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

Abstract

A proven therapeutic technique to overcome negative thoughts is to replace them with a more hopeful "*reframed thought*." Although therapy can help people practice and learn this *Cognitive Reframing of Negative Thoughts*, clinician shortages and mental health stigma commonly limit people's access to therapy. In this paper, we conduct a human-centered study of how language models may assist people in reframing negative thoughts. Based on psychology literature, we *define* a framework of seven linguistic attributes that can be used to reframe a thought. We develop automated metrics to *measure* these attributes and validate them with expert judgements from mental health practitioners. We collect a dataset of 600 situations, thoughts and reframes from practitioners and use it to train a retrieval-enhanced in-context learning model that effectively *generates* reframed thoughts and *controls* their linguistic attributes. To investigate what constitutes a "high-quality" reframe, we conduct an IRB-approved *randomized field study* on a large mental health website with over 2,000 participants. Amongst other findings, we show that people prefer highly empathic or specific reframes, as opposed to reframes that are overly positive. Our findings provide key implications for the use of LMs to assist people in overcoming negative thoughts.

1 Introduction

Negative thoughts are a natural part of human cognition. However, for people experiencing mental health challenges, such thoughts are often entrenched, automatic and emotionally triggering, making it difficult to overcome them in-the-moment (Beck, 1976). An evidence-based, well-established therapeutic intervention to overcome negative thoughts is *Cognitive Reframing*, in which a negative thought is replaced with a more hopeful "*reframed thought*," which offers an alternative perspective on one's situation (Beck, 1976). For example, imagine a person with a situation "*I'm submitting a research paper to ACL 2023*" has a thought "*This paper is going to get rejected*." A possible way to reframe this thought is to say "*This paper has some chance of getting accepted due to its novel methodology and potential impact*."

Psychotherapy research suggests that for a reframed thought to be effective, it must be (a) relatable to the individual, (b) helpful in overcoming the negative thought and (c) memorable to be accessible the next time a similar thought arises (Beck, 1976; Burns, 1980). However, understanding what characterizes a relatable, helpful and memorable reframe is challenging and unknown. Professional therapists can support people in coming up with such highly effective reframed thoughts. However, barriers like clinician shortages, lack of insurance coverage and stigma commonly limit access to therapists (Olfson, 2016; Sickel et al., 2014). NLP-based methods that assist individuals in reframing negative thoughts, in-the-moment, may provide scaffolding that is easier to engage with and that could be made widely accessible.

Prior NLP research has developed methods for a range of text reframing tasks like sentiment and empathy rewriting (Reif et al., 2022; Sharma et al., 2021) and more recently, positive reframing (Ziems et al., 2022). However, little is known about how to develop cognitive reframing methods that automatically generate relatable, helpful and memorable reframed thoughts.

In this paper, we conduct a study of how language models can be used to reframe negative thoughts (Figure 1). We study ways in which a negative thought can be reframed, how LMs can be utilized to perform this reframing and what types of reframes are preferred by people who experience negative thoughts.

First, in collaboration with clinical psycholo-

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

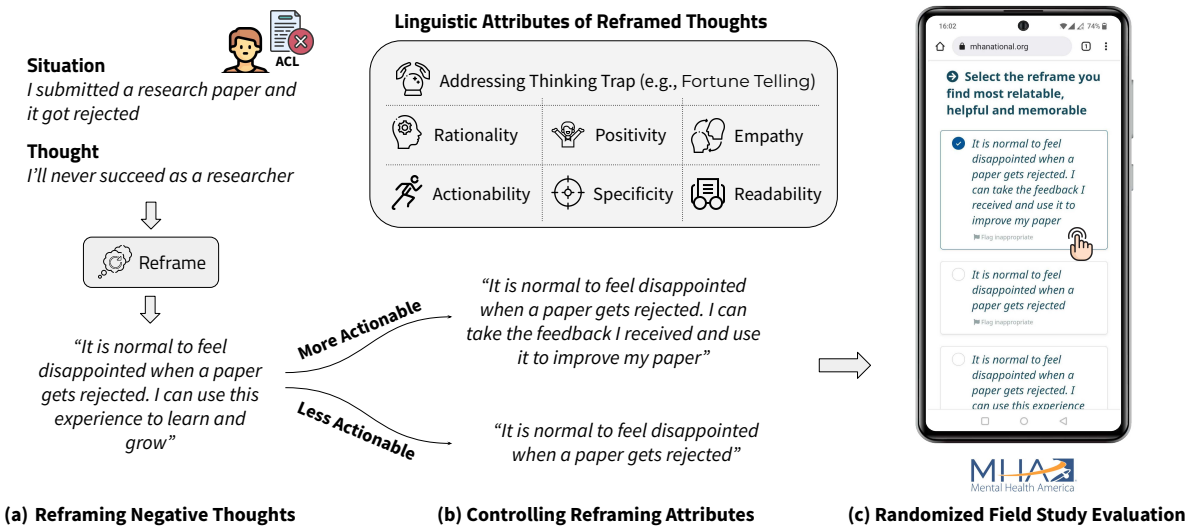


Figure 1: **(a)** We consider the task of reframing negative thoughts with different, more hopeful thoughts using LMs; **(b)** Different perspectives on a situation may result in different reframes. Here, we propose a framework of seven reframing attributes (see gray box). Given a reframed thought, we control each attribute (e.g., *actionability*) to generate reframes that score higher or lower on that attribute (e.g., *more or less actionable*); **(c)** We deploy this model on Mental Health America, a large U.S. national mental health website (bit.ly/changing-thoughts) and conduct a randomized field study with 2,067 participants. We suggest LM-generated reframes to MHA visitors and assess which reframing attributes are desirable and what constitutes a relatable, helpful and memorable reframe.

gists and mental health professionals, we develop a new conceptual framework for characterizing the ways in which a thought might be reframed. We synthesize the most prominent cognitive reframing processes used in therapy and define seven linguistic attributes of reframed thoughts: whether the reframe *addresses "thinking traps"* (faulty or distorted patterns of thinking), whether it is *rational*, *positive*, *empathic*, *actionable*, *specific* and *readable*. Building on prior research, we develop automated metrics to measure these attributes and establish construct validity by correlating them with judgements from mental health practitioners.

Next, to develop models for the cognitive reframing task, we collect and share[1] a dataset from mental health practitioners and clinical psychology graduate students. The dataset includes 600 situations and thoughts with expert-suggested reframes as well as annotations of the proposed reframing attributes. Using this dataset, we develop a retrieval-enhanced in-context learning method (Brown et al., 2020) to *generate* reframed thoughts and to *control* their linguistic attributes. We show that this method achieves the highest overlap with expert-suggested reframes and the highest relatability and helpfulness ratings based on evaluation from mental health experts, when compared to popular NLP baselines.

We investigate which reframing attributes are desirable and what constitutes a relatable, helpful and memorable reframe. In collaboration (and co-authorship) with mental health experts, and after appropriate ethical review, we deploy a month-long randomized field study on Mental Health America (MHA; a popular website that shares mental health resources and tools online), with 2,067 participants with informed consent. We ask MHA visitors to describe situations and negative thoughts they are experiencing and then suggest LM-generated reframed thoughts with varying linguistic attributes. We find that highly specific and highly empathic reframing is the most preferred and highly specific and actionable reframing is considered the most helpful and memorable. However, we find that reframes that are highly positive are less preferred. These findings provide key implications for cognitive reframing of negative thoughts and for the use of Human-LM interaction in this process.

