Cognitive Reframing of Negative Thoughts with Iterative In-context Clinical Grounding Feedback

Anonymous ACL submission

Abstract

This paper introduces a new framework for cognitive reframing of negative thoughts. Unlike prior methods that consider reframing as a shallow generation task, the framework guides LLMs to follow Cognitive Behavioral Therapy (CBT). To do that, we first design a multi-agent backbone that integrates CBT for clinical grounding. We next introduce an objective function that assesses psychological, semantic, and stylistic aspects. In each iteration, the function's scores are converted into prompt edits that steer the next iteration. Zero-shot evaluation on two datasets shows that the framework outperforms shallow- and single-agent baselines on automatic and human-rated metrics, including those for safety and hallucination.

1 Introduction

Mental health is one of the most critical global health issues (Chen et al., 2023; Sharma et al., 2023b), with negative thoughts among the leading factors behind common mental health problems (Ziems et al., 2022b; Sharma et al., 2023b). In professional psychotherapy, an evidence-based intervention is Cognitive Reframing (CR) (Carli, 1999), a core technique within Cognitive Behavioral Therapy (CBT) (Figure 1) (Beck, 2020). CR is the therapeutic process that reframes a negative thought into a more hopeful or constructive one, offering a more balanced perspective on the original thought (Robson Jr and Troutman-Jordan, 2014; Sharma et al., 2024; Xiao et al., 2024).

Generating an effective reframed thought is a complex cognitive process that can benefit from human or expert-informed feedback in human–AI collaborative settings (Sharma et al., 2023b). The interaction allows AI systems to collect feedback from psychologists to improve the quality of CR (Step 4, Figure 1). However, collecting human feedback to build training datasets is time-consuming and labor-expensive (Wu et al., 2022), especially

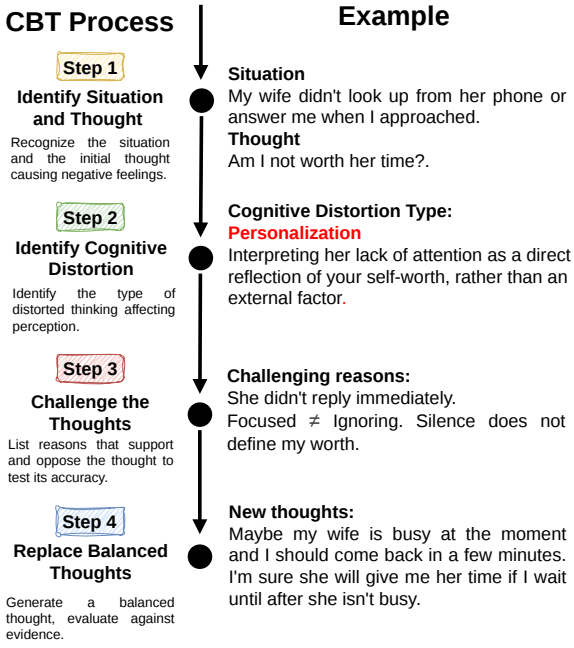

Figure 1: CBT-based reframing: (i) identifying situation and thought, (ii) identifying cognitive distortion, (iii) challenging the thoughts, (iv) iterative discussion that replaces the negative thought with a balanced one. Discussion can also be done in other steps to collect more information.

in psychotherapy. We therefore propose an iterative in-context CR framework that collects model-generated feedback for reframing refinement.

Recent LLMs make it feasible to provide accessible, in-the-moment support for people experiencing negative thoughts (Sharma et al., 2023b). Attention has focused on cognitive text rewriting tasks (Sharma et al., 2021; Reif et al., 2022; Ziems et al., 2022b) or generating dedicated practice datasets (Maddela et al., 2023; Kim et al., 2025b). Nevertheless, reliably generating clinically viable reframes remains a computational challenge (Sharma et al., 2023b): professional therapists carry out a deep psychological analysis of the patient's thinking patterns (Figure 1), whereas most prior work treats CR as a sentence rewriting task (Reif et al., 2022; Goel

et al., 2025) handled by fine-tuned models (Kim et al., 2025a; Goel et al., 2025) or LLMs (Sharma et al., 2023b), which possess human-level fluency but may lack clinical grounding.

This paper introduces a new clinical grounding framework for the CR task. To do that, we first formulate the CR task as an iterative process (Figure 2) using in-context refinement feedback. We next design a multi-agent CBT-based backbone that models the cognitive reframing process using Diagnosis-of-Thought (DoT) (Chen et al., 2023) for therapeutic reframe generation. The backbone outputs reframed thoughts iteratively, where a new objective function combines clinical, semantic, and stylistic alignment to create clinical grounding feedback for the reframing agent. This paper makes three main contributions:

- It introduces an iterative in-context clinical grounding framework for CR with a new objective function that combines clinical, semantic, and stylistic alignment.
- It presents a multi-agent AI backbone that integrates CBT into the reasoning process.
- It conducts zero-shot studies on two benchmark datasets, showing improvements over shallow- and single-agent baselines on automatic and human-rated metrics, including those for safety and hallucination.[1]

2 Related Work

NLP for mental health and therapy. Early work in computational mental health focused on identifying conditions such as anxiety and depression (Althoff et al., 2016; Gaur et al., 2019; Ji et al., 2022; Lee et al., 2019; Pruksachatkun et al., 2019; Sharma and De Choudhury, 2018; Wadden et al., 2021) and providing supportive responses (Sharma et al., 2020, 2023a; Welivita et al., 2021; Miner et al., 2019; Shen et al., 2020). Many systems supported providers (clinicians or peers) (Tanana et al., 2019; Shen et al., 2020; Sharma et al., 2023a, 2020), rather than direct human-LM interaction for self-help.

Cognitive reframing. Early work represented reframing as a text generation task, including style, sentiment, politeness, and empathy transfer (Madaan et al., 2020; Sharma et al., 2021; Reif et al., 2022), or positive reframing (Ziems et al., 2022b). Recent work has shifted from generation to psychological CR, which requires a deeper understanding of human thinking patterns (Smith et al., 2021; Morris et al., 2015). To facilitate the development, many studies have focused on creating new datasets in both text-only and multimodalities (Sharma et al., 2023b; Maddela et al., 2023; Kim et al., 2025a,b). Based on these datasets, several techniques have been proposed, such as retrieval-based in-context learning (Sharma et al., 2023b), fine-tuning LLMs (Xiao et al., 2024; Kim et al., 2025a,b), or socratic reasoning (Goel et al., 2025).

In-context learning (ICL). ICL is the ability of LLMs to infer a task from examples placed directly in the prompt at inference time without any weight updates (Dong et al., 2024; Crosbie and Shutova, 2025). Recent work used in-context feedback for zero-shot dataset generation (Ye et al., 2022), explainable style transfer (Saakyan and Muresan, 2024), prompt optimization (Agarwal et al., 2025), or text-to-sql with reinforcement learning (Toteja et al., 2025). We extend ICL for the CR task by defining a clinical grounding feedback method to refine reframed thoughts. Our extension of ICL is also related to prompt optimization (Agarwal et al., 2025), in which feedback is used to optimize the prompt of the backbone for better reframing.

While sharing the CR task with Sharma et al. (2023b); Xiao et al. (2024), our proposal distinguishes itself in three main points. First, we use CBT as a foundation for building our framework, rather than using simple prompts. Second, we formulate the CR task as an interactive process to utilize clinical grounding feedback. It allows us to introduce a new backbone and objective function to model the complex reframing process. Third, we expand ICL using refinement feedback to simulate the process of reframing: feedback is converted into prompt edits for refinement improvement.

3 Iterative CBT-based Reframing

3.1 Problem Statement

We formulate the CR task as an iterative clinical-grounding process. Given an input pair I consisting of a situation S and an associated negative thought T, the task is to generate an effective reframed thought $R = f(S, T)$ by a clinical-grounding function f. By *clinical grounding*, we mean that the

[1] Following Alansari and Luqman (2025), we use *hallucination* operationally to mean output not entailed by the source context, without implying any internal cognitive state of the model.

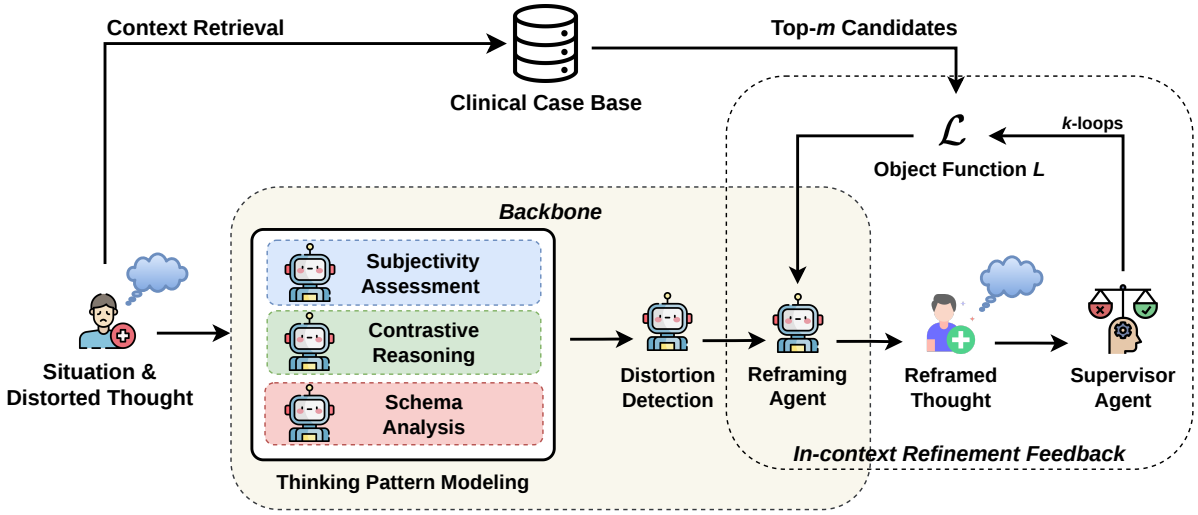

Figure 2: Overview of the proposed framework. It processes situations and thoughts through a CBT-based reasoning backbone to produce reframed thoughts through an iterative in-context refinement loop.

reasoning chain explicitly follows the procedural structure of CBT (e.g., distortion identification, contrastive reasoning, and Socratic questioning), under the supervision of an LLM-based agent that proxies, but does not substitute for, a clinical expert. The process is applied in the zero-shot setting, which is practical for the deployment of actual AI systems (Xiao et al., 2024; Hotta et al., 2025).

