COGNITIVE REFRAMING VIA LARGE LANGUAGE MODELS FOR ENHANCED LINGUISTIC ATTRIBUTES

Xiaomeng Wang* **Dharmendra Sharma*** **Dinesh Kumar***

ABSTRACT

Cognitive Reframing aims to reshape negative thoughts into more positive perspectives to enhance mental well-being. While previous research has highlighted the efficacy of Large Language Models (LLMs) for cognitive reframing, there has been limited focus on enhancing reframing quality across multiple linguistic attributes in the final output. We build **ReframeGPT**, which fills this gap by employing LLMs to generate and iteratively refine reframed thoughts. The results of our study outperform in helpfulness, empathy and rationality in GPT-4 evaluation.

1 INTRODUCTION

Cognitive Reframing is an effective treatment in mental health therapy, particularly for people with Cognitive Distortions (*CDs*). Prior NLP research has demonstrated the effectiveness of LLMs in cognitive reframing through fine-tuning (Maddela et al., 2023), few-shot learning (Ziems et al., 2022), and retrieval-enhanced in-context learning (Sharma et al., 2023). Sharma et al. (2023) introduced a reframing framework that incorporates multiple linguistic attributes (empathy, rationality, etc.) and explored enhancing reframed thoughts within a single attribute in one generation, but with limited attempts for multiple features. Given the sensitivity of mental health support, it is crucial to assess the reframes before presenting them to users. Inspired by the studies that utilize LLMs for task completion across multiple steps (Yao et al., 2022; Wu et al., 2023; Tang et al., 2024), we develop **ReframeGPT**, leveraging GPT-3 as a reasoner to generate and iteratively refine reframed thoughts across various features, aiming for a high-quality reframe. The results indicate superior performance of our model in sentiment improvement and empathy in automatic evaluation, as well as helpfulness, empathy, and rationality in GPT-4 evaluation.

2 METHOD

Dataset We experiment with *Cognitive Reframing (CF)* dataset (Sharma et al., 2023), including 300 situations, thoughts and two reframed thoughts per situation. We maintain the same 70:30 train-test split as in Cognitive Reframing (Sharma et al., 2023).

Method Given a situation and a negative thought S_i, the task is to generate a reframed thought R_i that meets specific criteria. The Prompt Manager M decides on the need for tools and selects an appropriate one from Reframing Models (RMs). The tools generate intermediate outputs I_i^j, including preprocessed results p_i, intermediate reframed thought r_i^n, and assessment results a_i^n. r_i^n is iteratively assessed and refined until it fulfills all criteria, recognized as the final output. The task is formalized as: $R_i = LLM(M(P), M(S_i), M(RMs), M(F(I_i^j)))$

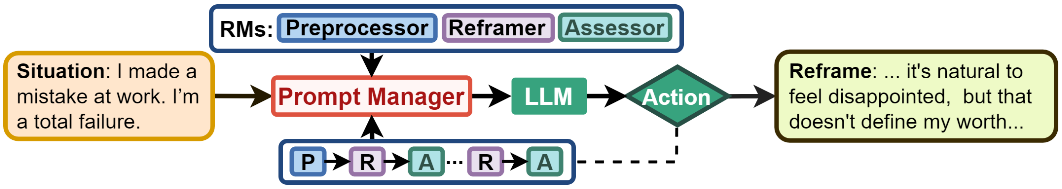

RMs:
Preprocessor
Reframer
Assessor
Situation: I made a mistake at work. I'm a total failure.
Prompt Manager
LLM
Action
Reframe: ... it's natural to feel disappointed, but that doesn't define my worth...
P
R
A
...
R
A

Figure 1: Pipeline of ReframeGPT.

---

*All authors are from the Faculty of Science and Technology, University of Canberra, Australia. For correspondence contact {cherry.wang, dharmendra.sharma, dinesh.kumar}@canberra.edu.au

Reframing Models (RMs)
1. Preprocessor Preprocessor involves:(1) identifying *CDs* using GPT-3 Curie fine-tuned on thinking traps in *CF* dataset; (2) detecting the user's emotion through a fine-tuned RoBERTa[1]; (3) determining the user's intention by prompting GPT-3.5-turbo as a psychologist. $p_i = RM_{(pre)}(S_i)$

2. Reframer a - Initial Reframing We adopt in-context learning reframing method in Sharma et al. (2023), utilizing RoBERTa embeddings to compute cosine similarity and prompting the top 5 similar examples to GPT-3 text-davinci-003. $r_i^1 = RM_{(Ref)}(S_i, p_i)$