2 Problem Definition and Goals

We work on the task of *Cognitive Reframing*. Given a situation $\mathbf{S_i}$ and a negative thought $\mathbf{T_i}$, the task is to generate a reframed thought $\mathbf{R_i}$.

Psychotherapy literature (Beck, 1976) highlights three desirable outcomes for a successful reframe:

[1]Our code and datasets are available at github.com/behavioral-data/Cognitive-Reframing.

(a) the reframed thought must be *relatable* to the individual, **(b)** it must *help* them overcome the negative thought and **(c)** it must be *memorable* the next time a similar negative thinking pattern emerges.

Here, we aim to understand what constitutes successful reframing and how language models can assist people in this process. Towards this goal, we characterize the linguistic attributes of reframed thoughts (§3), collect a dataset of situations, thoughts and reframes (§4), develop methods to generate reframes and to measure and control their attributes (§5; §6) and investigate which linguistic attributes are related to the reframing outcomes of relatability, helpfulness and memorability (§7).

3 Framework of Linguistic Attributes of Reframed Thoughts

We draw from clinical therapy practices and collaborate with mental health experts (some of whom are co-authors) to develop a framework of linguistic attributes of reframed thoughts. We illustrate these attributes with the following example for all reframes below – Situation: "*I submitted a research paper and it got rejected*;" Thought: "*I'll never succeed as a researcher*."

Addressing Thinking Traps. Negative thinking often falls into common patterns, called "*thinking traps*." Also called *cognitive distortions*, these include exaggerated and biased patterns of thinking which cause individuals to perceive reality inaccurately (Beck, 1976; Ding et al., 2022). Common thinking traps include: assuming what others think ("*Mind reading*"), thinking in extremes ("*All-or-nothing thinking*"), focusing on the worst-case scenario ("*Catastrophizing*"), trying to predict the future ("*Fortune telling*"), etc. See Appendix D for the full list.

A reframe may or may not directly address one or more of the thought's thinking traps. A reframe like "*I don't know what the future will bring*" directly addresses the thinking trap "*Fortune telling*," whereas a reframe like "*I will surely become a successful researcher*" does not address this thinking trap but rather continues to express it.

Rationality. Another strategy to reframe a thought is to reflect on evidence for and against it and reason about what these evidence imply (Beck, 1976). For example, the rejection of the paper is evidence of having the thought "*I'll never succeed as a researcher*." However, the evidence against this thought could be that acceptance or rejection of one paper does not make someone a failure, which may lead to a reframe "*Just getting one paper rejected doesn't define my failure*." A rational reframe is guided by such strong evidence whereas an irrational reframe is based on unrealistic assumptions.

Positivity. A reframe of a negative thought tries to emphasize the positive perspectives of the situation but different reframes may have different levels of positivity. An overly positive reframe like "*I'm going to win best paper awards for every paper from now on*" exaggerates the positive perspectives, which is likely to set the person up for disappointment rather than success (Dember and Penwell, 1980). On the other hand, a balanced response like "*I may or may not succeed, but I'll keep trying*" considers both positive and negative perspectives of the situation.

Empathy. It can be helpful to acknowledge the feelings caused by negative thoughts (Allen and Leary, 2010; Elliott et al., 2011). A reframe may express empathy or self-compassion by validating how one is feeling. E.g., "*It is okay to feel disappointed when a paper gets rejected*."

Actionability. To encourage pleasant emotions, one commonly used therapeutic approach is Behavioral Activation (Dimidjian et al., 2011; Burkhardt et al., 2021). This involves engaging in behaviors or actions that may help in overcoming negative thoughts. A reframe may suggest specific actions (e.g., "*I can take the feedback I received and use it to improve my paper*"), may not suggest specific actions but be actionable (e.g., "*I can use this experience to learn and grow*") or may not be actionable at all (e.g., "*I may or may not become a successful researcher*").

Specificity. A reframe may specifically address the situation and the thought (e.g., "*One paper rejection doesn't define my failure as a researcher*") or may be generic enough to be applicable to a wide range of negative situations and thoughts (e.g., "*I'm going to succeed*"). While a specific reframe may be more helpful in-the-moment, a generic reframe could be effective for recurring thoughts, which are frequently a result of the "core" beliefs that a person holds (Beck, 2005; David et al., 2009).

Readability. The linguistic reasoning capabilities of individuals may be different (e.g., across age groups or education levels) (Kaplan et al., 1995). Accordingly, a reframe may either be simple or complex to read (e.g., "*I'll do well in the future*" vs.

"*I'm resolute in my ambition to succeed*").

4 Data Collection

To facilitate computational methods for cognitive reframing, we collect a dataset of reframed thoughts, annotated with their linguistic attributes.

4.1 Curated Situations & Negative Thoughts

We start by curating data sources for situations and negative thoughts.

Thought Records Dataset (Burger et al., 2021). This dataset contains hypothetical and real-world situations, thoughts and emotional processes reported by crowdworkers on Amazon Mechanical Turk. We manually curate 180 pairs of diverse situations with negative thoughts from this dataset.

Mental Health America (MHA). Situations and thoughts from crowdworkers may not reflect the broad range of mental health challenges that people face in real-life. To incorporate more real-world situations and thoughts, we ran a survey on the MHA website (screening.mhanational.org). MHA visitors (who typically use the website for screening of mental illnesses) were asked to describe any negative thoughts and the associated situations they were struggling with. We manually curate 120 pairs of self-reported situations and thoughts to ensure broad coverage of relevant topics based on high diversity and manual filtering.

4.2 Annotation Task and Procedure

Reframing negative thoughts is a cognitively difficult process that requires practice and training, making crowdwork data collection approaches challenging. To ensure high-quality reframes and annotations, we recruit 15 current mental health practitioners and clinical psychology graduate students with significant practical experience in cognitive reframing.[2] For each (situation, thought) pair in our data source (§4.1), we ask them to (1) write two different reframed thoughts, (2) annotate the thinking traps addressed by each reframed thought and (3) compare the two reframes and choose the one that is more rational, more positive, more actionable, more empathic, more specific and more readable. In total, we collect 600 reframed thoughts with annotations on their linguistic attributes. We share this dataset publicly at github.com/behavioral-data/Cognitive-Reframing.

4.3 Ethics and Safety

Our data collection and randomized field study (§7) were designed and conducted after review of potential benefits and risks to participants in consultation and collaboration with mental health experts. Both studies were approved by the University of Washington's Institutional Review Board and informed participants about study purpose, risks and data collection. All participants were 18 or older, provided informed consent and were given access to a crisis hotline. We do not assess any clinical outcomes. See §10 for an extended discussion of ethical and safety considerations.

5 Method

We design automated metrics for measuring linguistic attributes (§5.1), develop methods to generate reframed thoughts (§5.2) and to control their attributes (§5.3).

5.1 Measuring Reframing Attributes

Addressing Thinking Traps. Given a situation $\mathbf{S_i}$, a negative thought $\mathbf{T_i}$ and a reframed thought $\mathbf{R_i}$, our goal is to identify the set of thinking traps addressed by the reframed thought. We approach this as a multi-label classification task, and fine-tune a GPT-3 model[3] on the expert-annotated thinking trap labels collected in §4.2.

