Does "Reasoning" with Large Language Models Improve Recognizing, Generating and Reframing Unhelpful Thoughts?

Yilin Qi[*♢] **Dong Won Lee**[*♣] **Cynthia Breazeal**[♣] **Hae Won Park**[♣]
[♢]Harvard University [♣]Massachusetts Institute of Technology

Abstract

Cognitive Reframing, a core element of Cognitive Behavioral Therapy (CBT), helps individuals reinterpret negative experiences by finding positive meaning. Recent advances in Large Language Models (LLMs) have demonstrated improved performance through reasoning-based strategies. This inspires a promising direction of leveraging the reasoning capabilities of LLMs to improve CBT and mental reframing by simulating the process of critical thinking, potentially enabling more effective recognition, generation and reframing of cognitive distortions. In this work, we investigate the role of various reasoning methods, including pre-trained reasoning LLMs, such as DeepSeek-R1, and augmented reasoning strategies, such as CoT (Wei et al., 2022) and self-consistency (Wang et al., 2022), in enhancing LLMs' ability to perform cognitive reframing tasks. We find that augmented reasoning methods, even when applied to "outdated" LLMs like GPT-3.5, consistently outperform state-of-the-art pretrained reasoning models such as DeepSeek-R1 (Guo et al., 2025) and o1 (Jaech et al., 2024) on recognizing, generating and reframing unhelpful thoughts.

1 Introduction

Cognitive Behavioral Therapy (CBT) (Beck, 1963) is one of the most widely used and well-supported approaches in psychotherapy (Fenn and Byrne, 2013). CBT focuses on both the process and content of thoughts, including core beliefs, assumptions, and automatic thoughts (Fenn and Byrne, 2013). Cognitive Reframing is central to CBT, helping individuals reinterpret negative experiences by critically reasoning through and aligning them with their belief systems to find purpose or positive meaning in adversity (Blum et al., 2012). Recent advancement in Large Language Models (LLMs) research have focused on reasoning, which stands out as a fundamental element of human intelligence that drives key processes like problem-solving, decision-making, and critical thinking (Huang and Chang, 2022). Furthermore, LLMs that incorporate reasoning in its pretraining phase or as a post-hoc augmentation procedure have shown significant improvement in performance across many tasks (Qiao et al., 2022).

In this paper, we investigate the extent to which reasoning can improve LLM's ability in Cognitive Reframing. We implement and evaluate three conditions of LLM reasoning on established cognitive reframing tasks, which include generating, recognizing, and reframing unhelpful thoughts. In addition, we propose a novel task of reframing thoughts conditioned on reframing strategies based on positive psychology (Harris et al., 2007). The reasoning conditions we evaluate include: (1) LLMs pre-trained specifically for reasoning; (2) LLMs augmented with state-of-the-art reasoning methods such as CoT (Wei et al., 2022), ToT (Yao et al., 2023), and self-consistency (Wang et al., 2022) and DoT (Chen et al., 2023); and (3) Non-reasoning LLMs that were not explicitly trained or augmented with reasoning capabilities. We find that reasoning-augmented models consistently outperform pre-trained reasoning models, suggesting that simply augmenting LLMs with reasoning strategies can provide strong performance gains on cognitive reframing tasks without the cost and complexity of pretraining explicitly for reasoning.

2 Related Work

Early AI Systems for Cognitive Reframing Early mental health chatbots and apps incorporated elements of Cognitive Reframing, but relied on scripted responses or simple AI (Hodson et al., 2024). Systems like the CBT-based chatbot Wysa could walk users through CBT-style prompts by using AI to select from pre-written therapist re-

[*]Equal Contribution.

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 62–69

sponses, but they lacked the flexibility to produce personalized new reframes (Hodson et al., 2024).

LLMs for Identifying and Reframing Unhelpful Thoughts Recent studies have begun leveraging LLMs to identify and reframe unhelpful thoughts in more flexible ways. Previous work explored LLM-assisted cognitive reframing by training a retrieval-augmented model to suggest alternative thoughts with controlled therapeutic attributes (Sharma et al., 2023). Others introduced a "Diagnosis of Thought" prompting technique that guides the model to separate facts from subjective interpretations and reason about evidence, significantly improving the detection of distorted thinking patterns while producing expert-approved explanatory rationales (Chen et al., 2023). These works demonstrate the feasibility of LLMs both in generating helpful reframed thoughts and in pinpointing unhelpful thinking.

