

Training Models to Generate, Recognize, and Reframe Unhelpful Thoughts

Mounica Maddela*
Georgia Tech &
Meta AI

Megan Ung*
Meta AI

Jing Xu
Meta AI

Andrea Madotto
Meta AI

Heather Foran
Klagenfurt University

Y-Lan Boureau
Meta AI

Abstract

Many cognitive approaches to well-being, such as recognizing and reframing unhelpful thoughts, have received considerable empirical support over the past decades, yet still lack truly widespread adoption in self-help format. A barrier to that adoption is a lack of adequately specific and diverse dedicated practice material. This work examines whether current language models can be leveraged to both produce a virtually unlimited quantity of practice material illustrating standard unhelpful thought patterns matching specific given contexts, and generate suitable positive reframing proposals. We propose PATTERNREFRAME, a novel dataset of about 10k examples of thoughts containing unhelpful thought patterns conditioned on a given persona, accompanied by about 27k positive reframes. By using this dataset to train and/or evaluate current models, we show that existing models can already be powerful tools to help generate an abundance of tailored practice material and hypotheses, with no or minimal additional model training required.

1 Introduction

Cognitive Behavioral Therapy (CBT) (Beck, 1963, 1976) is one of the most robustly validated approaches in psychology (Hofmann et al., 2012; David et al., 2018). A core pillar of CBT consists in identifying and reframing unhelpful ways of thinking. Low-intensity CBT interventions have shown promise in self-help formats (Shafran et al., 2021; Williams, 2001), yet a lack of sufficient practice material suited to people’s specific circumstances is a barrier to adoption (Helgadóttir et al., 2009).

Through prompting, control tokens, or adequate conditioning, modern language models can guide generation of language towards desired outcomes, such as conforming to a given persona (Zhang et al., 2018), style (Ziems et al., 2022), or level of confidence (Mielke et al., 2022). This makes them a potentially powerful practice aid for learning cog-

nitive reframing techniques. A major barrier is the lack of publicly available data. Most existing work in natural language processing (NLP) for CBT focuses on interactions between patients and mental health professionals, which are not publicly available (Mieskes and Stiegelmayr, 2018; Rojas-Barahona et al., 2018; Shreevastava and Foltz, 2021). Ziems et al. (2022) released the first public dataset for reframing tweets marked with a hashtag indicating stress, using known reframing techniques, but it does not specifically look at the categories of unhelpful thinking used in CBT, and uses existing tweets rather than allowing the generation of examples suited to a particular situation.

In this work, we propose¹ a novel dataset, PATTERNREFRAME, consisting in ~10k crowdsourced examples of thoughts containing ten classical types of unhelpful thought patterns (Burns, 1980), conditioned on personas, matched with crowdsourced proposals of reframing that do not exhibit the patterns. We introduce two controllable text-to-text generation tasks on the dataset: (1) generating and (2) reframing unhelpful thoughts, given a persona and pattern as the context. We also define a classification task to identify the unhelpful thought pattern, given a persona and a thought. We train and evaluate different fine-tuned and few-shot approaches for the tasks, and show that these approaches perform reasonably well on the tasks.

2 Related Work

2.1 NLP for Mental Health

Recent work has used linguistic features and pre-trained language models to identify mental health conditions such as anxiety (Owen et al., 2020; Shreevastava and Foltz, 2021; Fine et al., 2020),

¹The dataset and task have been released through the ParlAI framework (Miller et al., 2017) and are available at https://github.com/facebookresearch/ParlAI/tree/main/projects/reframe_thoughts

depression (Wolohan et al., 2018; Poświata and Perełkiewicz, 2022; Ji et al., 2022), schizophrenia (Jiang et al., 2020b; Mitchell et al., 2015; Sarioglu Kayi et al., 2017), and post-traumatic stress disorder (Coppersmith et al., 2015). Most of these works annotate social media posts to create datasets for the task, and then train and evaluate different classification models. Shreevastava and Foltz (2021) and Rojas-Barahona et al. (2018) created datasets for identifying unhelpful thoughts by annotating patient-therapist interactions and finetuned different pretrained models for the task. However, these datasets are not publicly available.

The closest work to ours is that of Ziems et al. (2022), which introduces a reframing task, releases a parallel corpus of reframed sentences, and uses controllable text generation models to reframe social media content from Twitter that was marked as expressing stress. However, the source social media material is not conditioned on personas, or focused on the classical unhelpful thought patterns from CBT. Our work introduces conditioning on personas and classical unhelpful thought patterns, and extends the reframing task to identifying and generating thoughts matching a given persona and unhelpful pattern.

2.2 Controllable Text Generation

Controllable text generation approaches using pretrained language models (PLMs) typically fall into four categories: (i) prompt-based methods that either construct templates for PLMs to complete (Jiang et al., 2020a; Schick and Schütze, 2021a,b) or finetune a task-specific layer to guide the generation (Li and Liang, 2021; Lester et al., 2021), (ii) finetuning methods that either use labelled data prepended with controlled attributes (Ziems et al., 2022; Fan et al., 2018; Martin et al., 2020; Ross et al., 2022) or define a task-specific reward function using reinforcement learning (Ziegler et al., 2019; Liu et al., 2020), (iii) post-processing methods that train discriminator models to guide the generation towards a specific criterion during decoding (Dathathri et al., 2019; Hua and Wang, 2020; Xu et al., 2020), and (iv) pretraining methods that pre-train PLMs from the start with different control tokens prepended to the input (Keskar et al., 2019). In our work, we experiment with prompt-based and finetuning methods.

3 Identifying and Reframing Unhelpful Thoughts

We use the ten categories of unhelpful thought patterns described in lay terms in a widely used CBT self-help book used for bibliotherapy (Burns, 1980). Table 1 lists these categories and provides examples for each category. For reframing unhelpful thoughts, we follow Ziems et al. (2022), who describe five reframing strategies based on positive psychology (Harris et al., 2007): (i) *Growth Mindset*: Focusing on learning from challenges and improving the skills needed to deal with a difficult situation; (ii) *Optimism*: Directing the attention towards the positive aspects of the situation and expressing gratitude while still acknowledging the negative aspects; (iii) *Impermanence*: Understanding that adversities are inevitable and temporary and focusing on accepting the situation; (iv) *Neutralizing*: Challenging unhelpful thoughts that are far from reality and replacing them with realistic neutral alternatives; (v) *Self-affirmation*: Reflecting on core values to ground oneself in a difficult situation. Note that other reframing strategies exist, such as “*being mindful*” (Robertson, 2012), or “*focusing on forgiveness and compassion*” (Gilbert, 2010). We provide the above five strategies only as a starting point, but crowd workers are free to use other strategies.

4 PATTERNREFRAME Dataset

4.1 Data Collection

We briefly explain the four-step data collection process used to crowdsource the dataset. We provide further data collection details and snapshots of the interface in Appendix A and B.

4.1.1 Task 1: Writing Unhelpful Thoughts

In order to generate unhelpful thoughts that match a diversity of contexts and situations, we use personas from the PERSONA-CHAT dataset (Zhang et al., 2018) as context for writing unhelpful thoughts. We give a persona and one of the ten unhelpful thought patterns to the crowdsource workers, and ask them to write sentences that both are consistent with the given persona, and exhibit the given unhelpful thought pattern.

