

Detection and Positive Reconstruction of Cognitive Distortion Sentences: Mandarin Dataset and Evaluation

Shuya Lin¹, Yuxiong Wang¹, Jonathan Dong², and Shiguang Ni^{1*}

¹Shenzhen International Graduate School, Tsinghua University

²École polytechnique fédérale de Lausanne

Abstract

This research introduces a Positive Reconstruction Framework based on positive psychology theory. Overcoming negative thoughts can be challenging, our objective is to address and re-frame them through a positive reinterpretation. To tackle this challenge, a two-fold approach is necessary: identifying cognitive distortions and suggesting a positively reframed alternative while preserving the original thought’s meaning. Recent studies have investigated the application of Natural Language Processing (NLP) models in English for each stage of this process. In this study, we emphasize the theoretical foundation for the Positive Reconstruction Framework, grounded in broaden-and-build theory. We provide a shared corpus containing 4001 instances for detecting cognitive distortions and 1900 instances for positive reconstruction in Mandarin. Leveraging recent NLP techniques, including transfer learning, fine-tuning pretrained networks, and prompt engineering, we demonstrate the effectiveness of automated tools for both tasks. In summary, our study contributes to multilingual positive reconstruction, highlighting the effectiveness of NLP in cognitive distortion detection and positive reconstruction.

1 Introduction

The rapid pace of modern life and the prevalence of high-pressure lifestyles have led to a rise in mental health issues among individuals. According to a report by the World Health Organization (WHO), depression stands out as a significant contributor to both physical and mental disorders globally (WHO, 2021). Moreover, in response to the escalating concern, the WHO recently launched a new Commission on Social Connection in November 2023, recognizing the urgency of combating loneliness as a critical health hazard (WHO, 2023). These indicators strongly suggest that within an

environment marked by increased instability and diminished personal agency, there is a heightened risk for individuals to experience negative emotions, closely linked with negative thought patterns. These patterns, identified as cognitive distortions (Beck, 1963), represent rigid and inaccurate ways of thinking. Such detrimental thoughts reinforce self-negation and perpetuate negative emotions (Rnic et al., 2016). These distorted thoughts, often irrational and reinforced over time, become deeply ingrained. Habitual adoption of such patterns makes them challenging to recognize. Consequently, they are profoundly destructive, when individuals perceive these distortions as truths.

Cognitive restructuring is the main method to work on cognitive distortions in a clinical setting (Dawes, 1964). The process can be broadly categorized in two key phases: First, recognizing negative thinking patterns and mitigating their impact, this process requires individuals to write down their negative thoughts and identify any underlying cognitive distortions in those thoughts. Then, the cognitive restructuring process is designed to help cultivate a more objective view of the situation.

Yet, self-examining and rectifying distorted thinking can prove challenging for individuals (Beck, 1979). Often, intervention and guidance by trained counselors are necessary to identify cognitive distortions and erroneous beliefs. Professional counselors can then follow with cognitive restructuring, drawing from their expertise to aid individuals in fostering new perspectives. This approach significantly relies on the counselor’s unique skills and individual style. However, due to the high price of psychological counseling (SingleCare Team, 2020) and the reluctance to speak with a therapist, most struggling people find it challenging to get the support they need. Even if they can afford the high cost of a consultation, they may have to wait months for an appointment (Mulraney et al., 2021). Therefore, in recent years, research

*Corresponding author: ni.shiguang@sz.tsinghua.edu.cn

has begun to turn its attention to Natural Language Processing (NLP) methods to assist individuals in reconstructing negative thoughts.

In recent years, the integration of NLP into the field of psychology has been gaining momentum. These NLP models, initially trained on large datasets for general purposes, are now being fine-tuned to address specific needs in mental health (Sharma et al., 2021; Zheng et al., 2021; Zhu et al., 2022; Chen et al., 2022; Maddela et al., 2023). The emerging success of NLP in mental health is exemplified by applications (Prochaska et al., 2021; Chiauzzi et al., 2023; Lin et al., 2023; Sharma et al., 2023a), demonstrating its potential for widespread, accessible psychological support. This development represents a significant advancement, offering a new avenue for cognitive restructuring and mental wellness, seamlessly connecting the high demand for mental health support with innovative, technology-driven solutions.

In this rapidly-evolving landscape, the scarcity of language-specific datasets poses a significant challenge. Currently, available datasets in the field of mental health NLP predominantly focus on the English language. Recognizing the linguistic and cultural intricacies inherent in mental health expressions, it becomes evident that dedicated datasets in other languages is essential for evaluating and fine-tuning NLP models.

In this study, we build and openly share the first Mandarin dataset¹ for both phases of the cognitive restructuring process, encompassing both cognitive distortion detection and positive reframing. Additionally, we benchmark different NLP models and training strategies for each task. Our main contributions are:

- Construction of a Chinese dataset for the detection of cognitive distortions, based on real psychology Q&A sentences and annotated by trained specialists.
- Construction of a Chinese dataset for positive reconstruction of cognitive-distorted sentences. Each sentence is reframed with five different strategies from the theory of positive psychology.
- Benchmark of cognitive distortion detection using pre-trained RoBERTa-wwm-ext network with fine-tuning or transfer learning with

¹<https://github.com/405200144/Dataset-of-Cognitive-Distortion-detection-and-Positive-Reconstruction/tree/main>

different readout strategies (linear, multilayer perceptron, LSTM), with the best results obtained with fine-tuning the whole pre-trained network.

- Benchmark of different approaches for positive reconstruction (P-Tuning, fine-tuning, prompt engineering) with both algorithmic and human evaluation, with the best results consistently across all metrics obtained with P-Tuning. We show also that among the different strategies possible in positive psychology, the best performing network predominantly chooses an optimistic strategy.

2 Related Works

Previous studies have proposed various approaches to build NLP pipelines on the detection, categorization of distortions, and positive reframing. We detail here the challenges regarding corpus annotation, in particular labeling consistency, and the need for multilingual research.

2.1 Cognitive Distortion Detection

The seminal classification of cognitive distortions (Beck, 1963) has been revisited by recent studies. For instance, a study (Shickel et al., 2020) used 15 categories from a psychological website to attempt classification. Similarly, Shreevastava and Foltz (2021) used a standard of 10 categories for classification, highlighting the complexity of corpus annotation for detection and categorization. Another study (Tauscher et al., 2023) conducted a classification task with 5 categories, utilizing over 7,000 text messages in a study involving individuals with serious mental illness and clinicians.