Therapeutic Frameworks and Prompt Engineering To further enhance LLM-based cognitive restructuring, researchers have applied explicit therapeutic frameworks and structured prompting. RESORT framework provides a series of psychologically grounded reappraisal instructions (Zhan et al., 2024). Similarly, the HealMe system integrated core CBT techniques into the prompt structure, systematically guiding the LLM to distinguish circumstances from feelings, brainstorm alternative perspectives, and develop empathetic, actionable new thoughts (Xiao et al., 2024).

3 Experiments

In this work, we investigate the contribution of reasoning methods in cognitive reframing. We utilize the PatternReframe dataset (Maddela et al., 2023), where each sample contains (1) a persona (i.e *"I enjoy gardening. My favorite drink is red wine. I work for a clothes retailer. I have one child."*), (2) unhelpful thought (i.e. *"My child wishes they had another sibling. I bet they think I'm a horrible parent for stopping at one child."*), (3) the unhelpful thinking pattern (i.e. *"Jumping to conclusions: mind reading"*), and (4) the reframed positive thought (i.e. *"My child wishes they had another sibling, but I'm grateful I can focus all my attention on one child."*) and the aligned reframe strategy (i.e. *"Optimism"*). The unhelpful thinking patterns as well as strategies used to reframe unhelpful thoughts are both grounded in psychology literature (David and Burns, 1980), (Harris et al., 2007). We sample a set of 1,000 examples from the dataset such that the occurence of each unhelpful thinking pattern is distributed uniformly (∼100 per category, e.g., Personalization, Catastrophizing) for use across all tasks.

3.1 Methods

We experiment with three conditions of LLM models and reasoning methods. For the purpose of this work, we define "reasoning" as any systematic process that guides a model's decision-making steps beyond simple input-output mappings. **(1) Non-Reasoning (NR)** models include those that have not been specifically trained for reasoning purposes. In our experiments, we focus on `GPT-3.5`, `GPT-4`, `GPT-4o`. On the other hand, we also consider **(2) Pretrained Reasoning (PR)** models that have been specifically trained for reasoning, these include `Llama-3.3`, `Deepseek-R1` (Guo et al., 2025), `GPT-o1` and `GPT-o3-mini`. Finally, to study the effects of modern reasoning methods and prevent confounding analysis due to data leakage, we utilize `GPT-3.5` as the base model, as other recent models' data cutoff date is beyond the data release date for PatternReframe (Jul 2023). We consider state-of-the-art **(3) Augmented Reasoning (AR)** methods described below:

Chain-of-Thought (CoT) (Kojima et al., 2022; Wei et al., 2022): supplies LLMs with step-by-step reasoning demonstrations instead of conventional input-output pairs. We focus on the popular technique of `zero-shot CoT`, where a simple prompt of "Let's think step by step" is prepended to the prompt to facilitate step-by-step thinking.

Self-Consistency (SC) (Wang et al., 2022): is a reasoning method based on the decoding strategy, self-consistency. Instead of selecting a single greedy path, it samples a diverse set of reasoning paths and determines the most consistent answer by marginalizing over these sampled paths.

Tree-of-Thought (ToT) (Yao et al., 2023): is a framework that enhances language models' problem-solving by exploring multiple reasoning paths structured as a tree. Each node represents a partial solution, and the model generates, evaluates, and searches through these "thoughts" using strategies like breadth-first (BFS) or depth-first search (DFS). In our experiments, we use DFS.

Diagnosis-of-Thought (DoT) (Chen et al., 2023): is the most relevant to our work and was previously proposed for the same task of cognitive distortion detection. The method diagnoses a patient's speech through three stages: subjectivity assessment to

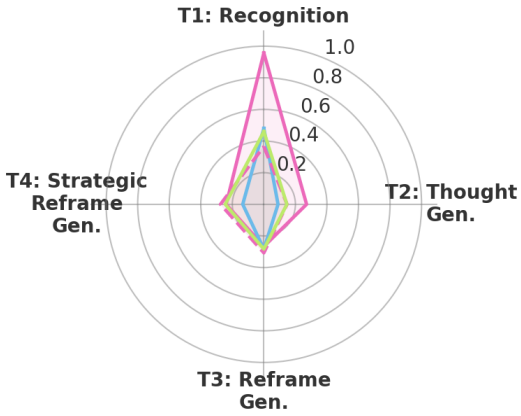

Figure 1: Performance for Representative Models in Each Class of Reasoning. Non-Reasoning Method —: GPT-4o; Pre-trained Reasoning Method —: o1; Reasoning-Augmented Method —: GPT-3.5 + DoT; - -: GPT-3.5 + Self-Consistency.

distinguish facts from thoughts, contrastive reasoning to explore reasoning processes that support or contradict the thoughts, and schema analysis to summarize cognitive schemas.