4.1.2 Task 2: Categorizing Unhelpful Thoughts

Unhelpful thoughts can exhibit multiple patterns, and the patterns themselves are overlapping rather

Unhelpful Thought Patterns and their distribution	Example Thoughts and their Rewrites that remove the pattern
Catastrophizing by giving greater weight to the worst possible outcome. (1024 thoughts / 2826 rewrites)	<i>My mom hasn't come home from work yet. I hope the store isn't getting robbed!</i> Rewrite: <i>My mom hasn't come home from work yet. She must have gotten swamped. I'll cook dinner now so it's ready when she gets home.</i>
Discounting the positive: experiences by insisting that they "don't count". (970 thoughts / 2680 rewrites)	<i>My restaurant is the most popular in my city, but that's just luck.</i> Rewrite: <i>My restaurant is the most popular in the city. I suppose all my hard work has paid off.</i>
Overgeneralization is making faulty generalizations from insufficient evidence. (983 thoughts / 2747 rewrites)	<i>My nephews didn't want to spend the weekend with me this week. I must not be as good of an aunt as I thought.</i> Rewrite: <i>My nephews didn't want to spend the weekend with me this week. They must be busy.</i>
Personalization is assigning a disproportionate amount of personal blame to oneself. (934 thoughts / 2544 rewrites)	<i>My sister was not happy with the makeup look I did for her. I am a bad artist.</i> Rewrite: <i>My sister was not happy with the makeup I did for her, next time I'll try something different.</i>
All-or-nothing is viewing things as either good or bad and nothing in-between. (952 thoughts / 2628 rewrites)	<i>The school christmas choir concert got canceled. This holiday season is ruined.</i> Rewrite: <i>Even though the choir concert got canceled there are still other fun activities to do on the holiday.</i>
Mental Filtering occurs when an individual dwells only on the negative details of a situation. (936 thoughts / 2562 rewrites)	<i>It's nice to enjoy the sea breeze when you live near the ocean but it's not worth it when you think of all the sand getting dragged into your home and all the tourists making so much noise at the beach.</i> Rewrite: <i>I am so fortunate to live where I can enjoy the sea breeze. Not everyone is this lucky.</i>
Mind Reading is inferring a person's probable (usually negative) thoughts from their behavior. (992 thoughts / 2688 rewrites)	<i>I auditioned for the surf team and the coach avoided me. I am sure it is because he does not like my skills.</i> Rewrite: <i>I auditioned for the surf team and the coach avoided me. I'm sure the coach always tries to appear neutral during try-outs.</i>
Fortune Telling is predicting outcomes (usually negative) of events. (997 thoughts / 2758 rewrites)	<i>I didn't make it to Yellowstone this year, I am never going to go to that park.</i> Rewrite: <i>I didn't get to go to Yellowstone this year, I will work extra hard and save up to definitely go next year!</i>
Should statements , where a person demands particular behaviors regardless of the realistic circumstances. (921 thoughts / 2413 rewrites)	<i>I prefer texting over phone calls. People should never call me and expect me to answer.</i> Rewrite: <i>Just because I like texting doesn't mean everyone needs to like it.</i>
Labeling and mislabeling is attributing a person's actions to their character rather than the situation. (960 thoughts / 2661 rewrites)	<i>I fell off my skateboard yesterday, I'm a terrible athlete.</i> Rewrite: <i>I fell off my skateboard yesterday, but even the best crash sometimes.</i>

Table 1: Examples of unhelpful thoughts and their reframed versions from our PATTERNREFRAME dataset. The thought pattern definitions are derived from Wikipedia.

than distinct (Burns, 1980). In order to capture this, as well as filter out low-quality crowdsourced data, we use a second crowdsourcing task requesting workers to label the previously generated thoughts. Workers are given a thought and the list of unhelpful patterns, and select all the patterns that appear in the thought. The annotators can choose a "None" option in case the thought is irrelevant or nonsensical. We collect five annotations for each thought, and discard the thoughts that are marked "None" by a majority of annotators.

4.1.3 Task 3: Reframing Unhelpful Thoughts

In a third task, we ask crowdworkers to rewrite thoughts containing unhelpful patterns, in a more helpful way, similar to the task in Ziems et al.

(2022). We give crowdworkers a thought and the persona and unhelpful pattern that were used to generate it, and ask them to rewrite the thought in a way that still aligns with the context, but does not contain the unhelpful pattern. We also show the five reframing strategies described in §3 to aid the workers in reframing the thoughts, and ask them to select what strategy they used, if any. Note that the strategies are only provided as suggestions, and the workers are free to reframe the thought in other appropriate ways. We collect three rewrites for each thought.

4.1.4 Task 4: Evaluating the Rewrites of Unhelpful Thoughts

Finally, we assess the quality of the rewrites as follows: workers are given a persona, unhelpful thought pattern, generated thought, along with three rewrites. They are asked to select which rewrites successfully remove the unhelpful pattern while not logically contradicting the source (following [Ziems et al. \(2022\)](#)). If worker selects a valid rewrite, we further ask them to identify which of the five proposed reframing strategies were used, if any. We collect five annotations for each set, and include only the rewrites that are marked as “valid” by a majority of annotators.

4.2 Data Quality

We use the Mephisto² and Amazon Mechanical Turk³ platforms to collect crowdsource data. We use the labeling tasks (2nd and 4th task) to select a pool of high-quality workers (that is, crowdsource workers whose generative work was validated by a majority of separate annotators in a separate labeling task), after first seeding the set of annotators through manual inspection of a first batch of data. We use only selected annotators for evaluation tasks (tasks 2 and 4). We first kept the generative text tasks (tasks 1 and 3) open to all workers. We expanded the list of selected workers after every iteration by adding new workers that had completed at least five generative text tasks with at least 80% of generated text validated through the evaluation tasks. We ended up with 524 qualified workers after nine rounds of the entire pipeline, where each iteration started with a batch of 500 thoughts. Once we gathered > 500 qualified workers, we restricted all the tasks to the selected pool. In the final dataset, we included only the annotations provided by these selected workers.

Along with the selected pool of workers, we also included onboarding tasks (details in §A) to ensure that the workers adequately understood the concept of reframing thoughts. Only the workers who passed the onboarding tasks were qualified to work on the actual tasks. We calculated inter-annotator agreement using Krippendorff’s Alpha, which was 0.355 for the second task and 0.454 for the fourth task.⁴

²<https://github.com/facebookresearch/Mephisto>

³Our crowdsourcing tasks pay workers well above minimum wage.

⁴We compute Krippendorff’s Alpha for the binary pattern-level judgments from the the second task and the binary

DP	46.2	3.8	0.6	4.2	2.6	15.4	7.7	6.7	5.3	4.0	3.3
PER	3.9	35.6	1.8	3.4	5.1	5.2	13.7	7.9	5.9	16.2	1.5
SS	2.0	2.2	74.9	1.8	1.6	2.7	3.6	2.2	2.6	1.5	4.8
A/N	8.2	4.7	0.5	18.0	13.2	13.7	14.9	4.3	15.9	4.5	2.2
CAT	2.8	2.5	0.2	3.9	45.1	3.7	7.3	5.6	26.6	1.6	0.7
MF	26.2	4.0	0.9	8.1	6.0	26.5	7.4	5.7	8.4	3.6	3.2
OG	3.6	4.9	0.6	9.2	13.9	6.1	21.4	5.5	28.9	4.9	1.0
MR	3.6	8.3	0.7	2.2	7.0	4.4	9.3	49.8	8.3	5.5	1.0
FT	3.9	3.5	0.7	5.0	22.9	5.0	11.5	6.4	37.0	3.2	1.0
LM	5.4	22.5	0.8	5.6	4.7	6.8	17.1	6.7	5.8	22.7	1.8
	Discounting positive (DP)	Personalization (PER)	Should statements (SS)	All or nothing (A/N)	Catastrophizing (CAT)	Mental filtering (MF)	Overgeneralization (OG)	Mind reading (MR)	Fortune-telling (FT)	Label & mislabel (LM)	None

Figure 1: Confusion matrix representing the distribution of unhelpful thoughts across different patterns in our dataset. Rows represent the patterns used to collect the unhelpful thoughts in the first task (§4.1.1). Columns represents the patterns chosen by annotators in the second task (4.1.2). As expected, some related patterns such as Discounting the positive (DP) and Mental filtering (MF) exhibit strong cross-labeling.

4.3 Data Analysis

4.3.1 Dataset Statistics

PATTERNREFRAME contains 9,688 thoughts and 26,507 reframed versions of thoughts. We split the dataset into training, validation, and test sets of respective sizes 1,920 / 961 / 6,807 for thoughts, and 5,249 / 2,623 / 18,635 for reframed thoughts. One thought can have up to three reframed versions, with an average of 2.74 rewrites / thought after filtering out lower-quality rewrites. The average word lengths of thoughts and rewrites are 19.1 and 23.9, respectively.

4.3.2 Analysis of Unhelpful Thought Patterns

Figure 1 shows the distribution of thoughts across different patterns in our dataset, with initial conditioning pattern (1st task) in rows and annotator identified patterns (2nd task) in columns. As expected, there is a high overlap among some related patterns, e.g., *Discounting the positive / Mental Filtering, Fortune Telling/ Catastrophizing, and Personalization / Labeling and Mislabeling*. *All or Nothing Thinking* is difficult to distinguish, and shows high overlap with many categories. *Mind Reading* and *Should Statement* show the lowest amounts of overlap with other patterns.

reframe-level judgements from the fourth task.

4.3.3 Analysis of Reframing Strategies:

Figure 2 shows the distribution of reframing strategies used to reframe the unhelpful thoughts in our dataset, among the five strategies proposed by Ziems et al. (2022). Here, we use the strategies identified by the workers in the fourth task of evaluating reframed thoughts. Most rewritten thoughts make use of one of the five strategies, with very few being labeled as "None." *Growth Mindset* and *Optimism* are the most commonly used reframing strategies, followed by *Neutralizing* and *Self-Affirmation*. *Optimism* is especially common for patterns that focus on the negative aspects of the situation such as *Discounting the positive* and *Mental Filtering*.

Discounting positive	16.7	4.1	13.6	45.4	16.3	3.8
Personalization	24.0	13.7	23.5	25.8	9.9	3.1
Should statements	21.5	8.0	20.6	31.9	12.0	6.0
All or nothing	23.6	14.1	19.8	30.1	9.8	2.5
Catastrophizing	18.9	9.8	26.9	27.4	12.8	4.0
Mental filtering	18.9	8.9	17.7	41.3	9.7	3.6
Overgeneralization	27.0	16.4	17.4	25.4	11.1	2.7
Mind reading	17.0	8.6	29.6	29.6	10.3	4.9
Fortune-telling	22.1	9.8	20.3	30.5	13.6	3.7
Label & mislabel	27.0	12.5	20.1	25.6	11.8	3.1
	Growth Mindset	Impermanence	Neutralizing	Optimism	Self-Affirmation	None

Figure 2: Matrix showing the distribution of reframing strategies across different unhelpful thought patterns. Rows represent the unhelpful thought patterns and columns represent the reframing strategies.

5 Models to Generate, Recognize, and Reframe Unhelpful Thoughts

We train and evaluate different models using our PATTERNREFRAME dataset on three tasks: generating, identifying, and reframing unhelpful thoughts – all conditioned on a given persona.

5.1 Generating Unhelpful Thoughts

5.1.1 Task and Data

Given a persona and an unhelpful thought pattern, the goal is to generate a thought that exhibits the given pattern and aligns with the persona. We formulate the task as a standard conditioned generation problem and optimize the maximum likelihood loss during training. We use the train, validation, and test splits described in §4.3.1.

5.1.2 Methods

We evaluate methods based on fine-tuning and few-shot learning. We fine-tune BART-large (Lewis et al., 2020), T5-large (Raffel et al., 2020), and R2C2-3B (Shuster et al., 2022) (a BART-based language model specialized in dialogues). For the input, we concatenate the persona and the unhelpful thought pattern texts using a special delimiter token. We also generate responses with GPT3.5 (Ouyang et al., 2022), a state-of-the-art language model trained to follow human instructions, as a 1-shot method. We generated thoughts for only 100 random inputs in the PATTERNREFRAME test set, since we had limited access to the API⁵ to GPT3.5 (text-davinci-002)⁶. We provide implementation details and examples of input prompts in Appendix D and E, respectively.

5.1.3 Automatic Evaluation

Following previous work on text reframing (Ziems et al., 2022; Chen et al., 2021), we report BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020), which capture the semantic similarity between the generated thought and the human reference. We also report distinct-1, and distinct-2 metrics to measure the diversity of the generations. Distinct-n (Li et al., 2016) calculates the ratio between the number of unique n-grams and the total number of n-grams in a generation.

Table 2 shows the automatic evaluation results for the task. All the models perform close to each other in terms of BLEU, BERTScore, and ROUGE. GPT3.5 generates lexically diverse rewrites with the best Distinct-n scores. We provide examples of system outputs in Table 3.

5.1.4 Human Evaluation

As automatic metrics often fail to fully capture human preferences in text generation tasks, we also perform human evaluation. We collect human ratings of 100 random thoughts from the test set. Similar to previous style transfer works (Ziems et al., 2022; Briakou et al., 2021; Rao and Tetreault, 2018), we evaluate the generated rewrites along three dimensions through Yes/No binary ratings: (i) fluency, which evaluates the readability of the generation, (ii) meaning preservation, which here verifies if the rewrite aligns with the given persona

⁵<https://openai.com/api/>

⁶In our experiments, we used text-davinci-002, since text-davinci-003 had not been released yet.

	Generating Unhelpful Thoughts					Reframing Unhelpful Thoughts				
	BLEU	ROUGE	BScore	Dist-1	Dist-2	BLEU	ROUGE	BScore	Dist-1	Dist-2
BART	25.3	23.9	89.0	0.021	0.087	69.7	53.1	93.5	0.034	0.223
T5	24.5	24.3	89.1	0.019	0.08	69.9	55.5	93.6	0.039	0.261
R2C2	25.5	24.1	89.2	0.023	0.1	70.0	55.0	93.7	0.036	0.235
GPT3.5†	24.9	19.2	88.1	0.196	0.586	51.5	41.2	91.7	0.204	0.633
Reference	100.0	100.0	100.0	0.044	0.304	100.0	100.0	100.0	0.041	0.309

Table 2: Automatic evaluation results on the PATTERNREFRAME test set. We report **BLEU**, **ROUGE**, BERTScore (**BScore**), Distinct-1 (**Dist-1**), and Distinct-2 (**Dist-2**) metrics. †We calculate metrics over 100 random generations because of our limited access to the GPT3.5 API (text-davinci-002).

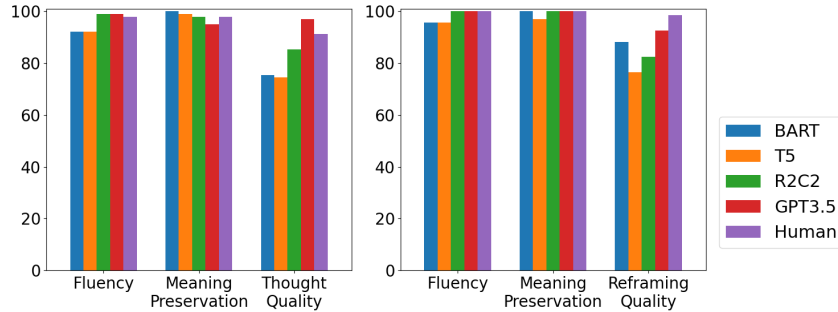


Figure 3: Human evaluation results for the tasks of generating (left) and reframing (right) unhelpful thoughts. Y-axis shows the percentage of outputs rated positively by at least five of the nine annotators.

and thought, and (iii) quality, which here evaluates if the generated thought exhibits the given unhelpful thought pattern. We collect 9 annotations for each system output and apply majority voting to extract the final annotation.⁷

Table 3 shows the percentage of outputs rated positively by at least five of the nine annotators. GPT3.5 outperforms all other approaches, including human references, in terms of fluency and quality. However, GPT3.5 shows the lowest (but still very high) meaning preservation score for generating thoughts. The other models have more difficulty including the unhelpful pattern (lower "thought quality" scores).

5.2 Classifying Unhelpful Thoughts

5.2.1 Task and Data

Given a persona and a thought, the goal is to classify them into one of the ten unhelpful thought patterns or "None", which indicates that the input thought does not contain any of the ten unhelpful patterns, or the thought does not align with the persona. We formulate the task as a multiclass classification problem with eleven categories.

We once again use the same train, validation, and

⁷We also provide results using a more stringent threshold of 7 out of 9 annotators rating positively, in Appendix F. The pattern of results is similar.

test splits described in §4.3.1. Note that the dataset contains only positive examples for the classification task, i.e., thoughts that align with a specific thought pattern and persona. For every positive example, we construct a negative example by randomly choosing one of the following options: (i) a thought from our dataset that belongs to the same pattern but a different persona. (ii) a dialog text from PERSONA-CHAT belonging to the same persona (but presumably not containing any unhelpful pattern), (iii) a dialog text from PERSONA-CHAT belonging to a different persona (and again, presumably not containing any unhelpful pattern). Thus, negative examples encompass neutral texts and misaligned thoughts and personas. We assign the category "None" to these examples. We have 3,834 train, 1,915 validation, and 13,572 test instances after augmenting the dataset with these examples.

5.2.2 Methods

We finetune RoBERTa (Liu et al., 2019) using the soft-label distribution obtained through the second task of our data collection pipeline (§4.1), where we asked multiple annotators to identify the patterns exhibited in a thought, and then normalized the votes across the patterns. We use a soft label distribution instead of single label because of the high overlap across patterns. We also perform

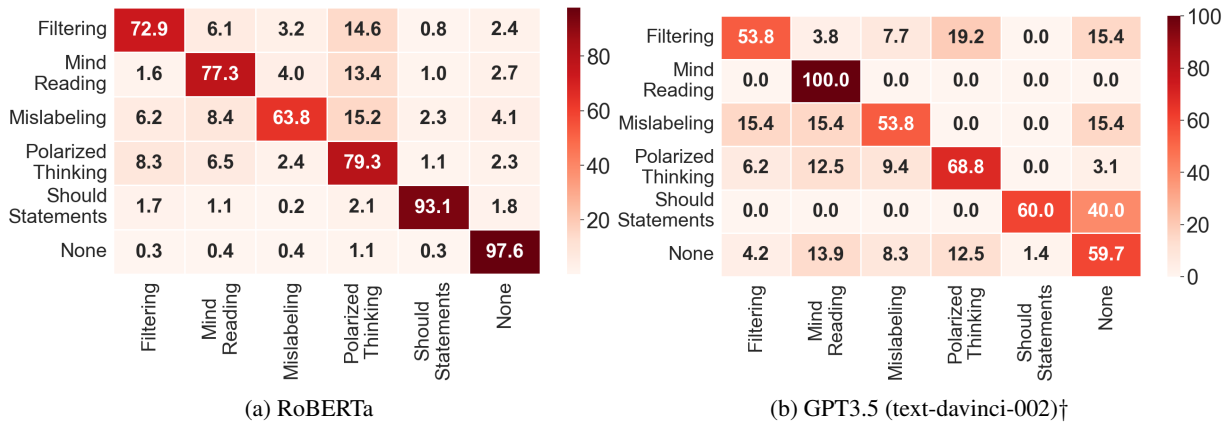


Figure 4: Confusion matrices for the unhelpful thoughts classification task on our classification test set. The rows represent true labels and the columns represent predicted labels. We clustered similar patterns for clearer interpretation. *Polarized Thinking* includes *Overgeneralization*, *Catastrophizing*, *All or Nothing Thinking*, and *Fortune Telling*. *Filtering* refers to *Mental Filtering* and *Discounting the positive*. *Mislabeled* encompasses *Personalization* and *Labeling and Mislabeled*. †We obtain outputs for only 100 random thoughts.

11-way, 1-shot classification using GPT3.5. We construct the input prompt using one example from each category (examples in §E) and classify 100 random inputs in the test set. We include further implementation details in Appendix D.

5.2.3 Evaluation

Figure 4 shows the confusion matrices for RoBERTa and GPT3.5 on the augmented version of the PATTERNREFRAME test set. Given that several unhelpful thinking patterns are closely related (for example, All or Nothing Thinking and Catastrophizing), we cluster the patterns using the KMeans algorithm (Lloyd, 1982) to group together patterns that were deemed close by the model⁸. RoBERTa performs well on all the categories (> 72%) except the *Mislabeled* category, which has a high overlap with the *Polarized Thinking* category. The *None* category has the highest performance, which shows that the classifier is able to differentiate neutral texts that do not contain any unhelpful pattern, or texts that are not aligned with the persona. 1-shot classification using GPT3.5 performs worse than fine-tuned RoBERTa. GPT3.5 has trouble distinguishing texts with and without unhelpful patterns and gets a low score for *None*. We also observed that 40% of the classification predictions changed for GPT3.5 after reordering the sequence of exam-

⁸We first constructed an aggregated vector for each pattern by averaging the 10-dimensional outputs of the classifier for the examples corresponding to each pattern on the validation set. We then clustered the ten 10-dimensional aggregated vectors into five clusters using the KMeans algorithm, with the number of clusters determined using the elbow method.

ples in the prompt, which shows that few-shot classification is not as reliable for this task, while still providing decent performance way above chance.

5.3 Reframing Unhelpful Thoughts

5.3.1 Task and Methods

Given a persona, an unhelpful thought pattern, and a thought exhibiting the given pattern, the goal is to reframe the thought in a way that still aligns with the persona and the context of the thought but does not contain the pattern. The reframing problem is similar to the generating one, except that the unhelpful thought is now a part of the input instead of the target. We use the same training, validation, and test splits for the reframing task (§4.3.1). We also evaluate the same approaches described in §5.1.2. For fine-tuned methods, we concatenate the persona, the pattern, and the thought texts with a special token. For few-shot methods, we construct a prompt similar to the one used for the generation task, but with the reframed statements (examples in Appendix E).

5.3.2 Automatic Evaluation

Table 2 shows the automatic evaluation results on the PATTERNREFRAME test set. We use the metrics described in §5.3.2 namely BLEU, ROUGE, BERTScore, and Distinct-n metrics. As each unhelpful thought can have up to 3 ground truth reframed versions, we take the maximum of the three scores and report the mean of these maxima. R2C2 performs the best in terms of BLEU and BERTScore. GPT3.5 again outperforms the

Generating Unhelpful Thoughts	
Persona	I like to chat with my friends. I like to read books that are in a series. I like to go to the movies. I love to go out to eat with my family.
Pattern	Catastrophizing
T5	I love to go out to eat with my family, but I'm sure we'll all get sick and die.
BART	I like to read books that are in a series, but I will never finish them.
R2C2	I love to go out to eat with my family, but one day we will all get food poisoning and die.
GPT3.5	I'm out of money to go out to eat with my family, maybe they won't want to hang out with me anymore.
Human	My friends didn't want to chat with me today. They must have all hate me and wish me ill.
Persona	I write children's books. I like to drink. I live alone. I shop online, too much. I don't believe in love.
Pattern	Should statements
T5	I write children's books, everyone should read them.
BART	I write children's books, everyone should read them.
R2C2	I write children's books. I should be published by now.
GPT3.5	I should be able to write the perfect book, drink moderately, be in a relationship, and not shop online so much.
Human	I shop online and I should make enough money to be able to afford anything and everything I want.
Reframing Unhelpful Thoughts	
Persona	My mother was a teacher. My favorite food is a salad. I enjoy nature. I teach a yoga class. I am single.
Pattern	Black-and-white or polarized thinking / All or nothing thinking
Thought	The date I went on last week was a total fail. I'm going to be single forever.
T5	The date I went on last week was a total fail. I will try harder next time.
BART	The date I went on last week was a total fail. Next time I will be more careful.
R2C2	The date I went on last week was a total fail. I'm going to be single forever.
GPT3.5	The date I went on last week wasn't a great match, but I'm open to meeting someone new and trying again.
Human	I had a bad date last week, thankfully I went on a nature walk and realized I have plenty of time to find a partner.
Persona	I was diagnosed with mild OCD in college. Once a year, I take a week off and head to Florida. I freeze up in social situations. Sometimes I feel like I've already been through life once before.
Pattern	Personalization
Thought	I was diagnosed with mild OCD in college, it's my fault and I deserve it.
T5	I was diagnosed with mild OCD in college. It's not my fault.
BART	I was diagnosed with mild OCD in college. I am working hard to overcome it.
R2C2	I was diagnosed with mild OCD in college. I'm glad to have a diagnosis so I can get the help I need.
GPT3.5	I was diagnosed with mild OCD in college, it's something I'm learning to manage.
Human	I was diagnosed with mild OCD in college. I've been seeing a therapist to get help managing it.

Table 3: Examples of system outputs for the tasks of generating and reframing unhelpful thoughts.

other models and the human references in terms of Distinct-1 and Distinct-2 scores, which indicates that the generations are lexically diverse. Table 3 provides examples of system outputs.

5.3.3 Human Evaluation

Figure 3 shows human evaluation results on 100 reframed thoughts generated by different models given the persona, the pattern type, and the unhelpful thought from our test set. Similar to the generating thoughts task, we evaluate the reframed thoughts along fluency, meaning preservation, and quality, where we ask the annotators if the reframed thought removes the given unhelpful pattern while being consistent with the initial thought. All models perform close to human reference in terms of fluency and meaning preservation. In fact, all the outputs of R2C2 and GPT3.5 are fluent and preserve meaning (that is, they generate statements that are not contradictory with the initial thought). For reframing quality, that is, removing the un-

helpful pattern, all models perform over 70%, but GPT3.5 performs the best. GPT3.5's superiority is even more marked when using the more stringent threshold of 7 out of 9 annotators rating positively in Appendix F.

Overall, the evaluation suggests that using modern models to produce reframing is a feasible approach, even with a small amount of data for fine-tuning. In particular, GPT3.5 performs remarkably well and very close to crowdsourced worker performance, only based on prompting.

6 Conclusion

In this work, we introduced a novel dataset, PATTERNSREFRAME, which contains (1) about 10k statements exhibiting unhelpful thought patterns, conditioned on a persona, and (2) multiple rewritten complementary thoughts that do not contain the initial unhelpful pattern, instead reframing the thought in a more constructive way.

Using this dataset to train or prompt various mod-

ern language models, we showed that this range of models can already be a powerful tool to generate, identify, and reframe unhelpful thoughts, conditioned on a persona. By releasing our dataset⁹, we hope to help practitioners of CBT draw from a richer, more diverse set of examples of unhelpful thought patterns and reframings. This would help address the important limitation of a lack of personalized and specific examples in existing datasets, when teaching cognitive techniques.

Future work will evaluate whether leveraging models to produce richer training material results in more robust learning and understanding of the types of unhelpful thought patterns in humans. This may serve as the basis for future psychological validation studies of the materials and support future studies of low-intensity self-help interventions.

7 Limitations

This work relied on previously published datasets to source personas on which to anchor the generated unhelpful thoughts, and thus shares the limitations of those datasets. In particular, they use English-language responses, written by workers located in the United States.¹⁰ While these workers are reasonably diverse (Moss et al., 2020), the examples generated may not reflect the thought patterns and personas across cultures and diverse populations. This data is also generated by people who are being paid, as opposed to people genuinely engaging about situations that matter to them. Besides the substance of the thoughts themselves, a more direct limitation is that the models generate only English, so would not be directly usable for speakers of other languages.

In addition, the data collected reflects the understanding of lay people, rather than trained clinical psychologists. While this makes the material more immediately relatable to other lay people, it is possible that the data do not capture what clinical psychologists would consider adequate illustrations of unhelpful patterns. Our data has been spot-checked by a CBT-trained clinical psychologist and found generally sound, but the entire material should undergo further validation.

Another limitation is that the models that we have tested are resource-intensive. In particular, the

⁹https://github.com/facebookresearch/ParlAI/tree/main/projects/reframe_thoughts

¹⁰Our crowdsourcing tasks pay workers well above minimum wage.

best-performing model, GPT3.5, is only available through a paid API.

8 Ethical considerations

While our work was developed to generate abundant data supporting work towards improving well-being, the negative statements it generates could be misused. The parallel data of unhelpful thoughts and their reframed versions can also be used to generate negative texts from neutral ones, by training systems with reframed versions as the input and unhelpful thoughts as the output. This risk of generating negative content from positive/neutral texts aligns with the risks of toxicity reduction and sentiment style transfer tasks.

Conversely, a different risk stems from over-eager use of our work. This work aims to examine the feasibility of generating ample practice material anchored on specific personas. We hope that releasing a large dataset of unhelpful thoughts and reframings will further research that will ultimately help practitioners, but there is a danger that people attempt to use the material as is, without the supervision of a trained professional, which could be harmful, as the material has not been tested with participants while monitoring adverse events such as increased anxiety or warped understanding of what unhelpful thoughts and useful reframings are.

References

- Aaron T. Beck. 1963. Thinking and Depression: I. Idiosyncratic Content and Cognitive Distortions. *Archives of General Psychiatry*.
- Aaron T. Beck. 1976. Cognitive therapy and the emotional disorders. international universities press.
- Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021. A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics.
- D.D. Burns. 1980. *Feeling Good: The New Mood Therapy*. A Signet book.
- Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2021. Controlled neural sentence-level reframing of news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.
- Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD

- on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation.
- Daniel David, Ioana Cristea, and Stefan G. Hofmann. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in Psychiatry*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Alex Fine, Patrick Crutchley, Jenny Blase, Joshua Carroll, and Glen Coppersmith. 2020. Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*.
- Paul Gilbert. 2010. An introduction to compassion focused therapy in cognitive behavior therapy. *International Journal of Cognitive Therapy*.
- Alex H. S. Harris, Carl E. Thoresen, and Shane J. Lopez. 2007. Integrating positive psychology into counseling: Why and (when appropriate) how. *Journal of Counseling & Development*.
- Fjóla Dögg Helgadóttir, Ross G Menzies, Mark Onslow, Ann Packman, and Sue O'Brian. 2009. Online cbt i: Bridging the gap between eliza and modern online cbt treatment packages. *Behaviour Change*, 26(4):245–253.
- Stefan Hofmann, Anu Asnaani, Imke Vonk, Alice Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*.
- Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.
- Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020a. How can we know what language models know? *Transactions of the Association for Computational Linguistics*.
- Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020b. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*.
- Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics.
- Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- 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.
- S. Lloyd. 1982. [Least squares quantization in pcm](#). *IEEE Transactions on Information Theory*, 28(2):129–137.

- Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.
- Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Margot Mieskes and Andreas Stieglmayr. 2018. Preparing data from psychotherapy for natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84. ACL.
- Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.
- Aaron J Moss, Cheskie Rosenzweig, Jonathan Robinson, and Leib Litman. 2020. Demographic stability on mechanical turk despite covid-19. *Trends in cognitive sciences*, 24(9):678–680.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.
- David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. 2020. Towards preemptive detection of depression and anxiety in Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.
- Rafał Poświata and Michał Perełkiewicz. 2022. OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*.
- Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.
- D. Robertson. 2012. *Build Your Resilience: CBT, mindfulness and stress management to survive and thrive in any situation*. Teach Yourself.
- Lina M. Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gašić. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.
- Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213.
- Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2017. Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*.
- Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.
- Timo Schick and Hinrich Schütze. 2021b. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Roz Shafran, Pamela Myles-Hooton, Sophie Bennett, and Lars-Göran Öst. 2021. [The concept and definition of low intensity cognitive behaviour therapy](#). *Behaviour Research and Therapy*, 138:103803.
- Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. Association for Computational Linguistics.
- Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge:

Modular search and generation for dialogue and prompt completion.

- Chris Williams. 2001. [Use of written cognitive behaviour therapy self-help materials to treat depression](#). *Advances in Psychiatric Treatment*, 7.
- JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zee-shan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*.
- Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences.
- Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text re-framing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

A Data Collection Details

A.1 Onboarding Tasks

We introduce two onboarding tasks to ensure that the crowdsource workers understood the concept of unhelpful thoughts and how to reframe them. The onboarding tasks were reviewed by a CBT-trained psychologist. We use one onboarding task for tasks 1 and 2 and another onboarding task for tasks 3 and 4 of the data collection pipeline. For the first onboarding task, we display an unhelpful thought pattern, one positive example that contains the pattern, and one negative example that does not, and ask the workers to select the positive one. We only allowed the workers that were able to identify the correct example for three out of four such instances. For the second onboarding task, we display an unhelpful thought pattern, a thought containing the pattern, one positive example that reframes the thought, and one negative example that does not. We only allow the workers that were able to identify the positive example in three out of four such instances.

B Data Collection Interface Snapshots

Write Examples of Cognitive Patterns
You are currently at task: 1 / 4
After completing each, click [NEXT] button, which will be enabled below.

Task Description
You will be given (1) a type of unhelpful thinking pattern and the definition of the pattern, (2) a character. Please write an example of how this character could have thoughts that match the given thinking pattern.

Example:
Pattern: Should statements
Definition: Expecting the world to be different than it is, demanding particular achievements or behaviors regardless of the realistic circumstances of the situation.
Character: Likes camping. Has 2 kids.
Example: The kids are not enjoying this camping trip, they should really be more grateful about the effort we put in planning week-end activities for them.

Pattern: Should statements
Definition: Expecting the world to be different than it is, demanding particular achievements or behaviors regardless of the realistic circumstances of the situation.
Character: i work in marketing , and i love it . i m hoping to find someone to marry me . i m an only child . i spend a lot of time on my hair and make up .

Please write an example of how this character could have thoughts that match the given thinking pattern.
Please enter your response here

Next

Figure 5: Data collection interface for the first task of the data collection pipeline, where crowdworkers are asked to write an unhelpful thought.

Label Examples of Unhelpful Thinking Patterns
You are currently at conversation: 1 / 4
After completing each, click [NEXT] button, which will be enabled below.

Task Description
You will be given (1) a character and (2) an example of an unhelpful thought by this character. The goal is to select all the unhelpful thinking patterns exhibited in the example. We will provide a list of unhelpful thinking patterns and their examples.

Examples:
Persona: I am an accountant. I like to play the guitar.
Unhelpful Thought Sample: I made a mistake while playing the guitar piece. I will never be able to play the song right.
Patterns exhibited: Overgeneralization

Persona Text: i won the spelling bee in third grade . i love to eat eggplant . i have been married five times . i smoke a pipe .

Sample Thought / Statement: I have been married five times. I'm such a bad spouse.

Select all the unhelpful thought pattern(s) that are exhibited in the sample.

- Catastrophizing:** Imagining the worst possible unlikely outcome. For example, *My kid is late from school. Perhaps she got run over by a car and is in a hospital.*
- Discounting the positive:** Rejecting positive experiences by insisting they "don't count" for some reason. For example, *This cake turned out really well! Beginner's luck for sure.*
- Labeling and mislabeling:** Associating a person's actions to their character instead of to an external attribute or circumstance. For example, *I missed a stitch again, I'm so stupid.*
- Mental filtering:** Dwelling only on the negative details of a situation and filtering out the positive aspects. For example, *I watched Game of Thrones from beginning to end but the last season ruined everything.*
- Jumping to conclusions: mind reading:** Inferring a person's possible or probable (usually negative) thoughts from their behavior and nonverbal communication. For example, *Nobody has commented on my drawing. I'm sure they hate it.*
- Jumping to conclusions: Fortune-telling:** Predicting outcomes (usually negative) of events. For example, *I'd love to try baking those muffins. But they'll turn out horrible.*
- Overgeneralization:** Making faulty generalizations from insufficient evidence. For example, *This drawing has bad perspective. I always mess up the lines.*
- Personalization:** Assigning personal blame disproportionate to the level of control a person realistically has in a given situation. For example, *The kids are having bad grades. It's because I'm a bad father.*
- Black-and-white or polarized thinking / All or nothing thinking:** Looking at life in all-or-nothing categories. Things are either good or bad and there is no in-between. For example, *The color of the tree in the corner of this drawing is off, this drawing is garbage.*
- Should statements:** Demanding particular achievements or behaviors regardless of the realistic circumstances of the situation. For example, *I should really try harder to improve faster on my fitness goals.*
- None of the above apply.** (The example does not contain any unhelpful pattern / is nonsensical / does not align with the persona.)

Next

Figure 6: Annotation interface for the second task of the data collection pipeline, where crowdworkers are asked to select the patterns exhibited by an unhelpful thought.

Reframing unhelpful thinking patterns

You are currently at task: 1 / 4
After completing each, click [NEXT] button, which will be enabled below.

Task Description

You will be given (1) a type of unhelpful thinking pattern, (2) a character, and (3) an example of how this character could have thoughts that match the given thinking pattern. Please rewrite the thoughts in a way that still aligns with the persona and the context of the unhelpful thought, but does not contain the unhelpful pattern.

Examples

Pattern: Catastrophizing
Definition: Catastrophizing: Giving greater weight to the worst possible outcome, however unlikely, or experiencing a situation as unbearable or impossible when it is just uncomfortable.
Character: Likes camping. Has 2 kids.
Unhelpful Thought: My kid is late from school. Perhaps she got run over by a car and is in a hospital.
Reframing (Answer): My kid is late from school. But it's only been 10 minutes, she might have stayed a bit longer to chat with her friends. If there had been an accident, someone would have called already.

Pattern: Catastrophizing
Catastrophizing: Giving greater weight to the worst possible outcome, however unlikely, or experiencing a situation as unbearable or impossible when it is just uncomfortable.

Character: i enjoy hiking and riding my bicycle . i am 29 years old . i'm a night owl . my favorite band is pearl jam . i live in a suburb of boston .

Unhelpful Thought: I am a night owl. One night I will get mugged and murdered while walking to the park.

Please rewrite the unhelpful thought in a way that still matches the context, but does not contain the unhelpful pattern. The following strategies are often helpful to produce better rewrites, please tick the box if you choose to use any.

- Growth Mindset (Reframe a challenging event as an opportunity to grow instead of dwelling on the setbacks).
- Impermanence (Say that bad things don't last forever, will get better soon, and/or that others have experienced similar struggles).
- Neutralizing (Challenge the negative or catastrophic possibilities and reframe it with a neutral possibility).
- Optimism (Focus and be thankful for the positive aspects of the current situation).
- Self-Affirmation (Say that the character can overcome the challenging event because of their strengths or values.)

Please enter your response here

Figure 7: Data collection interface for the third task of the data collection pipeline, where the crowdworkers are asked to reframe unhelpful thoughts.

Evaluating rewrites of unhelpful thinking patterns

You are currently at task: 1 / 4
After completing each, click [NEXT] button, which will be enabled below.

Task Description

You will be given (1) a type of unhelpful thinking pattern, (2) a character, (3) an example of how this character could have thoughts that match the given thinking pattern, and (4) upto 3 statements that rewrite the thought without the unhelpful pattern. Please select the statements that reframe the unhelpful thought in a way that still aligns with the persona and the context of the unhelpful thought, but does not contain the unhelpful pattern.

Examples

Pattern: Should statements
Definition: Expecting the world to be different than it is, demanding particular achievements or behaviors regardless of the realistic circumstances of the situation.
Character: Likes camping. Has 2 kids.
Unhelpful Thought: The kids are not enjoying this camping trip, they should really be more grateful about the effort we put in planning week-end activities for them.

Pattern: Should statements
Expecting the world to be different than it is, demanding particular achievements or behaviors regardless of the realistic circumstances of the situation.

Character: i like to play video games . my father used to be a mechanic . i have been working the same job for two years . i am a musician . my favorite band is hop along .

Unhelpful Thought: I like to play video games, everybody else should love it as much as me.

Please **select all the statements that reframe the unhelpful thought** in a way that still matches the context, but does not contain the unhelpful pattern.

- like to play video games, everyone has different tastes so it is ok that some people do not love it as much as me.
- love video games because I'm very good at them. It's my hobby. Other folks like other things, and that's OK by me!
- like to play video games, but I understand that some people do not like to.

For each of the selected statements, please **select the strategy** that you think was used to reframe the given unhelpful thought.

Reframe 1: I like to play video games, everyone has different tastes so it is ok that some people do not love it as much as me.

- Growth Mindset (Reframe a challenging event as an opportunity to grow instead of dwelling on the setbacks).
- Impermanence (Say that bad things don't last forever, will get better soon, and/or that others have experienced similar struggles).
- Neutralizing (Challenge the negative or catastrophic possibilities and reframe it with a neutral possibility).
- Optimism (Focus and be thankful for the positive aspects of the current situation).
- Self-Affirmation (Say that the character can overcome the challenging event because of their strengths or values.)
- None of the above

Reframe 2: I love video games because I'm very good at them. It's my hobby. Other folks like other things, and that's OK by me!

- Growth Mindset (Reframe a challenging event as an opportunity to grow instead of dwelling on the setbacks).
- Impermanence (Say that bad things don't last forever, will get better soon, and/or that others have experienced similar struggles).
- Neutralizing (Challenge the negative or catastrophic possibilities and reframe it with a neutral possibility).
- Optimism (Focus and be thankful for the positive aspects of the current situation).

Figure 8: Annotation interface for the fourth task of the data collection pipeline, where the crowdworkers are asked to evaluate the quality of the reframed thoughts.

C Evaluation Interface Snapshots

Persona Text: i hope to retire to florida . i own my own music store . my mother and father are both in the church choir . i played in a band for 17 years .

Unhelpful Thinking Pattern: Black-and-white or polarized thinking / All or nothing thinking (Looking at life in all-or-nothing categories. Either things are a success or a failure; either they are good or bad; there is no in-between, no good enough or partial success.)

Thought / Statement: I played in a band for 17 years. I will never be able to play again.

Is the thought fluent?

Yes

No

Does the thought align with the persona?

Yes

No

Does the thought contain the given unhelpful thinking pattern?

Yes

No

Figure 9: Annotation interface used to evaluate generated thoughts.

Persona Text: my husband is a pastor . i do not like to clean house . i have two children . my hair is curly and dark .

Unhelpful Thinking Pattern: Mental filtering (Filtering distortions occur when an individual dwells only on the negative details of a situation and filters out the positive aspects.)

Thought / Statement: My hair is curly and dark, everyone will makes jokes about me now

Reframe: My hair is curly and dark. It makes me unique.

Is the reframe fluent?

Yes

No

Does the reframe align with the persona and the context of the thought?

Yes

No

Does the reframe remove the given unhelpful thinking pattern expressed in the thought?

Yes

No

Figure 10: Annotation interface used to evaluate statements that reframe unhelpful thoughts.

D Implementation details

D.1 Generation Models

We finetuned the BART, T5, and R2C2 baselines using ParlAI¹¹. We used the BART_{large} (400M parameters), T5_{large} (770M parameters), and R2C2_{base} (2.7b parameters)¹² architectures. We used Adam optimizer (Kingma and Ba, 2014) and performed a hyperparameter search over learning rates 1e-05, 1e-06, 1e-07, and 1e-08. We used linear warmup of 100 steps and applied early stopping with a patience value of 5. We evaluated the validation set once in every 200 updates and truncated the input and the labels to 1000 tokens. We applied gradient clipping value of 1.0. We used a batch size of 32. During inference, we used beam search with beam size 10. We chose the best checkpoint during training based on the perplexity on the validation set. Each model takes around 1 hour to run on 8 NVIDIA Tesla V100 Volta 32GB GPUs.

D.2 Classification Models

For classification experiments, we finetuned the RoBERTa-large checkpoint from Huggingface¹³. We used Adam optimizer (Kingma and Ba, 2014), learning rate of 1e-05, with linear warmup of 100 steps. We trained the model for a maximum of 10 epochs. We evaluated on the validation set every 200 updates. We used a batch size of 16. We chose the best checkpoint during training based on the weighted F1 value on the validation set. The model takes around 1 hour to run on 1 NVIDIA Tesla V100 Volta 32GB GPU.

E GPT3.5 Prompt Examples

You will be given (1) a type of unhelpful thinking pattern and the definition of the pattern and (2) a character. Please write an example of how this character could have thoughts that match the given thinking pattern.

Persona: Likes camping. Has 2 kids.

Unhelpful Thinking Pattern: Discounting the positive (Rejecting positive experiences by insisting they "don't count" for some reason or other.)

Unhelpful Thought: My friends said they really enjoyed the camping trip I organized, but anyone could have done it.

Persona: i'm a business man. i love to sing. i'm a karate black belt. my wife has terminal cancer.