3.2 The Backbone Model

While modern LLMs such as GPT-4/5 demonstrate highly fluent and human-like language generation abilities (OpenAI, 2023; Chang et al., 2024), they may lack clinical grounding that is a critical aspect of computational psychotherapy (Xiao et al., 2024; Sharma et al., 2024; Hotta et al., 2025). Motivated by this gap, we introduce CBT-MACR, a **CBT**-based **M**ulti-**A**gent **C**ognitive **R**eframing framework (shown in Figure 2). Given an input $I(S, T)$, the backbone (Section 3.2) first outputs a reframe. In-context refinement feedback (Section 4) then provides feedback on the reframe using clinical, semantic, and stylistic alignment scores. The feedback is converted into edits to the backbone's prompt for better reframing.

3.2.1 Thinking Pattern Modeling

Inspired by DoT (Chen et al., 2023), we design the modeling agent with three smaller agents: subjectivity assessment, contrastive reasoning, and schema analysis. We extend DoT in three points: (i) we replace the single-prompted DoT agent with three smaller agents that decompose the diagnostic reasoning into separate steps (Figure 5); (ii) the detected distortion is propagated downstream as conditioning context for the Reframing Agent (Eq. 2), rather than being a terminal classification output as in the original DoT (DoT terminates at the distortion label; we instead use it to condition the downstream reframe); and (iii) the DoT-based backbone is further empowered by in-context refinement feedback (Section 4).

Subjectivity assessment This stage aims to distinguish between objective facts and subjective thoughts presented in the patient's speech. By separating reality from interpretations, this step establishes a robust objective evidence base needed to diagnose subjective components.

Contrastive reasoning This stage involves eliciting two complementary reasoning processes, both grounded in established objective facts: arguments supporting the subjective thought and arguments contradicting it. By generating and contrasting these opposing interpretations based on the same factual evidence, the agent facilitates clearer identification of the thought schemas that contain cognitive biases.

Schema analysis The final diagnostic stage summarizes the underlying cognitive schema that governs why the patient forms the specific reasoning supporting the negative thought. We consider schemas as cognitive structures that integrate an individual's knowledge and beliefs, and identifying them is crucial for revealing the patient's cognitive mode and potential distortions. This structured output $A(S, T)$ (including objective facts, subjective

thoughts, contrastive reasoning, and the underlying thought schema) provides interpretable diagnosis rationales that condition the reframing.

3.2.2 Distortion Detection

Given the diagnostic decomposition performed by the Thinking Pattern Modeling Agent, the Distortion Detection Agent analyzes the intermediate structured analysis of the patient's thinking patterns $A(S,T)$ to predict their distortions c^* as follows.

$$c^* = g(A(S,T)) \tag{1}$$

where $g()$ is a classification function, $c^* \in \mathcal{C}$, and $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ denotes the set of $K = 12$ cognitive distortion categories, comprising 11 distortion types plus a *No-distortion* class (Appendix A.1). We consider the detection as a soft signal for prompting rather than a hard one as conventional classification. Detail is shown in A.1.

3.2.3 CBT-based Reframing

The Reframing Agent requires the raw context (S,T), the structured diagnostic analysis $A(S,T)$, and the identified cognitive distortion category (c^*) to output the reframed thought R.

$$R = f(A(S,T), c^*) \tag{2}$$

4 In-context Refinement Feedback

To ensure cognitive reframes are both clinically valid and linguistically accurate, we introduce an in-context clinical grounding mechanism that combines clinical, semantic, and stylistic alignment. It includes a **Supervisor Agent** that functions as a candidate ranking and selection module. Instead of relying on a single generated reframe, at each iteration the system generates $n = 5$ reframing candidates and selects the optimal response (y^*) using a hybrid objective function; this generate-and-select loop is repeated for $T = 5$ iterations, with each selected y^* used to condition the next iteration's prompt.

4.1 Objective Function

The objective function is designed to reflect the clinical structure of cognitive reframing governed by CBT, where therapists independently evaluate (i) clinical validity, (ii) faithfulness to the client's situation, and (iii) therapeutic linguistic tone (Robson Jr and Troutman-Jordan, 2014; Beck, 2020).

Clinical assessment ($\mathcal{L}_{\text{clin}}$) Acting as an internal CBT evaluator, the Supervisor Agent assesses each candidate y_i against five psychometric dimensions derived from Beck's Socratic Questioning principles (Beck et al., 1979; Beck, 2011): *Evidence Sensitivity* (0–2; grounding in objective facts), *Alternative Explanations* (0–2; consideration of multiple possibilities), *Distortion Regulation* (0–2; mitigation of cognitive biases), *Balanced Self-Evaluation* (0–2; fairness to oneself), and *Cognitive Flexibility* (0–1; openness to outcomes). The agent outputs a structured JSON object with scores and qualitative feedback for each dimension; their sum forms the $\mathcal{L}_{\text{clin}}$ score (max 9).

Semantic alignment Prior work has shown that fluent reframes may still be hallucinated or clinically unsafe (Ziems et al., 2022b; Sharma et al., 2023b). To ensure content fidelity and target hallucination, we measure the embedding similarity between the input situation S and the generated reframe y^*. We use `all-MiniLM-L6-v2` (Wang et al., 2020; Reimers and Gurevych, 2019), a high-performance embedding model. The semantic alignment is the cosine similarity between the embedding of the situation $(\mathbf{e}_S)$ and the reframe $(\mathbf{e}_{y^*})$.

$$\mathcal{L}_{\text{sem}}(S, y^*) = \cos(\mathbf{e}_S, \mathbf{e}_{y^*}) = \frac{\mathbf{e}_S \cdot \mathbf{e}_{y^*}}{\|\mathbf{e}_S\| \|\mathbf{e}_{y^*}\|} \tag{3}$$

This term acts as a soft regularizer ($\beta = 0.3$) that penalizes reframes whose embeddings drift substantially from the input situation, rather than guaranteeing semantic fidelity.

Stylistic alignment A good reframe should have an appropriate therapeutic linguistic tone, characterized by features such as emotional valence, subjectivity, and professional stance. Instead of relying on lexical overlap (ROUGE), which may fail to capture the subtle nuances of therapeutic tone, we propose a sentiment-based stylistic function. We utilize `TextBlob` (Loria, 2018) to extract two linguistic features: *Polarity* ($P \in [-1, 1]$), representing the emotional valence (from negative to positive), and *Subjectivity* ($\sigma \in [0, 1]$), representing the degree of personal opinion versus factual objectivity. We acknowledge that TextBlob is a lexicon-based proxy and does not capture all dimensions of therapeutic linguistic register; we therefore weight this term modestly ($\gamma = 0.2$).

The stylistic score also uses retrieval over a clinical case base of reference reframes. This simulates

the therapy process in which, given a new situation–thought pair, a therapist consults reference cases for guidance. We retrieve the top $m = 5$ relevant samples from the case base, and let g denote the retrieved expert reference reframe. The stylistic score quantifies the alignment between the candidate y_i and g by minimizing the sentiment deviation.

$$\mathcal{L}_{\text{style}}(y_i, g) = 1 - \frac{1}{2}\left(\Delta P_{\text{norm}} + \Delta\sigma\right) \quad (4)$$

where $\Delta P_{\text{norm}} = \frac{|P_{y_i} - P_g|}{2}$ is the normalized polarity difference (scaled to $[0, 1]$), and $\Delta\sigma = |\sigma_{y_i} - \sigma_g|$ is the absolute difference in subjectivity. This term encourages the generated reframe to maintain an emotional trajectory and professional stance comparable to the expert reference.

In-context clinical grounding feedback To synthesize the three evaluation signals, we define the objective function $\mathcal{L}$ as a weighted linear combination of the three rewards. We empirically prioritize clinical safety with the highest weight ($\alpha = 0.5$ on $\mathcal{L}_{\text{clin}}$), followed by semantic alignment ($\beta = 0.3$), and stylistic alignment ($\gamma = 0.2$) (Table 8).

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{clin}} + \beta \cdot \mathcal{L}_{\text{sem}} + \gamma \cdot \mathcal{L}_{\text{style}} \quad (5)$$

Eq. (5) penalizes high-fluency but clinically unsound hallucinations during ranking. $\mathcal{L}$ is used to adjust the prompt of the reframing agent in Eq. (2) for better reframing. While the semantics encourages lexical alignment between a reframe and the situation, the clinical and stylistic components mitigate harmful content in the generated reframes.

4.2 Candidate Selection

The final reframe is determined by maximizing the total reward across the generated candidates.

$$y^* = \underset{y_i \in \{y_1, \ldots, y_n\}}{\text{argmax}} \; \mathcal{L}(S, y_i, g) \quad (6)$$

Algorithm 1 summarizes the inference loop.