3. Assessor We assess r_i^n on four attributes selected from the reframing framework, following a similar approach as in Sharma et al. (2023). Feedback a_i^n is provided based on predefined criteria. **(1) Addressing CDs** - Using a fine-tuned GPT-3 Curie on *CF* training set, we evaluate whether r_i^n addresses *CDs*. **(2) Rationality** - Measuring rationality with Reasoning Strength (*RS*), GPT-3.5-turbo generates sound and flawed explanations e for r_i^n, up to a depth of 2. A fine-tuned GPT-3 Curie labels e as 'sound' or 'flawed'. *RS* is calculated using the token probability of generating the label, evaluating if *RS* is greater than zero, indicating a rational reframe. $RS(r_i^n, S_i) = \mathbb{E}[P(e_{d=1,sound})] * (1 + RS(e_{d=2})) - \mathbb{E}[P(e_{d=1,flawed})] * (1 + RS(e_{d=2}))$ **(3) Positivity** - Using a RoBERTa-based sentiment classifier (Barbieri et al., 2020), we determine if the sentiment improves. **(4) Empathy** - A fine-tuned RoBERTa-based empathy classifier (Sharma et al., 2020) predicts the empathy level from 0 to 6, assessing whether it exceeds 3. $a_i^n = RM_{(ass)}(S_i, r_i^n)$

4. Reframer b - Refining We first prepare retrieval data from *CF* training set. Reframes not addressing *CDs* are revised. For other attributes, GPT-4 labels 'lower' or 'higher' levels for each paired reframe of S_i. During refining, we retrieve 8 most similar examples with paired 'lower' and 'higher' (or not addressed and addressed *CDs*) reframes and randomly select 5 to guide GPT-3 text-davinci-003 (Sharma et al., 2023) with Assessor feedback. $r_i^{n+1} = RM_{(ass)}(S_i, r_i^n, p_i, a_i^n)$.

3 EXPERIMENTS AND RESULTS

Baselines (1) Cognitive Reframing (Cog.R) (Sharma et al., 2023): Retrieves top 5 similar examples and prompt GPT-3 text-davinci-003. (2) Fine-tuning the training set on GPT-3 text-davinci-002.

Model	**Automatic**						**GPT-4**		
	BLEU	R-1	BScore	△TextBlob	EL	RS	Help.	Emp.	Rat.
Fine-tuning	0.187	0.213	**0.877**	0.027	2.73	**0.01**	3.7	3.5	3.833
Cog.R	**0.197**	**0.222**	**0.877**	0.037	4.33	-0.005	4.267	3.967	4.433
ReframeGPT	0.180	0.207	0.875	**0.053**	**5.01**	0.009	**4.633**	**4.667**	**4.833**
Reference	1	1	1	0	3.05	0.002	3.683	3.167	3.783

Table 1: Automatic and GPT-4 Evaluation. ROUGE-1 (R-1), BertScore (BScore), Empathy Level (EL), Reasoning Strength (RS, depth = 2), Helpfulness (Help.), Empathy (Emp.), Rationality(Rat.).

Results (1) Automatic Evaluation: We utilize BLEU (Papineni et al., 2002), ROUGE-1 (Chin-Yew, 2004), BERTScore (Zhang et al., 2019), △Textblob (to measure sentiment change) (Loria et al., 2018), *RS* (for rationality), and empathy level. Our model excels in △Textblob and empathy level but scores lower in BLEU, R-1, and BScore, possibly due to iterative refinement introducing word choice variations. (2) GPT-4 Evaluation: Score 10 randomly selected outputs from 1 to 5 (three times) regarding helpfulness (effectiveness in overcoming negative thoughts), empathy (level of conveyed empathy), and rationality (logical coherence). Our model outperforms in all aspects.

4 CONCLUSION

Our study has demonstrated the efficacy of employing a LLM as a reasoner in cognitive reframing. Future work will focus on exploring the relationship among the linguistic attributes and involve experts in data annotation and evaluation.

---

[1] twitter-roberta-base-emotion-multilabel-latest (`https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion-multilabel-latest`)

URM STATEMENT

We acknowledge that all authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. arXiv preprint arXiv:2010.12421, 2020.

Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, 2004, 2004.