However, these studies have revealed certain limitations. The study by Shreevastava and Foltz (2021), for instance, reported only a 61% internal consistency in detecting distortions. Shickel et al. (2020) encountered challenges in accurately classifying non-distorted cases. Furthermore, research of Tauscher et al. (2023) had a low performance in the F1 score of merely 0.62. Importantly, a recent study (Maddela et al., 2023) showcased the effectiveness of fine-tuned language model for cognitive distortion detection, achieving scores above 0.9 thanks to a large crowdsourced dataset. Similarly, Wang et al. (Wang et al., 2023) released a Chinese dataset in 2023, achieving a 0.73 F1 score, also leveraging a large crowdsourced dataset. However,

a limitation of both studies is that the cognitive distortion sentences used were artificially created for the specific task rather than being derived from genuine conversations, potentially limiting the models' real-world applicability.

To sum up, recent research findings underscore the ongoing collective effort on the detection of cognitive distortions using language models. This fast-evolving landscape highlights the potential of large language models on this particular task, with a particular emphasis on the construction of a specialized dataset. We emphasize the importance of the detection task over classification due to its higher practical applicability. In a conversation, distortion detection may then activate the rest of the cognitive restructuring framework. Notably, there is currently no pertinent dataset available in the Chinese context, based on real-world dialogues, annotated by trained specialists.

2.2 Text Reframing

Earlier works in text style transfer primarily focus on altering semantic meaning, encompassing tasks such as rewriting biased articles (Ma et al., 2020; Reid et al., 2020), modifying proper nouns (Van den Bercken et al., 2019), and style transfer on comments from social media (Sudhakar et al., 2019).

A study (Ziems et al., 2022), introduced the notion of employing text reconstruction to foster a positive outlook, utilizing stress-tagged tweets from Twitter. However, their reconstructing strategies were extracted from previous research (Harris et al., 2007) focused more on forgiveness and spirituality in counselling practice. Instead, we believe that applying principles from the broaden-and-build theory can bolster the theoretical foundation of current and future works on the positive reframing task. Moreover, their approach uniformly rewrites all collected data, overlooking that all sentences do not necessarily present cognitive distortion. Their strategy was used by another study (Maddela et al., 2023), which created around thirty thousand English sentences with unhelpful thought patterns and their rewritten counterpart. Nevertheless, they focus more on creating and reconstructing cognitive distortion texts, not detecting them from existing real-life data.

Furthermore, Sharma et al. (2023a) introduced seven additional reconstruction strategies and curated a dataset comprising 1,077 samples labeled with thirteen cognitive distortion categories, along-

side a collection of 600 actively reconstructed corpora. While this study used English data only, they emphasized the importance of considering multiple languages beyond one. Based on this dataset, a followup study (Sharma et al., 2023b) developed a human language model interactive system, employing language models to assist individuals through cognitive reconstruction stages. While both studies underscore the detrimental effects of cognitive distortions, clarity on what kind of methods they used for distortion identification remains elusive.

3 Positive Reconstruction Framework

This section presents five positive reconstruction strategies based on the broaden-and-build theory which we describe in this section. Utilizing principles derived from this theory can enhance the theoretical underpinnings of both present and forthcoming endeavours on the positive reframing task. These strategies attempt to reframe negative thoughts from a positive perspective, fundamentally reorganizing rigid and harmful cognitive distortion beliefs.

The theory posits (Fredrickson, 2001, 2013) that positive emotions broaden an individual's scope of attention, cognition, and action. Additionally, positive emotions can undo the effects of sustained negative emotions, known as the undoing hypothesis. They construct positive psychological traits, providing immediate and long-term benefits to individuals. Our study will propose positive reconstruction strategies based on the key positive emotions outlined in the theory, as shown in Table 1.

Optimistic Strategy. According to the theory, joy refers to the feeling of happiness when unexpected good things happen in life and the anticipation of good things happening. Hope, on the other hand, represents an optimistic outlook towards the future. The difference between hope and joy lies in the fact that hope exists in stressful situations, where despite worries about the worst-case scenario, there is a simultaneous expectation for a better future. Therefore, we define the **optimistic strategy** as being hopeful for the future without disregarding nor denying the negative circumstances. For example, "I am so tired. Although I've been working so hard, I'll never succeed.," is a sentence that contains cognitive distortion, using optimistic strategy to reconstruct it can be: "This process is truly exhausting, but I've put in a lot of effort and accumulated a lot. Continuing to persevere will

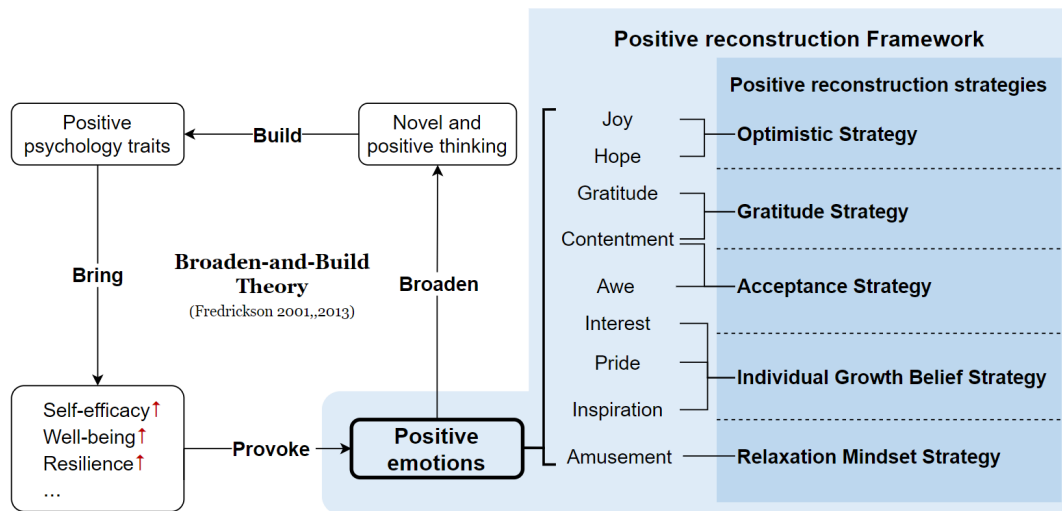


Figure 1: Broaden-and-build-theory (Fredrickson, 2001, 2013) and how positive reconstruction can initiate this virtuous cycle

Table 1: Positive reconstruction strategies and associated positive emotions.

Strategy	Description	Emotions
Optimism	To be hopeful for the future while not overlooking or denying negative circumstances, shifting the focus towards more positive aspects.	joy, hope
Gratitude	Developing a positive perspective on surrounding events and expressing gratitude.	thanksgiving, contentment
Individual Growth	The belief that individuals can grow and have infinite potential within themselves.	interest, pride, inspiration
Relaxation Mindset	Reducing the perceived harm of an event and relaxing tense negative emotions by understanding the limitations of negative experiences.	amusement
Acceptance	Embracing one's imperfections and acknowledging others' abilities or efforts.	awe, contentment

yield positive results.”