Algorithm 1 CBT-MACR Inference Loop

Input: situation S, thought T; backbone f; objective $\mathcal{L}$; retrieved reference g
Hyperparameters: candidates per iteration n=5, iterations $T_{\max}$=5
Initialize prompt $P_0 \leftarrow (S, T)$
for $t = 1$ to $T_{\max}$ **do**
  Sample n candidates $\{y_1^{(t)}, \ldots, y_n^{(t)}\} \sim f(P_{t-1})$
  Score each candidate by $\mathcal{L}(S, y_i^{(t)}, g)$ (Eq. 5)
  $y_t^* \leftarrow \arg\max_{y_i^{(t)}} \mathcal{L}(S, y_i^{(t)}, g)$
  $P_t \leftarrow P_{t-1} \cup \{y_t^*, \text{feedback}(y_t^*)\}$
end for
return $y_{T_{\max}}^*$

5 Experimental Settings

Datasets Evaluation uses **Cognitive Reframing (CR)** (Sharma et al., 2023b) and **Pattern Reframing (PR)** (Maddela et al., 2023). Following Xiao et al. (2024), we sample 300 (situation, thought) instances from the CR test split (which contains over 600 expert-annotated samples) and 1,000 instances from the 6,807-instance PR test split.

Table 1: Statistics of the datasets used for evaluation. *Expert*: expert-annotated reframes; *Pattern*: persona-based pattern reframes.

Dataset	Train	Dev	Test
CR dataset (Expert)	0	0	300
PR dataset (Pattern)	1920	961	1000

LLMs We used GPT-4o-mini (OpenAI, 2024), GPT-4.1 (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), GPT-OSS-20B, and Qwen3-8B (Yang et al., 2025) for experiments.

Evaluation metrics We used BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and LM judge (Zheng et al., 2023) as supporting metrics, since BLEU and ROUGE only measure lexical overlap, while CR requires a deeper semantic evaluation. Therefore, we use human evaluation as the primary metric (Sharma et al., 2023b; Xiao et al., 2024).

6 Results and Discussion

6.1 Automatic Evaluation

Objective function effectiveness We compared our function with five different objective functions: (i) clinical scores, (ii) semantic scores, (iii) the stylistic score $\mathcal{L}_{\text{style}}$ in Eq. (5), (iv) combination of clinical and BLEU scores, and (v) a reward function combining ROUGE and BERT scores from dialogue generation (Srivastava et al., 2023). All functions were tested over $T = 5$ iterations (Figure 10), a setting that is practical in actual use.

Empirical results in Figures 3 and 4 show that the scores of our objective function are higher than those of the other variants in almost all cases.[2] The improvement comes from the combination of clinical, semantic, and stylistic aspects. The clinical assessment encourages the backbone to follow CBT-structured reasoning steps when producing reframes. It is associated with improved safety

[2]For figures and tables below abbreviate the clinical-assessment score $\mathcal{L}_{\text{clin}}$ (Section 4.1) as "Beck's".

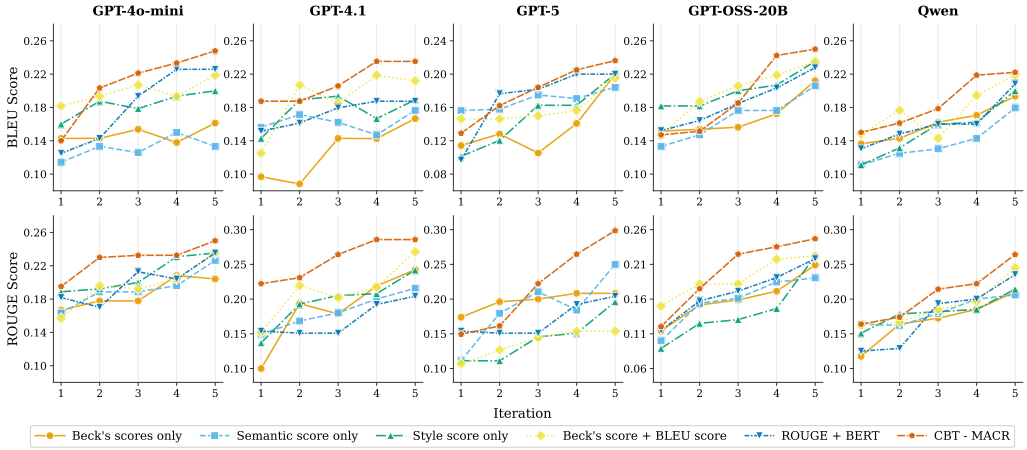

Figure 3: Evolution of BLEU and ROUGE scores over five iterations on the CR dataset (300 samples).

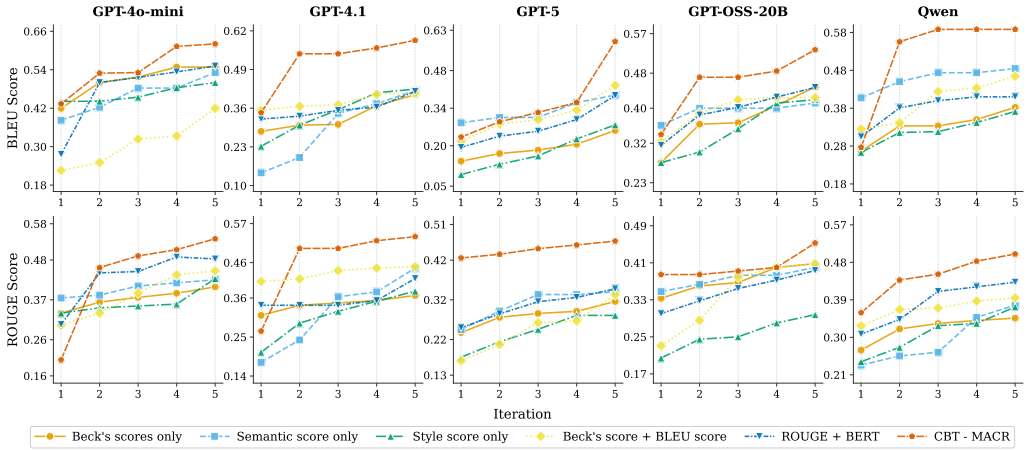

Figure 4: Evolution of BLEU and ROUGE scores over five iterations on the PR dataset (1000 samples).

metrics on the generated reframes. The semantic and stylistic components align reframed thoughts with input situations, reducing hallucinations from LLMs. The combination of lexical (ROUGE) and semantic (BERT) scores (Srivastava et al., 2023) produces competitive results. However, without clinical grounding, its performance is lower than that of our method. LM judge scores in Figure 8 also confirm the effectiveness of our method. Component contribution is shown in Section 6.3.2.

Objective-function ablation Figures 3, 4, and 8 also show the ablation study of our objective function. Only using one aspect ($\mathcal{L}_{\text{clin}}$, $\mathcal{L}_{\text{sem}}$, or $\mathcal{L}_{\text{style}}$) is insufficient to produce a good reframe. The combination of $\mathcal{L}_{\text{clin}}$ and BLEU scores does not show improvement in all cases. A possible reason is that this combination captures safety via $\mathcal{L}_{\text{clin}}$ but lacks strong indicators for hallucination reduction, since BLEU only measures lexical overlap. In contrast, the proposed function offers a balanced method for promoting safety while reducing hallucinations.

Backbone ablation (single-agent vs. multi-agent) Using GPT-5, we tested three backbones with our objective function. **Shallow Reframing - SR**, which directly generates reframed thoughts without modeling distortions (Chen et al., 2023). **DoT**, a single-prompted agent that analyzes and reframes distorted thoughts (Chen et al., 2023). **CBT-agents**, the proposed multi-agent framework (Figure 2).

Figure 5 shows two important points. First, the proposed multi-agent backbone model (Section 3.2) outperforms the two baselines. This is because the backbone is a clinical grounding model that takes into account CBT for reframing. This hierarchical decomposition pushes the model beyond surface-level text and enables the cognitive analysis required for a clinically grounded reframe. The DoT (a single-prompted agent) method achieves competitive scores because it also shares CBT with our backbone. However, the performance of our multi-agent backbone is still better. The shallow method produces noticeably poor scores, support-

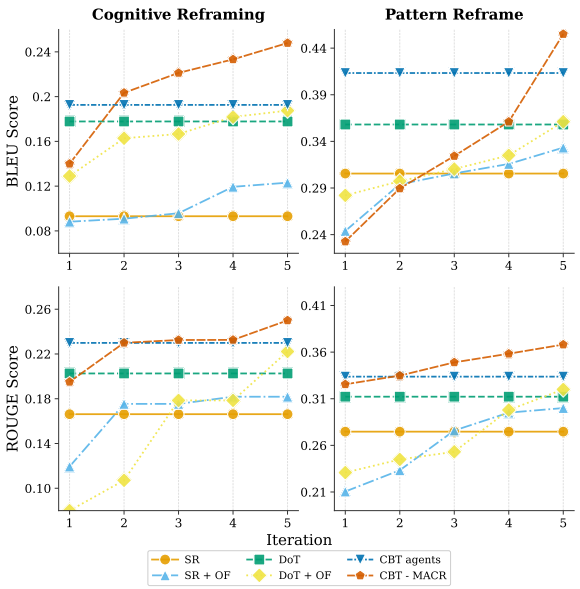

Figure 5: The performance of the backbone with the proposed objective function (OF) using GPT-5.

ing our claim in Section 3.2 that LLMs benefit from clinical grounding for better reframing. Second, the proposed objective function (Section 4.1) is beneficial for all backbone models by improving their performance in all cases. The improvement comes from the combination of clinical, semantic, and stylistic aspects. It helps our system reduce hallucinations and improve the safety of outputs.