Steven Loria et al. textblob documentation. Release 0.15, 2(8):269, 2018.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. Training models to generate, recognize, and reframe unhelpful thoughts. arXiv preprint arXiv:2307.02768, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. arXiv preprint arXiv:2009.08441, 2020.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. Cognitive reframing of negative thoughts through human-language model interaction. arXiv preprint arXiv:2305.02466, 2023.

Daniel Tang, Zhenghan Chen, Kisub Kim, Yewei Song, Haoye Tian, Saad Ezzini, Yongfeng Huang, and Jacques Klein Tegawende F Bissyande. Collaborative agents for software engineering. arXiv preprint arXiv:2402.02172, 2024.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

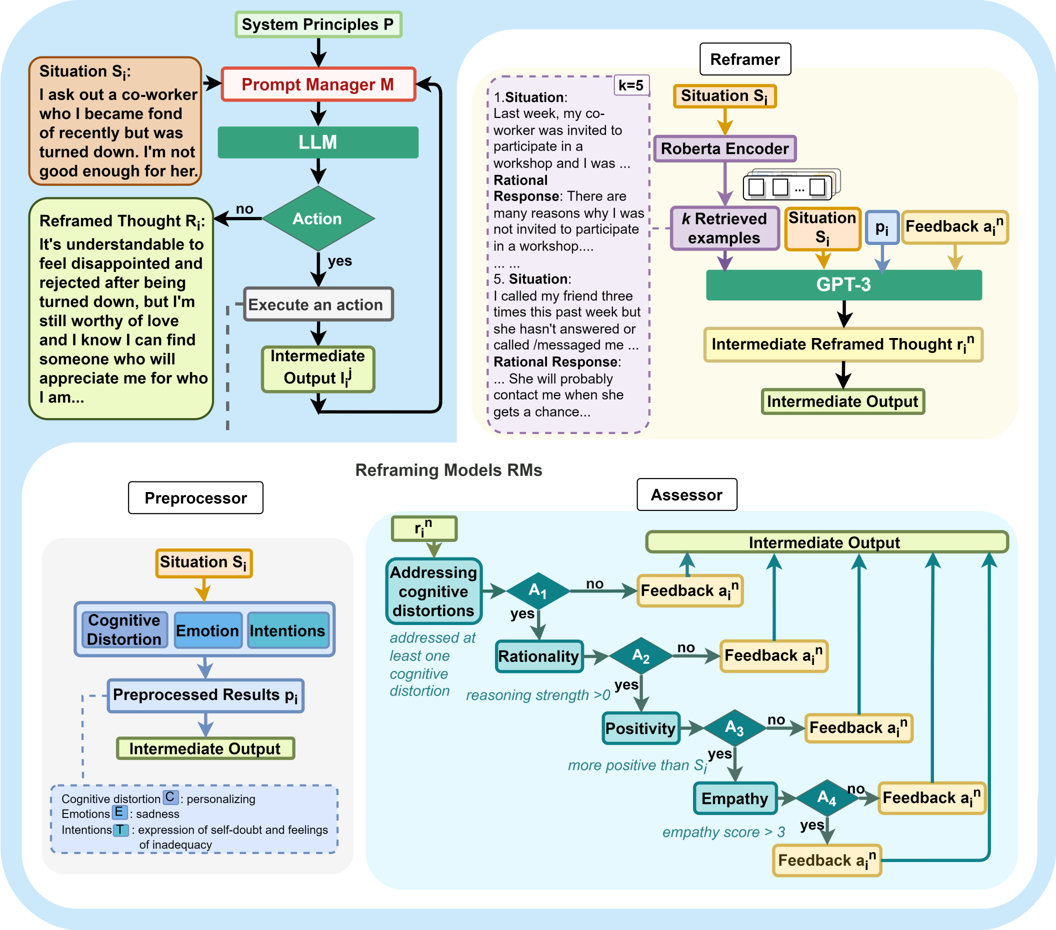
Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. Inducing positive perspectives with text reframing. arXiv preprint arXiv:2204.02952, 2022.

A APPENDIX

Framework Overview Figure 2 and 3 provide an overview of ReframGPT Framework and an example of how the system iteratively refined the reframed thought. In the framework, System Principle P establishes rules, such as using RMs to perform tasks instead of generating reframes directly.

Parameters & Evaluation Prompt Table 2. provides parameters and GPT-4 evaluation prompt.

Case Study Table 3. provides examples of outputs for generating reframed thoughts.