4 Tasks & Results

To evaluate the effectiveness of varying conditions of modern LLM reasoning methods, we incorporate the following tasks: (1) recognizing unhelpful thought patterns, (2) generating unhelpful thoughts, and (3) generating reframes of unhelpful thoughts, in line with the proposed tasks from PatternReframe (Maddela et al., 2023). Given the advances of instruction tuning and alignment (Ouyang et al., 2022), we propose a novel (4)-th task: generating strategic reframes of unhelpful thought, strictly enforcing the reframe of the unhelpful thought to be aligned to a specific reframing strategy. The performance of representative models from each condition (PR, NR, AR) are shown in Fig. 1, where we find that simple augmented reasoning methods perform well across all tasks, and obtain massive performance gains for the task of unhelpful thought pattern recogntion.

Task 1: Recognition of Unhelpful Thought Patterns assesses whether LLMs can recognize the unhelpful thinking pattern given a description of the persona and the unhelpful thought. An example prompt for this task can be found in App. B.1. We conduct an automatic performance evaluation using F1-score, accuracy, precision, and recall from prior literature (Maddela et al., 2023). The results for Task 1 are presented in Table 1. While pretrained reasoning (PR) methods generally outperform non-reasoning (NR) methods, a simple augmentation of the GPT-3.5 model with DoT (AR) achieves a remarkable performance across all metrics, outperforming the strongest pre-trained reasoning models, i.e. DeepSeek-R1 and o1, by a big margin of $\sim$ 40% in accuracy scores. Notably, DoT is specifically tailored for the task of cognitive distortion detection, which aligns directly with the set-up of Task 1. These results imply that, in recognizing unhelpful thought patterns, minimally adapting LLMs with task-aligned augmented reasoning methods can significantly surpass the performance of general-purpose reasoning models. However, while not requiring extensive fine-tuning, AR methods like DoT are the most computationally expensive, as reflected by their high token usage (see Fig. 2).

Model	Acc.	Precision	Recall	F1
(NR) GPT-3.5	0.425 ± 0.037	0.457 ± 0.055	0.362 ± 0.034	0.346 ± 0.048
(NR) GPT4	0.504 ± 0.018	0.529 ± 0.024	0.459 ± 0.005	0.435 ± 0.021
(NR) GPT4o	0.597 ± 0.037	0.532 ± 0.034	0.478 ± 0.014	0.460 ± 0.028
(PR) Llama-3.3	0.558 ± 0.025	0.556 ± 0.034	0.528 ± 0.032	0.527 ± 0.039
(PR) o1	0.560 ± 0.040	0.550 ± 0.048	0.490 ± 0.020	0.480 ± 0.036
(PR) o3-mini	0.549 ± 0.029	0.558 ± 0.054	0.510 ± 0.046	0.493 ± 0.047
(PR) Deepseek-R1-70B	0.527 ± 0.047	0.522 ± 0.041	0.480 ± 0.037	0.479 ± 0.041
(AR) GPT3.5 + CoT	0.395 ± 0.052	0.41 ± 0.057	0.391 ± 0.040	0.358 ± 0.053
(AR) GPT3.5 + DoT	**0.956 ± 0.011**	**0.959 ± 0.011**	**0.959 ± 0.008**	**0.957 ± 0.011**
(AR) GPT3.5 + SC	0.419 ± 0.036	0.479 ± 0.028	0.371 ± 0.023	0.366 ± 0.027
(AR) GPT3.5 + ToT	0.434 ± 0.018	0.515 ± 0.050	0.415 ± 0.025	0.417 + 0.028

Table 1: Task 1 – Recognition of Unhelpful Thought Patterns. Accuracy, Precision, Recall, F1 reported.