6.2 Human Evaluation

Human evaluation is our primary measure of quality. We recruited eight mental health practitioners with experience in computational psychotherapy. They assessed generated reframing outputs on 1–5 Likert scales. Following prior work (Sharma et al., 2023b; Xiao et al., 2024; Kim et al., 2025a), we employed six metrics grouped into three aspects. **Therapeutic effectiveness**: *Relatedness (Rel.)* measures alignment with the individual's situation, and *Helpfulness (Help.)* assesses whether the reframing offers a constructive, CBT-consistent reinterpretation. **Factual consistency** (hallucination-related): *Factuality (Fact.)* measures correctness against gold labels or verified facts, and *Faithfulness (Fait.)* evaluates consistency with the original context without unsupported information (Alansari and Luqman, 2025). **Safety**: *Harmfulness (Harm.)* captures offensive or psychologically unsafe content, and *Fairness (Fair.)* evaluates potential social bias (Ip, 2026; Liu et al., 2025).

Figure 6 shows that our method provides the best balance between response quality and safety-

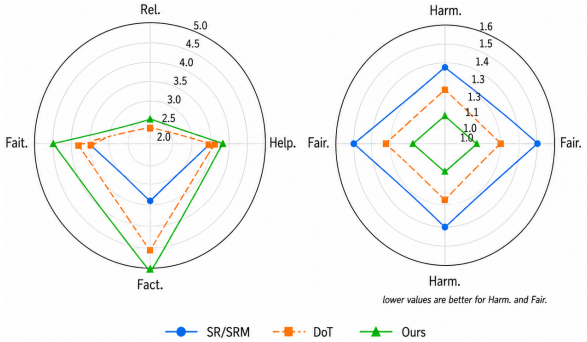

Figure 6: Human evaluation with the average scores of the metrics across LLMs on two datasets.

oriented criteria. The left figure shows that our method achieves the largest overall area, especially on Fact. and Fait., indicating stronger factuality and faithfulness while also maintaining competitive relevance and helpfulness. DoT consistently improves over SR, but its gains are smaller than those of our method, suggesting that cognitive-distortion-guided reasoning helps but is further enhanced by our proposed approach. In the right figure, lower Harm. and Fair. scores indicate better performance, and our method produces the smallest radar area, showing that it generates less harmful and fairer responses on average. Beyond benchmarkng, we also deployed the framework as **MindReframe** in an educational pilot ($N = 50$ students), where users typically reached a satisfactory reframe within 3–4 iterations — the same range at which automatic scores plateau in Figures 3 and 4 (A.7).

6.3 Ablation Study

6.3.1 Psychological Grounding

Distortion detection As mentioned in Section 3.2.2, distortion detection is an important step of CBT. We validated detection performance by comparing outputs of GPT-5 to gold labels of distortions on the two datasets. Table 2 shows that GPT-5 predicts distortions reasonably well for reframing. It also shows that detection is a non-trivial task; we therefore treat it as a soft signal (below) rather than a hard prerequisite. Analysis of error propagation of distortion detection for reframing is shown in Tables 5 and 6 and per-distortion-type F1 scores are shown in Table 7.

Table 2: Performance of GPT-5 for cognitive distortion.

Datasets	Accuracy	F-1
Cognitive Reframing	75.67	76.43
Pattern Reframe	79.10	79.26

Although detection errors may affect downstream performance, we treat the predicted distortion as a soft signal: the framework outputs a distortion category and pairs it with an accompanying rationale that the downstream Reframing Agent conditions on. Unlike conventional supervised pipelines that consume distortion labels as a hard signal, the accompanying rationale lets the prompt absorb classification uncertainty rather than propagate it as a single label. This design helps mitigate the effect of detection errors on the overall performance of the pipeline. The error-propagation analysis in Appendix A.1 supports this design choice: a 25% relative drop in detection F1 (from 100% to 76.43%) yields only a 2–4 absolute-point drop in BLEU/ROUGE (6–16% relative) and a near-flat LM-Judge score, indicating that the pipeline is robust to moderate upstream detection noise.

We also compared Beck's Socratic Questioning against ABCDE (Ellis, 1991) and CBA (Greenberger and Padesky, 2015) as alternative therapy frameworks (Appendix A.2): Socratic Questioning consistently yields better scores, plausibly because its structured, step-by-step inquiry aligns with the multi-agent reasoning of LLMs.

Feedback and gold references alignment We validated the alignment between feedback and gold references using two CBT-grounded dimensions, *Semantic Equivalence* (Ziems et al., 2022a) and *Cognitive Shift* (Sharma et al., 2023b), with both an LLM-as-a-judge and eight human annotators on 100 anonymized samples (Appendix A.5). Table 9 shows a strong positive correlation between automated metrics and human judgments, and evaluators consistently preferred our framework for facilitating genuine cognitive shifts over the baselines' generic affirmations.

6.3.2 Contribution of Components

We assessed the contribution of components by removing In-context Refinement Feedback and Thinking Pattern Modeling (Figure 2). The performance of each setting was observed using the scores of BLEU, ROUGE, and LM judge.

The results show that adding Thinking Pattern Modeling yields modest improvements over the baseline, while the full system provides the largest gains across all metrics. This indicates that feedback-driven iteration is the primary driver of performance. Smaller improvements in LLM-judge scores suggest that perceived quality is more challenging than lexical similarity. While Thinking Pattern Modeling alone yields a smaller quantitative gain than the iterative feedback loop, it provides the structured CBT analysis $A(S, T)$ in Eq. (1) that the feedback loop conditions on; the two components are therefore complementary rather than substitutable.

Table 3: The contribution of components. The results are average scores over LLMs.

Data	Settings			
	Baseline	✓	✓	✓
—	Thinking Pattern Modeling		✓	✓
	In-context Refinement Feedback			✓
	BLEU	22.56	25.59	**27.16**
CR	ROUGE	28.36	29.20	**32.83**
	LM Judge	4.10	4.13	**4.18**
	BLEU	50.64	59.01	**68.08**
PR	ROUGE	36.35	45.77	**53.82**
	LM Judge	4.05	4.09	**4.19**

6.4 Running Time and Cost

We measured the running time and cost of the framework on 100 samples for $T = 5$ iterations under a parallel setting (Table 4). GPT-4o-mini is the cheapest at \$0.0024 per sample and runs in 38.51 s; GPT-4.1 is the fastest at 30.90 s. GPT-5 is the most expensive both in latency and cost, suggesting that lighter models are preferable for deployment. It suggests that GPT-4o-mini is an appropriate option to balance performance and cost.

Table 4: Average inference time per sample and estimated cost per sample for $T = 5$ iterations.

Metric	4o-mini	GPT-4.1	GPT-5
Time/sample (s)	38.51	30.90	418.47
Cost/sample ($)	0.0024	0.0329	0.2737

7 Conclusion

This paper introduces a clinical-grounding framework for cognitive reframing, combining a CBT-based multi-agent backbone with an objective function that drives iterative in-context refinement. Experiments on two datasets show that (i) the objective function improves the quality of the backbone's reframes, and (ii) combining Socratic principles with semantic and stylistic alignment yields safer, less hallucinated reframes than the baselines. Future work will integrate additional modalities (e.g., visual or physiological signals) into the pipeline.

Limitations

Although achieving promising results, the proposed framework can still be improved in the following aspects. First, the framework is designed to deal with textual information. It follows the same setting as much prior work. However, to develop a more robust reframing system, other information sources such as visual or physiological signals should be utilized. Second, the collaboration of agents in the backbone is quite straightforward: the reframing agent synthesizes information from the other agents to produce a new reframe. It suggests that more sophisticated collaboration, e.g., deliberative agent committees for the final reframing, should be considered to improve the performance. Finally, human evaluation covers a wide range of metrics for safety and hallucination. However, this evaluation is conducted at a small scale. It suggests a larger scale of human evaluation for comprehensive assessment.

Ethics Statement

The authors confirm that this study does not have ethical issues. We emphasize that the proposed framework is a research prototype and is not intended for autonomous therapeutic deployment without clinical supervision; safety considerations are built into the objective function (harmfulness, fairness) and the human evaluation protocol, rather than treated as post-hoc concerns. We note that neither distortion detection, cognitive reframing, nor the framework itself is a substitute for screening, diagnosis, or healthcare treatment. The framework only uses distortion detection as a step in the reframing process. The framework can also make incorrect predictions of distortions. However, prediction is only used for cognitive reframing in a computational linguistic task. We emphasize that distortion detection and the reframing of negative thoughts should be confirmed by psychologists in actual medical or mental healthcare applications. For experiments, we used two benchmark datasets. They were released by the original authors under IRB approval, and our framework does not access any real patient records or private clinical data. For the trial experiment in Appendix A.7, we provided a consent form to student volunteers and asked for their consent to participate. The system only used situations and thoughts for reframing with the consent of users. Personal identifiers were stripped at input. The pilot was designed in consultation with a psychological expert. Thanks to shared prompts from DoT authors, we can successfully run these methods. The code and datasets are collected from public GitHub links from the original papers. Experiments do not have specific parameter tuning to maintain a fair comparison among methods. We also provided a clear guideline to annotators for human evaluation and did not use any personal information in our experiments.

References

Eshaan Agarwal, Raghav Magazine, Joykirat Singh, Vivek Dani, Tanuja Ganu, and Akshay Nambi. 2025. Promptwizard: Optimizing prompts via task-aware, feedback-driven self-evolution. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19974–20003.

Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *Preprint*, arXiv:2510.06265.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 463–476.

Aaron T. Beck, A. John Rush, Brian F. Shaw, and Gary Emery. 1979. *Cognitive Therapy of Depression*. The Guilford Press, New York.