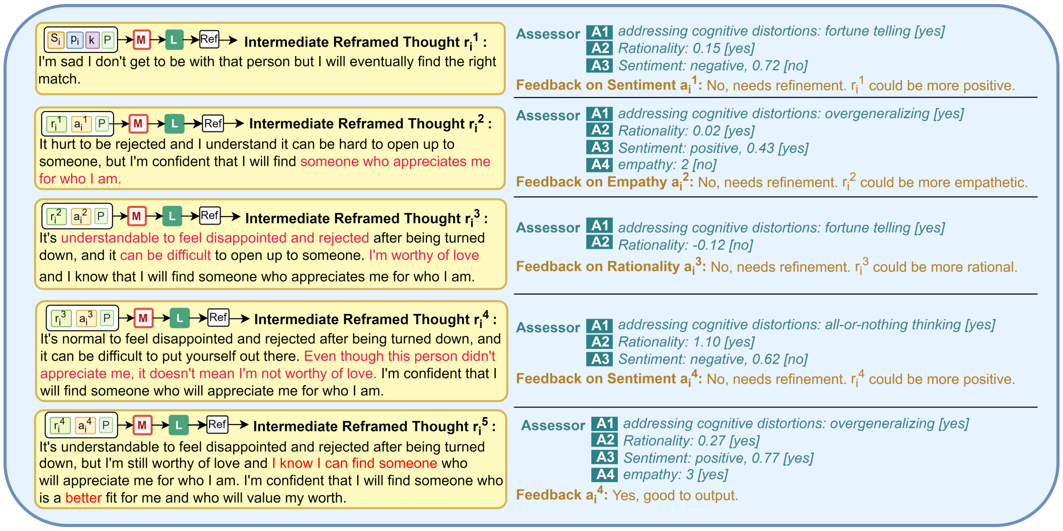
Situation S_i: I ask out a co-worker who I became fond of recently but was turned down. I'm not good enough for her.

Reframed Thought R_i: It's understandable to feel disappointed and rejected after being turned down, but I'm still worthy of love and I know I can find someone who will appreciate me for who I am...

1.**Situation**: Last week, my co-worker was invited to participate in a workshop and I was ...
Rational Response: There are many reasons why I was not invited to participate in a workshop....
... ...
5. **Situation**: I called my friend three times this past week but she hasn't answered or called /messaged me ...
Rational Response: ... She will probably contact me when she gets a chance...

Cognitive distortion C : personalizing
Emotions E : sadness
Intentions T : expression of self-doubt and feelings of inadequacy

Figure 2: Overview of ReframeGPT.



Intermediate Reframed Thought r_i^1 :
I'm sad I don't get to be with that person but I will eventually find the right match.

Assessor
A1 *addressing cognitive distortions: fortune telling [yes]*
A2 *Rationality: 0.15 [yes]*
A3 *Sentiment: negative, 0.72 [no]*
Feedback on Sentiment a_i^1: No, needs refinement. r_i^1 could be more positive.

Intermediate Reframed Thought r_i^2 :
It hurt to be rejected and I understand it can be hard to open up to someone, but I'm confident that I will find someone who appreciates me for who I am.

Assessor
A1 *addressing cognitive distortions: overgeneralizing [yes]*
A2 *Rationality: 0.02 [yes]*
A3 *Sentiment: positive, 0.43 [yes]*
A4 *empathy: 2 [no]*
Feedback on Empathy a_i^2: No, needs refinement. r_i^2 could be more empathetic.

Intermediate Reframed Thought r_i^3 :
It's understandable to feel disappointed and rejected after being turned down, and it can be difficult to open up to someone. I'm worthy of love and I know that I will find someone who appreciates me for who I am.

Assessor
A1 *addressing cognitive distortions: fortune telling [yes]*
A2 *Rationality: -0.12 [no]*
Feedback on Rationality a_i^3: No, needs refinement. r_i^3 could be more rational.

Intermediate Reframed Thought r_i^4 :
It's normal to feel disappointed and rejected after being turned down, and it can be difficult to put yourself out there. Even though this person didn't appreciate me, it doesn't mean I'm not worthy of love. I'm confident that I will find someone who will appreciate me for who I am.

Assessor
A1 *addressing cognitive distortions: all-or-nothing thinking [yes]*
A2 *Rationality: 1.10 [yes]*
A3 *Sentiment: negative, 0.62 [no]*
Feedback on Sentiment a_i^4: No, needs refinement. r_i^4 could be more positive.