Gratitude Strategy. In the theory, positive emotion gratitude refers to the feeling that arises when an individual receives something beneficial to themselves. Contentment, on the other hand, occurs when an individual cherishes their current circumstances. Both can be seen as perspectives of gratitude towards the present situation, aiding individuals in cultivating a positive outlook on surrounding events, known as the **gratitude strategy**. Using the previously distorted statement as an example, applying the gratitude strategy could rephrase it as, "Despite the hardship, I'm truly grateful for the opportunity to be part of this task. I've never experienced such high-intensity work before. Now I have more experience in dealing with pressure."

Individual Growth Strategy centers on the core belief that individuals can grow over time and have

infinite potential within themselves. This links to key positive emotions in the theory such as interest, pride, and inspiration. Interest fosters an impulse for exploration and learning, allowing individuals to immerse themselves in novel experiences, expanding the self and facilitating personal growth. Pride often arises when individuals receive recognition, instilling a belief in their ability to achieve greater success in the future. Inspiration occurs when individuals witness others' commendable actions or outstanding performances, leading to a surge in self-motivation and a drive to surpass oneself, thus contributing to personal growth. These three positive emotions all contribute to fostering the belief that individuals can improve themselves. Employing this strategy to reconstruct statements with distortions, such as "I procrastinate too much and never finish anything." can be transformed into

"I can work on improving my time management skills." This shift in mindset has been proven to enhance students' academic performance (Yeager et al., 2014; Dweck and Yeager, 2019).

Relaxation Mindset Strategy. Understanding the limitations, impermanence, or commonality of errors reduces their perceived harm, thus easing tense negative emotions. For instance, when someone fails an exam, they might have thoughts like, "I failed the exam because I'm too stupid, I must be an idiot." However, the truth is that most people experience failures to some extent, and many have failed exams or even faced academic setbacks. According to the theory, amusement occurs when people encounter errors that pose small harm. Thus, a positive reconstruction to make the harmful error less severe could be, "I hope to perform better, but everyone makes mistakes; it's not just me who didn't pass."

Acceptance Strategy. According to the theory, contentment occurs when individuals feel comfortable within themselves or experience harmony with their external environment, it can be seen as self-acceptance. On the other hand, admiration arises when individuals encounter people or things that impress them, and a sense of awe motivates individuals to absorb and accept new experiences or individuals. Therefore, the **acceptance strategy** involves being able to calmly accept oneself without belittling others due to one's strengths or feeling inferior due to one's weaknesses.

4 Dataset construction

Previous research in cognitive distortion classification has revealed the challenges in achieving robust outcomes, necessitating ample data and high-quality annotation. Hence, we describe here the collection and annotation of a Chinese corpus with binary labels for the presence or absence of cognitive distortion, as well as positive reframing with five different strategies. All collected and annotated data are shared online¹.

4.1 Data Collection

The corpus labeling utilizes one specialized open-source dataset, the Chinese psychological Q&A dataset PsyQA (Sun et al., 2021), which is related to psychological issues and tends to contain statements with cognitive distortions.

¹<https://github.com/405200144/Dataset-of-Cognitive-Distortion-detection-and-Positive-Reconstruction/tree/main>

This study aims to determine if the initial sentence input by the topic initiator is distorted, rather than the subsequent responses or discussions. The length of the sentences is restricted to between 10 and 66 characters to avoid excessively short or long and meaningless sentences.

4.2 Annotation

Two postgraduate students with a background in psychology annotated 4,001 pieces of corpus for cognitive distortion detection based on the Burns 10 classification standard (Burns, 1999) (Appendix B), 1129 of which were labeled as cognitive distortions, and internal consistency is 88%.

The rewriting task was posted on the college student social platform. A total of 85 undergraduates and postgraduates participated in the pre-writing test. Finally, 42 participants passed the test and participated in the writing task, including 17 males and 25 females, from different majors, with an average age of 22.7. Each writer is requested to compose five distinct positive strategy reconstructions for one cognitive distortion sentence and complete 5 to 10 sets. In total, this process resulted in the collection of 1,900 sentences.

4.3 Data Quality

In addition to using a pre-test to ensure that writers clearly understand the writing task (Appendix C), we also conducted a 5-class classification experiment, a sentiment assessment, and a manual evaluation to verify the quality of the writing. In the five-classification experiment, we split the training and test sets in a 4:1 ratio, using Chinese pre-trained model RoBERTa-wwm-ext (Cui et al., 2021), an enhanced version of BERT (Kenton and Toutanova, 2019), as the pre-trained model. The accuracy metric reached 82.85% indicating that writers can compose different positive sentences based on various strategies. Furthermore, we utilized SnowNLP² for sentiment positivity analysis of the corpus. The results indicated a notable uptrend for the manually written content. The average sentiment score of the original sentences was 0.72, while that of the manually written content was 0.96. In the manual evaluation, the average scores for handwritten submissions all exceeded 4 on a 5-point scale (Table 4).

²<https://github.com/isnowfy/snownlp>

	Acc	F1	R	P
Linear	0.858	0.742	0.701	0.788
MLP	0.863	0.775	0.808	0.744
LSTM	0.874	0.781	0.769	0.793
Fine-tuning	0.896	0.822	0.821	0.824

Table 2: **Results of Cognitive Distortion Detection** with linear, multilayer perceptron (MLP), LSTM readouts, and fine-tuning the pretrained RoBERTa-wwm-ext network, measured by Accuracy (Acc), F1 score (F1), Recall (R), and Precision (P)

5 Cognitive Distortion Detection

In our study addressing the detection of cognitive distortions, we trained a binary classification model using a corpus of 4,001 entries annotated by individuals with a background in psychology. The sentences are fed into the neural network for a binary classification task to detect cognitive distortions. The dataset was randomly divided into a training set and a test set in a 4:1 ratio. Given the imbalanced nature of our dataset, where only approximately 28% of the data represented distorted cognitions, we utilized focal loss to enhance the model’s performance (with hyperparameters $\alpha = 0.25$ and $\gamma = 2$). We trained for 50 epochs with the AdamW optimizer and a learning rate equal to 10^{-5} , determined using a coarse grid search.

We employ RoBERTa-wwm-ext as pre-trained model (Cui et al., 2021). Two main strategies have been implemented to explore the performance of this language model for our particular task. First, transfer learning freezes all the weights in the pre-trained RoBERTa-wwm-ext network and augments it with a linear layer, a multilayer perceptron, or an LSTM model on the sequential RoBERTa-wwm-ext output. This last strategy is inspired by previous studies (Tan et al., 2022; Sirisha and Bolem, 2022) which have shown the potential of hybrid RoBERTa-LSTM for sentiment analysis. The multilayer-perceptron has 2 hidden layers of dimensions 512 and 256, with dropout rate at 0.2, and the LSTM hidden dimension is set to 256.

The second strategy is fine-tuning, in which we tune the weights of the whole RoBERTa-wwm-ext network with a linear classification readout for our classification task. This is more computationally-demanding and one has to be careful about catastrophic forgetting but it enables the network to learn better embeddings for the task at hand.