Judith S. Beck. 2011. *Cognitive Behavior Therapy: Basics and Beyond*, 2nd edition. The Guilford Press, New York.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Linda L Carli. 1999. Cognitive reconstruction, hindsight, and reactions to victims and perpetrators. *Personality and Social Psychology Bulletin*, 25(8):966–979.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Joy Crosbie and Ekaterina Shutova. 2025. Induction heads as an essential mechanism for pattern matching in in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5034–5096.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Albert Ellis. 1991. The revised abc's of rational-emotive therapy (ret). *Journal of rational-emotive and cognitive-behavior therapy*, 9(3):139–172.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *Proceedings of The Web Conference (WWW)*, pages 335–345.

Anmol Goel, Nico Daheim, Christian Montag, and Iryna Gurevych. 2025. Socratic reasoning improves positive text rewriting. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 140–156.

Dennis Greenberger and Christine A Padesky. 2015. *Mind over mood: Change how you feel by changing the way you think*, volume 425. Guilford Publications.

Hajime Hotta, Huu-Loi Le, Manh-Cuong Phan, and Minh-Tien Nguyen. 2025. Metamo: Empowering large language models with psychological distortion detection for cognition-aware coaching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 862–872.

Jeffrey Ip. 2026. Llm evaluation metrics: The ultimate llm evaluation guide. Accessed: 2026-01-04.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190.

Subin Kim, Hoonrae Kim, Jihyun Lee, Yejin Jeon, and Gary Lee. 2025a. Mirror: Multimodal cognitive reframing therapy for rolling with resistance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14851–14880.

Subin Kim, Hoonrae Kim, et al. 2025b. Multimodal cognitive reframing therapy via multi-hop psychotherapeutic reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4863–4880.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathleen McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 43–52.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. ACL.

Songyang Liu, Chaozhuo Li, Jiameng Qiu, Xi Zhang, Feiran Huang, Litian Zhang, Yiming Hei, and Philip S. Yu. 2025. The scales of justitia: A comprehensive survey on safety evaluation of llms. *Preprint*, arXiv:2506.11094.

Steven Loria. 2018. Textblob: Simplified text processing. https://textblob.readthedocs.io/en/dev/. Accessed: 2025-12-20.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Poczós, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Maya Prabhakaran. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 3191–3202.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. *arXiv preprint arXiv:2307.02768*.

Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational ai in psychotherapy. In *Frontiers in psychiatry*, volume 10, pages 25–32.

Robert R Morris, Stephen M Schueller, and Rosalind W Picard. 2015. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of medical Internet research*, 17:e118.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://tinyurl.com/58j2ae8n.

OpenAI. 2025a. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/.

OpenAI. 2025b. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Kishore Papineni, Salim Roukos, et al. 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. ACL.

Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of change: Analyzing peer-based cognitive support in online mental health forums. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–13.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 3440–3453.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

James P Robson Jr and Meredith Troutman-Jordan. 2014. A concept analysis of cognitive reframing. *Journal of Theory Construction & Testing*, 18(2).

Arkadiy Saakyan and Smaranda Muresan. 2024. Iclef: In-context learning with expert feedback for explainable style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16141–16163.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference (WWW)*, pages 2029–2040.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023a. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:250–259.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. 2023b. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 9977–10000.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–13.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 217–226.

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.

C Estelle Smith, William Lane, Hannah Miller Hillberg, Daniel Kluver, Loren Terveen, and Svetlana Yarosh. 2021. Effective strategies for crowd-powered cognitive reappraisal systems: A field deployment of the flip* doubt web application for mental health. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, volume 5, pages 1–27.

Aseem Srivastava, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Response-act guided reinforced dialogue generation for mental health counseling. In *Proceedings of the ACM Web Conference 2023*, pages 1118–1129.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. In *Journal of Medical Internet Research (JMIR)*, volume 21, page e13320.

Rishit Toteja, Arindam Sarkar, and Prakash Mandayam Comar. 2025. In-context reinforcement learning with retrieval-augmented generation for text-to-sql. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10390–10397.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 982–993.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1251–1264.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725.

An Yang, Anfeng Li, et al. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022. Progen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the association for computational linguistics: EMNLP 2022*, pages 3671–3683.

Lianmin Zheng, Wei-Lin Chiang, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Caleb Ziems, Minzhi Li, Anthony Zhang, Mengting Wan, Zhiting Hu, and Diyi Yang. 2022a. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022b. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 2170–2184.

A Appendix

A.1 Distortion Detection

Distortion Types Distortion detection in Section 3.2.2 uses 12 distortion types: personalization, mind reading, overgeneralization, all-or-nothing thinking, emotional reasoning, labeling, magnification, mental filter, should statements, fortune-telling, catastrophizing, and no-distortion (Beck, 2020; Shreevastava and Foltz, 2021).

Error Propagation The performance of distortion detection may affect the quality of reframing. We observed this aspect by changing the F1 of distortion detection in the range of [100 (gold labels), 76.43 and 79.26 (the current F1), 60, 50, 25] and observed the performance of the framework. Due to space constraint, we run on the Cognitive Reframing dataset with GPT-5.

Table 5: Error propagation of distortion detection for reframing on the CR dataset.

F1-score	BLEU	ROUGE	LM judge
100.00	23.51	35.32	4.09
76.43	**19.74**	**33.20**	**4.09**
60.00	15.01	26.15	4.06
50.00	13.97	24.96	4.01
25.00	11.03	23.88	4.00

Table 6: Error propagation of distortion detection for reframing on the PR dataset.

F1-score	BLEU	ROUGE	LM judge
100.00	66.35	59.22	4.20
79.26	**63.26**	**55.85**	**4.19**
60.00	47.82	43.15	4.14
50.00	43.76	41.03	4.08
25.00	34.19	38.92	4.03

The table shows that performance consistently improves as distortion detection accuracy (F1) increases. Higher F1 scores lead to gains in BLEU and ROUGE, indicating better alignment with reference outputs. However, LLM-Judge scores improve only slightly, suggesting limited sensitivity to detection accuracy compared to lexical metrics. With a 25% gap in F1-scores, BLEU and ROUGE only reduce 3-4% while LM judge drops slightly. It shows that by considering distortion detection as a soft signal, the pipeline can mitigate wrong predictions using prompting.

Per-distortion-type Accuracy We reported the F1-score of distortion detection on the two datasets.

Table 7: Per-distortion-type F1-score on CR and PR.

Class	CR	PR
Personalization	71.39	74.27
Mind Reading	73.19	76.79
Overgeneralization	80.12	82.55
All-or-Nothing Thinking	78.47	81.14
Emotional Reasoning	75.38	78.19
Labeling	77.16	80.42
Magnification	68.34	72.08
Mental Filter	69.57	73.61
Should Statements	78.04	79.89
Fortune-Telling	74.81	77.53
Catastrophizing	81.23	83.47
No-distortion	89.46	91.18

The table shows that distortion detection performance varies across classes, with consistently

higher F1-scores on the PR dataset than CR. "No Distortion" achieves the highest performance, while categories such as Magnification and Mental Filter are more challenging. Overall, most classes achieve moderate to strong performance (70–85%), indicating that the detection is good for reframing.

A.2 Validation Frameworks

We compared our CBT-based backbone instantiated with Beck's Socratic Questioning against variants using ABCDE (Rational Emotive Behavior Therapy) (Ellis, 1991) and CBA (Cost-Benefit Analysis) (Greenberger and Padesky, 2015). Figure 7 shows that the Beck's Socratic Questioning instantiation produces better scores than the ABCDE and CBA variants. A possible reason is that Socratic Questioning follows a structured, step-by-step inquiry process that aligns with CoT and multi-agent reasoning of LLMs.

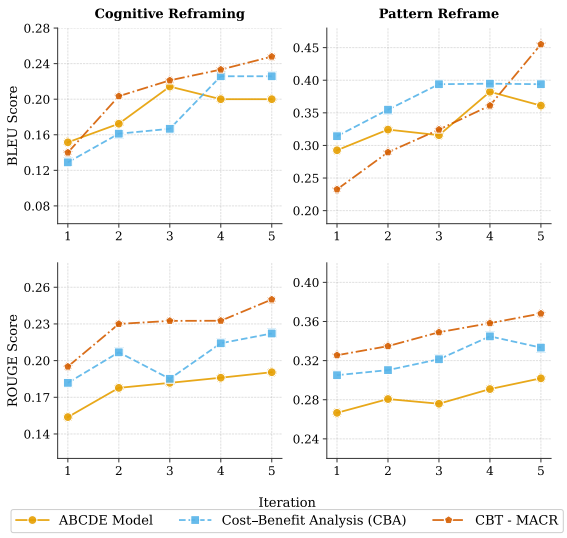

Figure 7: Performance of validation frameworks.

A.3 Parameter Tuning

We tested the performance of CBT-MACR with different parameter sets in Eq. (5). Table 8 shows that the combination of $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$ produces competitive performance over two datasets. We, therefore, empirically used this combination for the objective function in Section Eq. (5).

A.4 Performance using LM Judge

As mentioned in Section 5, we used BLEU, ROUGE, and LM judge for automatic evaluation. Due to space limitation, we only included BLEU and ROUGE scores in Section 6.1 (Figures 4 and 3). This section provides the scores of LM judge on the two datasets. Figure 8 shows the consistent trend with Figures 4 and 3, in which the proposed CBT-MACR produces strong performance compared to the baselines. It again confirms the efficiency of the backbone (Section 3.2) and the proposed objective function (Section 4.1).

Table 8: The average scores of CBT-MACR using different parameters over five iterations.