Model	ROUGE	BScore	mE5 Sim.
(NR) GPT-3.5	0.150 ± 0.084	0.874 ± 0.017	0.842 ± 0.039
(NR) GPT4	0.145 ± 0.093	0.876 ± 0.018	0.844 ± 0.040
(NR) GPT4o	0.146 ± 0.091	0.876 ± 0.018	0.845 ± 0.039
(PR) Llama-3.3	0.139 ± 0.064	0.867 ± 0.015	0.851 ± 0.034
(PR) o1	0.090 ± 0.070	0.823 ± 0.191	0.850 ± 0.030
(PR) o3-mini	0.121 ± 0.057	0.858 ± 0.013	0.850 ± 0.027
(PR) Deepseek-R1-70B	0.142 ± 0.081	0.873 ± 0.017	0.841 ± 0.038
(AR) GPT3.5 + CoT	0.147 ± 0.085	0.872 ± 0.017	0.843 ± 0.038
(AR) GPT3.5 + DoT	**0.271 ± 0.186**	**0.899 ± 0.031**	**0.884 ± 0.052**
(AR) GPT3.5 + SC	0.147 ± 0.085	0.874 ± 0.017	0.844 ± 0.039
(AR) GPT3.5 + ToT	0.146 ± 0.085	0.873 ± 0.017	0.841 ± 0.042

Table 2: Task 2 – Generation of Unhelpful Thought. ROUGE, BertScore, mE5 (Wang et al., 2024) embedding similarity scores reported.

Task 2: Generation of Unhelpful Thought assesses how well LLMs can generate an unhelpful thought given a persona and unhelpful thought pattern as shown in App. B.2. For automatic performance evaluation on this task, we report the ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and a sentence similarity metric using the multilingual-e5-large-instruct embedding model (Wang et al., 2024) – one of the top-5 best performing embedding models for retrieval on the

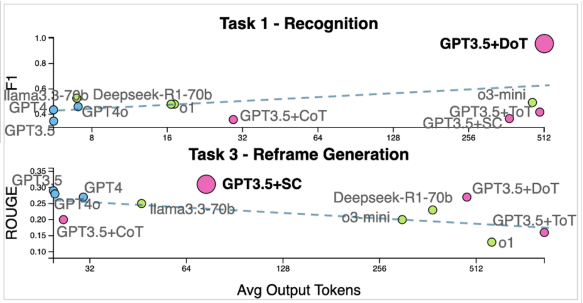

Figure 2: Output Tokens compared to Performance for each method across Tasks 1,3 (●: Reasoning-Augmented models; ●: Non-Reasoning models; ●: Pre-trained reasoning models). As indicated by the best performing model, encoded with a larger circle, we find that Reasoning-Augmented models can outperform Pre-trained reasoning models. - - -: Linear Regression fit on average output tokens to performance. We observe a positive linear relationship for the task of recognition and a negative relationship for reframe generation.

MTEB benchmark (Enevoldsen et al., 2025). As seen in Table 2, the non-reasoning `GPT-3.5` model augmented with `DoT` (AR) again emerges as the best-performing variant across all metrics, outperforming the strongest pre-trained reasoning model `Deepseek-R1` by 0.138 in ROUGE score. To further clarify, `DoT` is specifically designed for the detecting cognitive distortion types, not the generation of unhelpful thought. This surprising result extends the findings from Task 1, reinforcing the idea that task-related reasoning strategies not only outperform general pretrained reasoning models but can also generalize well to adjacent tasks within the same domain.

Task 3: Reframing of Unhelpful Thought is used to assess how well LLMs can generate a reframe of the persona's unhelpful thought given a persona, an unhelpful thought, and the unhelpful thinking pattern. An example is shown in App. B.3. As displayed in Table 3, we find that augmented reasoning (AR) methods again outperform all pre-trained reasoning (PR) and non-reasoning (NR) methods. Specifically, `GPT-3.5` augmented with `Self-Consistency` is the best-performing variant for the task of Reframe Generation. This may be attributed to the nature of the task, which likely benefits from exploring diverse reasoning paths to produce varied yet coherent reframes. Moreover, this AR method offers a noticeable reduction in computational cost compared to other high-performing variants (see Fig. 2), making it an effective and efficient choice for this task. The `Self-Consistency`-augmented `GPT-3.5` model exhibits this favorable trend across Tasks 2, 3, and 4 (see App. A).