Rationality. Rationality is the quality of being guided by reasons (Damielson et al., 2004). Here, we operationalize rationality of a reframed thought $\mathbf{R_i}$ as its *reasoning strength* and ask the following two questions: (1) What might be the reasoning behind $\mathbf{R_i}$?; (2) Are the reasons sound? To understand the reasoning behind $\mathbf{R_i}$, we develop an *abductive explanation* based method (Peirce, 1974; Bhagavatula et al., 2020; Jung et al., 2022). For a given ($\mathbf{S_i}$, $\mathbf{T_i}$), we use a language model to generate (a) the most plausible explanations that *support* $\mathbf{R_i}$ and (b) the most plausible explanations that *refute* it. Moreover, to check if the explanations are sound, we recursively generate explanations behind the explanations to test their reasoning strength (Appendix E). Let $sup(\cdot)$ be a generator function that generates explanation *supporting* a reframe and let

[2] For recruitment, we advertised our study through university mailing lists and newsletter of a mental health organization. Recruited experts were paid @ 37.5 USD / hr.

[3] We use `text-davinci-003` as our GPT-3 model for all experiments in this paper.

$ref(\cdot)$ be a generator function that generates explanation *refuting* a reframe. Then, we recursively define reasoning strength $RS(\mathbf{S_i}, \mathbf{T_i}, \mathbf{R_i})$ as

$$\big(\mathrm{P}(\mathbf{R_i} = \texttt{sound} | \mathbf{S_i}, \mathbf{T_i}) \times \mathbb{E}_{r \sim sup(\cdot)}[RS(\mathbf{S_i}, \mathbf{T_i}, r)]\big)$$
$$- \big(\mathrm{P}(\mathbf{R_i} = \texttt{flawed} | \mathbf{S_i}, \mathbf{T_i}) \times \mathbb{E}_{r \sim ref(\cdot)}[RS(\mathbf{S_i}, \mathbf{T_i}, r)]\big)$$

To design the explanation generator functions, $sup(\cdot)$ and $ref(\cdot)$, we leverage in-context learning (Brown et al., 2020). In collaboration with mental health experts, we design 10 demonstration examples of situations, thoughts and reframed thoughts with explanations that support ("*This reframed thought is sound because...*") and refute ("*This reframed thought is flawed because...*") a particular reframe. We use these examples to prompt GPT-3. Moreover, to estimate the probabilities $\mathrm{P}(\mathbf{R_i} = \texttt{sound})$ and $\mathrm{P}(\mathbf{R_i} = \texttt{flawed})$, we use the token probability of generating "*sound*" and "*flawed*" respectively, given $\mathbf{S_i}, \mathbf{T_i}, \mathbf{R_i}$ and the text "*This reframed thought is*" as input to GPT-3.[4]

Positivity. To measure the positivity of the generated reframed thought, we use a RoBERTa-based sentiment classifier fine-tuned on the TweetEval benchmark (Barbieri et al., 2020).

Empathy. To measure empathy, we build upon the empathy classification model presented in Sharma et al. (2020b). This RoBERTa-based model leverages a theoretically-grounded framework of empathy consisting of three empathy communication mechanisms (emotional reactions, interpretations and explorations) and predicts empathy levels in mental health conversations on a scale from 0 to 6. Here, we further fine-tune this model on the domain of reframed thoughts through a manually labeled dataset of 300 reframed thoughts with empathy labels (labeled by one author with expertise in empathy in mental health context).

Actionability. To measure actionability, we hypothesize that an actionable reframe is one that either (1) *suggests a concrete action* or (2) does not suggest a concrete action but is *easy to act upon*.

We cast action concreteness as a binary classification task: given reframe $\mathbf{R_i}$, predict $contains\_action(\mathbf{R_i}) \in \{0, 1\}$. We make few-shot predictions by prompting GPT-3 with 10 examples of reframed thoughts paired with actionability ratings from §4.2 (details in Appendix A.1).

To determine the ease with which $\mathbf{R_i}$ can be acted upon, we examine the next set of actions entailed by $\mathbf{R_i}$. Our hypothesis is that a *diverse* next action set may indicate ambiguity which might be less actionable, whereas a *coherent* next action set may indicate clarity which might be more actionable. Here, we instruct GPT-3 to generate $k = 5$ next action candidates given a reframed thought (instruction prompting; zero-shot). We compute the next action coherence — denoted $next\_action\_coherence(\mathbf{R_i})$ — by embedding each of the k action candidates using RoBERTa (Liu et al., 2019) and computing the average pairwise cosine similarity. Higher similarity indicates greater coherence among the possible next actions. Our overall actionability measurement is defined as $contains\_action(\mathbf{R_i}) + next\_action\_coherence(\mathbf{R_i})$.

Specificity. Following prior work (Xu et al., 2018; Sharma et al., 2021), we measure specificity using sentence embedding similarity between the reframed thought $\mathbf{R_i}$ and the concatenation of the situation $\mathbf{S_i}$ and the thought $\mathbf{T_i}$ (using RoBERTa embeddings (Liu et al., 2019)).

Readability. We employ the commonly used Coleman-Liau Index (CLI) metric (Coleman and Liau, 1975) which assesses readability based on the character and word structure within a sentence. The Coleman-Liau Index is calculated as $0.0588L - 0.296S - 15.8$, where L: average number of letters per 100 words; S is the average number of sentences per 100 words.

5.2 Reframe Generation

In-context learning methods can learn to generalize NLP tasks from a handful of examples (*few-shot learning*) or from hand-written instructions alone (*instruction prompting*) (Brown et al., 2020). However, through a qualitative analysis of 100 manually written situations and thoughts, we found that a simple in-context learning method with a fixed set of examples often failed to appropriately reframe situations and thoughts for which no relevant in-context examples were provided (e.g., someone with anxiety having "*racing thoughts*").

To appropriately reframe thoughts related to a range of situations and thoughts, we develop a retrieval-based in-context learning method (Liu et al., 2022b). For each situation $\mathbf{S_i}$ and negative thought $\mathbf{T_i}$, we retrieve k-similar examples from our dataset (§4). We first encode situations and thoughts using RoBERTa embeddings. Then, we retrieve k examples, $\{(s_1, t_1), ..., (s_k, t_k)\}$,

[4] We experimented with different alternatives for "*sound*" and "*flawed*" and observed similar results.

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. *:$p < 0.05$; **:$p < 0.001$; ***:$p < 10^{-5}$.

from our dataset based on the top-k values of $cosine\_sim(concat(s,t), concat(\mathbf{S_i}, \mathbf{T_i}))$. We choose $k = 5$ (Appendix A.3).

5.3 Controlling Linguistic Attributes of Generated Reframes

While our proposed method allows us to generate a single reframe, it does not directly give us control over its linguistic attributes beyond mimicking the retrieved examples (§3). Here, we intend to vary the linguistic attributes of the reframes.

A reframed thought may or may not address a thinking trap in the original thought $\mathbf{T_i}$. Here, we generate two reframes $\mathbf{R_i}^{(tt,\mathrm{Y})}$ and $\mathbf{R_i}^{(tt,\mathrm{N})}$, one that addresses the thinking traps in $\mathbf{T_i}$ and another that does not address it.[5] We extract two separate sets of in-context examples from our dataset – those that address at least one thinking trap and those that do not (as collected in §4). We use those examples to prompt GPT-3 to generate $\mathbf{R_i}^{(tt,\mathrm{Y})}$ and $\mathbf{R_i}^{(tt,\mathrm{N})}$.