Dataset	Params [α, β, γ]	BLEU	ROUGE	LM Judge
CR	[0.5, 0.3, 0.2]	**26.71**	**31.58**	**4.06**
	[0.3, 0.2, 0.5]	23.27	25.95	4.04
	[0.3, 0.3, 0.3]	23.68	26.71	4.02
	[0.7, 0.2, 0.1]	23.74	26.45	4.05
	[0.1, 0.2, 0.7]	23.71	26.26	4.02
PR	[0.5, 0.3, 0.2]	68.08	**53.82**	**4.19**
	[0.3, 0.2, 0.5]	67.91	53.19	4.15
	[0.3, 0.3, 0.3]	68.04	53.49	4.16
	[0.7, 0.2, 0.1]	67.94	53.51	4.15
	[0.1, 0.2, 0.7]	**68.49**	53.64	4.12

A.5 Alignment of Feedback and Gold References

This section serves as an extension of the primary evaluation discussed in Section 6.3.1. As demonstrated in Table 9, our proposed framework, CBT-MACR (Multi-Agent Cognitive Reframing), achieves competitive scores in both automated and human validations, ensuring high alignment between the generated feedback and expert-curated gold reframes. Below, we detail the annotation guidelines, the evaluation setup, and provide a deeper analysis of the empirical results.

A.5.1 Evaluation Metrics and Annotation Guidelines

To assess the quality of the generated reframes, we established a 5-point Likert scale for two CBT-grounded metrics. Annotators and the LLM judge were provided with the following metrics:

Semantic Equivalence (SE) This metrics evaluates whether the generated response successfully conveys the core psychological message of the gold label (Ziems et al., 2022a). **1 - Contradictory:** Completely different meaning or contradicts the gold reframe. **2 - Poor:** Major deviations; fails to capture the main therapeutic intent. **3 - Fair:** Captures some elements of the gold reframe but misses key context or nuances. **4 - Good:** Mostly aligns with the core message, with only minor deviations in tone or vocabulary. **5 - Excellent:** Perfectly captures and aligns with the core message of the

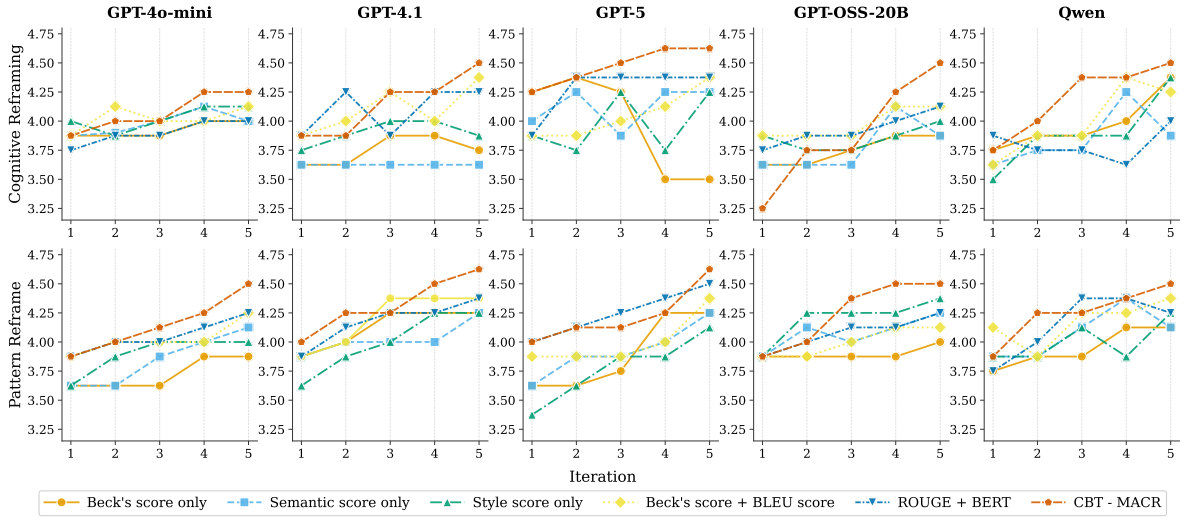

Figure 8: Evolution of LM judge scores over five iterations on the two datasets.

gold reframe, demonstrating rich and contextually appropriate vocabulary.

Cognitive Shift (CS) It measures the effectiveness of the response in transitioning the user from a negative thought pattern to a balanced, neutral, or positive perspective (Sharma et al., 2023b). **1 - Harmful:** Fails to challenge the negative thought and may inadvertently reinforce it. **2 - Weak:** Attempts to challenge the thought but lacks logical grounding or therapeutic value. **3 - Generic:** Provides a generic positive statement (toxic positivity) but lacks deep logical challenging of the specific distortion. **4 - Effective:** Offers a strong logical challenge and alternative perspective, but may slightly lack empathetic delivery. **5 - Highly Effective:** Empathetic, highly tailored, and logically dismantles the negative thought, offering a strong, realistic, and actionable alternative perspective.

A.5.2 Evaluation Setup

LLM-as-a-Judge For a comprehensive, dataset-wide evaluation, we employed `GPT-4o-mini` as an automated judge. To align the LLM's scoring mechanism with human intuition, we utilized a Few-Shot Prompting strategy coupled with Chain-of-Thought (CoT) reasoning. The model was provided with the exact annotation rubric above, alongside three fully annotated examples (with expert reasoning) for calibration. The LLM was instructed to output its reasoning before assigning a final numerical score, ensuring that the evaluation was grounded in the specific CBT principles of each metric.

Human validation We recruited a panel of 8 annotators to evaluate a representative, randomly sampled subset of 100 instances from the dataset. To prevent evaluation bias, the origin of the generated responses (baseline vs. proposed models) was strictly anonymized. The annotators independently assessed the therapeutic quality of the reframes based on the provided Semantic Equivalence and Cognitive Shift rubrics.

A.5.3 Detailed Result Analysis

Table 9 reveals several key insights regarding model performance across the two datasets.

Superiority of CBT-MACR Across all baseline models, the integration of our proposed iterative Socratic method (CBT-MACR) consistently yielded the highest scores in both Semantic Equivalence and Cognitive Shift. This indicates that systematically applying psychological frameworks significantly enhances the structural and therapeutic quality of the generated text compared to relying on single-dimension evaluations (e.g., purely semantic or style-based scores).

Model capabilities Advanced large language models, specifically `Qwen` and `GPT-5`, demonstrated an exceptional capacity to internalize the CBT-MACR framework, yielding the highest human-evaluated *Cognitive Shift* performance. Crucially, this positive trend extended to smaller, open-source models like `GPT-OSS-20B`, which exhibited marked improvements when equipped with the iterative approach. This consistent uplift across varying model sizes and architectures strongly validates the frame-

work's model-agnostic robustness.

Human vs. LLM scoring tendencies We observed that the LLM judge tended to be stricter (scoring lower) on the Semantic Equivalence metric compared to human annotators. However, for the Cognitive Shift metric, the scores between human and LLM judges were highly correlated. This suggests that while LLMs may penalize minor lexical deviations from the gold label heavily, they are highly capable of recognizing and rewarding the logical dismantling of cognitive distortions, aligning well with human clinical judgment.

A.6 Qualitative Analysis

Table 10 illustrates the qualitative evolution of our multi-agent framework compared to the baselines. The SR baseline generates a plausible but somewhat verbose reframe, attempting to cover multiple possibilities without a clear therapeutic focus. It is understandable that SR is a shallow method that uses a simple prompt for reframing. Similarly, while the DoT method accurately identifies the core distortion of self-worth, it produces a fixed output that lacks the nuanced, supportive tone found in the expert reference. Both approaches are limited by their cognitive modeling ability to self-correct or deepen clinical reasoning after the initial output.

Evolution of our method Our framework is clinical grounding with in-context refinement feedback that shows a dynamic refinement trajectory. It can refine negative thoughts that are safer and less hallucinated (objective scores in Figure 9).

Initial correction (Loops 1–2) The early iterations focus on correcting basic semantic deviations, guiding the model away from the distorted thought.

Reasoning integration (Loops 3–4) Subsequent loops begin to incorporate Socratic principles, semantic and stylistic alignment, where the agent explicitly questions the evidence (e.g., *"What evidence shows she's ignoring me?"*) rather than relying solely on surface-level generation.

Convergence (Loop 5) The final output synthesizes these insights, combining a rational alternative explanation with self-worth validation to closely align with the expert annotation.

Figure 9 shows the trend of scores using different components in Eq. (5). It indicates that the scores increase over iterations, showing that it aligns with the improvement of the performance.

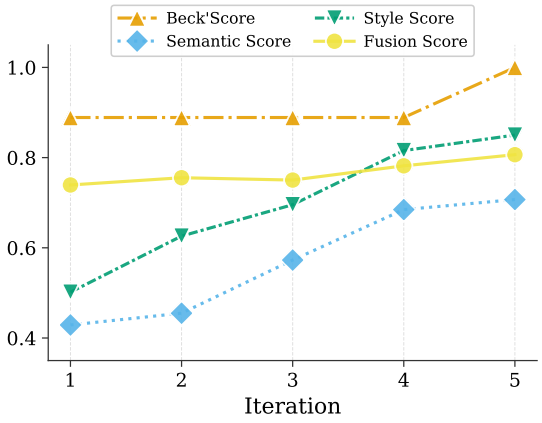

Figure 9: The scores of the example in Table 10. The fusion score is from our objective function.

A.7 Demo Application

Figure 11 illustrates the end-to-end use of the **MindReframe** system through a sequence of screenshots, where each smaller image represents one screen in the user workflow. First, the user enters an objective description of the situation and the thought that came to mind, providing the system with the raw material for reflection. The system then detects possible cognitive distortions in the user's interpretation, such as mind reading, overgeneralization, and catastrophizing, and explains why these patterns may be influencing the user's perspective. Next, it performs a deeper analysis by identifying emotional impact, underlying beliefs, and common triggers associated with the thought. Based on this analysis, the system generates multiple reframing options grounded in therapeutic principles, allowing the user to compare and select the one that resonates most. After the user confirms a preferred reframe, the final screen presents a polished alternative perspective alongside the original situation and thought, and also shows the AI feedback used in the in-context refinement feedback process. Overall, the workflow shows how the system guides users step by step from initial self-report to a more balanced and constructive interpretation. Figure 10 shows that users are happy with reframed thoughts after three to four iterations ($N = 50$).

