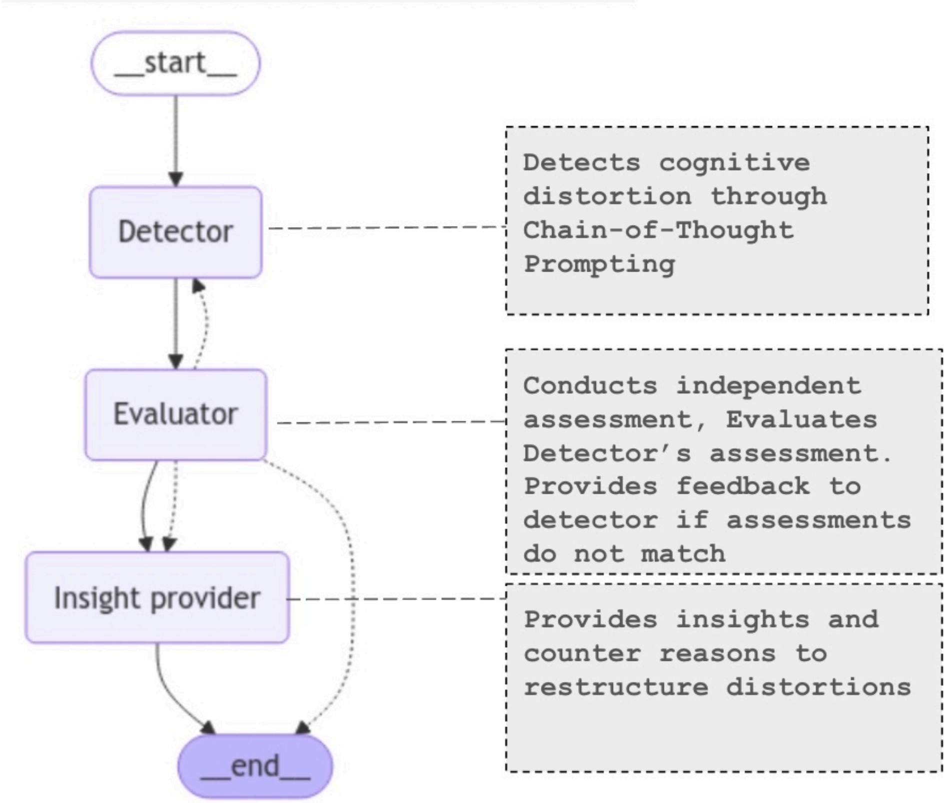
__start__
Detector
Evaluator
Insight provider
__end__
Detects cognitive distortion through Chain-of-Thought Prompting
Conducts independent assessment, Evaluates Detector's assessment. Provides feedback to detector if assessments do not match
Provides insights and counter reasons to restructure distortions

Figure 1: Agentic workflow implemented in Langgraph

3.2. Detector

The detector checks the entry if a cognitive distortion is present or not through the 'chain of thought' prompting through (i) assessing the situation of the person, (ii) Questions 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 Watson’s assessment. It outputs both the assessment and if there is a match or not. If unmatched, it provides feedback to the detector to revise. It 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 illustarte the working through the example below, (i) The entry is first recived 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 best “counter reason” theme mentioned in the prompt to incorporates 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 (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 ,Äì 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. **detected distortion:** All-or-nothing thinking\nreason: 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	['**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?"]

4.1 Limitations

One of the main limitations of this study is that it needs further clinical validation with a clinician annotated dataset to further confirm the efficacy of the system. Lack of open and verified clinical dataset prevented a quantitative assessment of the performance of this prototype.

4.2 Future work

The multi-agent network needs to be further made more complex to yield more natural conversational responses. It can be extended to be an interactive journal where the user can interact and reason further with the agent. The scope for improving AI system performance through feedback from users 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. 2023. 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 (August 2023), 88. https://doi.org/10.1186/s12955-023-02178-y

[2] Haaga Da, Dyck Mj, and Ernst D. 1991. 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. 2021. Detecting Cognitive Distortions from Patient-Therapist Interactions. In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, 2021. Association for Computational Linguistics, Online, 151–158. 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/Insight-Box