Intermediate Reframed Thought r_i^5 :
It's understandable to feel disappointed and rejected after being turned down, but I'm still worthy of love and I know I can find someone who will appreciate me for who I am. I'm confident that I will find someone who is a better fit for me and who will value my worth.

Assessor
A1 *addressing cognitive distortions: overgeneralizing [yes]*
A2 *Rationality: 0.27 [yes]*
A3 *Sentiment: positive, 0.77 [yes]*
A4 *empathy: 3 [yes]*
Feedback a_i^4: Yes, good to output.

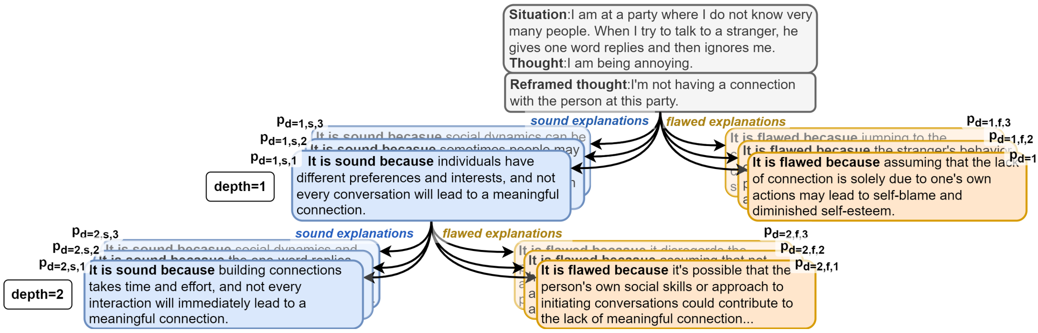
Figure 3: Example of Iterative Refinement.

System LLM & Reframer	GPT3: text-davinci-003 top_p: 0.6 frequency_penalty: 0 presence_penalty: 0
GPT-4 Evaluation	You will be given a situation and thought with its five cognitive reframes. Please provide an evaluation of each reframe. Assign a score from 1 to 5 based on each of the following criteria: helpfulness, empathy, and rationality. You are sensitive to the nuances of expressions and you need evaluate them with different scores based on their performance. 1.Helpfulness: Assess the efficacy of the reframed thought in overcoming negative thoughts.Reflect on whether the output provides constructive and supportive elements. 2.Empathy: Gauge the perceived level of empathy conveyed in the reframed thought. Evaluate the sensitivity and understanding expressed towards the user's emotions. 3. Rationality: Evaluate the logical coherence and soundness of the reframed thought. Consider whether the output provides a rational and reasoned perspective. Scoring Guidelines: 1-2: Poor (Significant issues with unhelpful, lacks empathy, or lacks rationality) 3: Average (Room for improvement in one or more aspects) 4: Good (Well-written, helpful, empathetic, or rational) 5: Excellent (Exceptional in terms of helpfulness, empathy, or rationality).

Table 2: Parameters & Prompt

Empathy Level The RoBERTa-based empathy classifier is fine-tuned on the training dataset from *CF* and *PATTERNREFRAME* (Maddela et al., 2023), where empathy levels are labeled by GPT-4 based on the definition provided in the theoretical framework of expressed empathy (Sharma et al., 2020), including emotional reactions, interpretation and exploration with no (level: 0), weak (level: 1), and strong (level: 2) communication. The sum of these score is our final empathy level.

Reasoning Strength Figure 2. shows a depth of two explanations for reasoning strength. To calculate RS, we first prompt GPT-3.5-turbo with 10 examples to generate three sound $e_{d=1,sound}$ and three flawed $e_{d=1,flawed}$ explanations of r_i^n. Leveraging a fine-tuned GPT-3 Curie to label these explanations as 'sound' or 'flawed' and select the explanation with the highest token probability to proceed the second depth. We compute the difference in average token probabilities between the labeled sound and flawed explanations and follow this equation to calculate RS: $RS(r_i^n, S_i) = \mathbb{E}[P(e_{d=1,sound}|S_i, r_i^n)] * (1 + RS(e_{d=1,sound}, r_i^n, S_i)) - \mathbb{E}[P(e_{d=1,flawed}|S_i, r_i^n)] * (1 + RS(e_{d=1,flawed}, r_i^n, S_i))$



Situation:I am at a party where I do not know very many people. When I try to talk to a stranger, he gives one word replies and then ignores me.
Thought:I am being annoying.

Reframed thought:I'm not having a connection with the person at this party.

It is sound because individuals have different preferences and interests, and not every conversation will lead to a meaningful connection.