Results are presented in Table 2. We see that the

best performing model is obtained with fine-tuning. The improvement is quite significant compared to transfer learning methods, which means that the embeddings by RoBERTa-wwm-ext are probably suboptimal for the cognitive distortion classification. Among the different transfer learning methods, using more complex readout layers improves the performance as we observe that the linear readout layer presents the worst performance.

6 Positive Reconstruction

The emergence of large-scale language models has revolutionized various tasks, enabling their completion through prompt engineering (Meskó, 2023; Liu and Chilton, 2022). This led us to wonder whether the performance of prompt engineering parallels or surpasses that of fine-tuning for positive reconstruction generation.

6.1 Experimental Setup

Compared to the binary classification for cognitive distortion detection, the task at hand involves conversational Large Language Models for text generation. We leverage the recent developments with open-source models and APIs to benchmark 4 different approaches:

- P-tuning (V2) of ChatGLM-6B (Du et al., 2022)
- Fine-tuning of ChatGPT-3.5 Turbo-1106 (Ouyang et al., 2022)
- Prompt engineering with ChatGLM-6B
- Prompt engineering with ChatGPT-3.5

6.1.1 Prompt engineering

We provided each model with one set of examples shown in Appendix D. We installed ChatGLM-6B by following their instructions on Github³, and accessed the API⁴ from the free version of ChatGPT. Prompt engineering offers the advantage that it can be used with closed models, even in situations where access is limited to model outputs rather than weights. Moreover, as it is based on generalist language models, it does not require a parallel (or supervised) dataset, only the current sentence is required.

³<https://github.com/THUDM/ChatGLM-6B>

⁴<https://platform.openai.com/docs/api-reference>

Table 3: **Algorithmic evaluation results of positive reconstruction** measured by ROUGE-1, ROUGE-L, BLEU, BERTScore, and Sentiment analysis with SnowNLP, for P-tuning (V2) of ChatGLM-6B, Fine-tuning of ChatGPT3.5, and prompt engineering of both models.

Method	Model	Auto evaluation				
		ROUGE-1	ROUGE-L	BLEU	BERTScore	Sentiment
Fine-tune	ChatGLM-6B	31.89	27.14	10.67	32.25	0.91
	ChatGPT3.5	30.00	25.70	10.68	29.94	0.98
Prompt	ChatGLM-6B	24.69	19.04	6.70	23.51	0.93
	ChatGPT3.5	28.73	23.69	9.98	28.09	0.93

6.1.2 Training ChatGLM-6B Based on P-Tuning (V2)

P-Tuning v2 (Liu et al., 2022) is based on Deep Prompt Tuning (Lester et al., 2021; Qin and Eisner, 2021) designed for generation and knowledge probing. For a large language model with 300M to 10B parameters, P-Tuning v2 can achieve similar results as Fine Tuning by introducing trainable "prompts" to each layer in the language model, but it only requires the tuning of 0.1% to 3% of the parameters, which greatly reduces the cost of training.

ChatGLM-6B is trained with a batch size of 1 and global training steps of 3,000, and a learning rate of 0.02 is applied. We randomly split the datasets in an 8:1:1 ratio for training, validation, and testing. The maximum input and target token lengths are set to 64, tailored for concise and focused input-output relationships. Additionally, 4-bit quantization is employed, optimizing the model for efficient computation. Optimization on an NVIDIA RTX A4000 GPU (16 GB) took approximately 5 hours.

Table 4: **Human evaluation results** in the preservation of meaning (M), positivity (P), and an overall appreciation (O) of the reconstructed sentences. Comparison between P-Tuning (V2) of ChatGLM-6B (GLM), Fine-tuning of ChatGPT3.5 (GPT3.5), prompt engineering of both models, and ground truth.

Method	Model	Human evaluation		
		M	P	O
Fine-tune	GLM	4.08	3.88	3.87
	GPT3.5	3.05	3.25	2.95
Prompt	GLM	2.92	3.29	2.87
	GPT3.5	3.57	3.79	3.60
Ground Truth		4.04	4.17	4.02

6.1.3 Training ChatGPT 3.5 Turbo-1106 Based on Fine-tuning

For the fine-tuning process of ChatGPT 3.5 Turbo-1106, we harnessed OpenAI's existing API. According to OpenAI⁵, their fine-tuning API enhances few-shot learning by training on a significantly larger set of examples than can be accommodated within the prompt. This approach enables users to achieve improved results across a wide array of tasks. Once a model has been fine-tuned, users will not need to provide as many examples in the prompt.

6.2 Evaluation

Our research involved assessing the performance of our models in terms of their semantic similarity to the original text. Following other sentence reconstruction work (Ziems et al., 2022; Sharma et al., 2023a), we used several metrics for this evaluation: BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2019). Given that each cognitive distortion sentence had five reconstructed ground truth annotations, we documented these independently and selected the highest results to put in the Table 3. Additionally, we determined the sentiment score using SnowNLP.

We conducted a human evaluation to obtain more insights about the ground truth and model-generated sentences. Each participant was asked to rate 50 sentences on a five-point Likert scale based on three aspects: the preservation of meaning, the positivity, and the overall appreciation of the sentences. A total of 29 evaluation feedbacks from non-research-related personnel were received.

6.3 Results and Analysis

In the automatic sentiment evaluation results (Table 3), it was found that prompt results generated by

⁵<https://platform.openai.com/docs/guides/fine-tuning> (Accessed December 2023)

Table 5: **Algorithmic evaluation results** of sentences generated by ChatGLM-6B P-Tuning (V2) compared with the ground truth of five different strategies.

Strategy	R-1	R-L	BLEU	BScore
Optimism	31.89	27.14	10.67	32.25
Gratitude	29.82	25.16	10.11	30.56
Growth	31.44	26.40	10.35	31.73
Relaxation	30.46	25.29	9.97	29.36
Acceptance	29.69	24.31	9.05	30.24

ChatGPT-3.5 (GPT-3.5) achieved the highest score of 0.98. However, the difference between the four models was minimal, and all showed a significant improvement in positivity compared to the negative original sentences 0.72. On the other hand, in the aspect of semantic similarity, ChatGLM-6B (GLM) P-Tuning scored the best results across all four metrics, followed by GPT-3.5 fine-tuning, with GLM prompt performing the worst. These results indicate that fine-tuning large language models performs better than prompt engineering for positive reconstruction tasks. As mentioned earlier, this underscores the significance of possessing a parallel dataset for the process of supervised fine-tuning. Interestingly, fine-tuning yields better results than prompt engineering, a finding that diverges from the conclusion presented in (Maddela et al., 2023). This underscores the importance of tailoring approaches to individual tasks.