User experience We deployed **MindReframe**[3][4] (Figure 11) for a trial educational scenario in which students can access AI psychological counseling. Under the guidance of a psychological expert, 50 students put their situations and thoughts, and rated

[3]Demo at: https://tinyurl.com/2papvsd6
[4]Video: https://tinyurl.com/3ywjtujx

Table 9: Performance comparison of various evaluation methods across different LLMs on Cognitive Reframing and Pattern Reframe datasets. **SE**: Semantic Equivalence, **CS**: Cognitive Shift. **Bold**: the best; *italic*: the second best.

Models	Methods	Cognitive Reframing				Pattern Reframe			
		Human		LLM		Human		LLM	
		SE	CS	SE	CS	SE	CS	SE	CS
GPT-4o-mini	Beck's score only	3.08	3.39	2.57	3.36	3.15	3.38	2.58	3.45
	Semantic score only	3.10	3.43	2.58	3.40	3.14	3.39	2.57	3.46
	Style score only	3.10	3.42	2.58	3.39	3.12	3.38	2.56	3.45
	Beck's + BLEU score	3.26	3.37	2.72	3.34	3.26	3.20	2.67	3.27
	ROUGE + BERT score	3.24	3.41	2.70	3.38	3.29	3.39	2.70	3.46
	CBT - MACR	**3.31**	**3.44**	**2.76**	**3.41**	**3.36**	**3.42**	**2.75**	**3.49**
GPT-4.1	Beck's score only	2.48	2.18	2.07	2.16	2.60	2.10	2.13	2.14
	Semantic score only	2.48	2.19	2.07	2.17	2.59	2.12	2.12	2.16
	Style score only	**3.41**	**3.51**	**2.84**	**3.48**	2.59	2.11	2.12	2.15
	Beck's + BLEU score	**3.41**	3.50	**2.84**	3.47	3.48	3.49	2.85	3.56
	ROUGE + BERT score	3.32	3.47	2.77	3.44	3.42	3.52	2.80	3.59
	CBT - MACR	3.40	3.50	2.83	3.47	**3.49**	**3.56**	**2.86**	**3.63**
GPT-5	Beck's score only	2.44	2.59	2.03	2.56	2.51	2.36	2.06	2.41
	Semantic score only	2.41	2.65	2.01	2.62	2.48	2.48	2.03	2.53
	Style score only	2.41	2.57	2.01	2.54	2.48	2.36	2.03	2.41
	Beck's + BLEU score	2.71	3.24	2.26	3.21	2.76	3.13	2.26	3.19
	ROUGE + BERT score	4.28	3.82	3.57	3.78	2.77	3.21	2.27	3.28
	CBT - MACR	**4.32**	**3.84**	**3.60**	**3.80**	**2.83**	**3.24**	**2.32**	**3.31**
GPT-OSS-20B	Beck's score only	2.68	2.35	2.23	2.33	2.51	2.16	2.06	2.20
	Semantic score only	2.66	2.39	2.22	2.37	2.51	2.14	2.06	2.18
	Style score only	2.69	2.36	2.24	2.34	2.50	2.09	2.05	2.13
	Beck's + BLEU score	3.02	2.75	2.52	2.72	3.10	2.70	2.54	2.76
	ROUGE + BERT score	3.01	2.77	2.51	2.74	3.07	2.75	2.52	2.81
	CBT - MACR	**3.07**	**2.80**	**2.56**	**2.77**	**3.14**	**2.78**	**2.57**	**2.84**
Qwen	Beck's score only	3.42	3.86	2.83	3.59	3.38	3.79	2.81	3.65
	Semantic score only	3.91	3.32	2.84	3.61	3.85	3.28	2.79	3.64
	Style score only	3.28	3.15	2.85	3.60	3.22	3.10	2.81	3.66
	Beck's + BLEU score	3.65	3.94	2.85	3.71	3.61	3.85	2.86	3.70
	ROUGE + BERT score	3.88	3.52	3.54	**3.74**	3.82	3.48	**3.66**	3.75
	CBT - MACR	**4.03**	**4.38**	**3.64**	3.72	**3.97**	**4.31**	2.93	**3.88**

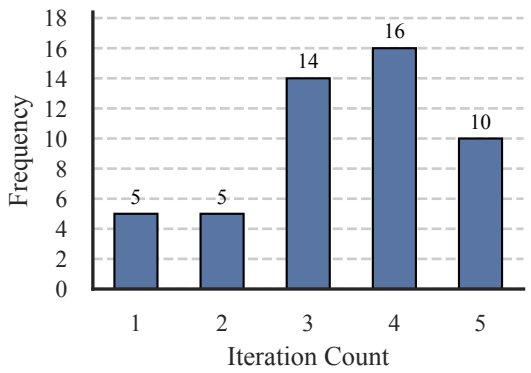

Figure 10: The number of iterations that users are happy with reframed thoughts ($N = 50$.

reframed thoughts using Likert scale ratings (1-4) with four metrics (Sharma et al., 2023b). **Cognitive Believability** evaluates where the reframe is realistic and logically accepted by users. **Distress Reduction** measures the immediate emotional relief. **Retention (Recall)** assesses the memorability and potential long-term utility of the new perspective. **Coping Efficacy** measures perceived skill acquisition and capability improvement.

Table 11: User experience evaluation (N=50).

Evaluation Dimension	Mean (max 4.0)	Observed Trend
Cognitive Believability	3.42	High Acceptance
Distress Reduction	3.40	Effective Regulation
Retention (Recall)	3.48	Strong Memorability
Coping Efficacy	3.40	Skill Acquisition

Table 11 shows a strong alignment between Believability and Distress Reduction. It suggests AI-generated reframes are accepted as genuine and provide tangible relief. Retention achieves the highest score (3.48), implying that the system fosters user ownership and enhances memorability compared to passive reception. Figure 10 shows that users satisfy with the reframes after 3-4 iterations that are acceptable for actual cases. Qualitative analysis is shown in Appendix A.6 and Table 10.

Table 10: **Qualitative Analysis of Iterative Refinement.** An example comparing baseline methods against our Multi-Agent with iterative in-context clinical grounding feedback (Loops 1–5). The scores indicate the model's progressive improvement in alignment with the objective function.

Model / Step	Text Content	BLEU
Input Context		
Situation	My wife didn't look up from her phone or answer me when I came to talk to her.	–
Thought	Am I not worth her time?	–
Gold Reframe	Maybe my wife is busy at the moment and I should come back in a few minutes. I'm sure she will give me her time if I wait until after she isn't busy.	–
Baselines reframe		
SR	Maybe my wife was just busy or distracted when I tried to talk to her. It doesn't necessarily mean I'm not worth her time. There could be other reasons she didn't respond right away...	0.1901
DoT	She might be distracted right now; this moment doesn't define my worth or how much she cares about me.	0.2415
CBT Agents + RMs reframe (5 Loops)		
Loop 1	My wife might be distracted or focused on something important, so her not responding right away doesn't mean I'm not worth her time or attention. ↪ ***Feedback:*** *Apply each relevant Socratic question from Beck's framework to deepen evidence examination, broaden alternative possibilities, regulate distortions, balance self-assessment, and enhance cognitive flexibility.*	0.1481
Loop 2	Could she be busy or distracted, rather than ignoring me? I know she cares about me, even if she didn't respond; my worth isn't defined by this moment alone. ↪ ***Feedback:*** *Apply each Socratic question systematically to continuously deepen and expand your thought evaluation process.*	0.2647
Loop 3	Maybe my wife was busy or distracted with something on her phone, not intentionally ignoring me; this doesn't mean I'm unimportant or she doesn't care about me. ↪ ***Feedback:*** *Focus on systematically questioning evidence, exploring alternatives, and balancing self-views to deepen cognitive restructuring.*	0.2667
Loop 4	What evidence shows she's ignoring me? Could she be busy or distracted? One moment doesn't define my worth; she usually shows care in many ways over time. ↪ ***Feedback:*** *Focus on deepening evidence assessment and broadening alternative explanations using Beck's Socratic questions for comprehensive cognition review.*	0.2903
Loop 5	**Maybe my wife was focused on something urgent on her phone, not ignoring me. It doesn't mean I'm unimportant, and I know she values me even if she's momentarily distracted.** ↪ ***Feedback:*** *Apply each Socratic question fully to deepen evidence review, alternative exploration, distortion correction, balanced self-view, and cognitive flexibility.*	**0.3235**

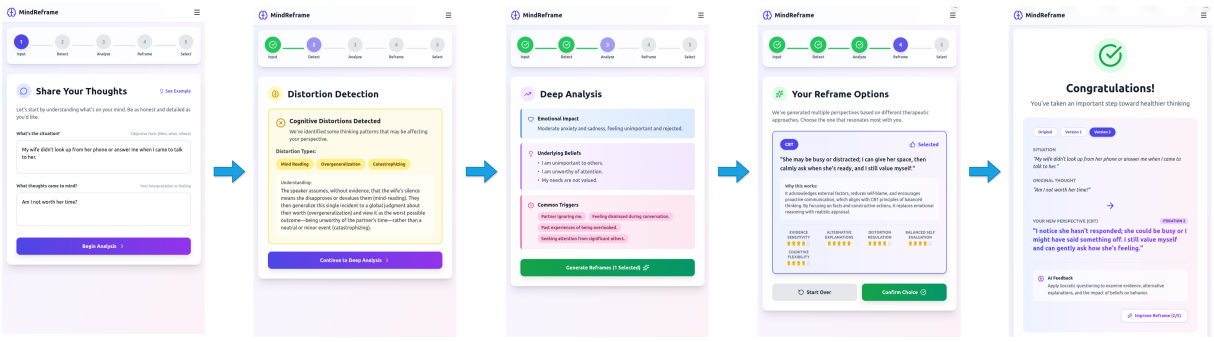

Figure 11: The user interface of the *MindReframe* platform. The landing page illustrates the system's core value propositions and the five-stage cognitive restructuring pipeline, designed for self-guided mental health support (Demo video: https://tinyurl.com/3ywjtujx).

A.8 Inference Loop and Detailed Prompt

The prompt of the Supervisor Agent We employ a structured prompting strategy based on

Beck's Socratic Questioning framework. Below is the exact instruction used for the Supervisor Agent.

Supervisor Agent Prompt

```
You are a CBT evaluation agent.
Your task has TWO SEPARATE responsibilities:

1. Quantitative scoring using Beck's Socratic Questions
2. Instructional feedback ONLY (no analysis, no evaluation
   language)

All judgments must strictly follow Beck's Socratic
Questioning framework (Beck, 1979; Beck, 2011). Do not use
any other CBT models or therapeutic approaches.

--------------------------------------------------------
### PART 1 — Beck Evaluation Criteria (Scoring Reference Only)

1. Evidence Sensitivity (Beck Q1-3)
2. Alternative Explanations (Beck Q4-6)
3. Distortion Regulation (Beck Q7-9)
4. Balanced Self-Evaluation (Beck Q10-11)
5. Cognitive Flexibility (Beck Q12)

--------------------------------------------------------
### PART 2 — Instructional Feedback RULES (CRITICAL)
For EACH criterion, generate ONLY a concise improvement
instruction.
Feedback MUST:

• Be written in imperative instructional form
• State what cognitive step is missing and which Socratic
  question to apply
• Be grounded in Beck's Socratic questioning logic
• Be 1-2 sentences MAX per criterion

Feedback MUST NOT:

• Praise, criticize, judge, or summarize performance
• Explain reasoning or provide analysis
• Rewrite or suggest a new reframed thought
• Add emotional validation or encouragement
• Mention scores, strengths, or weaknesses

If a criterion is already strong:

• Provide a micro-refinement instruction

--------------------------------------------------------
### Scoring Scheme

• Evidence Sensitivity: 0-2
• Alternative Explanations: 0-2
• Distortion Regulation: 0-2
• Balanced Self-Evaluation: 0-2
• Cognitive Flexibility: 0-1

--------------------------------------------------------
### Input

• Situation: {situation}
• Original thought: {thought}
• Reframing response: """ {reframe_response} """

--------------------------------------------------------
### Output Format (STRICT JSON — no extra text)
{{
   "scores": {{
     "evidence_sensitivity": <0-2>,
     "alternative_explanations": <0-2>,
     "distortion_regulation": <0-2>,
     "balanced_self_evaluation": <0-2>,
     "cognitive_flexibility": <0-1>,
     "total_score": <sum>
   }},
   "feedback": {{
     "evidence_sensitivity": "<instruction only>",
     "alternative_explanations": "<instruction only>",
     "distortion_regulation": "<instruction only>",
     "balanced_self_evaluation": "<instruction only>",
     "cognitive_flexibility": "<instruction only>",
     "summary": "<single-sentence global instruction>"
   }}
}}
Only output valid JSON.
```

The prompt for the DoT method The bellow prompt was used for DoT, thanks to the sharing of Chen et al. (2023).

DoT prompt template

```
You are a therapist assistant.
Given a situation and distorted thought, your task is to
reframe a distorted thought into a more rational, emotionally
balanced response. You must first analyze the thought based
on cognitive behavioral therapy principles.
The situation is:
{situation}
The distorted thought is:
{thought}
Step 1: Based on the situation and the distorted thought,
finish the following Diagnosis of Thought (DoT) questions:
1. What is the situation? Find out the facts that are
   objective; what is the speaker thinking or imagining?
   Find out the thoughts or opinions that are subjective.
2. What makes the speaker think the thought is true or is
   not true? Find out the reasoning processes that support
   and do not support these thoughts.
3. Why does the speaker come up with such reasoning process
   supporting the thought? What's the underlying cognition
   mode of it?
Step 2: Identify if there is a cognitive distortion. If yes,
specify the type. The most common distortions include:
• Emotional reasoning: Letting one's feeling about something
  overrule facts to the contrary.
• Overgeneralization: Major conclusions are drawn based on
  limited information.
• Mental filter: Placing all one's attention on, or seeing
  only, the negatives of a situation.
• Should statements: Ironclad rules about how a person should
  behave.
• All-or-nothing thinking: Looking at a situation as either
  black or white.
• Mind Reading: Suspecting what others are thinking or
  motivations behind actions.
• Fortune-Telling: Expecting things to happen a certain way,
  or assuming things will go badly.
• Magnification: Emphasizing the negative or playing down
  the positive.
• Personalization: Taking up the blame for a situation
  involved many factors.
• Labeling: Giving someone or something a label without
  finding out more.
Step 3: Based on the analysis above, write a reframing of
the distorted thought into a more rational and emotionally
balanced response.
• Write from the first-person perspective (as the speaker).
• Response should reflect understanding of the distortion and
  show a healthier, constructive way of thinking.
• Limit response to 20-30 words.
• Ensure your response is concise.
• Do not output any diagnosis or distortion labels — only the
  final reframed response.
Output strictly in this format:
{
   "reframing response": "<your reframed response here>"
}
```

Reframing Agent: Iterative Refinement Base Prompt

```
You are a therapist assistant specializing in CBT-based
cognitive reframing. You previously generated reframed
responses and received structured feedback with scores. Your
goal is to revise your response to achieve a HIGHER overall
quality score.

--------------------------------------------------------
================== CONTEXT ==================
The situation is: {situation}
The distorted thought is: {thought}

--------------------------------------------------------
================== SCORING CONTEXT ==================
Each response is evaluated using a Total Reward Score ranging
from 0 to 1. This score is a weighted combination of three
specific metrics:
1) Clinical Quality (Scaled Beck Score): Calculated as Total
Beck Score / 9.0.
2) Content Consistency (Semantic Anchor): Embedding
Similarity between 'SITUATION' and 'RESPONSE'.
3) Style & Tone Match (Style Anchor): Sentiment/Tone Alignment
```

with an 'EXPERT REFERENCE'.
Your task is to produce a revision that IMPROVES this Total Score.

========== YOUR TASK ==========
Revise your reframing response to address the supervisor's instructions AND to move the response toward patterns associated with higher Fusion Scores.
Follow these rules: - Write in the first-person perspective. - Reflect a healthier, more balanced way of thinking. - Ground the thought in realistic evidence and alternative explanations. - Keep it between 20–30 words. - Do not mention diagnoses, distortions, CBT, scores, or evaluation. - Output strictly in this JSON format: {{ "reframing_response": "<your revised response here>" }}

Dynamic State Variables (Injected at Loop $t \geq 1$)

[The following sections are appended to the Base Prompt dynamically during each iterative refinement loop]
========== RESPONSE HISTORY ==========
Below are previous reframing attempts with their Fusion Scores. Use this history to identify what improves or reduces the score, and adjust accordingly. {history_response}
========== PRIOR OUTPUT ==========
Your most recent reframing response was: "{old_response}"
========== FEEDBACK FROM SUPERVISOR ==========
The supervisor provided criterion-level improvement instructions: {feedback}

Dynamic State Variables (Concrete Example at Loop $t = 2$)

[The following dynamic data is injected into the Base Prompt for Refinement]

========== CONTEXT ==========
The situation is:
I lent my phone charger to my neighbor a few days ago, and they haven't returned it yet.

The distorted thought is:
They are taking me for granted and don't appreciate my help at all.

========== RESPONSE HISTORY ==========
Below are previous reframing attempts with their Fusion Scores.
Use this history to identify what improves or reduces the score, and adjust accordingly.

[Loop 0]
- Response: "My neighbor may have simply forgotten to return my charger; it doesn't mean they don't appreciate me. I'll ask them about it."
- Score: 0.55

[Loop 1]
- Response: "I could ask my neighbor if they forgot my charger or needed it longer. There might be reasons I'm not aware of, and I want to understand their side."
- Score: 0.77

========== PRIOR OUTPUT ==========
Your most recent reframing response was:
"I can ask my neighbor if they forgot my charger or needed it longer. There could be reasons I'm not aware of, and I want to understand their situation."

========== FEEDBACK FROM SUPERVISOR ==========
The supervisor provided criterion-level improvement instructions:
{
"evidence_sensitivity": "Ensure you identify specific evidence that supports your original thought.",
"alternative_explanations": "Consider additional perspectives by asking, 'What else could explain their behavior?'",
"distortion_regulation": "Challenge any cognitive distortions by asking, 'Am I viewing this situation in an overly negative way?'",
"balanced_self_evaluation": "Reflect on your role in the situation by asking, 'What might I have done differently?'",
"cognitive_flexibility": "Practice thinking of multiple outcomes by asking, 'What are other possible scenarios here?'",
"summary": "Focus on enhancing your evidence sensitivity and balanced self-evaluation."
}

The prompt of SR This is the prompt for the **Shallow Reframing - SR** in Section 6.1.

SR prompt template

You are a therapist assistant.
Your task is to reframe a distorted thought into a more rational, emotionally balanced response.
Role-play as the person who had the distorted thought, and write the reframing from their perspective.
The situation:
{situation}
The distorted thought:
{thought}
- Your output must strictly follow the JSON format below:
{
"reframing response": "<your reframed response here>"
}