Moreover, a reframed thought may have high or low rationality, positivity, empathy, actionability, specificity and readability values. For these six attributes, given a reframe $\mathbf{R_i}$ and a linguistic attribute a, we generate two *additional* reframes $\mathbf{R_i}^{(a,\mathrm{H})}$ and $\mathbf{R_i}^{(a,\mathrm{L})}$, one that scores higher on attribute a and another that scores lower on it (e.g., higher or lower actionability). To accomplish this, recall that each (situation, thought) pair from §4.2 is annotated with two reframes and that the reframes are compared along each linguistic attribute. For a human-annotated instance j, let $\mathbf{R_j}^{*(a,\mathrm{H})}$ and $\mathbf{R_j}^{*(a,\mathrm{L})}$ be the reframes judged to be high and low on attribute a, respectively. To generate $\mathbf{R_i}^{(a,\mathrm{H})}$ from $\mathbf{R_i}$, we prompt GPT-3 with in-context examples $\{\mathbf{R_j}^{*(a,\mathrm{L})} \rightarrow \mathbf{R_j}^{*(a,\mathrm{H})}\}_{j=1}^{k}$, using $k = 5$.

[5] If a thought exhibits multiple thinking traps, we check if the reframe addresses at least one of them.

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.

Similarly, to generate $\mathbf{R_i}^{(a,\mathrm{L})}$ from $\mathbf{R_i}$, we prompt GPT-3 with examples $\{\mathbf{R_j}^{*(a,\mathrm{H})} \rightarrow \mathbf{R_j}^{*(a,\mathrm{L})}\}_{j=1}^{k}$.

6 Experiments and Results

We assess the construct validity of proposed linguistic attributes (§6.1) and evaluate the performance of the reframe generation model (§6.2).

6.1 Construct Validity of Linguistic Attributes

We validate our proposed linguistic attribute measures by correlating them with the human judgments of mental health experts, as obtained in §4.2. We find a strong Pearson correlation for addressing thinking traps (0.680***) and actionability (0.647***), a moderate correlation for rationality (0.448**), positivity (0.550***), empathy (0.575***) and specificity (0.427**), and a weak correlation for readability (0.331*) (Table 1).[6]

6.2 Reframe Generation Performance

We use both automatic and human evaluation to assess the performance of our proposed reframe generation model as developed in §5.2.

Experimental Setup. We use top-p sampling with $p = 0.6$ for text generation (Holtzman et al., 2020). We split the 600 expert-annotated examples (§4) into train and test using a 70:30 split.

Baselines. **(1)** *Retrieval Only* – For a test input, we retrieve the training set example with the highest cosine similarity based on RoBERTa embeddings; **(2)** *Positive Reframing* – We reuse the BART-based positive reframing model from Ziems et al. (2022); **(3)** *DialoGPT* – GPT-2 adapted to dialogue (Zhang et al., 2020); **(4)** *T5* – Text-to-text transfer LM (Raffel et al., 2020);[7] **(5)** *GPT-3 Only* – We randomly

[6] *:$p < 0.05$;**:$p < 0.001$;***:$p < 10^{-5}$

[7] Training DialoGPT and T5 on 600 samples only may be challenging. Here, we use an *overgeneration* strategy –

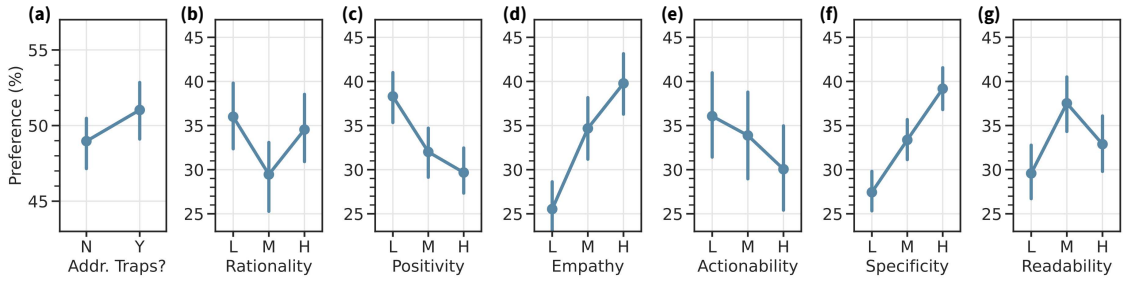


Figure 2: **Which linguistic attributes of reframed thoughts do people prefer?** For a given situation and thought from MHA visitors, we show them multiple LM-generated reframes with variance across a randomly selected attribute (e.g., low, medium and high actionability). We find that highly empathic 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.

retrieve 5 examples from our training set and use them to prompt GPT-3.

Automatic Evaluation. We examine the semantic similarity between the model outputs and the ground truth reframings in the above-created test split. We use BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-L (Lin, 2004) and the BERTScore (Zhang et al., 2019b) metrics. We find that our proposed model has an 11.2% higher BLEU score and 9.7% higher ROUGE scores than the next best-performing baselines – GPT-3 Only and Positive Reframing (Table 2).

Human Evaluation. We assess the two key reframing outcome metrics of *relatability* (how relatable would a reframed thought be) and *helpfulness* (how helpful would a reframed thought be in overcoming negative thoughts). We recruit three mental health practitioners. We ask them to rate the models' outputs on test set examples based on their reliability and helpfulness on a 1 to 5 scale. We find that our proposed model achieves the highest relatability and helpfulness ratings (Table 2). Surprisingly, the Positive Reframing method showed the least helpfulness and low relatability, indicating that just reframing negative thoughts based on positivity may not be highly relatable and helpful.

7 Randomized Field Study on a Large Mental Health Platform

Next, we deploy our model on a large mental health platform (§7.1) and study what types of reframes do people prefer (§7.2) and what characterizes relatable, helpful and memorable reframes (§7.3).

Starting from our collected data samples, we utilize the pattern replication capabilities of GPT-3 to generate 10,000 more examples, similar to Liu et al. (2022a).

7.1 Model Deployment

We try to understand how our proposed cognitive reframing model may assist people who experience negative thoughts. After careful assessment of ethical and safety considerations, active collaboration with mental health experts and clinical psychologists (some of whom are co-authors) and IRB approval, we deploy our model on Mental Health America (MHA), a large mental health website that provides mental health resources and tools to millions of users (bit.ly/changing-thoughts). We conduct a month-long randomized field study with 2,067 MHA visitors as participants. After choosing to use our model and after consenting to participate, MHA visitors described their situation and the thoughts they were struggling with. Next, they were shown multiple model-generated reframed thoughts in random order, asked to select the reframed thought they find most relatable, helpful and memorable and finally evaluate the selected reframed thought based on relatability, helpfulness and memorability (See Appendix F).

7.2 What types of reframed thoughts do people prefer?

To understand which reframing attributes people prefer, we suggest multiple LM-generated reframes which vary across our attribute values. Given a situation and thought, we start by generating one reframed thought using our model. Next, we randomly select an attribute (e.g., actionability) and vary the first reframe based on it (e.g., to generate two additional reframes with higher or lower actionability) using our proposed controllable text generation method (§5.3). Figure 2 reveals key differences between the linguistic attributes of reframes that people select and prefer:

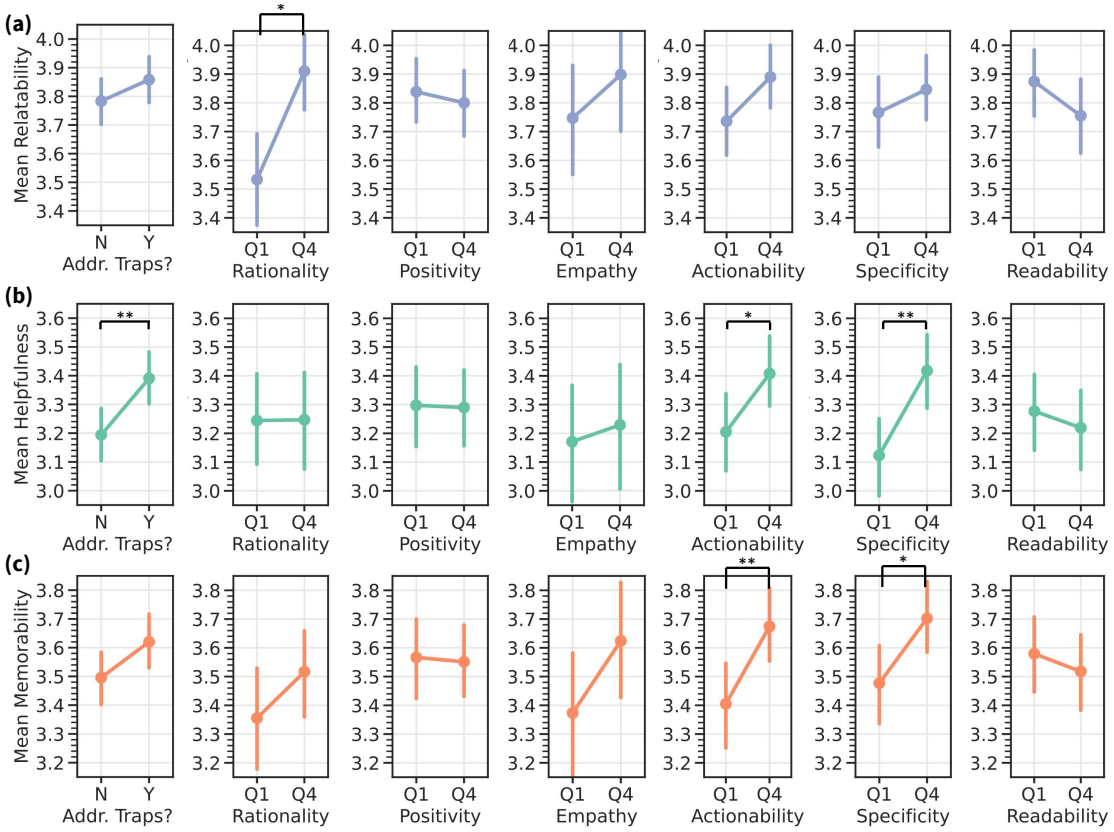


Figure 3: **Which linguistic attributes are associated with desired cognitive reframing outcomes?** For a given situation and thought, we show one LM-generated reframe to MHA participants and ask them to rate it on relatability, helpfulness and memorability on a 1 to 5 scale. For each linguistic attribute, 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. *:$p < 0.05$;**:$p < 0.01$.

(1) Highly empathic and specific reframings are preferred more. We find that highly empathic reframes are preferred 55.7% more frequently than reframes with lower empathy (39.7% vs. 25.5%; $p < 10^{-5}$); highly specific reframes are preferred 43.1% more frequently than reframes with lower specificity (39.2% vs. 27.4%; $p < 10^{-5}$). Prior work has shown the importance of empathy and less "templated" responses in mental health support conversations (Sharma et al., 2020b; Althoff et al., 2016). Here, we show that empathy and specificity of LM-generated reframes may support people in reframing negative thoughts.

(2) Overly positive reframes are preferred less. On the other hand, reframes with high positivity are preferred 22.7% less frequently than reframes with lower positivity (29.6% vs. 38.3%; $p < 10^{-5}$). This may be because adopting an overly positive reframed thought may be challenging for individuals who are already experiencing emotionally triggering negative thoughts (Dember and Penwell, 1980).

Participants also prefer medium-readability reframes over very simple or very complex reframes, perhaps because their language is balanced for a wider audience.

7.3 How do the linguistic attributes of reframed thoughts relate to the desired outcomes of cognitive reframing?

We assess what characterizes a reframe that is relatable, helpful and memorable. We show a *single* model-generated reframe to the participants and ask them to rate it on a 5-point Likert scale (Likert, 1932) with regards to the three outcome measures (1: Strongly Disagree; 5: Strongly Agree). We do not provide participants in this experiment with a choice of multiple reframes to avoid any selection effects (§7.2). Figure 3 offers key insights on which attributes of reframed thoughts are related to different desired outcomes:

(1) Reframes that are more rational are more *relatable*. We find that reframes that have higher

rationality are 10.8% more relatable than lower rationality reframes (3.91 vs. 3.53; $p < 0.05$). This may be because higher rationality reframes, by definition, are more likely to be based on reasons and are less likely to make unrealistic assumptions, making them easier to relate to.

(2) Reframes that address thinking traps and are more actionable and specific are more *helpful*. Reframes that address thinking traps are 6.3% more helpful than reframes that do not address them (3.39 vs. 3.19; $p < 0.01$). Such reframes specifically challenge the cognitive biases in thinking patterns (e.g., "*Fortune Telling*"; Appendix D), which has shown to be more effective in dealing with negative thoughts in psychotherapy research (Beck, 1976; Burns, 1980). Moreover, we find that reframes with higher actionability are 6.6% more helpful than lower actionability reframes (3.41 vs. 3.20; $p < 0.05$) and reframes with higher specificity are 9.6% more helpful than lower specificity reframes (3.42 vs. 3.12; $p < 0.01$).

(3) Reframes that are more actionable and more specific are more *memorable*. We find that reframes with higher actionability are 7.9% more memorable than lower actionability reframes (3.67 vs. 3.40; $p < 0.01$) and reframes with higher specificity are 6.3% more memorable than lower specificity reframes (3.70 vs. 3.48; $p < 0.05$).

8 Related Work

Several Human-LM interaction tools for mental health assist support providers, e.g., clinicians (Tanana et al., 2019; Shen et al., 2020) or peers (Sharma et al., 2023). Our work provides insights on how Human-LM interaction may directly support people struggling with mental health challenges through cognitive reframing. Computational work on cognitive reframing has relied on small-scale crowdsourcing studies (Smith et al., 2021; Morris et al., 2015). Our work develops scalable methods for cognitive reframing and conducts a randomized field study on a large mental health platform. Prior text reframing research has developed methods for related tasks including style, sentiment, politeness and empathy transfer (Reif et al., 2022; Madaan et al., 2020; Sharma et al., 2021) as well as positive reframing (Ziems et al., 2022). Our work develops text-reframing methods for cognitive reframing and demonstrates that linguistic attributes of addressing thinking traps, rationality, actionability, specificity and readability are critical to high-quality reframes. More broadly, our work relates to the growing body of research in NLP for mental health and psychological well-being (Althoff et al., 2016; Sharma and De Choudhury, 2018; Gaur et al., 2019; Lee et al., 2019; Miner et al., 2019; Pendse et al., 2019; Pérez-Rosas et al., 2019; Pruksachatkun et al., 2019; Yang et al., 2019; Zhang et al., 2019a; Jaidka et al., 2020; Saha and Sharma, 2020; Sharma et al., 2020a,b; Wadden et al., 2021; Welch et al., 2020; Zhang and Danescu-Niculescu-Mizil, 2020; Lahnala et al., 2021; Lin et al., 2022; Naseem et al., 2022; Pérez-Rosas et al., 2022; Shah et al., 2022; Shen et al., 2022; Stewart et al., 2023).

9 Conclusion

In this paper, we conducted a study of how Human-Language Model Interaction may support humans in the cognitive reframing of negative thoughts. We define a framework of seven linguistic attributes of cognitive reframing, develop automatic metrics to measure these attributes and validate their measurements with mental health experts. We collect and share a dataset of 600 situations, thoughts and reframes from mental health experts and use it to train a retrieval-enhanced in-context learning model based on GPT-3. We deploy this model on the Mental Health America website and conduct a randomized field study with 2,067 participants. We find that people struggling with negative thoughts prefer reframes that are highly empathic or specific, but do not prefer reframes that are highly positive.