Additionally, we also observed that GLM P-Tuning outperformed GPT-3.5 fine-tuning, a finding that aligns with a previous study (Hu et al., 2023), which found that in certain specialized tasks, smaller language models, such as LLaMA-13B and LLaMA-7B, demonstrated superior performance compared to larger counterparts like ChatGPT. At the same time, it is important to point out here that GPT-3.5 Fine-Tuning results show limited improvement over Prompt Engineering. This aligns with the observation in (Lester et al., 2021) that fine-tuning is relatively effective only for models with less than 10 billion parameters.

We present in Table 5 an algorithmic evaluation of the sentences generated by the best performing model, ChatGLM-6B with P-Tuning (v2), to compare them with the different strategies employed. This makes use of our dataset which also classifies each positive reconstruction according to its reframing strategy. We also observed that the outputs of each model were more similar to the op-

timistic ground truth as shown in Table 5. Given that the proportions of the five strategies in the fine-tune training set were similar, this result may suggest that large language models are inherently more likely to generate statements similar to the optimistic strategy, or that the training corpus with different strategies contains more sentences that resemble the optimistic strategy. This call for further exploration and research to analyze the bias in the outputs of language models.

The results of the human evaluation (Table 4) were similar to the automatic assessment. Among the four models, GLM P-Tuning performed the best, even exceeding the manually written scores in terms of preserving the original meaning, while GLM prompt was the worst. For GPT-3.5, automatic evaluation metrics are higher with the fine-tuning strategy, while the prompt engineering version performs better for human evaluation. The reason for this discrepancy is also something we hope to explore further.

7 Discussion

In conclusion, this study demonstrated the efficacy of NLP techniques in detecting and reconstructing cognitive distortions. We introduced a Positive Reconstruction Framework, grounded in positive psychology theory, and developed a dataset based on its five strategies. This framework was applied to fine-tune large language models, and its performance was compared with that of prompt engineering. Our research indicates that fine-tuning these models is more effective than prompt engineering.

In future work, a key focus will be conducting extensive user testing experiments to determine whether sentences reconstructed with a positive approach can effectively alleviate negative emotions or stimulate positive ones in users. Additionally, exploring the vast potential of prompt engineering to identify the most effective prompts for optimal resource utilization will be a priority. Also, further research into how the size of large language models influences their fine-tuning performance in specific domain tasks will be crucial, offering a comprehensive understanding of their application in practical scenarios.

Ethics

The data used for this study was obtained from an online public psychological Q&A, which contains some overly negative sentences that may have a

negative impact on annotators. To mitigate this potential risk, we made sure that all contributors were 18 years or older and none of them would be affected by any negative statements before starting the annotation task. Those with mental health issues, who felt they might be susceptible to such issues, or who could be influenced by negative emotions severely, were not permitted to participate in the data construction tasks.

Limitations

While this study tackles the problem of language-specific datasets and algorithms, it should be noted that our focus was primarily on the Chinese language. Future research is needed to include a more diverse range of languages, thereby providing a more comprehensive understanding of cognitive restructuring across different linguistic and cultural contexts.

Furthermore, it's important to acknowledge that this work is still far from any immediate application. While we have made strides in constructing the Mandarin dataset and benchmarking various NLP models, future studies with more refined human evaluation, investigating biases and risks, are required before it can directly benefit individuals seeking mental health support.

Additionally, it's worth noting that some of the NLP models utilized in this study are only accessible through paid APIs, limiting their accessibility to researchers and practitioners without the necessary resources.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback and suggestions. This work is based on work supported by the Chinese National Social Sciences Foundation (No. BBA210042).

References

- Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.
- Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.
- David D Burns. 1999. *The Feeling Good Handbook: The Groundbreaking Program with Powerful New Techniques and Step-by-Step Exercises to Overcome Depression, Conquer Anxiety, and Enjoy Greater Intimacy*. Penguin.
- Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074.
- Emil Chiauzzi, Athena Robinson, Kate Martin, Carl Petersen, Nicole Wells, Andre Williams, Mary Margaret Gleason, et al. 2023. A relational agent intervention for adolescents seeking mental health treatment: Protocol for a randomized controlled trial. *JMIR Research Protocols*, 12(1):e44940.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Robyn Mason Dawes. 1964. Cognitive distortion. *Psychological Reports*, 14(2):443–459.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.
- Carol S Dweck and David S Yeager. 2019. Mindsets: A view from two eras. *Perspectives on Psychological Science*, 14(3):481–496.
- Barbara L Fredrickson. 2001. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American psychologist*, 56(3):218.
- Barbara L Fredrickson. 2013. Positive emotions broaden and build. In *Advances in experimental social psychology*, volume 47, pages 1–53. Elsevier.
- Alex HS Harris, Carl E Thoresen, and Shane J Lopez. 2007. Integrating positive psychology into counseling: Why and (when appropriate) how. *Journal of Counseling & Development*, 85(1):3–13.
- Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv e-prints*, pages arXiv–2304.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- 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*, pages 3045–3059.

- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Shuya Lin, Lingfeng Lin, Cuiqin Hou, Baijun Chen, Jianfeng Li, and Shiguang Ni. 2023. Empathy-based communication framework for chatbots: A mental health chatbot application and evaluation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 264–272.
- Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. **P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.
- Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. Powertransformer: Unsupervised controllable revision for biased language correction. pages 7426–7441.
- Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather M Foran, and Y-Lan . 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of Medical Internet Research*, 25:e50638.
- Melissa Mulraney, Cindy Lee, Gary Freed, Michael Sawyer, David Coghill, Emma Sciberras, Daryl Efron, and Harriet Hiscock. 2021. How long and how much? wait times and costs for initial private child mental health appointments. *Journal of Paediatrics and Child Health*, 57(4):526–532.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- 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 of the Association for Computational Linguistics*, pages 311–318.
- Judith J Prochaska, Erin A Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study. *Journal of medical Internet research*, 23(3):e24850.
- Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212.
- Pryzant Reid, Martinez Richard Diehl, Dass Nathan, S Kurohashi, Jurafsky Dan, and Yang Diyi. 2020. Automatically neutralizing subjective bias in text. In *AAAI*.
- Katerina Rnic, David JA Dozois, and Rod A Martin. 2016. Cognitive distortions, humor styles, and depression. *Europe's journal of psychology*, 12(3):348.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2021. A computational approach to understanding empathy expressed in text-based mental health support. In *EMNLP*.
- Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023a. **Cognitive reframing of negative thoughts through human-language model interaction**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.
- Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2023b. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. *arXiv e-prints*, pages arXiv–2310.
- Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE.
- 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*, pages 151–158.
- SingleCare Team. 2020. Mental health survey 2020. <https://www.singlecare.com/blog/news/mental-health-survey/>. Accessed: 2023-3-16.
- Uddagiri Sirisha and Sai Chandana Bolem. 2022. Aspect based sentiment & emotion analysis with roberta, lstm. *International Journal of Advanced Computer Science and Applications*, 13(11).
- Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. “transforming” delete, retrieve, generate approach for controlled text style transfer. pages 3269–3279.

- Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. pages 1489–1503.
- Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. 2022. Roberta-istm: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525.
- Justin S Tauscher, Kevin Lybarger, Xiruo Ding, Ayesha Chander, William J Hudenko, Trevor Cohen, and Dror Ben-Zeev. 2023. Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric services*, 74(4):407–410.
- Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.
- Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023. C2d2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160.
- WHO. 2021. Depression. <https://www.who.int/news-room/fact-sheets/detail/depression>. Accessed: 2021-12-16.
- WHO. 2023. Who launches commission to foster social connection. <https://www.who.int/news/item/15-11-2023-who-launches-commission-to-foster-social-connection>. Accessed: 2023-11-23.
- David Scott Yeager, Rebecca Johnson, Brian James Spitzer, Kali H Trzesniewski, Joseph Powers, and Carol S Dweck. 2014. The far-reaching effects of believing people can change: implicit theories of personality shape stress, health, and achievement during adolescence. *Journal of personality and social psychology*, 106(6):867.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.
- Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824.
- Ling Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-party empathetic dialogue generation: A new task for dialog systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–307.
- Caleb Ziem, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing positive perspectives with text reframing. pages 3682–3700.

A Positive Reconstruction

Table 6: A model comparison example of positive reconstruction among the same original text measured by ROUGE-1(R-1), ROUGE-L(R-L), BLEU, BERTScore, Sentiment analysis, and human evaluation with the different models, along with English translations and associated scores. GLM-6B and GPT3.5 represent ChatGLM-6B and ChatGPT3.5, respectively.

<i>Original Text: I'm not sure what to do either;</i>								
<i>I find myself getting angry easily every day and taking it out on others.</i>								
<i>我也不知道怎么办，每天都容易发脾气，牵怒到别人。</i>								
Model	Algorithmic evaluation					Human evaluation		
	R-1	R-L	BLEU	BERTScore	Sentiment	Meaning	Positivity	Overall
GLM-6B P-Tuning (V2)	41.43	34.81	6.91	43.88	0.93	4.38	3.46	3.85
<p><i>Although I easily lose my temper and take my anger out on others every day, I believe I can try to change and become a person with stable motions.</i></p> <p><i>虽然每天都容易发脾气、牵怒到别人，但我相信我可以试着改变，变成情绪稳定的人。</i></p>								
GPT3.5 Fine-Tuning	34.31	24.97	7.37	35.87	0.99	3.69	3.46	3.46
<p><i>Although I easily lose my temper every day, everyone develops from immaturity, and I can continue to improve my emotional control and communication skills.</i></p> <p><i>虽然每天都容易发脾气，但是每个人都是从不成熟发展起来的，我可以继续提高自己的情绪控制和沟通能力。</i></p>								
GLM-6B Prompt	24.66	16.41	7.16	28.79	1.00	2.85	3.77	3.00
<p><i>Learn to control your emotions and try to reduce the frequency of losing your temper. If you find it difficult to control your emotions, consider techniques like deep breathing or meditation to relax. You could also try talking to a close friend or family member, sharing your feelings and emotions to seek their advice and support.</i></p> <p><i>学会控制自己的情绪，尽可能减少发脾气的次数。如果无法控制自己的情绪，可以尝试深呼吸、冥想等方式来放松自己。也可以尝试与亲密的朋友或家人谈心，倾诉自己的感受和情绪，获得他们的建议和支持。</i></p>								
GPT3.5 Prompt	38.97	33.15	10.44	43.44	0.99	3.54	3.54	3.80
<p><i>I've realized that I've been quick to anger lately. I will try to alleviate my emotions by learning relaxation techniques, such as practicing yoga or meditation. I am committed to managing my emotions better and avoiding taking my anger out on others.</i></p> <p><i>我意识到自己最近容易发脾气，我会尝试通过学习放松的方法，比如做瑜伽或者练习冥想来缓解自己的情绪。我会努力控制自己的情绪，避免牵怒到别人。</i></p>								

B Cognitive Distortion Detection Annotation

This task was conducted by two postgraduate students with a background in psychology. The goal was to distinguish whether texts contain any distortion according to 10 Cognitive Distortion type of Burns (1999), annotators gave the label 0 as non-distortion, and 1 as distortion.

Table 7: 10 Cognitive Distortion type of Burns (1999)

Distortion type	Description
Emotional Reasoning	Reasoning base on emotions, believing “I feel that way, so it must be true”
Overgeneralization	Drawing conclusions with limited and often un negative experience.
Mental Filter	Focusing only on limited negative aspects and not the excessive positive ones.
Should Statement	Expecting things or personal behavior should be a certain way.
All or Nothing	Binary thought pattern. Considering anything short of perfection as a failure.
Mind Reading	Concluding that others are reacting negatively to you, without any basis in fact.
Fortune Telling	Predicting that an event will always result in the worst possible outcome.
Magnification	Exaggerating or Catastrophizing the outcome of certain events or behavior.
Personalization	Holding oneself personally responsible for events beyond one’s control.
Labeling	Attaching labels to oneself or others (ex: “loser”, “perfect”).

C Data Collection Interface Snapshots and English translation

Before conducting the positive rewrite annotation task we set a pre-test to ensure that annotators can understand the task correctly.

负面想法在脑中浮现，往往会让人深信不疑，造成情绪上更加低落，因此若是能用积极的视角来重新阐述相同的句子，将有可能降低此负向信念带来的影响。
在本次任务中我们锁定了**5种积极策略**，作为重构策略，**请仔细阅读以下重构策略说明。**

乐观策略：为对未来充满希望，同时不忽视或否定消极状况的一种视角
例句：“我真的好累好担心，我已经拼命努力工作了，但是我永远不可能成功” → “这个过程真的让我好累，但是我付出了很多也累积了很多，继续坚持一定会有好结果的。”

感恩策略：对当下状况的一种感恩视角，此视角可以帮助个人对周围的事务产生正向的价值观念。
例句：“我真的好累好担心，我已经拼命努力工作了，但是我永远不可能成功” → “虽然很辛苦，但还是很感谢有参与这个任务的机会，之前我从没经历过如此高强度的工作。”

人格成长信念策略：此策略的核心认为人可以随着时间而成长并且相信自身有无限的潜力。
例句：“我实在是太拖延了，什么事都做不完” → “我还可以好好培养时间管理能力。”

放松心态策略：核心观念在于理解负向经历的有限性、无常性及普遍性，通过此种认知的建立将事件的危害性降低，从而放松紧绷的负向情绪。
例句：“我考试不及格，一定是我太笨了，跟智障一样。” → “我希望自己能考得更好，但每个人都有犯错的时候，不是只有我会不及格。”

接纳策略：可以分为对自己的接纳以及对别人的接纳，对自身的接纳，是真实的认识自己并接受自己的不完美。另一方面，对别人的接纳包含敬佩能让自己折服的人事物，或是接纳遇到的新事物。亦即使自身能够坦然地接受自己，不因自身的优点而轻视别人也不因自身的缺点而自卑的一种视角。
例句：“我考试不及格，一定是我太笨了，跟智障一样。” → “我知道这次的考试我搞砸了，是我不太擅长的科目，或许我可以跟考高分的同学请教解题方式。”

When negative thoughts appear in the mind, people often believe them deeply, causing them to become even more depressed. Therefore, if the same sentence can be rephrased from a positive perspective, it will be possible to reduce the impact of this negative belief. In this task, we utilized 5 positive strategies as reconstruction methods. Please read the following descriptions of reconstruction strategies carefully.