Model	ROUGE	BScore	mE5 Sim.
(NR) GPT-3.5	0.287 ± 0.130	0.904 ± 0.020	0.902 ± 0.032
(NR) GPT4	0.270 ± 0.119	0.900 ± 0.019	0.906 ± 0.02
(NR) GPT4o	0.283 ± 0.136	0.904 ± 0.021	0.904 ± 0.032
(PR) Llama-3.3	0.247 ± 0.102	0.895 ± 0.017	0.901 ± 0.031
(PR) o1	0.126 ± 0.042	0.865 ± 0.136	0.886 ± 0.033
(PR) o3-mini	0.203 ± 0.087	0.888 ± 0.016	0.890 ± 0.030
(PR) Deepseek-R1-70B	0.228 ± 0.102	0.894 ± 0.019	0.897 ± 0.032
(AR) GPT3.5 + CoT	0.196 ± 0.121	0.885 ± 0.023	0.872 ± 0.050
(AR) GPT3.5 + DoT	0.267 ± 0.126	0.899 ± 0.019	0.898 ± 0.032
(AR) GPT3.5 + SC	**0.307 ± 0.135**	**0.907 ± 0.019**	**0.906 ± 0.032**
(AR) GPT3.5 + ToT	0.160 ± 0.099	0.870 ± 0.024	0.859 ± 0.046

Table 3: Task 3 – Reframing of Unhelpful Thought

Model	ROUGE	BScore	mE5 Sim.
(NR) GPT-3.5	**0.272 ± 0.129**	**0.902 ± 0.019**	**0.901 ± 0.032**
(NR) GPT4	0.238 ± 0.105	0.895 ± 0.018	0.902 ± 0.029
(NR) GPT4o	0.245 ± 0.124	0.897 ± 0.019	0.900 ± 0.032
(PR) Llama-3.3	0.208 ± 0.087	0.887 ± 0.016	0.895 ± 0.029
(PR) o1	0.134 ± 0.031	0.825 ± 0.173	0.809 ± 0.038
(PR) o3-mini	0.184 ± 0.082	0.884 ± 0.015	0.886 ± 0.030
(PR) Deepseek-R1-70B	0.203 ± 0.091	0.888 ± 0.017	0.892 ± 0.031
(AR) GPT3.5 + CoT	0.200 ± 0.112	0.888 ± 0.019	0.881 ± 0.040
(AR) GPT3.5 + DoT	0.239 ± 0.106	0.895 ± 0.018	0.895 ± 0.031
(AR) GPT3.5 + SC	**0.275 ± 0.127**	**0.903 ± 0.020**	**0.903 ± 0.031**
(AR) GPT3.5 + ToT	0.166 ± 0.109	0.870 ± 0.029	0.854 ± 0.046

Table 4: Task 4 – Strategic Reframing of Unhelpful Thought

Task 4: Strategic Reframing of Unhelpful Thought We introduce a novel task that extends Task 3, aiming to evaluate how effectively large language models (LLMs) can generate a reframe of the persona's unhelpful thought *aligned to a specific reframe strategy* (Harris et al., 2007). This task specifically measures the alignment and instruction-tuning capabilities of LLMs in Cognitive Reframing, which is particularly important in CBT practices, where the intervention used is chosen and tailored to the specific formulation of the individual (Fenn and Byrne, 2013). An example of the task implementation is shown in App. B.4. The results for Task 4 are shown in Table 4. Surprisingly, we find that the non-reasoning (NR) version of `GPT-3.5` and its `Self-Consistency`-augmented (AR) variant display the strongest but similar performance over other methods. In addition, overall performance on Task 4 is lower than Task 3. These two results combined indicate that even the most advanced pretrained and augmented reasoning (PR, AR) models lack sufficient alignment to be able to generate mental reframes that are strictly aligned to specific reframe strategies. These findings call for further research on alignment and controllable generation methods for LLMs to be effectively and reliably used for CBT applications.

Limitations and Ethical Considerations

While our work explores the potential of LLMs with reasoning augmentation strategies to improve performance on cognitive reframing tasks, several limitations remain. First, the evaluation relies predominantly on automatic metrics, which may not fully capture the nuanced, subjective quality of cognitive reframing, an area that often requires human interpretation and sensitivity to context. Although our experiments show that models augmented with reasoning techniques outperform larger pretrained reasoning models on aggregate metrics, the high standard deviations reported in some tasks (e.g., Task 2) raise concerns about the consistency and statistical significance of these findings. Future work should incorporate robust statistical testing and, where possible, human-in-the-loop evaluations to validate and interpret these results more thoroughly.