10 Ethics Statement

Intervention in high-risk settings such as mental health necessitates ethical considerations related to safety, privacy and bias. There is a possibility that, in attempting to assist, AI may have the opposite effect on people struggling with mental health challenges. Here, in active collaboration and consultation with mental health professionals and clinical psychologists, we took several measures to minimize these risks.

IRB Approval. We obtained approval from the University of Washington's Institutional Review Board for both our data collection (IRB ID STUDY00015882) as well as the randomized field study (IRB ID STUDY00016783). Our organization requires all research personnel who conduct human subjects research to complete human subjects protection training using the online CITI

course. The graduate students conducting these studies were certified by our IRB.

Informed Consent from Participants. We obtained informed consent from all participants in our randomized field study (Appendix H). All participants were 18 years of age and older. Participants were informed that they will be interacting with an AI-based model that automatically generates reframed thoughts and is not monitored by a human. Also, they were informed about the possibility that some of the generated content may be upsetting or disturbing.

Crisis Resources. We made it very explicit that the model should not be used as a "cry for help" outlet and should not be used in cases of suicidal ideation and self-harm. Also, we provided two crisis resources – Crisis Text Line (crisistextline.org) and 988 Suicide and Crisis Lifeline (988lifeline.org) – to our participants at the start of the study.

Safety Measures. To minimize harmful LM-generated reframings, we filtered out any response that contained suicidal ideation or self-harm-related words or phrases. For this, we created a list of 50 regular expressions (e.g., to identify phrases like "*feeling suicidal*", "*wish to die*", "*harm myself*") using suicidal risk assessment lexicons such as Gaur et al. (2019). An LM-generated response that matched any of the regular expressions was filtered out and not shown to the participants. Also, participants were given an option to flag inappropriate reframing suggestions through a "Flag inappropriate" button (Appendix C).

Privacy. We did not collect any privately identifiable information in our randomized field study and removed any user identifiers before conducting our data analysis. All research data was stored within a separate secure computing environment and only trained research personnel were provided access to data. The situations and thoughts collected in §4.1 went through an anonymization process, where we manually removed any user identifiers and replaced any specific identifiable information including locations, names, etc. with their more general version, following Matthews et al. (2017).

11 Limitations

We conducted our randomized field study on a single platform (Mental Health America) and in a single language (English). However, MHA is a particularly popular source for mental health resources with over ten million yearly visitors.

In addition, we note that a range of socio-cultural factors might influence how negative thoughts should be reframed and how LMs assisting this process should be developed. Conducting studies on specific communities, including underrepresented communities and minorities, was beyond the scope of this research. Ensuring equitable access of these tools and adapting them to various socio-cultural contexts requires further investigation.

Not all cognitive reframing implementations elicit situations, but we believed it was essential for making the reframe personally relatable. In the future, when an individual uses the system for multiple situations and thoughts, it would be interesting to study how their context can be learned more effectively over time. Due to privacy concerns, we presently do not gather information to link multiple sessions. However, with appropriate ethical considerations and user consent, this approach may be beneficial.

Our focus in this paper was primarily on creating an intervention that is effective in-the-moment. This was motivated by recent clinical psychology research that suggests that such single-session, in-the-moment interventions can lead to significant positive long-term mental health outcomes (Schleider et al., 2022). To integrate a partial longer-term perspective, we assessed the memorability of a reframe, which may be essential for future utility. Nevertheless, evaluating long-term outcomes is critical and forms an important future research direction. Finally, we emphasize that our study does not investigate short-term or long-term clinical outcomes.

Acknowledgements

We are grateful to the mental health practitioners and clinical psychology graduate students for data annotation, as well as the MHA visitors for participating in our field study. We thank the anonymous reviewers and the UW Behavioral Data Science Group members for their suggestions and feedback. We also thank Justin Evans for their assistance in model deployment, Xiang Lorraine Li for their input on data collection and Sebastin Santy for their input on the tool interface. T.A., A.S. and I.W.L. were supported in part by NSF grant IIS-1901386, NSF CAREER IIS-2142794, NSF grant CNS-2025022, NIH grant R01MH125179, Bill & Melinda Gates Foundation (INV-004841), the Of-

fice of Naval Research (#N00014-21-1-2154), a Microsoft AI for Accessibility grant, and a Garvey Institute Innovation grant.

References

Ashley Batts Allen and Mark R Leary. 2010. Self-compassion, stress, and coping. *Social and personality psychology compass*.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *EMNLP Findings*.

Aaron T Beck. 1976. *Cognitive therapy and the emotional disorders.* International Universities Press.

Judith S Beck. 2005. *Cognitive therapy for challenging problems: What to do when the basics don't work*. Guilford Press.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *ICLR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Franziska Burger, Mark A Neerincx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: Extracting schemas from thought records. *PloS one*.

Hannah A Burkhardt, George S Alexopoulos, Michael D Pullmann, Thomas D Hull, Patricia A Areán, and Trevor Cohen. 2021. Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. *JMIR*.

David D Burns. 1980. Feeling good: Thenew mood therapy. *New York*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*.

Peter Damielson, Robert Audi, Cristina Bicchieri, et al. 2004. *The Oxford handbook of rationality*. Oxford University Press, USA.

Daniel David, Steven Jay Lynn, and Albert Ellis. 2009. *Rational and irrational beliefs: Research, theory, and clinical practice*. Oxford University Press.

William N Dember and Larry Penwell. 1980. Happiness, depression, and the pollyanna principle. *Bulletin of the Psychonomic Society*.

Sona Dimidjian, Manuel Barrera Jr, Christopher Martell, Ricardo F Muñoz, and Peter M Lewinsohn. 2011. The origins and current status of behavioral activation treatments for depression. *Annual review of clinical psychology*.

Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. 2011. Empathy. *Psychotherapy*.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *WWW*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Kokil Jaidka, Niyati Chhaya, Saran Mumick, Matthew Killingsworth, Alon Halevy, and Lyle Ungar. 2020. Beyond positive emotion: Deconstructing happy moments based on writing prompts. In *ICWSM*.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *EMNLP*.

Carole A Kaplan, Anne E Thompson, and Sheila M Searson. 1995. Cognitive behaviour therapy in children and adolescents. *Archives of disease in childhood*.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *ACL-IJCNLP Findings*.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathleen McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered mental health stigma in masked language models. In *EMNLP*.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022a. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Poczós, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *ACL*.

Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *CHI*.

Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in psychiatry*.

Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *JMIR*.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *WWW*.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*. Harvard University Press.

Sachin R Pendse, Kate Niederhoffer, and Amit Sharma. 2019. Cross-cultural differences in the use of online mental health support forums. *CSCW*.

Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, et al. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *EMNLP*.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *ACL*.

Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *CHI*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *ACL*.

Koustuv Saha and Amit Sharma. 2020. Causal factors of effective psychosocial outcomes in online mental health communities. In *ICWSM*.

Jessica L Schleider, Michael C Mullarkey, Kathryn R Fox, Mallory L Dobias, Akash Shroff, Erica A Hart, and Chantelle A Roulston. 2022. A randomized trial of online single-session interventions for adolescent depression during covid-19. *Nature Human Behaviour*.

Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *CSCW*.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *ICWSM*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW*.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP*.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *CHI*.

Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *ACL*.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *SIGDIAL*.

Amy E Sickel, Jason D Seacat, and Nina A Nabors. 2014. Mental health stigma update: A review of consequences. *Advances in Mental Health*.

C Estelle Smith, William Lane, Hannah Miller Hillberg, Daniel Kluver, Loren Terveen, and Svetlana Yarosh. 2021. Effective strategies for crowd-powered cognitive reappraisal systems: A field deployment of the flip* doubt web application for mental health. *CSCW*.

Ian Stewart, Charles Welch, Lawrence An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2023. Expressive interviewing agents to support health-related behavior change: A study of covid-19 behaviors. *JMIR formative research*.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *JMIR*.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *ICWSM*.

Charles Welch, Allison Lahnala, Verónica Pérez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *ACL*.

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *CHI*.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *ACL*.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019a. Finding your voice: The linguistic development of mental health counselors. In *ACL*.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. In *ACL*.

A Method

A.1 Linguistic Attributes of Reframed Thoughts

We provide additional detail on the approaches described in §5.1.

Actionability. As described in §5.1, we measure actionability using: $contains\_action(\mathbf{R_i})$, and $next\_action\_coherence(\mathbf{R_i})$.

For $contains\_action(\mathbf{R_i})$, our few-shot in-context learning approach proceeds as follows. Using the reframed thoughts that were annotated as high or low actionable in our collected data (§4.2), we manually create 10 demonstration examples. If a reframed thought contains an action, we ask GPT-3 to extract the action from it. Otherwise, we ask it to generate the text "*No Action*". Appendix A.2 shows examples. We then use these 10 demonstrations as in-context examples, followed by the reframe $\mathbf{R_i}$ which we aim to classify. If GPT-3 predicts an action for $\mathbf{R_i}$, we assign $contains\_action(\mathbf{R_i}) = 1$; else we assign 0.

For $next\_action\_coherence(\mathbf{R_i})$, we instruct GPT-3 to generate $k = 5$ possible next actions given a reframed thought. Given ($\mathbf{S_i}$, $\mathbf{T_i}$, $\mathbf{R_i}$), let $\mathbf{A_i} = a_{i1}, a_{i2}, ..., a_{ik}$ be the generated set of next actions. Let $emb(\cdot)$ denote RoBERTa embeddings. Then, we define $next\_action\_coherence(\mathbf{R_i})$ as the average cosine similarity between $emb(a_i)$ and $emb(a_j)$ for all $a_i, a_j \in \mathbf{A_i}$.

A.2 Action Generation Prompt

We use the following prompt template for extracting actions through GPT-3:

> **Statement**: "My bank card could be in many different places and I want to check them first before making any conclusions"
> **Proposed Action**: "Check bank card."
>
> **Statement**: "I cancelled that trip because I had to. It hurts to have done so but it was the right thing"
> **Proposed Action**: None

Also, we use the following instruction prompt for generating the next set of actions through GPT-3: "*Suggest 5 actions that the person could take based on the following statement:*"

A.3 Hyperparameter Choices for our Proposed Retrieval-Enhanced In-Context Learning Method

For the number of examples to retrieve, we experimented with $k = 1, 5, 10$ and 20 and found $k = 5$ to generate the most effective reframed thoughts, based on a qualitative assessment of 100 manually written situations and thoughts.

B Reproducibility

Codes and datasets created in the paper can be found at https://github.com/behavioral-data/Cognitive-Reframing under an academic, attribution-only license. The use of existing artifacts was consistent with their intended use. For GPT-3 based models, we will use the OpenAI library. For other deep learning models, we train and them on two NVIDIA Titan RTX GPUs. We use the evaluate python library (pypi.org/project/evaluate) for measuring BLEU and ROUGE scores and scipy for statistical tests.

C Flagged Reframes

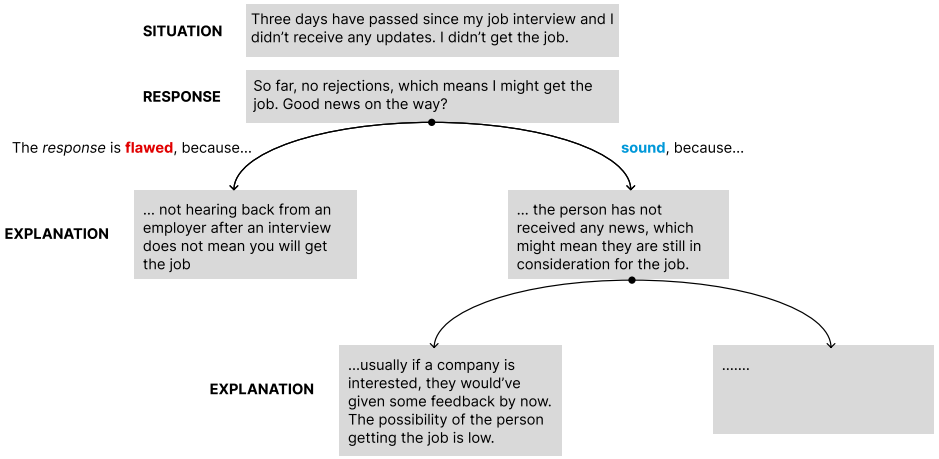
There were 32 reframing suggestions out of 5,760 which were flagged (0.56%). 19 of them were generic (59%). 5 of them made incorrect assumptions about the person's situation (16%). And 8 of them may not have been relatable to the person (25%). Importantly, we did not find any flagged reframes that were harmful or unsafe, which is critical in these scenarios. In future, exploring ways to create more personalized reframes could help avoid generic, assumptive or less relatable reframes.

D List of Thinking Traps

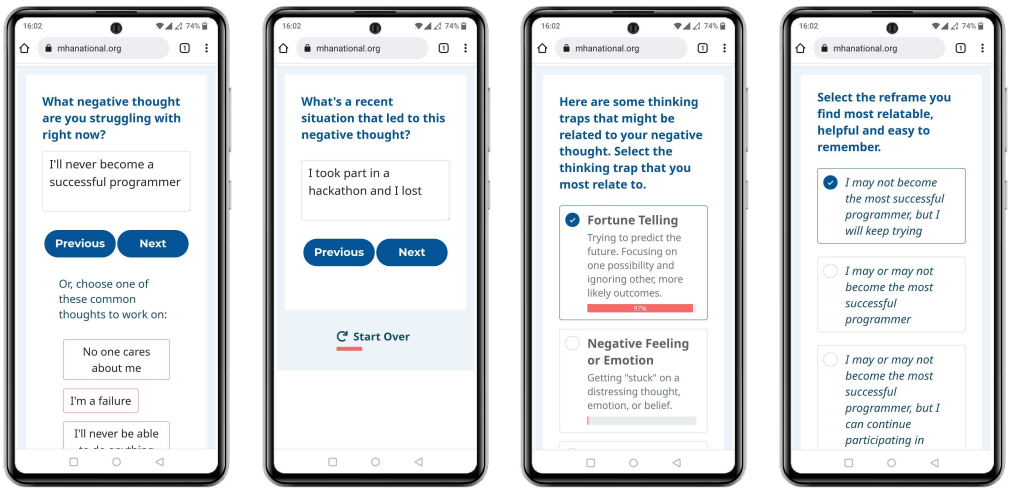
Thinking Traps	Description	Example
All-or-Nothing Thinking	Thinking in extremes.	If it isn't perfect, I failed. There's no such thing as "good enough".
Overgeneralizing	Jumping to conclusions based on one experience.	They didn't text me back. Nobody ever texts me back.
Labeling	Defining a person based on one action or characteristic.	I said something embarrassing. I'm such a loser.
Fortune Telling	Trying to predict the future. Focusing on one possibility and ignoring other, more likely outcomes.	I'm late for the meeting. I'll make a fool of myself.
Mind Reading	Assuming you know what someone else is thinking.	She didn't say hello. She must be mad at me.
Emotional Reasoning	Treating your feelings like facts.	I woke up feeling anxious. I just know something bad is going to happen today.
Should Statements	Setting unrealistic expectations for yourself.	I shouldn't need to ask for help. I should be independent.
Personalizing	Taking things personally or making them about you.	He's quiet today. I wonder what I did wrong.
Disqualifying the Positive	When something good happens, you ignore it or think it doesn't count.	I only won because I got lucky.
Catastrophizing	Focusing on the worst-case scenario.	My boss asked if I had a few minutes to talk. I'm going to get fired!
Comparing and Despairing	Comparing your worst to someone else's best.	My niece's birthday party had twice the amount of people
Blaming	Giving away your own power to other people.	It's not my fault I yelled. You made me angry!
Negative Feeling or Emotion	Getting "stuck" on a distressing thought, emotion, or belief.	I am feeling lonely.

E Example Illustrating Our Rationality Measurement

Figure 4: To measure reasoning strength, we generate two explanations for each reframe – one for why it might be sound; another for why it may be flawed. To check if the explanations themselves are well-reasoned, we recursively generate explanations for the explanations. Here, we choose a recursive tree depth of 3. Also, at every step, we generate three explanations in favour of a reframe and three explanations against it.



F Randomized Field-Study Interface



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

Figure 5: Illustration of the interface used for our randomized field-study: (**a**) Participant starts by writing the negative thought they are struggling with in the moment; **(b)** We ask the participant to describe a recent situation that may have led to their thought; **(c)** An AI model identifies possible cognitive distortion(s) in the thought. Participant selects the cognitive distortions that they most relate to; **(d)** An AI model generates and suggests three different reframed thoughts that may help overcome negative thinking and the associated cognitive distortion. Participant selects the reframe they find the most relatable, helpful and memorable. Some of the instructions provided to the participants, including informed consent and evaluation, have been omitted from this illustration for brevity.

G Data Collection Instructions

Figure 6: Instructions shown during data collection with mental health experts. Continued on the next page (1/3).

Cognitive Restructuring

Study Goals

The goal of this study is to collect a dataset for cognitive restructuring.

Definitions

Situation Anything that happens to the person or the circumstance that the person finds themselves in (e.g., "*My boss walked past me the hallway without saying hello*")

Thought What goes through the person's mind in the situation (e.g., "*Why are they angry with me?*").

Thinking Traps Thinking traps, also called *cognitive distortions*, are **exaggerated, biased, and irrational** thoughts which cause individuals to perceive reality inaccurately.

Thinking Trap Categories Categories of thinking traps include assuming what others think ("Mind Reading"), thinking in extremes ("All-or-nothing thinking"), focusing on the worst-case scenario ("Catastrophizing"), focusing only on the bad ("Disqualifying the positive"), etc.

Example Thinking Trap

Situation: *My boss walked past me the hallway without saying hello*

Thought: *Why are they angry with me?*

Thinking Trap: `Mind Reading`

Figure 7: Instructions shown during data collection with mental health experts. Continued on the next page (2/3).

Here, I'm reading my boss's mind and assuming that they are upset with me. I can't figure this out unless I ask them.

Thought Response A thought response is self-talk (conversation with oneself) that tries to challenge the thinking trap in the original thought

Example Thought Responses

Situation: *My boss walked past me the hallway without saying hello.*

Thought: *Why are they angry with me?*

Response 1: *I have no way of figuring out what they might be thinking. Maybe they had a lot on her mind*

Response 2: *They are the most wonderful person I know. They must not have noticed me.*

Response 3: *They might be mad at me, but atleast they didn't say anything.*

Cognitive Restructuring Cognitive Restructuring is a process that helps people notice thinking traps in their thoughts and respond rationally to them.

Study Steps

In this study, you will perform 20 cognitive restructuring tasks. In each task, you will be shown a situation and a thought. You will be asked to identify thinking traps in the thought and write and rate thought responses.

Note: The use of "*Both are similar*" option (wherever applicable) is discouraged. Use it only when the two responses are truly identical and there is nothing to distinguish the two.

Figure 8: Instructions shown during data collection with mental health experts (3/3).

Content Warning

This study contains situations and thoughts including but not limited to self-harm and suicidal ideation, which may be disturbing. If you have any questions or concerns, please send us an email. Should you have a strong negative reaction to some of the content, you can reach a crisis counselor at crisis text line or by texting HOME to 741741.

If you have questions about your rights as a research participant, or wish to obtain information, ask questions or discuss any concerns about this study with someone other than the researcher(s), please contact the Human Subjects Division at xxx.

H Consent Form Used in the Randomized Field Study on MHA

Figure 9: Consent form shown to the MHA visitors. Continued on the next page (1/2).

Terms of Use

This tool uses artificial intelligence to generate reframed thoughts and is part of a research study.

Purpose: The purpose of the study is to understand how digital tools can help people recognize thinking traps and practice reframing negative thoughts.

Procedure: You will be asked to describe a thought and a situation you are struggling with. You will then identify potential "thinking traps" (or cognitive distortions) in the thought and reframe it in a way that is more positive, realistic, or helpful. Finally, you will be asked to take an *optional* demographic survey, which can be skipped as preferred. The tool is expected to take ~5 minutes to complete.

Benefits: By using this tool, you may learn about thinking traps. You will practice identifying them and reframing negative thoughts and situations. However, there is no guarantee that the tool will help you reframe your thoughts.

Data Collection and Sharing: We will not ask you for your name or any identifiable personal information. Usage data will be made unidentifiable to the best of our extent, will be analyzed to improve the tool, and may be shared and used for future research.

Risks: Talking about situations and thoughts you are struggling with may be disturbing to you and may bring up negative emotional reactions. In addition, the tool uses artificial intelligence to generate reframed thoughts. Appropriate steps have been taken to avoid harmful reframes, but there is a possibility that the generated content might be upsetting to you. Also, the *optional* demographic survey asks for information that may be sensitive and could make you feel uncomfortable (e.g., "*What are the main things contributing to your mental health problems right now?*"). This

Figure 10: Consent form shown to the MHA visitors (2/2).

tool is not being actively monitored by a human and should not be used as a "cry for help" outlet. Should you have a strong negative reaction to some of the content, you can text MHA to 741741 or call or text 988.

Participation: Participation in this study is completely voluntary. You will not receive any payment for participation. You can refuse participation or stop participating at any time without penalty or loss of benefits to which you are otherwise entitled.

Contact Us: If you have questions or concerns about this research, or if you think you have been harmed from being in the study, please email us at XXX. If you have questions about your rights as a research participant, you can call Human Subjects Division at XXX.

☐ By ticking this box, you are agreeing to use this tool. You are also confirming that you are at least 18 years old. Be sure that questions about the tool have been answered and that you understand what you are being asked to do. You may contact us if you think of a question later. You are free to stop using the tool at any time. To save a copy of this consent form, you can use this link.

Previous Next

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 11

☑ A2. Did you discuss any potential risks of your work?
Section 10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4; Section 5

☑ B1. Did you cite the creators of artifacts you used?
Section 4; Section 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix B

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix B

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 10

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 11

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4; Section 6

C ☑ Did you run computational experiments?

Section 6; Section 7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5; Section 6; Appendix A.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6; Section 7

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4; Section 7

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix D; Appendix E; Appendix F

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4; Section 7

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 4.3; Section 10

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 4.3; Section 10

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 4