Optimism Strategy: A perspective that provides hope for the future without ignoring or negating negative situations.
Example: “I am so tired. Although I’ve been working so hard, I’ll never succeed.” → “This process is truly exhausting, but I’ve put in a lot of effort and accumulated a lot. Continuing to persevere will yield positive results.”

Gratitude Strategy: A perspective of gratitude towards current situations can help foster positive values towards one’s surroundings.
Example: “I am so tired. Although I’ve been working so hard, I’ll never succeed.” → “Despite the hardship, I’m truly grateful for the opportunity to be part of this task. I’ve never experienced such high-intensity work before. Now I have more experience in dealing with pressure.”

Individual Growth Strategy: At the core of this strategy is the belief that individuals have the potential to grow over time without limits.
Example: “I procrastinate too much and never finish anything.” → “I can work on improving my time management skills.”

Relaxation Mindset Strategy: The key is to understand that negative experiences are finite, impermanent, and universal. This realization helps to reduce the harmfulness of events and alleviate negative emotions.
Example: “I failed the exam because I’m too stupid, I must be an idiot.” → “I hope to perform better, but everyone makes mistakes; it’s not just me who didn’t pass.”

Acceptance Strategy: Acceptance can be divided into two types: acceptance of oneself and acceptance of others. Acceptance of oneself means having a deep understanding of oneself and accepting one’s imperfections. On the other hand, acceptance of others involves appreciating people and things that inspire you, or being open to new experiences. This perspective allows you to accept yourself calmly, without looking down on others because of your own strengths, or feeling inferior because of your own weaknesses.
Example: “I failed the exam because I’m too stupid, I must be an idiot.” → “I did not do well on the exam, and I struggled with this subject. Perhaps I can ask my high-scoring classmates for help with this problem.”

Figure 2: Description of the 5 positive reconstruction strategies, this was shown in pre-test and the official rewriting task.

Two types of questions were presented in the pre-test. The first part consisted of 7 questions. We asked participants to choose the strategies they thought were most likely to be used by providing one negative and one positive sentence. Followed by answering three questions on choosing the most qualified rewritten sentence from four options in the second part. The goal was to ensure that the revised sentence conveyed the original meaning while avoiding overgeneralization. Participants who answered correctly to more than 8 questions can participate in the writing task.

第一部份:
请选出重构语句所使用的策略(7题)

* 2. 原句:
我现在无法好好的工作, 不想从事现在的工作, 从内心觉得这份工作没前途, 但是有没有勇气去改变。
重构句:
我现在虽然无法好好工作而且逐渐丧失勇气, 但我知道一路以来的付出都有累积, 一定会有回报。

- 乐观策略
- 感恩策略
- 人格成长信念策略
- 放松心态策略
- 接纳策略

* 3. 原句:
我没有什么特别的经历。
重构句:
我的经历也许跟大多数人类类似, 但这并不代表我的人生没有价值, 我可以尝试放松, 并专注于自己喜欢的事情来培养自己的独特性。

- 乐观策略
- 感恩策略
- 人格成长信念策略
- 放松心态策略
- 接纳策略

Part one:
Please select the strategy used to reconstruct the statement (7 questions)

2. Original sentence:
I can't work well now, and I don't want to do my current job. I feel from the bottom of my heart that this job has no future, but I don't have the courage to change it.
Reconstruct the sentence:
Although I can't work well now and am gradually losing my courage, I know that my efforts along the way have been accumulated and will definitely pay off.

- Optimism Strategy
- Gratitude Strategy
- Individual Growth Strategy
- Relaxation Mindset Strategy:
- Acceptance Strategy

3. Original sentence:
I don't have any exceptional experiences.
Reconstruct the sentence:
Although my experiences may be similar to others, my life still holds value. By taking time to relax and pursue my passions, I can cultivate my uniqueness.

- Optimism Strategy
- Gratitude Strategy
- Individual Growth Strategy
- Relaxation Mindset Strategy:
- Acceptance Strategy

Figure 3: Example of distinguishing different strategies in pre-test.

第二部份:
重构句子除了有积极策略之外, **保留句子本意, 不随意更改句子的意思, 不过于泛化**也是很重要的一部份, 如果一个重构句可以适用于多个情境, 此为过度泛化。
请选出最保留原意且不会过于泛化的重构句(3题)

* 9. 请选出最保留原意且**不会过于泛化**的重构句。
原句:
总是控制不住的想到可怕的事, 没办法做到不去想, 很难受害怕。

- 虽然我有负面想法, 但我相信未来会变得更好, 我可以通过努力学习 and 改变自己的思维方式, 逐渐减少这种情况。
- 虽然我有时想法很负面, 但我相信我可以成长并变得更好, 我会花时间学习和发展自己, 以克服这种负面情绪。
- 我常常会想到可怕的事情, 但是我相信只要我开始行动, 我就有能力突破这个困境, 或许我可以尝试一些转移注意力的练习, 并在控制不住时使用。
- 我接受自己的状况, 这是我的一部分, 我不会因此而贬低自己, 也不会用这个缺点去评判别人, 我会尽力接受并处理这种情况。

Part two:
In addition to having a positive strategy when reconstructing a sentence, it is also important to **retain the original meaning** of the sentence, **not to change the meaning of the sentence arbitrarily, and not to overgeneralize**. If a reconstructed sentence can be applied to multiple situations, this is overgeneralization.
Please select the reconstructed sentence that best retains the original meaning and is not overgeneralize. (3 questions)

9. Please select the reconstructed sentence that best retains the original meaning and is **not overgeneralize**.
Original : *I can't stop thinking about terrible things and it's making me feel very uncomfortable and scared.*

- Although I experience negative thoughts, I believe I can improve my future through hard work and changing my mindset.*
- Although I tend to have negative thoughts, I believe that I can improve myself. I am willing to put time and effort into learning and growing to overcome this negativity.*
- I often have negative thoughts, but I believe taking action can help me break through. I plan to try distraction exercises when I struggle to control myself.*
- I acknowledge and accept my condition as a part of my identity. I will refrain from belittling myself or using this weakness to pass judgment on others. Instead, I will make an effort to cope with the situation to the best of my ability.*

Figure 4: Example of selecting the most qualified rewritten sentence in pre-test.