Unhelpful Thinking Pattern: Discounting the positive (Rejecting positive experiences by insisting they "don't count" for some reason or other.)

Unhelpful Thought:

Table 4: Example GPT3.5 prompt for the task of generating unhelpful thoughts.

You will be given a type of unhelpful thinking pattern, a character, and an example of how this character could have thoughts that match the given thinking pattern. Please rewrite the thoughts in a way that still aligns with the persona and the context of the unhelpful thought, but does not contain the unhelpful pattern.

Persona: Likes camping. Has 2 kids.

Unhelpful Thinking Pattern: Overgeneralization (Someone who overgeneralizes makes faulty generalizations from insufficient evidence. Even if something bad happens only once, it is expected to happen over and over again.)

Unhelpful Thought: My younger kid has gotten bad grades at his maths test this week. He'll never be good at maths.

Reframe: My younger kid has gotten bad grades at his maths test this week. It's been a few times but hopefully we can figure out a way to help him get better.

Persona: i obsess over working out and being the best . i got a scholarship for playing soccer . its important for my instagram posts to look like i am having fun . i try to eat healthy or i don't eat at all .

Unhelpful Thinking Pattern: Overgeneralization (Someone who overgeneralizes makes faulty generalizations from insufficient evidence. Even if something bad happens only once, it is expected to happen over and over again.)

Unhelpful Thought: My future college team lost another game, I will never become a good athlete playing for them.

Reframe:

Table 5: Example GPT3.5 prompt for the task of reframing unhelpful thoughts.

¹¹<https://www.parl.ai/docs/index.html>

¹²<https://parl.ai/docs/zoo.html#r2c2-base-2-7b>

¹³<https://github.com/huggingface/transformers>

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: The kids have stopped paying attention to how we can pitch the tent. They will never learn.
 Unhelpful Thinking Pattern: Jumping to conclusions: Fortune-telling

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: The kids are not enjoying this camping trip, they should really be more grateful about the effort we put in planning week-end activities for them.
 Unhelpful Thinking Pattern: Should statements

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: My kid is late from school. Perhaps she got run over by a car and is in a hospital.
 Unhelpful Thinking Pattern: Catastrophizing

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: This camping trip was a catastrophe. Sure the weather was gorgeous and the kids had a lot of fun, but the waterfall always had many people ruining the photos we wanted to take.
 Unhelpful Thinking Pattern: Mental filtering

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: I like camping with my kids. We had a lot of fun the other weekend.
 Unhelpful Thinking Pattern: None

Persona: Likes camping. Has 2 kids. Unhelpful Thought: The kids are having bad grades. It's because I'm a bad father.
 Unhelpful Thinking Pattern: Personalization

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: My younger kid has gotten bad grades at his math test this week. He'll never be good at math.
 Unhelpful Thinking Pattern: Overgeneralization

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: My friends said they really enjoyed the camping trip I organized, but anyone could have done it.
 Unhelpful Thinking Pattern: Discounting the positive

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: My kids are being very silent. I am sure it's because they really hate me for taking them on this camping trip.
 Unhelpful Thinking Pattern: Jumping to conclusions: mind reading

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: I didn't manage to light up the fire for the camp today, I'm such a useless outdoors person.
 Unhelpful Thinking Pattern: Labeling and mislabeling

Persona: Likes camping. Has 2 kids.
 Unhelpful Thought: One of the 5 trails we planned to do on this trip is closed to the public. This trip is ruined.
 Unhelpful Thinking Pattern: Black-and-white or polarized thinking / All or nothing thinking

Persona: i'm a woman . i've several children . we have a dog . we live in a rural area . my parents are still married .
 Unhelpful Thought: congratulations ! have you graduated college ? i am attending the university of michigan in the fall .
 Unhelpful Thinking Pattern:

Table 6: Example GPT3.5 prompt for the task of identifying unhelpful thoughts.

F Results with 7 over 9 agreement

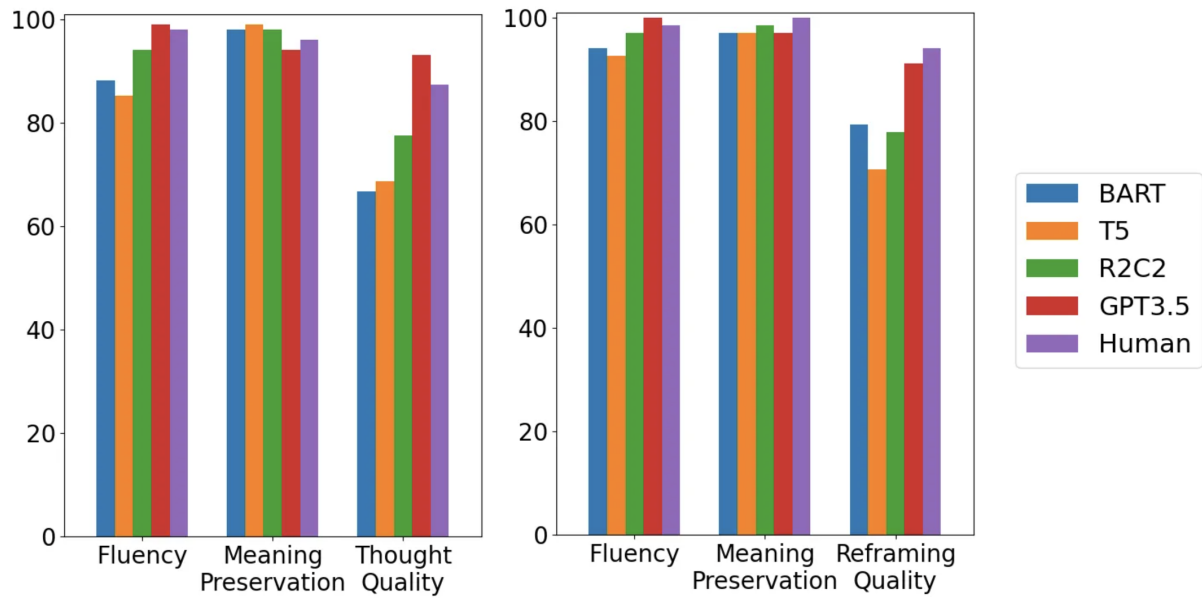


Figure 11: Human evaluation results for the tasks of generating (left) and reframing (right) unhelpful thoughts. Y-axis shows the percentage of outputs rated positively by at least seven of the nine annotators.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
7
- A2. Did you discuss any potential risks of your work?
8
- A3. Do the abstract and introduction summarize the paper’s main claims?
Abstract + introduction (1) + Conclusion (6)
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

4 describes the data we collect, 5 the models we use and finetune, with more details in the appendix

- B1. Did you cite the creators of artifacts you used?
all citations given across sections 4, 5, and the appendix
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
all the models and data we use have been open-sourced for academic use and are widely used with that intent
- 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)?
Sections 7 and 8 reiterate that the use of our data is intended as early research
- 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?
We specify in section 4 that we are not collecting any personal information.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We describe limitations of the dataset in Section 7, in particular language and culture limitations.
- 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.
In section 4.

C Did you run computational experiments?

5

- 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 D

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?

Appendix D

- 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 5

- 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.)?

Not applicable. We use our own set-up for evaluation

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

Section 4 and Appendix

- 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 B and C

- 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

- 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 + appendix B and C: crowdsource workers are not providing personal information

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The data collection protocol is reviewed by internal reviewers but not subject to an IRB as there is no sensitive data or personal information

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

Described in section 7