It is flawed because assuming that the lack of connection is solely due to one's own actions may lead to self-blame and diminished self-esteem.

It is sound because building connections takes time and effort, and not every interaction will immediately lead to a meaningful connection.

It is flawed because it's possible that the person's own social skills or approach to initiating conversations could contribute to the lack of meaningful connection...

sound explanations
flawed explanations
$p_{d=1,s,3}$
$p_{d=1,s,2}$
$p_{d=1,s,1}$
$p_{d=1,f,3}$
$p_{d=1,f,2}$
$p_{d=1,f,1}$
depth=1
It is sound becasue social dynamics can be
It is sound becasue sometimes people may
It is flawed becasue jumping to the
It is flawed becasue the stranger's behavior
sound explanations
flawed explanations
$p_{d=2,s,3}$
$p_{d=2,s,2}$
$p_{d=2,s,1}$
$p_{d=2,f,3}$
$p_{d=2,f,2}$
$p_{d=2,f,1}$
depth=2
It is sound becasue social dynamics and
It is sound becasue the one word replies
It is flawed becasue it disregards the
It is flawed becasue assuming that not

Figure 4: Explanations in Reasoning Strength (depth=2).

Situation 1	I posted a new piece of artwork online and people started comparing it with artwork created by others. I shouldn't have posted it.
GPT.Finetune	I feel vulnerable sharing my art. I'm proud of it and want to share it but I also want to protect it.
Cog.Reframe	It's normal to compare artwork and it's okay if people don't like my work. I should be proud of my work and take constructive criticism as an opportunity to grow and learn.
ReframeGPT	I am proud of my artwork and I am also proud to share it with others. Everyone has their own unique style and I should not be afraid to share it. It's normal to feel discouraged when faced with comparisons, but I can use it as an opportunity to learn and develop my own style. I can also use it as a chance to appreciate the work of others and take constructive criticism to improve my own work.
Reference 1	I am proud to have posted my work.
Reference 2	I can still be proud of my work. It is totally up to me whether I should post it or not.
Situation 2	An mturk requester rejected my task and I wasn't sure why because I work very hard on my tasks. Being new it affected my approval rating more negatively.I'm not smart enough to succeed at mturk.
GPT.Finetune	I am smart enough to succeed at Mturk. It's okay that I made mistakes in the beginning and I can improve from here.
Cog.Reframe	Rejections are part of the process and it doesn't mean I'm not smart enough to succeed. I can learn from this experience and use it to improve my work and approval rating.
ReframeGPT	It is understandable to feel frustrated when my work is rejected, but I need to remember that rejection on mturk can be due to various factors beyond my control. I can look for guidance and feedback to help me succeed on mturk.
Reference 1	I am learning still and can improve. Just because this one task did not go as well as I would like does not mean I cannot improve and do better in the future.
Reference 2	It seems like there was some miscommunication. It doesn't mean that I do not have the skills to do well with mturk. I should reach out to see if I can get more clarity on why my task was rejected.
Situation 3	I got into a minor car crash.I messed up big time.
GPT.Finetune	I made a mistake, but I can learn from this and it won't happen again.
Cog.Reframe	Accidents happen and this was a minor one. I can take responsibility for my mistake and work to ensure it doesn't happen again.
ReframeGPT	This car accident was an unfortunate situation. I need to take responsibility for my part in the accident and learn from this experience. I will take the necessary steps to address any legal or insurance implications and use this experience as an opportunity to learn and grow.
Reference 1	I will drive more carefully next time.
Reference 2	I only scratched the bumper. I can get that fixed easily.
Situation 4	A week ago me and my GF were trying to find a store in a nearby town but we got lost.We won't be able to find the store before it closes.
GPT.Finetune	We will find the store. We just have to be more patient.
Cog.Reframe	We can try to find the store again another day, or find something else to do that would be fun and enjoyable.
ReframeGPT	We may not make it to the store in time, but that doesn't mean we have failed. We can take this opportunity to explore other options, and if we don't find anything else, we can come back another day and try again. In the meantime, we can still enjoy the journey and appreciate the time spent together.
Reference 1	I am feeling anxious and overwhelmed. Let me take a moment to refocus so we can work together to find the right way to go to get to the store on time.
Reference 2	It is possible the store may close but being lost is not purposeful. If the store closes before we can get there we can just figure out another option to get what we need to get through the night.

Table 3: Case Study

* empathy (emotional reactions), rationality