接下来请您完成10个句子的5种积极重构

* 1. 控制不住大脑思维, 明知是错却无力摆脱, 人生的苦都储存在这个脑壳里。做一个情绪的奴隶, 做一个环境的傀儡, 是极度可怜可悲的。

乐观策略重构
请输入

感恩策略重构
请输入

人格成长信念策略重构
请输入

放松心态策略重构
请输入

接纳策略
请输入

Next, please complete 5 positive reconstructions of 10 sentences.

Unable to control thinking, knowing that something is wrong but unable to get rid of it, all the suffering in life is stored in the brain. It is extremely pitiful and pitiful to be a slave of emotions and a puppet of the surroundings.

Optimism Strategy
please enter

Gratitude Strategy
please enter

Individual Growth Strategy
please enter

Relaxation Mindset Strategy
please enter

Acceptance Strategy
please enter

Figure 5: Example of positive rewriting task.

积极重构语料评估

推荐使用**电脑**填写:

负面想法在脑中出现, 往往会让人深信不疑, 造成情绪上更加低落, 因此若是能用**积极的视角来重新阐述相同的句子**, 将有可能降低此负向信念带来的影响。
好的积极改写必须**保留句子本意, 不否定句子的意思**, 并以**积极的视角来重新阐述**。

请根据原句为改写句打分。

相关性: 句子原意保留度, 指保留了原句提到的所有关键信息, 与关键信息的语意保持一致, 不否定原句的消极状态。

积极度: 是否以积极视角阐述原句

改写效果: 对改写效果的总体评价

* 1. 原句:
总是考不好, 记不住, 然后就焦虑, 无法正常上课, 已经去过医院了, 但感觉也没什么用。

	相关性 <i>Preservation</i>	积极度 <i>Positivity</i>	改写效果 <i>Overall effect</i>
改写1: 我明白考试不好会带来焦虑和不便, 但我也努力克服这些困难。我已经去过医院, 但感觉没什么用, 所以我想尝试其他方法来改善我的状况。或许我可以寻求专业帮助, 比如心理医生, 来更好地理解应对自己的情绪和问题。	★★★★★	★★★★★	★★★★★
改写2: 我理解自己的情绪失控可能是一个困扰我的问题, 但我也意识到这并不意味着我不能克服它。我会尝试通过冥想, 深呼吸, 参加运动等方式来控制自己的情绪, 同时也接受自己情绪失控的事实, 并努力改善它。	★★★★★	★★★★★	★★★★★
改写3: 我总是考不好, 记不住, 然后就焦虑, 无法正常上课, 但幸好我发现了这个状况, 并且想要改变。感谢这种经历可以让我体会焦虑的感受, 让我更加注重记忆和记忆深刻性, 我相信通过合理的调整 and 改变, 我一定会变得更好。	★★★★★	★★★★★	★★★★★
改写4: 虽然总是考不好, 记不住, 然后就焦虑, 无法正常上课, 已经去过医院了, 但我觉得能再多试几次, 我可以改变现状。	★★★★★	★★★★★	★★★★★
改写5: 虽然我总是考不好, 记不住, 导致焦虑, 无法正常上课, 即使去过医院也感觉没有太大改变, 但我相信未来会变得更好, 我会继续努力, 相信自己的能力, 坚持下去, 每一次努力都是对成功的一步靠近。	★★★★★	★★★★★	★★★★★

Positive reconstruct corpus evaluation

It is recommended to use a **computer** to fill in the form.

When negative thoughts appear in the mind, people often believe them deeply, causing them to become even more depressed. Therefore, **if the same sentence can be rephrased from a positive perspective, it will be possible to reduce the impact of this negative belief.**

A good positive rewording must **retain the original meaning of the sentence, not negate the meaning of the sentence, and reformulate it from a positive perspective.**

Please rate the rewritten sentences based on the original sentence.

Preservation : The retention of the original meaning of the sentence refers to retaining all the key information mentioned in the original sentence, keeping the semantics of the key information consistent, and not negating the negative status of the original sentence.

Positivity : Whether to elaborate on the original sentence from a positive perspective.

Overall appreciation : General opinion of the reconstructed sentences.

1. Original sentence:
I always failed in exams and couldn't remember, and then I became anxious and couldn't attend classes normally. I've already seen the doctor, but it didn't seem to be helpful.

Rewrite 1:
I understand that doing poorly on exams can cause anxiety and inconvenience, but I will work hard to overcome these difficulties. I've seen the doctor but nothing different so I want to try other things to improve my condition. Perhaps I could seek professional help, such as a psychiatrist, to better understand and cope with my emotions and problems.

Rewrite 2:
I understand that my lack of emotional control may be an issue that bothers me, but I also realize that that doesn't mean I can't overcome it. I try to control my emotions through meditation, deep breathing, and exercise. I acknowledge that my emotions can be difficult to control, and I am actively working to improve them.

Rewrite 3:
I always failed in exams and couldn't remember, and then I became anxious and couldn't attend classes normally. But fortunately, I discovered this situation and wanted to make a change. Thanks to this experience, I can experience the feeling of anxiety and make me pay more attention to memory and memory profundity. I believe that through reasonable adjustments and changes, I will definitely get better.

Rewrite 4:
Although I have always struggled with exams and have difficulty retaining information, leading to anxiety and difficulty attending classes, even tried medical attention, I believe that with continued effort, I can improve my situation.

Rewrite 5:
Although I have always struggled with exams and have difficulty retaining information, leading to anxiety and difficulty attending classes, not much changed after medical treatment, I believe that the future will get better, and I will continue to work hard, believe in my abilities, and persevere. Every effort is a step closer to success.

Figure 6: Example of Human Evaluation Interface Snapshots and English translation.

We used a free questionnaire platform⁶ to set up pre-tests, rewrite tasks and evaluate questionnaires.

⁶<https://www.wenjuan.com/>

D Example Prompt for the Task of Reframing Negative Thoughts

Prompt Content: [我需要你帮助我完成心理积极的重写任务，首先我会给你几个例子：输入和输出。输出是我希望你为我重写的内容。（这些是例子，之后我只会给你输入），例子：

“I need you to help me with a mental positive rewriting task, first I'll give you a couple examples: input and output. Output is what I want you to rewrite for me. (These are examples, after that I will only give you the INPUT), examples:”

+

{input1: output1},{input2: output2},{input3: output3}, {input4: output4} ,{ input5: output5}

+

“Please help me rewrite the following sentence with reference to these previous examples, only one answer is needed:”

+

{ input:one sentence from test set} 请参考前面的例子帮我重写下面的句子，只需要一个答案：

Figure 7: The translated English prompt example used for text reconstruction based on single turn conversation: (same for both ChatGPT3.5 and ChatGLM-6B). Specifically, we give five examples based on the five strategies, and finally we give the intended reconstruction sentence at the end of the prompt. We implement prompt engineering using the test dataset.