Another limitation relates to the dataset composition. Our use of uniform sampling from the PatternReframe dataset (Maddela et al., 2023) may not adequately reflect real-world distributions of cognitive distortions. As a result, model performance might be overestimated on rare reframing patterns and underestimated on more prevalent ones encountered in practical mental health applications. Moreover, the additional strategy-aligned reframing task we introduced, while conceptually valuable, requires further validation of clinical relevance and complexity compared to existing tasks.

Given the sensitive nature of cognitive reframing as an intervention commonly used in mental health contexts, deploying LLMs for such tasks carries significant ethical implications. Incorrect or poorly framed outputs could inadvertently harm vulnerable users by reinforcing negative thoughts or offering inappropriate advice. Since our work does not incorporate feedback from mental health professionals, these risks may not be adequately identified or mitigated. Future work should engage domain experts to co-design and evaluate model outputs for clinical safety and cultural sensitivity. Safeguards against misuse should also be implemented to prevent models from being used to generate harmful or manipulative reframing content. Additionally, the broader societal impacts of deploying reasoning-augmented LLMs in mental health settings should be considered, including issues of accessibility, bias, and cultural appropriateness.

At present, LLM-based systems for cognitive reframing are most accessible to users in technologically advanced and resource-rich settings, while under-resourced or marginalized communities who may have the greatest need for affordable and accessible mental health support might be less able to leverage these technologies effectively. To avoid exacerbating existing health disparities, future research should actively consider how to make these tools accessible and effective for a diverse range of users, including those in low-resource settings or non-Western contexts.

Acknowledgements

We would like to thank our colleagues and collaborators for their invaluable input and feedback. We would like to thank Dr. Hae Won Park and Dr. Cynthia Breazeal for their support in this work.

References

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. Archives of general psychiatry, 9(4):324–333.

S. Blum, M. Brow, and R.C. Silver. 2012. Coping. In V.S. Ramachandran, editor, Encyclopedia of Human Behavior (Second Edition), second edition edition, pages 596–601. Academic Press, San Diego.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. arXiv preprint arXiv:2310.07146.

BURNS David and MD Burns. 1980. Feeling good: The new mood therapy. NY: Signet Books. Chin, Richard, pages 42–3.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria

Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. arXiv preprint arXiv:2502.13595.

Kristina Fenn and Majella Byrne. 2013. The key principles of cognitive behavioural therapy. InnovAiT, 6(9):579–585.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

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.

Nathan Hodson, Simon Williamson, et al. 2024. Can large language models replace therapists? evaluating performance at simple cognitive behavioral therapy tasks. JMIR AI, 3(1):e52500.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13641–13660, Toronto, Canada. Association for Computational Linguistics.

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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. arXiv preprint arXiv:2212.09597.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. Healme: Harnessing cognitive reframing in large language models for psychotherapy. arXiv preprint arXiv:2403.05574.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36:11809–11822.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C Ong. 2024. Large language models are capable of offering cognitive reappraisal, if guided. arXiv preprint arXiv:2404.01288.

