2024.01.31 LabSeminar

Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction

송형우

HUMAN
CENTERED
COMPUTING
LABORATORY

Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction

Ashish Sharma[♠] **Kevin Rushton**[◇] **Inna Wanyin Lin**[♠] **David Wadden**[♣]
Khendra G. Lucas[◇] **Adam S. Miner**[♥♡] **Theresa Nguyen**[◇] **Tim Althoff**[♠]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Mental Health America [♣]Allen Institute for Artificial Intelligence
[♥]Department of Psychiatry and Behavioral Sciences, Stanford University
[♡]Center for Biomedical Informatics Research, Stanford University
{ashshar,althoff}@cs.washington.edu

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 9977–10000

HUMAN CENTERED COMPUTING LABORATORY

url: https://aclanthology.org/2023.acl-long.555/

Table of Contents

Negative Thoughts

- nature part of human cognition

- but for people with mental health challenges
  - entrenched, automatic and emotionally triggering
  - difficult to overcome

Cognitive Reframing

- replacing with a more hopeful "reframed thought"

- 3 factors for reframing to be effective
  1. relatable to the individual
  2. helpful in overcoming the negative thought
  3. memorable to be accessible the next time a similar thought arises

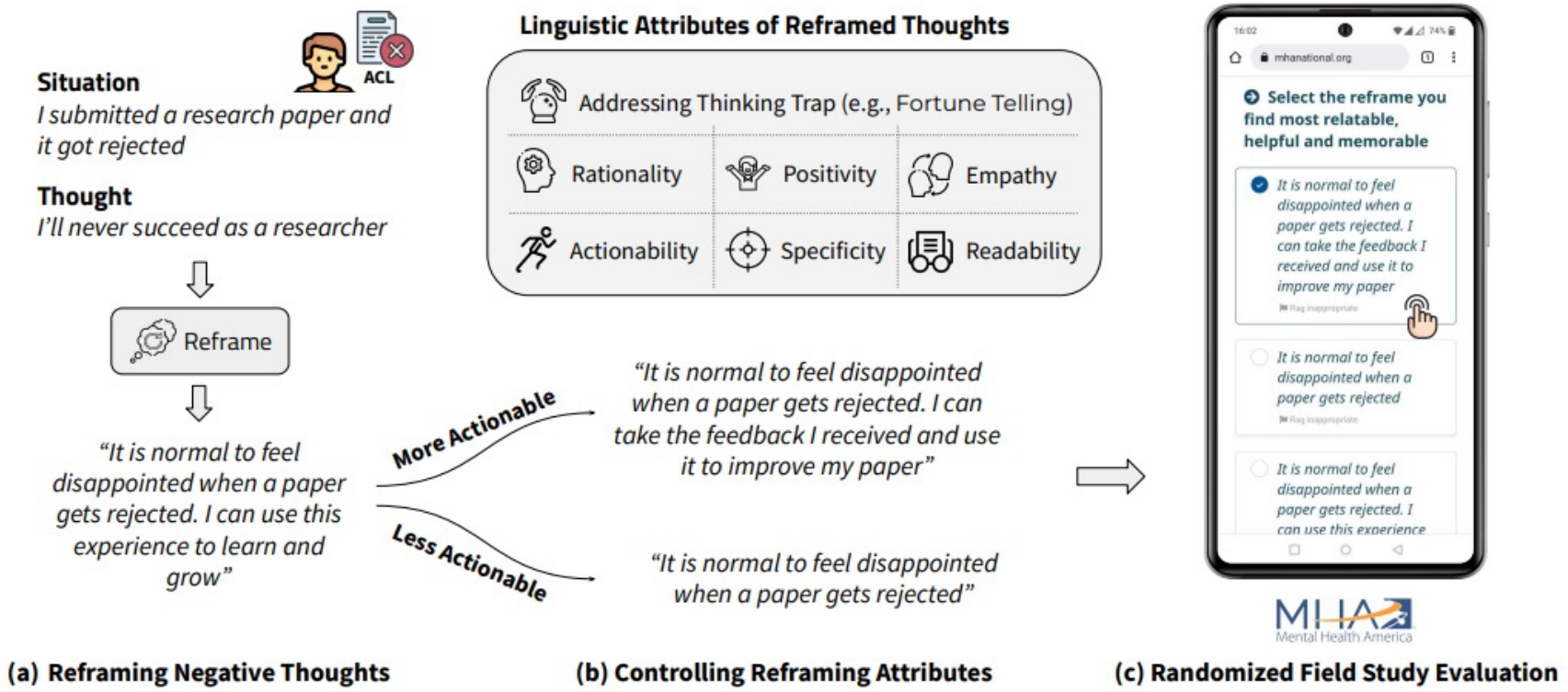
Background

Problem Definition → Goals

- Understanding what characterizes a relatable, helpful and memorable reframe is challenging and unknown
  → Understand what constitutes successful reframing

- Clinician shortages, lack of insurance coverage and stigma commonly limit access to therapists
  → Understand how LM can assist people in this process

Background - Overview



(a) Reframing Negative Thoughts **(b) Controlling Reframing Attributes** **(c) Randomized Field Study Evaluation**

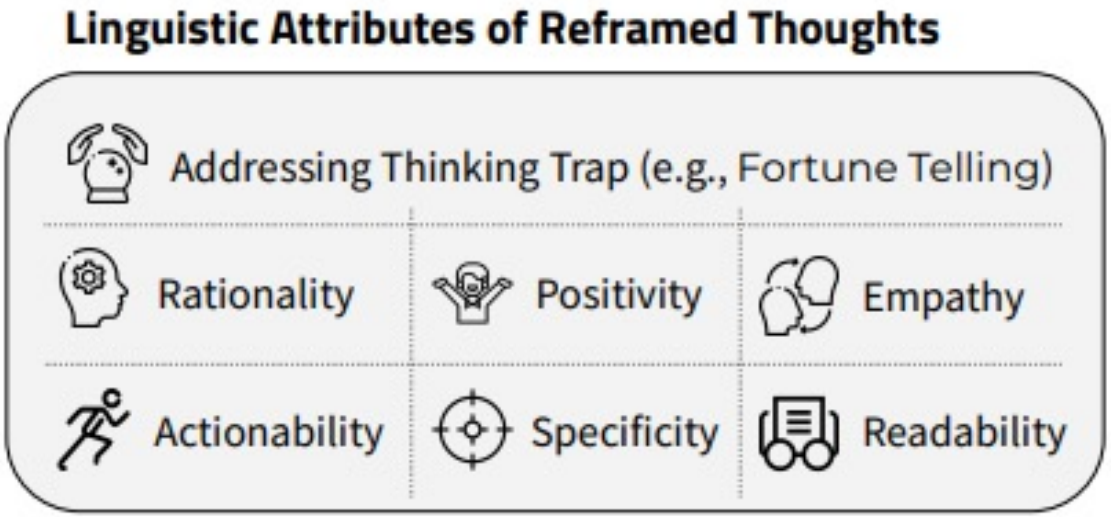
HUMAN
CENTERED
COMPUTING
LABORATORY

Aims of the first part

1. Characterize the linguistic attributes of reframed thoughts
2. Collect a dataset of situations, thoughts and reframes
3. Develop methods to generate reframes and to measure and control their attributes

4. Assess the validity of proposed linguistic attributes and evaluate the performance of the reframed generation model

First Part (1) Characterize the linguistic attributes of reframed thoughts



From

1. Clinical therapy practices
2. Collaborate with mental health experts

Datasets

- Thought Records Dataset
  - Amazon Mechanical Turk
  - 180 pairs of situations - thoughts
- Mental Health America
  - MHA
  - 120 pairs of situations - thoughts

Annotation

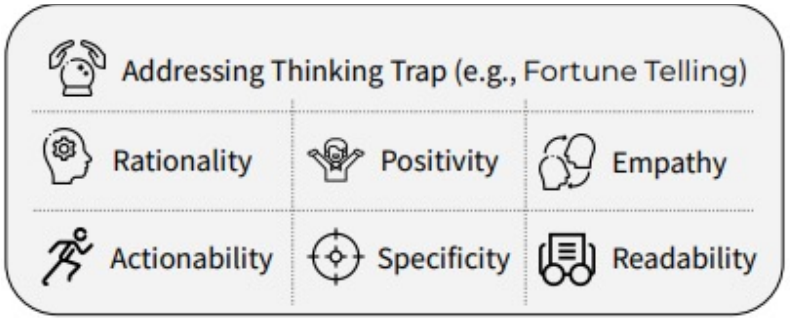
- 15 current mental health practitioners and clinical psychology graduate students with significant practical experience

    1. Write two different reframed thoughts
    2. Annotate the thinking traps addressed by each reframed thought
    3. Compare the two reframes and choose the one that is more rational, more positive, more actionable, more empathic, more specific, and more readable

- 600 reframed thoughts with annotations on their linguistic attributes

Automated metrics for measuring linguistic attributes

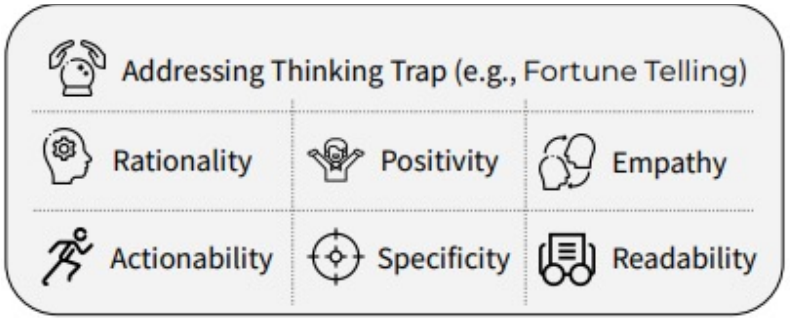
- Addressing Thinking Traps
  - multi-label classification
- Rationality
  - recursive reasoning strength : GPT3, k=10
    - the most plausible explanations that support Ri
    - the most plausible explanations that refute Ri
- Positivity
  - a RoBERTa-based sentiment classifier fine-tuned on the TweetEval benchmark



Automated metrics for measuring linguistic attributes

- Empathy
  - empathy classification model
    - emotional reactions, interpretations, explorations
    - trained upon 300 reframed thoughts
  - empathy levels on a scale from 0 to 6

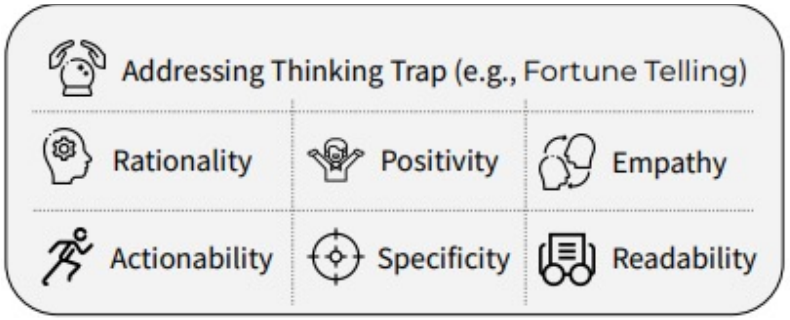
- Actionability
  - contains_action(Ri) + next_action_coherence(Ri)
  - contains_action
    - whether suggested concrete action
  - next_action_coherence
    - GPT-3, k=5 → next action candidates



Automated metrics for measuring linguistic attributes

- Specificity
  - sentence embedding similarity
    - the reframed thought Ri
    - the concatenation of the situation Si and the thought Ti

- Readability
  - Coleman-Liau Index (CLI) metric : 0.0588L − 0.296S − 15.8
    - L : average number of letters per 100 words
    - S : average number of sentences per 100 words



Methods to generate reframed thoughts

- Retrieval-based in-context learning method
  - For each situation Si and negative thought Ti, k-similar examples from the dataset
  - RoBERTa embedding
  - k=5, top-5 values of cosine_sim(concat(s,t),concat(Si,Ti))

Controlling literature attributes of generated reframes

- GPT-3, in-context examples, K=5

- For each attributes

$$Rj*(a,L) \rightarrow Rj*(a,H) \;////\; Rj*(a,H) \rightarrow Rj*(a,L)$$

The construct validity of proposed linguistic attributes

- correlating them with the human judgments of mental health experts

Attribute	Pearson Correlation
Addressing Thinking Traps	0.680***
Rationality	0.448**
Positivity	0.550***
Empathy	0.575***
Actionability	0.647***
Specificity	0.427**
Readability	0.331*

Table 1: Correlation of our proposed attribute measures by with human judgments from mental health experts. *:$p < 0.05$; **:$p < 0.001$; ***:$p < 10^{-5}$.

First Part (4) Assess the validity of proposed linguistic attributes and evaluate the performance of the reframed generation model

The performance of the reframe generation model

- top-p sampling , p=0.6
- 600 expert-annotated examples → 7:3

Model	Automatic				Human	
	BLEU	R-1	R-L	BScore	Rel.	Help.
Retrieval Only	21.6	18.8	14.2	86.7	2.58	3.14
Pos. Reframing	24.4	23.6	17.6	87.6	2.67	2.40
DialoGPT	22.5	17.4	13.5	86.3	2.49	3.21
T5	24.9	23.4	17.8	87.2	2.51	3.30
GPT-3 Only	25.0	23.9	18.0	88.3	2.97	3.98
Our Model	**27.8**	**26.0**	**19.9**	**88.6**	**3.10**	**4.11**

Table 2: Automatic and Human Evaluation Results. R-1: ROUGE-1; R-L: ROUGE-L; BScore: BertScore; Rel.: Relatability; Help.: Helpfulness.

An aim of the second part

- Investigate which linguistic attributes are related to the reframing outcomes of relatability, helpfulness and memorability

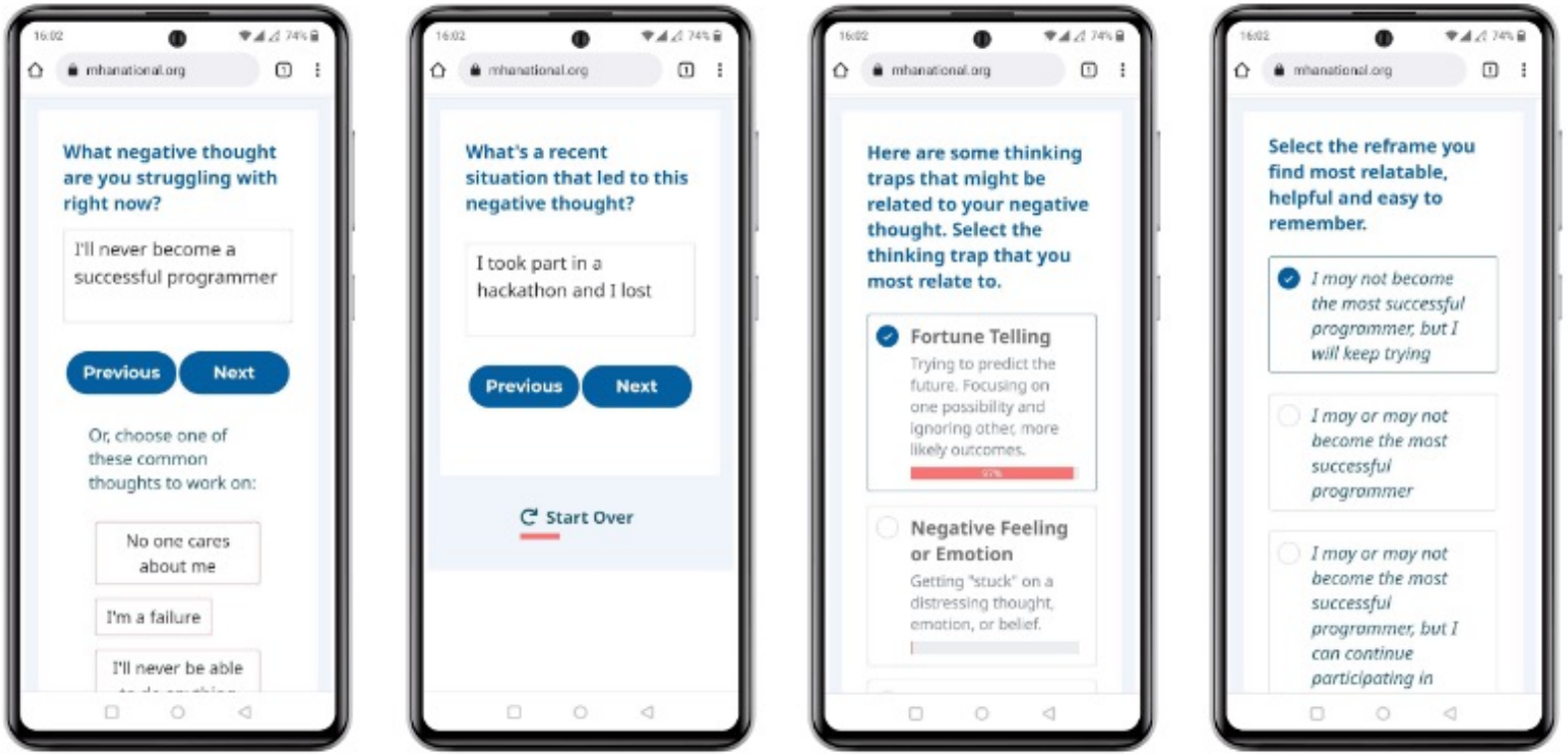
Second Part

Randomized Field Study on a Large Mental Health Platform

- Mental Health America
  - mental health website that provides mental health resources and tools to millions of users (bit.ly/changing-thoughts)

- 2,067 MHA visitors as participants
  - MHA visitors described their situation and the thoughts they were struggling with

Second Part



(a) Thought **(b) Situation** **(c) Cognitive Distortions** **(d) Reframed Thoughts**

HUMAN
CENTERED
COMPUTING
LABORATORY

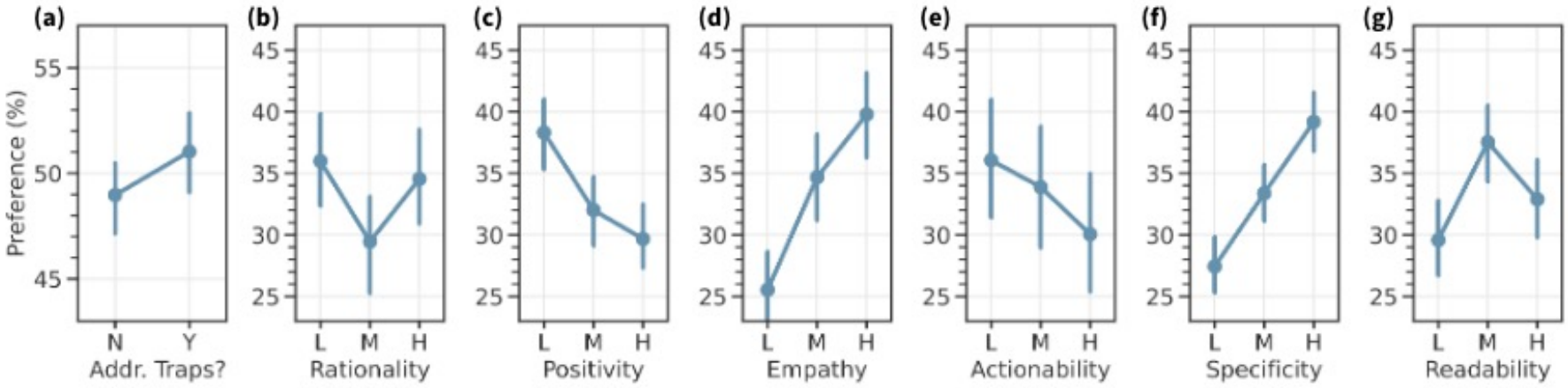


Figure 2: **Which linguistic attributes of reframed thoughts do people prefer?** For a given situation and thought from MHA visitors, we show them multiple LM-generated reframes with variance across a randomly selected attribute (e.g., low, medium and high actionability). We find that highly empathic and highly specific reframings are more preferred. On the other hand, reframes with high positivity are less preferred. N: No; Y: Yes; L: Low; M: Medium; H: High. Error bars in any figure represent 95% bootstrapped confidence intervals.

Second Part

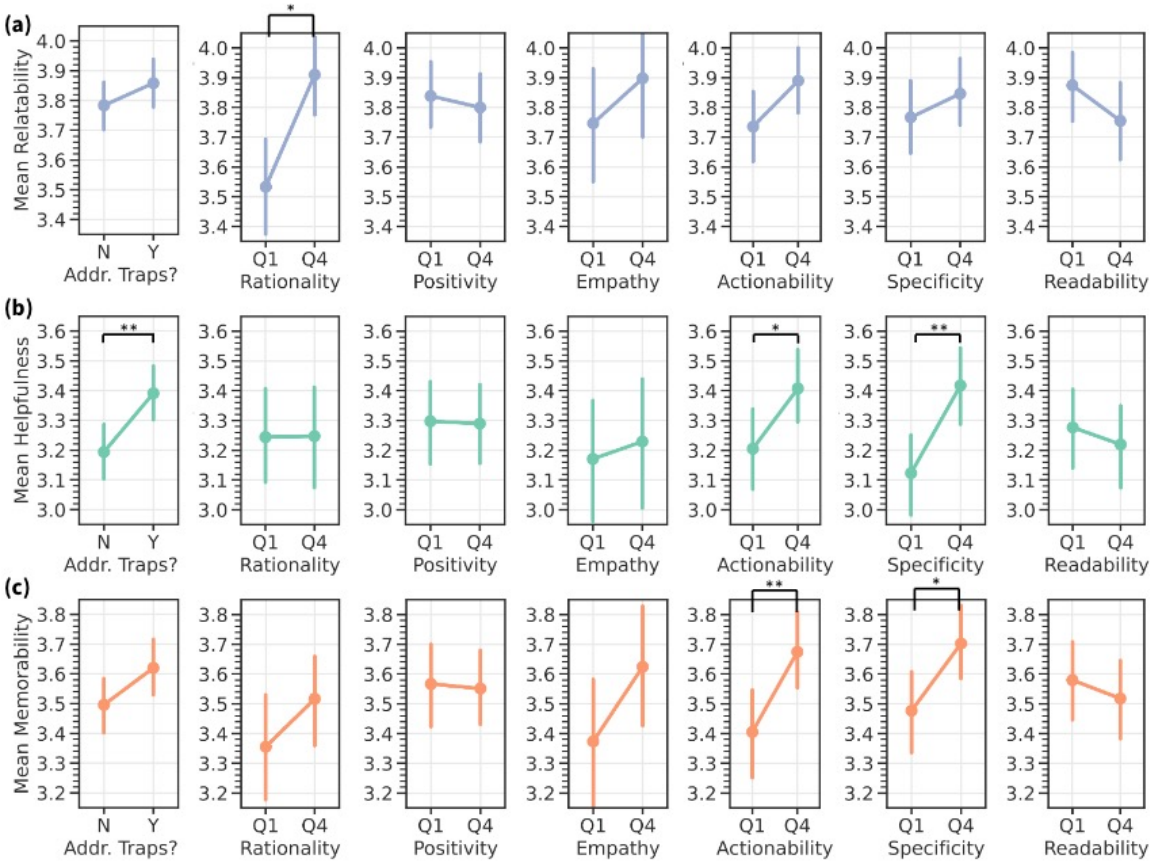


Figure 3: **Which linguistic attributes are associated with desired cognitive reframing outcomes?** For a given situation and thought, we show one LM-generated reframe to MHA participants and ask them to rate it on relatability, helpfulness and memorability on a 1 to 5 scale. For each linguistic attribute, we compare the first (Q1) and the fourth quartile (Q4). We find that **(a)** reframes that have higher rationality are more relatable; **(b)** reframes that address thinking traps, have higher actionability or higher specificity are more helpful; **(c)** reframes that have higher actionability or higher specificity are more memorable. *:$p < 0.05$;**:$p < 0.01$.

Conclusion

- Prefer
  - highly **empathic or specific**
- Do not prefer
  - highly **positive**

- More relatable
  - more **rational**
- More helpful
  - addressing thinking traps
  - more **actionable** and **specific**
- More memorable
  - more **actionable** and **specific**

Conclusion

- Limitation
  - Did not investigate clinical outcomes
  - Socio-cultural factors should be considered
    - specific communities
  - Evaluating long-term outcomes is critical
    - Although, there is a study that suggests that single-session, in-the-moment interventions can lead to significant positive long-term mental health outcomes

Thanks

HUMAN
CENTERED
COMPUTING
LABORATORY