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

A Relationship Between Output Tokens and Performance

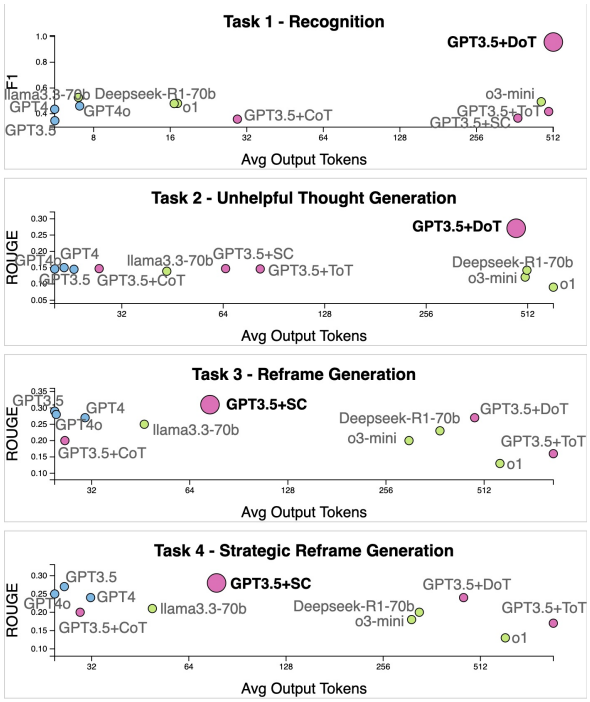

Figure 3: Output Tokens compared to Performance for each method across Tasks 1, 2, 3, 4 (●: Reasoning-Augmented models; ●: Non-Reasoning models; ●: Pre-trained reasoning models). As indicated by the best performing model, encoded with a larger circle, we find that Reasoning-Augmented models can outperform Pre-trained reasoning models.

B Prompts Used

The reframing strategy definitions:

- "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.

The unhelpful thinking pattern definitions:

- "Catastrophizing": by giving greater weight to the worst possible outcome.
- "Discounting the positive": experiences by insisting that they "don't count".
- "Overgeneralization": making faulty generalizations from insufficient evidence,
- "Personalization": assigning a disproportionate amount of personal blame to oneself.
- "Black-and-white or polarized thinking / All or nothing thinking": viewing things as either good or bad and nothing in-between.
- "Mental filtering": occurs when an individual dwells only on the negative details of a situation.
- "Jumping to conclusions: mind reading": inferring a person's probable (usually negative) thoughts from their behavior.
- "Jumping to conclusions: Fortune-telling": predicting outcomes (usually negative) of events.
- "Should statements": a person demands particular behaviors regardless of the realistic circumstances.
- "Labeling and mislabeling": attributing a person's actions to their character rather than the situation.
- "None": the thought does not contain any unhelpful pattern / is nonsensical / does not align with the persona.

B.1 Task 1 Example Prompt (Zeroshot)

You will be given a persona and an unhelpful thought conditioned on the persona. Your goal is to identify the unhelpful thinking pattern that the unhelpful thought falls into.

The unhelpful thinking patterns are defined as: *Pattern Definitions*.

Given a persona and an unhelpful thought, please identify the most appropriate unhelpful thinking pattern. In your response, include only the identified unhelpful thinking pattern from the categories above.

Persona: *Persona*
Unhelpful Thought: *Thought*
Unhelpful thinking pattern:

B.2 Task 2 Example Prompt (Zeroshot)

You will be given a persona and an unhelpful thinking pattern. Your goal is to generate an unhelpful thought that matches the given thinking pattern and the persona.

The unhelpful thinking patterns are defined as: *Pattern Definitions*.

Given a persona and an unhelpful thinking pattern, generate a corresponding unhelpful thought. Contain only the generated unhelpful thought in your response.

Persona: *Persona*
Unhelpful thinking pattern: *Pattern*
Unhelpful thought:

B.3 Task 3 Example Prompt (Zeroshot)

You will be given a persona, an unhelpful thought conditioned on the persona, and the unhelpful thinking pattern the thought falls into. Your goal is to reframe the unhelpful thought such that it aligns with the persona and context but does not contain the unhelpful thinking pattern.

The unhelpful thinking patterns are defined as: *Pattern Definitions*.

Given a persona, an unhelpful thought, and the unhelpful thinking pattern, please generate a reframed thought. Contain only the reframed thought in your response.

Persona: *Persona*
Unhelpful Thought: *Thought*
Unhelpful thinking pattern: *Pattern*
Reframing Strategy: *Strategy*
Reframed Thought:

B.4 Task 4 Example Prompt (Zeroshot)

You will be given a persona, an unhelpful thought conditioned on the persona, the unhelpful thinking pattern that the unhelpful thought falls into, and the reframing strategy used to reframe the thought. Your goal is to reframe the unhelpful thought to be aligned with the reframing strategy while still being aligned with the persona and the context of the unhelpful thought, but without containing the unhelpful thinking pattern.

The reframing strategies are defined as: *Strategy Definitions*.
The unhelpful thinking patterns are defined as: *Pattern Definitions*.

Given an example of a persona, an unhelpful thought, the unhelpful thinking pattern, and the reframing strategy used, please generate a reframed thought. Contain only the reframed thought in your response.

Persona: *Persona*
Unhelpful Thought: *Thought*
Unhelpful thinking pattern: *Pattern*
Reframing Strategy: *Strategy*
Reframed Thought: