MENTAL-R1: ALIGNING LLM REASONING FOR MENTAL HEALTH ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Mental health problems such as anxiety, depression, and suicide remain urgent global challenges, where timely and accurate assessment is critical for effective intervention. Recently, large language models have been explored for mental health assessment. However, existing general-purpose post-training methods do not align with the cognitive processes of human assessment, which may lead to unreliable reasoning outcomes. To bridge this gap, we propose Cognitive Relative Policy Optimization (CRPO), a reinforcement learning framework tailored for the mental health domain. CRPO extends group relative policy optimization by integrating stage-dependent uncertainty modeling into the policy optimization process. Specifically, we introduce a stage-wise entropy regularization mechanism that encourages broad exploration in early reasoning phases and progressively enforces confident decision-making in later stages, mimicking the human cognitive shift from uncertainty to certainty. In addition, inspired by cognitive appraisal theory, we formalize cognitive reasoning stages, thereby guiding theory-grounded interpretable inference. Experiments on 8 mental health datasets show that CRPO achieves an average improvement of 10.4 percentage points in weighted F1-score over the best reinforcement learning baseline. Furthermore, the CRPO-trained model Mental-R1 demonstrates clear advantages compared with existing large language models on reasoning-intensive cases, suggesting that CRPO enhances reasoning capabilities for mental health assessment.

1 INTRODUCTION

Mental health problems such as depression and suicidal behavior have become global burdens for human well-being. According to the World Health Organization, more than 720,000 people die by suicide each year, which is equivalent to one person every 43 seconds (World Health Organization, 2023). Early assessment of mental health conditions is therefore crucial for timely intervention and prevention. Mental Health Assessment (MHA) focuses on identifying individuals' mental conditions, including loneliness (Jiang et al., 2022; Wang et al., 2024), depression (Sampath & Durairaj, 2022; Naseem et al., 2022), stress (Wang et al., 2020; 2022), anxiety (Owen et al., 2020; Yu et al., 2023), and suicide risk (Cao et al., 2019; 2021) from their textual statements.

Recently, large language models (LLMs) have emerged as a promising paradigm for mental health assessment (Shi et al., 2025; Xu et al., 2024; Yang et al., 2024; Hu et al., 2025b; Ravenda et al., 2025; Rohei et al., 2026; Zhai et al., 2025) due to their strong generalization ability across diverse natural language understanding tasks (Naveed et al., 2025; Zhao et al., 2025; Hu et al., 2025a;c; Chen et al., 2026). Prior studies primarily use standard supervised fine-tuning (SFT) (Gao et al., 2025) or reinforcement learning (RL) (Kumar et al., 2025) to adapt LLMs for mental health tasks. However, these "generalist" post-training methods often fail to reflect the real-world assessment process of mental health professionals, limiting their reliability in healthcare applications.

In real-world assessment, mental health professionals often seek to understand an individual's condition by reconstructing their underlying cognitive process (Beck, 2020; Persons, 2012; Kuyken et al., 2011; Eells, 2022). Two characteristics of this reasoning process are particularly important. First, the assessment process follows the natural dynamics of human cognition (Elstein et al., 1978; Higgs et al., 2024; Garb, 1998). At early stages, mental health professionals tend to collect observations and explore possible signals with relatively high uncertainty. As more contextual in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

formation is considered, they gradually refine their interpretation and move toward a more confident assessment. This uncertainty-to-certainty transition reflects a fundamental property of human cognitive reasoning (Gold & Shadlen, 2007; Clark, 2013). Second, such assessments typically follow theory-grounded cognitive stages (Wright et al., 2017). Consistent with the cognitive-behavioral framework and the ABC model (Ellis, 1962; Beck, 1979), the process often begins by identifying potential stimuli or life events that may contribute to psychological distress. It then analyzes how the individual cognitively appraises these events (Lazarus & Folkman, 1984) and how such appraisals lead to affective or behavioral reactions. Finally, the individual's mental state is inferred.

Inspired by these characteristics of cognitive process reconstruction, we propose Cognitive Relative Policy Optimization (CRPO), a reinforcement learning framework that aligns LLM reasoning with human cognitive dynamics and theory-grounded cognitive stages.

To model the cognitive dynamics, we introduce stage-wise entropy regularization, a stage-dependent uncertainty control mechanism integrated into the policy optimization objective. Unlike many reinforcement learning approaches (Li et al., 2024; Hu et al., 2025d; Guo et al., 2025) that apply a uniform exploration strategy throughout the reasoning process, our method explicitly modulates entropy across reasoning stages. Early reasoning stages are encouraged to maintain higher entropy to promote diverse exploration, while later stages gradually reduce entropy to guide the model toward more confident conclusions. This design enables a principled stage-aware exploration-conclusion trade-off that mirrors the uncertainty-to-certainty transition in human cognition.

To capture the theory-grounded reasoning stages, we draw inspiration from cognitive appraisal theory (Lazarus & Folkman, 1984; Ellsworth, 1991; Watson & Spence, 2007), a classic psychological framework that explains the internal cognitive processes underlying human mental responses. Based on this theory, we formalize a set of reasoning stages, including stimulus, primary appraisal, secondary appraisal, reaction, and mental state. We operationalize this framework during training by designing a format reward that encourages outputs consistent with these reasoning stages.

In summary, CRPO bridges the reasoning gap between generalist LLM post-training methods and real-world mental health assessment practice, thereby improving performance. The main contributions of this paper are threefold:

- **Cognitive-Inspired Reinforcement Learning.** We propose Cognitive Relative Policy Optimization (CRPO), a reinforcement learning framework that aligns LLM reasoning with human cognitive dynamics. Our primary algorithmic contribution is stage-wise entropy regularization, which translates the "uncertainty-to-certainty" cognitive shift into a stage-dependent uncertainty modulation mechanism within the policy optimization objective. Furthermore, we formalize theory-grounded reasoning stages inspired by cognitive appraisal theory to support interpretable inference.
- **Extensive Empirical Validation.** Experiments on 8 mental health datasets show that CRPO consistently outperforms existing post-training baselines, yielding an average improvement of 10.4 percentage points in weighted F1-score. Furthermore, the CRPO-trained model Mental-R1 demonstrates an advantage of approximately 15.6 percentage points over the best-performing LLM on complex samples, suggesting that CRPO effectively enhances model's reasoning capabilities for mental health.
- **Transparent Benchmark.** The evaluation of mental health assessment can be hindered by the mixed accessibility of existing benchmarks, where datasets are not uniformly open. To facilitate this area, we construct a transparent benchmark based entirely on publicly accessible datasets. By systematically comparing modern RL-based and LLM baselines, we provide a solid foundation for advancing interdisciplinary research in AI and healthcare.

2 Related Work

2.1 Mental Health Assessment

Mental health assessment in computational research has primarily focused on identifying conditions such as depression (Sampath & Durairaj, 2022; Naseem et al., 2022; Fisher et al., 2026), stress (Wang et al., 2023; 2025; Ikae et al., 2026), anxiety (Owen et al., 2020; Yu et al., 2023; Hidayat et al., 2025), and suicide risk (Cao et al., 2019; KINA et al., 2026) from textual data. These

Table 1: Comparison of our CRPO-trained Mental-R1 with existing mental health–focused LLMs. Mental-R1 algorithmically introduces cognition-aligned uncertainty dynamics and theory-grounded cognitive reasoning stages.

Model	Training	Uncertainty Dynamics	Response Trajectory	Benchmark Datasets	Str	Anx	Dep	Sui	Lon
Mental-QLM (Shi et al., 2025)	SFT	Not modeled	Answer + explanation	5 public	✓	✗	✓	✗	✗
Mental-LLM (Xu et al., 2024)	SFT	Not modeled	Answer + explanation	5 public	✓	✗	✓	✓	✗
Mental-Llama (Yang et al., 2024)	SFT	Not modeled	Answer + explanation	5 public	✓	✗	✓	✓	✓
Mental-GLM (Zhai et al., 2025)	SFT	Not modeled	Answer + explanation	2 public	✗	✗	✗	✓	✗
Mental-R1 (Ours)	CRPO (RL)	Cognition-aligned exploration → certainty	Cognitive stages + answer	8 public	✓	✓	✓	✓	✓

Str: Stress; Anx: Anxiety; Dep: Depression; Sui: Suicide; Lon: Loneliness.

tasks are typically formulated as classification problems. Recent studies have applied large language models (LLMs) to mental health assessment (Yulianti et al., 2025; Nanda et al., 2024; Jin et al., 2025). Xu et al. (2024) evaluated multiple LLMs on textual mental health prediction tasks, showing that their instruction-tuned models such as Alpaca and FLAN-T5 substantially outperform prompt-based baselines. Yang et al. (2024) enhanced interpretability by constructing an explanation dataset through ChatGPT generation, and fine-tuned LLaMA-2 to jointly improve prediction and explanation quality. Shi et al. (2025) proposed a lightweight 0.5B-parameter model with dual LoRA modules and data pruning, achieving competitive results on benchmark datasets with much lower resource requirements. However, these works do not align with human cognitive dynamics and lack theory-grounded staged reasoning. To address these limitations, our work explicitly introduces a stage-wise entropy regularization strategy that guides LLM reasoning from early-stage exploration to final certainty, and integrates cognitive appraisal theory to define structured reasoning stages.

2.2 Reinforcement Learning for Reasoning

Reinforcement learning (RL) has recently become a central paradigm for enhancing the reasoning ability of large language models (Lightman et al., 2023; Shao et al., 2024; Guo et al., 2025; Zhang et al., 2026; Yue et al., 2025; Zhang et al., 2025). A representative approach is reinforcement learning from human feedback, i.e., RLHF (Ouyang et al., 2022), which fine-tunes models with preference data to align their outputs with human judgments. In parallel, Direct Preference Optimization, i.e., DPO (Rafailov et al., 2023) reformulates preference learning into a supervised objective, enabling more stable and efficient optimization. Beyond preference learning, recent work has explored how RL can directly improve reasoning quality. Group Relative Policy Optimization, i.e., GRPO (Guo et al., 2025) has been proposed as an efficient alternative for training LLMs on reasoning tasks, using group-based relative rewards to enhance sample efficiency and stability. Other studies further extend this line of research: RL Tango (Zha et al., 2026) employs joint generator-verifier training to bolster robustness. DAPO (Yu et al., 2025) introduces decoupled clipping strategies alongside dynamic sampling to optimize training stability and signal quality. However, these approaches do not reflect how human cognition works, where early reasoning tends to be exploratory and later reasoning becomes more decisive. Our CRPO framework builds on this line by introducing a stage-wise entropy regularization strategy, which explicitly modulates exploration and certainty across reasoning stages, thereby aligning LLM reasoning more closely with human cognition.

3 Cognitive Relative Policy Optimization (CRPO)

In this section, we present Cognitive Relative Policy Optimization (CRPO), a reinforcement learning framework specifically designed for mental health assessment. An overview of the proposed CRPO framework is illustrated in Figure 1.

3.1 Preliminary

CRPO is built on Group Relative Policy Optimization (GRPO), a critic-free reinforcement learning framework designed for reasoning-oriented generation that estimates advantages via relative comparisons among a group of sampled outputs (Guo et al., 2025). This relative optimization paradigm provides a robust foundation for our CRPO, upon which stage-aware uncertainty regularization can

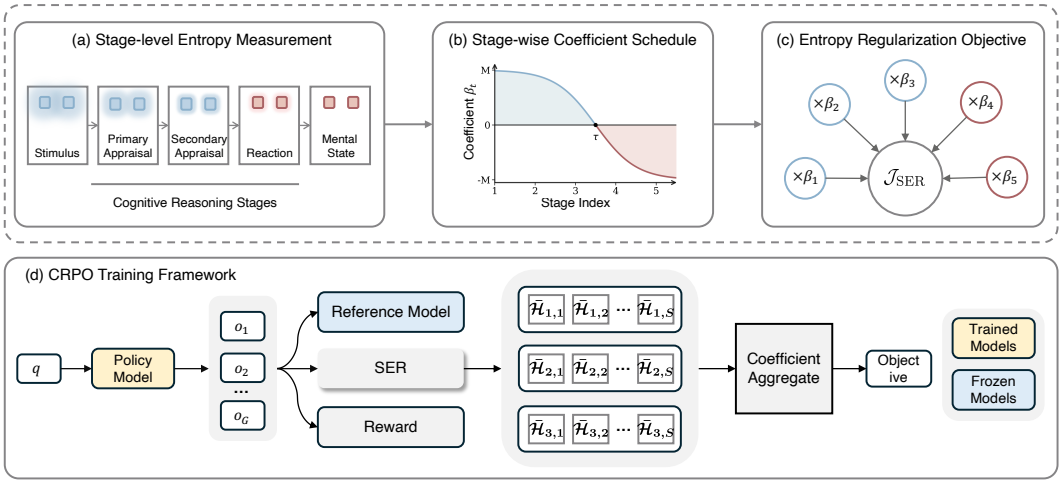

Figure 1: Bottom: Overall illustration of the proposed CRPO reinforcement learning framework. The objective is derived from reward-based advantages and the Stage-wise Entropy Regularization (SER) module, with a reference model used for stability. Top: Detailed illustration of SER. The process begins by measuring stage-level entropy for each reasoning stage. This entropy is then modulated by stage-wise coefficients derived from a hyperbolic tangent schedule. Finally, the stage-level entropy and stage-wise coefficients are combined to form the SER objective.

be incorporated. Formally, for each prompt q, a group of G outputs $\{o_1, o_2, \ldots, o_G\}$ is sampled from the old policy $\pi_{\theta_{old}}$. Each o_i corresponds to one independently generated completion sampled from the same prompt. The training objective to be maximized is defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{k=1}^{|o_i|}\Big[\min\big(\rho_{i,k}(\theta)A_i, \text{clip}(\rho_{i,k}(\theta), 1-\epsilon, 1+\epsilon)A_i\big) - \lambda D_{KL}(\pi_\theta \| \pi_{\text{ref}})\Big]. \tag{1}$$

Here, $\rho_{i,k}(\theta)$ is the current-to-old policy probability ratio for token $o_{i,k}$ conditioned on $(q, o_{i,<k})$, and A_i is the reward of completion o_i standardized by the group reward mean and standard deviation. The KL divergence is evaluated at the same token context against a fixed reference policy π_{ref}. The coefficients ϵ and λ control clipping and KL regularization, respectively.

3.2 Stage-wise Entropy Regularization

While GRPO provides a stable optimization foundation, it treats all reasoning stages uniformly and therefore does not align with the uncertainty evolution of human multi-stage reasoning. Cognitive theory suggests that certainty rarely remains constant across stages; instead, reasoning typically begins with exploratory uncertainty and gradually develops into confident conclusions (Gold & Shadlen, 2007; Clark, 2013). To address this limitation, we propose Stage-wise Entropy Regularization (SER), which modulates the model's uncertainty across the reasoning stages. The guiding idea is intuitive: allow high uncertainty in early reasoning stages, and gradually encourage lower uncertainty and greater confidence in later stages, where clarity and decisiveness are required.

Stage-level entropy. SER modulates uncertainty across reasoning stages to mimic the human uncertainty evolution process. To this end, we quantify the policy's uncertainty within each reasoning stage using a stage-level entropy. For a generated output o_i and reasoning stage t, the stage-level entropy is defined as the average entropy over all tokens belonging to that stage:

$$\bar{\mathcal{H}}_{i,t}(\theta) = \frac{1}{|\mathcal{N}_{i,t}|}\sum_{k\in\mathcal{N}_{i,t}}\mathcal{H}_{i,k}(\theta), \tag{2}$$

where i indexes a generated output in the sampled group, $\mathcal{N}_{i,t}$ is the set of token positions in o_i that belong to stage t, and $|\mathcal{N}_{i,t}|$ is the number of tokens assigned to that stage. Here, $\mathcal{H}_{i,k}(\theta)$ denotes

the token entropy at position k:

$$\mathcal{H}_{i,k}(\theta) = -\sum_{w \in V} \pi_\theta(w \mid q, o_{i,<k}) \log \pi_\theta(w \mid q, o_{i,<k}), \tag{3}$$

where w is a token in the vocabulary V, q denotes the input prompt, and $o_{i,<k}$ is the partial sequence of o_i generated before position k.

The stage-level entropy characterizes the policy's uncertainty within a reasoning stage, with higher values indicating more exploratory reasoning and lower values indicating more confident reasoning. Normalizing by $|\mathcal{N}_{i,t}|$ removes the influence of variable stage lengths, so that the contribution of each stage to the regularization term is governed by its stage-wise coefficient rather than by how many tokens it contains.

Based on the stage-level entropy, the overall entropy regularization term is defined as a weighted sum over all reasoning stages and all generated outputs:

$$\mathcal{J}_{\text{SER}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{S} \beta_t \, \bar{\mathcal{H}}_{i,t}(\theta), \tag{4}$$

where G denotes the number of generated outputs sampled for each prompt, S is the total number of reasoning stages, and β_t is a stage-specific coefficient that controls both the magnitude and direction of entropy regularization at stage t. The outer average over G ensures that the regularization term is applied consistently across all sampled completions, while the inner summation aggregates the contributions from different reasoning stages. This formulation aligns the uncertainty regularization process with the group-based optimization paradigm of GRPO, treating each generated completion as an independent carrier of stage-wise uncertainty information.

Stage-wise coefficient schedule. To achieve the intended uncertainty-to-certainty progression, we set β_t according to a hyperbolic tangent schedule:

$$\beta_t = -M \cdot \frac{e^{t-\tau} - e^{-(t-\tau)}}{e^{t-\tau} + e^{-(t-\tau)}}, \tag{5}$$

where M controls the maximum magnitude of the entropy regularization term, t denotes the reasoning stages index in the cognitive reasoning stages, and τ specifies the transition point where β_t changes sign, i.e., the stage where the regularization term shifts from positive values to negative values. This design provides a smooth progression:

- Early stages ($t < \tau$): receive positive β_t to increase entropy, thereby encouraging diverse exploration.
- Later stages ($t > \tau$): receive negative β_t to suppress entropy, thereby enforcing more deterministic reasoning.

Overall objective. The overall training objective to be maximized is a combination of group relative policy optimization objective and the stage-wise entropy regularization objective:

$$\begin{aligned}\mathcal{J}_{\text{CRPO}}(\theta) =& \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{k=1}^{|o_i|} \Big[\min \big(\rho_{i,k}(\theta) A_i, \text{clip}(\rho_{i,k}(\theta), 1-\epsilon, 1+\epsilon) A_i \big) - \lambda D_{KL}(\pi_\theta \| \pi_{\text{ref}}) \Big] \\ &+ \frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{S} \beta_t \left(\frac{1}{|N_{i,t}|} \sum_{k \in N_{i,t}} \Big[- \sum_{w \in V} \pi_\theta(w \mid q, o_{i,<k}) \log \pi_\theta(w \mid q, o_{i,<k}) \Big] \right).\end{aligned} \tag{6}$$

We summarize the overall training procedure in Algorithm 1.

3.3 Cognitive Reasoning Stages

SER operates on a multi-stage reasoning trajectory and therefore requires an explicit decomposition of reasoning into different stages. To this end, CRPO imposes a structured set of Cognitive Reasoning Stages inspired by cognitive appraisal theory (Lazarus & Folkman, 1984; Ellsworth, 1991;

Watson & Spence, 2007), a foundational framework in psychology that explains how individuals construct mental states through subjective evaluations of potential factors. The central principle of this theory is that mental states emerge through staged cognitive appraisals, rather than directly from stimuli. When confronted with a stimulus, a person engages in a sequence of appraisal processes: first, assessing the nature and significance of the event, and second, evaluating whether sufficient resources are available to cope with it. These appraisals in turn shape the individual's affective and behavioral responses. An illustrative example is provided in Appendix A. Building on this theory, we design a structured set of cognitive reasoning stages:

- Stimulus stage: Identify the root stimulus, i.e., the event, situation, or object explicitly or implicitly described in the input text.
- Primary appraisal stage: Infer the individual's initial evaluation of the stimulus, such as whether it is threatening, positive, or irrelevant.
- Secondary appraisal stage: Determine whether the individual perceives sufficient resources or coping ability to handle the situation.
- Reaction stage: Summarize the likely affective and behavioral responses elicited by the appraisals.
- Mental state stage: Conclude the probable mental health condition or state (e.g., stress, depression, or anxiety) resulting from the preceding reasoning process.

We denote the stages as $c = \langle c_t \rangle_{t=1}^{S}$ with $S = 5$. To ensure this reasoning structure in LLM outputs, we design a system prompt and a format reward integrated into reinforcement learning.

3.4 Reward Design

The format reward encourages the output to follow the cognitive reasoning stages. We assign $r_f(o) = +\alpha$ if the following conditions are met: 1) the output contains the outer tags '<think>' and '<answer>' exactly once each, and in the correct order; 2) within '<think>', the stage tags of c appear exactly once each, in the predefined order, and with non-empty content; and 3) the content within the '<answer>' section is non-empty. Otherwise, $r_f(o) = -\alpha$.

The answer reward encourages the final prediction to match the ground truth. We assign $r_a(o, y) = +1$ if the predicted answer matches the ground truth y, and -1 otherwise.

Combined reward. For a completion o associated with dataset d and ground-truth label y, the total reward is defined as a weighted sum of the two rewards:

$$r(o, d, y) = r_f(o) + \eta_{d,y}\, r_a(o, y), \tag{7}$$

where $\eta_{d,y} > 0$ controls the contribution of answer correctness. We set $\eta_{d,y} = \sqrt{w_{d,y} w_d}$, where

$$w_{d,y} = \frac{1/n_{d,y}}{\frac{1}{C_d}\sum_{j=1}^{C_d}(1/n_{d,j})}, \qquad w_d = \frac{1/n_d}{\frac{1}{D}\sum_{m=1}^{D}(1/n_m)}. \tag{8}$$

Here, $n_{d,y}$ is the number of training samples with label y in dataset d, n_d is the training size of dataset d, C_d is its number of classes, and D is the number of datasets. These coefficients are computed from training-set statistics and held fixed, with the square root smoothing extreme values. Note that $\eta_{d,y}$ matters only when both format and answer rewards vary within a group, where correct answers outrank format compliance iff $\eta_{d,y} > \alpha$; this prioritizes answer correctness for rarer classes and smaller datasets in early training. Otherwise, being constant within a group, it cancels under group-wise advantage normalization and becomes inactive once format compliance stabilizes.

4 Experiment

We evaluate our method through the following questions. Details are included in Appendix C.

Q1: How does CRPO compare with other RL methods, and how effective are its key components?

Q2: How does CRPO-trained Mental-R1 compare with other LLMs?

Q3: How does CRPO improve the model's reasoning ability?

Q4: How does the SER affect entropy and its schedule affect performance?

Table 2: Comparison with different critic-free RL methods and ablation results (Weighted F1-score, mean $\pm$ std, in %). Our CRPO outperforms SFT and RL training baselines across all eight datasets.

Method	DATD Anxiety D1	RSD Suicide D2	DepSeverity Depression D3	LT-EDI Depression D4	SDCNL Suicide D5	Dreaddit Stress D6	FIG Loneliness D7	LID Loneliness D8
Optimize methods								
SFT	$57.13_{\pm 0.76}$	$30.50_{\pm 1.14}$	$59.41_{\pm 0.68}$	$44.89_{\pm 1.46}$	$48.67_{\pm 3.92}$	$73.46_{\pm 1.44}$	$80.61_{\pm 1.64}$	$18.06_{\pm 1.77}$
RLOO	$57.06_{\pm 2.03}$	$22.34_{\pm 2.04}$	$63.67_{\pm 2.08}$	$38.77_{\pm 1.68}$	$50.38_{\pm 1.22}$	$76.34_{\pm 1.05}$	$86.26_{\pm 0.58}$	$17.24_{\pm 1.41}$
ReMax	$57.42_{\pm 1.16}$	$32.19_{\pm 2.83}$	$74.73_{\pm 0.21}$	$49.24_{\pm 0.98}$	$46.07_{\pm 0.79}$	$62.36_{\pm 0.57}$	$85.77_{\pm 1.06}$	$17.85_{\pm 3.65}$
Reinforce++	$49.42_{\pm 2.13}$	$21.81_{\pm 1.68}$	$75.95_{\pm 0.18}$	$47.61_{\pm 0.85}$	$49.94_{\pm 0.71}$	$66.06_{\pm 0.75}$	$86.01_{\pm 1.20}$	$18.68_{\pm 4.36}$
DAPO	$58.26_{\pm 1.29}$	$26.19_{\pm 3.34}$	$72.28_{\pm 2.92}$	$49.69_{\pm 0.91}$	$52.26_{\pm 0.86}$	$73.78_{\pm 1.13}$	$84.62_{\pm 1.03}$	$21.66_{\pm 1.17}$
GRPO	$54.26_{\pm 1.68}$	$28.34_{\pm 3.69}$	$75.88_{\pm 1.16}$	$47.40_{\pm 1.00}$	$53.05_{\pm 1.42}$	$72.60_{\pm 0.88}$	$84.91_{\pm 1.01}$	$19.37_{\pm 1.84}$
Ablation								
w/o SER	$56.31_{\pm 2.60}$	$42.05_{\pm 2.54}$	$76.78_{\pm 0.46}$	$53.97_{\pm 1.76}$	$56.93_{\pm 1.20}$	$75.39_{\pm 1.07}$	$87.46_{\pm 0.52}$	$29.42_{\pm 1.36}$
w/o CRS	$59.67_{\pm 1.57}$	$46.78_{\pm 3.30}$	$77.65_{\pm 1.91}$	$54.10_{\pm 1.67}$	$60.27_{\pm 1.18}$	$75.98_{\pm 1.28}$	$88.32_{\pm 1.53}$	$34.38_{\pm 4.27}$
CRPO	$\mathbf{63.62_{\pm 1.03}}$	$\mathbf{50.14_{\pm 4.72}}$	$\mathbf{78.86_{\pm 1.16}}$	$\mathbf{57.14_{\pm 0.97}}$	$\mathbf{61.75_{\pm 1.40}}$	$\mathbf{80.56_{\pm 0.57}}$	$\mathbf{91.57_{\pm 0.61}}$	$\mathbf{38.39_{\pm 1.66}}$

Table 3: Comparison with LLMs (Weighted F1-score, mean $\pm$ std, in %). CRPO-trained Mental-R1 outperforms all the baseline LLMs across eight mental health assessment datasets.

Model	DATD Anxiety D1	RSD Suicide D2	DepSeverity Depression D3	LT-EDI Depression D4	SDCNL Suicide D5	Dreaddit Stress D6	FIG Loneliness D7	LID Loneliness D8
Domain and public models								
Mentallama	$53.78_{\pm 1.79}$	$22.88_{\pm 5.33}$	$41.95_{\pm 1.54}$	$21.49_{\pm 1.05}$	$48.50_{\pm 2.12}$	$75.33_{\pm 1.03}$	$81.30_{\pm 1.01}$	$14.28_{\pm 1.49}$
Mental-GLM	$59.89_{\pm 1.28}$	$37.51_{\pm 2.45}$	$53.63_{\pm 0.33}$	$30.36_{\pm 1.30}$	$55.79_{\pm 1.26}$	$76.22_{\pm 0.36}$	$80.40_{\pm 0.50}$	$18.88_{\pm 1.30}$
Gemma-2-SFT	$54.33_{\pm 3.33}$	$29.48_{\pm 3.08}$	$56.01_{\pm 0.68}$	$40.23_{\pm 1.69}$	$49.82_{\pm 0.77}$	$73.08_{\pm 0.43}$	$83.58_{\pm 0.53}$	$20.33_{\pm 0.83}$
Llama-3.1-SFT	$50.78_{\pm 3.24}$	$26.52_{\pm 2.89}$	$63.67_{\pm 1.43}$	$39.92_{\pm 1.66}$	$45.17_{\pm 0.82}$	$70.90_{\pm 0.68}$	$77.46_{\pm 0.77}$	$15.82_{\pm 1.95}$
Industry-level models								
DeepSeek-V3	$52.21_{\pm 1.82}$	$41.99_{\pm 3.57}$	$49.14_{\pm 0.69}$	$35.60_{\pm 1.48}$	$48.53_{\pm 1.04}$	$71.61_{\pm 1.14}$	$78.23_{\pm 1.34}$	$11.84_{\pm 1.51}$
DeepSeek-R1	$56.91_{\pm 2.43}$	$43.16_{\pm 3.98}$	$62.57_{\pm 0.55}$	$26.40_{\pm 1.22}$	$52.75_{\pm 0.98}$	$72.97_{\pm 1.22}$	$84.56_{\pm 1.48}$	$11.63_{\pm 0.93}$
GPT-3.5	$52.78_{\pm 2.91}$	$29.87_{\pm 3.83}$	$50.47_{\pm 1.52}$	$22.36_{\pm 2.45}$	$42.33_{\pm 1.32}$	$68.46_{\pm 1.06}$	$75.41_{\pm 1.69}$	$20.99_{\pm 0.56}$
GPT-4o	$46.88_{\pm 2.31}$	$35.69_{\pm 3.70}$	$52.74_{\pm 0.97}$	$33.23_{\pm 2.16}$	$51.52_{\pm 1.18}$	$71.38_{\pm 0.76}$	$74.51_{\pm 0.99}$	$11.26_{\pm 1.38}$
GPT-5	$56.98_{\pm 1.28}$	$43.08_{\pm 1.65}$	$62.19_{\pm 0.79}$	$36.18_{\pm 0.84}$	$51.18_{\pm 1.73}$	$72.86_{\pm 1.84}$	$81.38_{\pm 0.96}$	$22.68_{\pm 0.85}$
Mental-R1	$\mathbf{63.62_{\pm 1.03}}$	$\mathbf{50.14_{\pm 4.72}}$	$\mathbf{78.86_{\pm 1.16}}$	$\mathbf{57.14_{\pm 0.97}}$	$\mathbf{61.75_{\pm 1.40}}$	$\mathbf{80.56_{\pm 0.57}}$	$\mathbf{91.57_{\pm 0.61}}$	$\mathbf{38.39_{\pm 1.66}}$

4.1 DATASETS

We adopt 8 different human-annotated datasets to comprehensively evaluate our method, each corresponding to a unique mental health assessment task, as summarized in Table 4. Stress state prediction: Dreaddit (Turcan & McKeown, 2019). Anxiety state prediction: DATD (Owen et al., 2020). Depression level classification: LT-EDI (Sampath & Durairaj, 2022). Depression severity prediction: DepSeverity (Naseem et al., 2022). Suicidal ideation detection: SDCNL (Haque et al., 2021). Suicide risk severity classification: RSD (Zheng et al., 2025). Loneliness state prediction: FIG (Jiang et al., 2022). Loneliness intensity prediction: LID (Katsman et al., 2025). To the best of our knowledge, this evaluation includes the most extensive collection of publicly accessible human-annotated mental health assessment datasets compared with prior work, offering a more comprehensive and transparent benchmark for future research.

4.2 Q1: COMPARISON WITH OTHER RL METHODS

We compare CRPO with supervised fine-tuning and a diverse set of reinforcement learning baselines, which include critic-free methods RLOO (Ahmadian et al., 2024), ReMax (Li et al., 2024), Reinforce++ (Hu et al., 2025d), GRPO (Guo et al., 2025), and DAPO (Yu et al., 2025). As reported in the upper portion of Table 2, CRPO attains the highest performance across all eight datasets. Averaged across datasets, CRPO outperforms the strongest reinforcement learning baseline DAPO by 10.4 points in Weighted F1, and improves over SFT by 13.6 points, respectively. The gains are especially pronounced on the RSD and LID datasets. These results substantiate the effectiveness and generality of our CRPO framework for mental-health–oriented reasoning tasks.

Ablation results are presented in the lower portion of Table 2. Removing SER leads to the largest average performance drop, confirming the critical role of SER in aligning the reasoning process with human cognitive dynamics, thereby improving overall model performance. For the w/o Cognitive Reasoning Stages (CRS) variant, we remove the five cognitive stages to allow free-form reasoning. The reasoning content is divided into $S = 5$ contiguous segments of approximately equal length, and SER is applied to these segments with the same schedule. The resulting performance decline supports the benefit of explicit cognitive stage structure over free-form reasoning.

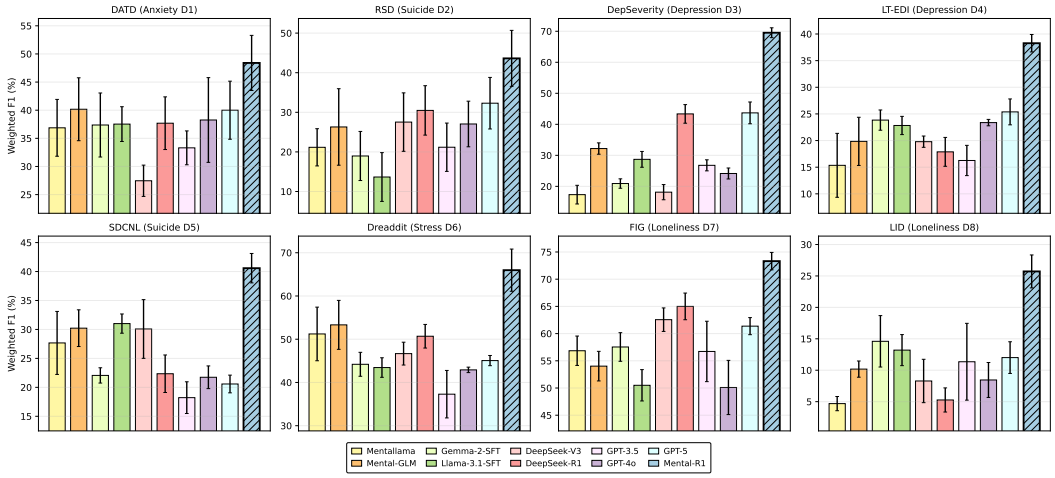

Figure 2: Reasoning-focused evaluation results (Weighted F1-score, mean ± std, in %). CRPO-trained Mental-R1 consistently outperforms all LLM baselines on reasoning-intensive samples.

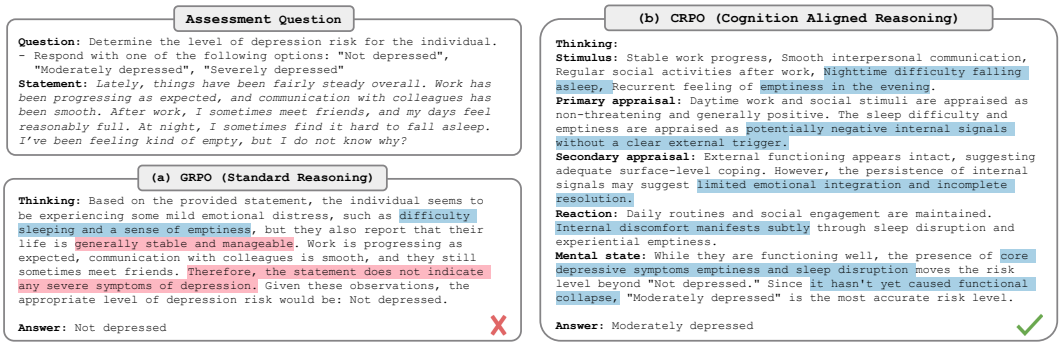

Figure 3: Comparison of reasoning outputs. (a) Standard reasoning produced by a GRPO-trained model. (b) Cognition-aligned reasoning generated by CRPO-trained Mental-R1, which correctly predicts the annotated depression risk.

4.3 Q2: Comparison with LLMs

We conduct a comprehensive comparison between CRPO-trained Mental-R1 and three groups of LLM baselines: mental health–specific large language models such as Mentallama-13B (Yang et al., 2024) and Mental-GLM (Zhai et al., 2025), open-source LLMs including Gemma-2-9B (Team et al., 2024) and Llama-3.1-8B (Dubey et al., 2024) that are supervised fine-tuned on the same training data for fairness, and large proprietary models such as DeepSeek-V3 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025), GPT-3.5 (OpenAI, 2023), GPT-4o (Openai, 2024), and GPT-5 (OpenAI, 2025). The results in Table 3 show that Mental-R1 consistently achieves best performance across all eight datasets. On average, Mental-R1 outperforms the best mental health–specific baseline by 13.6 points, open-source baseline by 14.4 points, and proprietary baseline by 11.9 points in weighted-F1. Mental-R1 delivers particularly strong gains on the LT-EDI depression detection dataset. These results demonstrate that our method enhances the model's ability for mental health assessment.

4.4 Q3: Reasoning Ability

We use GPT-5 as an external evaluator to score the reasoning demand of each test sample and retain the top 20% within each of the eight datasets. This evaluation focuses on samples judged to require more complex reasoning. As shown in Figure 2, Mental-R1 consistently outperforms all baselines by a substantial margin, achieving an average improvement of approximately 15.6 weighted F1 points

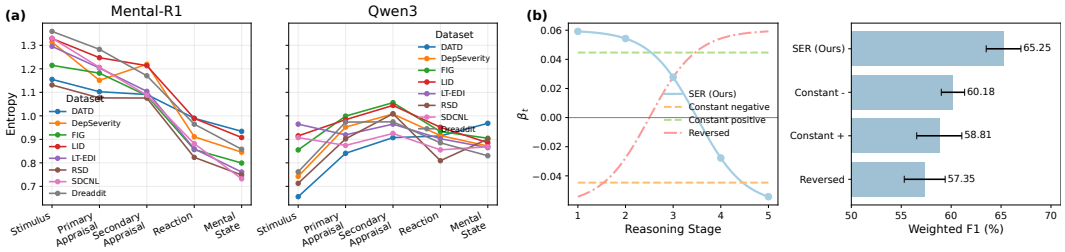

Figure 4: Analysis of SER. (a) Comparison of reasoning entropy across stages. We plot the average entropy for each reasoning stage across all test sets. Compared with vanilla Qwen3, CRPO-trained Mental-R1 shows a clear downward trend. (b) Visualization of uncertainty scheduling strategies and their corresponding performance. Left: different scheduling patterns across reasoning stages. Right: average weighted F1-score with standard deviation. The proposed SER follows a smooth transition from exploration to confidence and achieves the best performance.

over the strongest baseline, with particularly large gains on DepSeverity and LID. These results support the effectiveness of CRPO on reasoning-intensive mental health assessment samples.

Figure 3 illustrates how CRPO reshapes reasoning compared with standard GRPO. The input describes stable external functioning alongside subtle internal cues, including sleep disturbance and emptiness. GRPO places more emphasis on positive external functioning and predicts the risk "Not depressed." In contrast, CRPO initially considers both external functioning and internal signals, then progressively focuses on the internal cues. By distinguishing preserved external functioning from unresolved internal distress, it arrives at the risk prediction "Moderately depressed." This example illustrates how the cognitive stages organize competing cues and is consistent with the intended reasoning progression: considering multiple interpretations early and consolidating relevant evidence before reaching a final assessment.

4.5 Q4: Entropy Dynamics and Scheduling in SER

To examine whether SER produces the intended uncertainty dynamics, we compare the average token entropy at each reasoning stage for CRPO-trained Mental-R1 and the vanilla Qwen3 base model across all eight test sets. As shown in Figure 4(a), the entropy profile of Mental-R1 exhibits a gradual decline as the reasoning progresses, whereas the vanilla model shows relatively flat distribution throughout. This pattern is consistent with SER's intended progression from greater uncertainty in early stages toward more confident predictions in later stages.

We further compare our schedule with constant positive, constant negative, and reversed schedules (Figure 4(b)). Constant positive and negative coefficients encourage higher and lower entropy throughout reasoning, respectively. For a controlled comparison, their magnitudes equal the mean absolute value of β_t, while the reversed schedule uses the same coefficients in reverse order. Our schedule achieves the highest average weighted F1. Its advantage over constant schedules supports varying entropy regularization across stages. The performance decline under the reversed schedule further suggests that the direction of this variation matters, supporting the design of encouraging exploration early and reducing uncertainty as reasoning converges.

5 Conclusion

In this paper, we propose Cognitive Relative Policy Optimization (CRPO) to align large language model reasoning with real-world mental health assessment practice. CRPO combines stage-wise entropy regularization with theory-grounded reasoning stages to encourage early-stage exploration and late-stage certainty, mirroring the human cognitive progression from uncertainty to certainty. Across eight benchmark datasets, CRPO consistently outperforms multiple reinforcement learning baselines, and its trained model, Mental-R1, further surpasses strong LLM baselines, especially on reasoning-intensive samples. More broadly, this work suggests that incorporating cognition-aware uncertainty optimization can substantially improve model performance, offering a promising direction for developing human-aligned large reasoning models.

AI USE STATEMENT

In this work, we used generative AI to score the reasoning demand of test samples in order to select the evaluation subset (Section 4.4). We have not used generative AI tools for any other tasks with required disclosure under the ICLR 2027 policy. Additionally, we used generative AI tools to edit the manuscript for readability and clarity. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12248–12267. Association for Computational Linguistics, 2024.

Aaron T Beck. *Cognitive therapy and the emotional disorders*. Penguin, 1979.

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

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1718–1728, 2019.

Lei Cao, Huijun Zhang, Xin Wang, and Ling Feng. Learning users inner thoughts and emotion changes for social media based suicide risk detection. *IEEE Transactions on Affective Computing*, 14(2):1280–1296, 2021.

Sully F Chen, Anton Alyakin, Andreas Seas, Eunice Yang, Joanne J Choi, Jin Vivian Lee, Amelia L Chen, Pranav I Warman, Rochelle T Bitolas, Robert J Steele, et al. Llm-assisted systematic review of large language models in clinical medicine. *Nature medicine*, pp. 1–8, 2026.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Tracy D Eells. *Handbook of psychotherapy case formulation*. Guilford Publications, 2022.

Albert Ellis. Reason and emotion in psychotherapy. 1962.

Phoebe C Ellsworth. Some implications of cognitive appraisal theories of emotion. 1991.

Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, 1978.

Hadar Fisher, Nigel M Jaffe, Kristina Pidvirny, Anna O Tierney, Mia S Vaidean, Poorvesh Dongre, and Christian A Webb. Language-based detection of depression with machine learning: systematic review and meta-analysis. *npj Digital Medicine*, 2026.

Boyan Gao, Xin Wang, Yibo Yang, and David A. Clifton. Optimization inspired few-shot adaptation for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rZ2nSt1X58.

Howard N Garb. *Studying the clinician*. American Psychological Association (APA), 1998.

JI Gold and MN Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30: 535–574, 2007.

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

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction. In *International Conference on Artificial Neural Networks*, pp. 436–447. Springer, 2021.

Rohmad Hidayat, Kemas M Lhaksmana, and Iis Kurnia Nurhayati. Language anxiety detection in english texts using bert and svm. In *2025 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 357–362. IEEE, 2025.

Joy Higgs, Gail M Jensen, Stephen Loftus, Franziska V Trede, and Sandra Grace. *Clinical Reasoning in the Health Professions E-Book: Clinical Reasoning in the Health Professions E-Book*. Elsevier Health Sciences, 2024.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*, 2025a.

He Hu, Yucheng Zhou, Qianning Wang, Yingjian Zou, Chiyuan Ma, Juzheng Si, Jianzhuang Liu, Zitong Yu, Laizhong Cui, and Fei Ma. From pattern recognizers to personalized companions: A survey of large language models in mental health. 2025b.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025c.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*, 2025d.

Catherine Ikae, Souhir Ben Souissi, Jannic Stefan Bieri, Thomas Jörg Müller, Margarithe Charlotte Feuz-Schlunegger, and Christoph Golz. A scoping review of natural language processing for detecting work-related stress among health professionals. *Discover Computing*, 29(1):14, 2026.

Yueyi Jiang, Yunfan Jiang, Liu Leqi, and Piotr Winkielman. Many ways to be lonely: fine-grained characterization of loneliness and its potential changes in covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pp. 405–416, 2022.

Yu Jin, Jiayi Liu, Pan Li, Baosen Wang, Yangxinyu Yan, Huilin Zhang, Chenhao Ni, Jing Wang, Yi Li, Yajun Bu, et al. The applications of large language models in mental health: scoping review. *Journal of Medical Internet Research*, 27(1):e69284, 2025.

Yael Katsman, Hilly Segal, and Yarden Kamienney. Loneliness-intensity. `https://huggingface.co/datasets/yael-katsman/Loneliness-Causes-and-Intensity`, 2025.

Erol KINA, Jin-Ghoo Choi, Abid Ishaq, Rahman Shafique, Monica Gracia Villar, Eduardo Silva Alvarado, Isabel de la Torre Diez, and Imran Ashraf. Suicide ideation detection using social media data and ensemble machine learning model. *International Journal of Computational Intelligence Systems*, 2026.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

Willem Kuyken, Christine A Padesky, and Robert Dudley. *Collaborative case conceptualization: Working effectively with clients in cognitive-behavioral therapy*. Guilford Press, 2011.

Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Mehul Nanda, Diana Inkpen, and Aroldo Dargel. Detecting multiple mental health disorders with large language models. In *2024 28th International Conference Information Visualisation (IV)*, pp. 252–257. IEEE, 2024.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM web conference 2022*, pp. 2563–2572, 2022.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025.

OpenAI. Gpt3.5 turbo fine-tuning and api updates. `https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/`, 2023.

Openai. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

OpenAI. Gpt-5 is here. `https://openai.com/gpt-5/`, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. Towards preemptive detection of depression and anxiety in twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pp. 82–89, 2020.

Jacqueline B Persons. *The case formulation approach to cognitive-behavior therapy*. Guilford Press, 2012.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. Are llms effective psychological assessors? leveraging adaptive rag for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8975–8991, 2025.

Muhammad Sadiq Rohei, Kasturi Dewi Varathan, Shivakumara Palaiahnakote, and Nor Badrul Anuar. Review of predictive techniques for detecting mental disorders from user-generated content on social media. *PeerJ Computer Science*, 12:e3559, 2026.

Kayalvizhi Sampath and Thenmozhi Durairaj. Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science*, pp. 136–151. Springer, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiayu Shi, Zexiao Wang, Jiandong Zhou, Chengyu Liu, Poly ZH Sun, Erying Zhao, and Lei Lu. Mentalqlm: A lightweight large language model for mental healthcare based on instruction tuning and dual lora modules. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Elsbeth Turcan and Kathleen McKeown. Dreaddit: A reddit dataset for stress analysis in social media. *EMNLP-IJCNLP 2019*, pp. 97, 2019.

Ning Wang, Sanchit Goel, Stephanie Ibrahim, Varsha D Badal, Colin Depp, Erhan Bilal, Koduvayur Subbalakshmi, and Ellen Lee. Decoding loneliness: Can explainable ai help in understanding language differences in lonely older adults? *Psychiatry research*, 339:116078, 2024.

Xin Wang, Huijun Zhang, Lei Cao, and Ling Feng. Leverage social media for personalized stress detection. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2710–2718, 2020.

Xin Wang, Lei Cao, Huijun Zhang, Ling Feng, Yang Ding, and Ningyun Li. A meta-learning based stress category detection framework on social media. In *Proceedings of the ACM Web Conference 2022*, pp. 2925–2935, 2022.

Xin Wang, Huijun Zhang, Lei Cao, Kaisheng Zeng, Qi Li, Ningyun Li, and Ling Feng. Contrastive learning of stress-specific word embedding for social media based stress detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5137–5149, 2023.

Xin Wang, Ling Feng, Huijun Zhang, Lei Cao, Kaisheng Zeng, Qi Li, Yang Ding, Yi Dai, and David Clifton. Mise: Meta-knowledge inheritance for social media-based stressor estimation. In *Proceedings of the ACM on Web Conference 2025*, pp. 1866–1876, 2025.

Lisa Watson and Mark T Spence. Causes and consequences of emotions on consumer behaviour: A review and integrative cognitive appraisal theory. *European Journal of marketing*, 41(5/6): 487–511, 2007.

World Health Organization. Suicide. `https://www.who.int/news-room/fact-sheets/detail/suicide`, 2023. Accessed: 2025-09-09.

Jesse H Wright, Gregory K Brown, Michael E Thase, and Monica Ramirez Basco. *Learning cognitive-behavior therapy: An illustrated guide*. American Psychiatric Pub, 2017.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pp. 4489–4500, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yu, Qi Li, and Xiaoqian Liu. Automatic anxiety recognition method based on microblog text analysis. *Frontiers in Public Health*, 11:1080013, 2023.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

Eka Putri Yulianti, Yossie Susanti Eka Putri, Budi Anna Keliat, and Achmad Nizar Hidayanto. Feelings behind words: A systematic review on how effective is nlp-based assessment for mental health diagnosis in human studies. *International Journal of Medical Informatics*, pp. 106129, 2025.

Kaiwen Zha, Zhengqi Gao, Maohao Shen, Zhang-Wei Hong, Duane Boning, and Dina Katabi. Rl tango: Reinforcing generator and verifier together for language reasoning. *Advances in Neural Information Processing Systems*, 38:119283–119313, 2026.

Wei Zhai, Nan Bai, Qing Zhao, Jianqiang Li, Fan Wang, Hongzhi Qi, Meng Jiang, Xiaoqin Wang, Bing Xiang Yang, and Guanghui Fu. Mentalglm series: Explainable large language models for mental health analysis on chinese social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13599–13614, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:59849–59887, 2026.

Liang Zhao, Bokai Chen, Weiwei Zheng, Lingyan Zhou, and Xiaojun Ding. Cyberconfucius: An interactive confucian philosophical counseling system for mental health. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 343–347, 2025.

Shouwen Zheng, Yingzhi Tao, and Taiqi Zhou. Rsd-15k: A large-scale user-level annotated dataset for suicide risk detection on social media. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, pp. 190–196. IEEE, 2025.

Algorithm 1: Cognitive Relative Policy Optimization (CRPO)

Input: Training set $\mathcal{D}$; current policy π_θ; sampling policy $\pi_{\theta_{\text{old}}}$; reference policy π_{ref}; group size G; cognitive stages $\{c_t\}_{t=1}^{S}$; entropy schedule parameters M and τ.
Output: Optimized policy π_θ.
for *each training iteration* **do**
  Sample a mini-batch of prompts $\{q\}$ from $\mathcal{D}$;
  for *each prompt* q **do**
    Sample a group of outputs $\{o_1, \ldots, o_G\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$;
    for *each output* o_i **do**
      Compute reward $r_i = r_f(o_i) + \eta_{d,y} r_a(o_i, y)$ using format and answer rewards;
      Segment o_i into cognitive stages $\{c_t\}_{t=1}^{S}$ via stage tags;
      for *each stage* t **do**
        Identify token indices $N_{i,t}$ corresponding to stage c_t;
        Compute token entropy $\mathcal{H}_{i,k}(\theta)$ for each $k \in N_{i,t}$;
        Compute stage-level entropy $\bar{\mathcal{H}}_{i,t}(\theta) = \frac{1}{|N_{i,t}|} \sum_{k \in N_{i,t}} \mathcal{H}_{i,k}(\theta)$;
    Compute advantages $A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}$;
  Compute stage-wise entropy coefficients $\beta_t = -M \cdot \frac{e^{t-\tau} - e^{-(t-\tau)}}{e^{t-\tau} + e^{-(t-\tau)}}, \quad t = 1, \ldots, S$;
  Compute entropy regularization $\mathcal{J}_{\text{SER}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{S} \beta_t \bar{\mathcal{H}}_{i,t}(\theta)$;
  Compute the objective $\mathcal{J}_{\text{GRPO}}(\theta)$ using $\rho_{i,k}(\theta)$ and A_i;
  Update θ by maximizing $\mathcal{J}_{\text{CRPO}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) + \mathcal{J}_{\text{SER}}(\theta)$;
return π_θ

A An Illustrative Example of Cognitive Appraisal Theory

For example, consider public speaking as a stimulus. In the primary appraisal, the individual may perceive it as threatening due to potential embarrassment or negative evaluation. During the secondary appraisal, they evaluate their coping resources and abilities, such as preparation, prior experience, personal skills, or audience support. If these are judged insufficient, the reaction may involve anxiety, fear, or avoidance; if viewed as adequate, the response may instead involve confidence, focus, or even excitement. This appraisal process shows the process of human inner cognition when facing potential triggers in daily life.

B Dataset Details

Stress state prediction involves determining whether an individual is experiencing stress: Dreaddit (Turcan & McKeown, 2019). Anxiety state prediction aims to identify whether a person expresses signs of anxiety or depression: DATD (Owen et al., 2020). Depression level classification assigns individuals to non-depressed, moderately depressed, or severely depressed categories: LT-EDI (Sampath & Durairaj, 2022). Depression severity prediction further refines this by classifying individuals into minimal, mild, moderate, or severe cases: DepSeverity (Naseem et al., 2022). Note that DepSeverity is built upon the same textual corpus as Dreaddit but annotated for a different task. Suicidal ideation detection predicts whether an individual is at risk of suicide: SDCNL (Haque et al., 2021). Suicide risk severity classification assesses individuals across four levels of suicide risk, including indicator, ideation, behavior, and attempt: RSD (Zheng et al., 2025). For RSD, we retain only the first post encountered for each individual in the original dataset, as multiple posts from the same individual may share the same risk label and contain overlapping content. Loneliness state prediction identifies whether an individual is experiencing loneliness: FIG (Jiang et al., 2022). Loneliness intensity prediction categorizes an individual's loneliness into four ordered intervals, representing increasing levels of loneliness severity: LID (Katsman et al., 2025).

C Experiment Details

The learning rate is set to 5×10^{-6}. The SER parameters M and τ are fixed to 0.06 and 3.5, respectively, the KL coefficient is set to 0.01, and the format reward magnitude α is set to 0.5. Training

Table 4: Summary of the eight human-annotated datasets used in this study. The datasets span multiple mental disorders and task formulations, providing a comprehensive benchmark for evaluating performance on mental health assessment.

Dataset	Disorder	Task	#Samples	#Classes	Label Space
Dreaddit	Stress	State prediction	3,553	2	No, Yes
DATD	Anxiety	State prediction	1,050	2	No, Yes
LT-EDI	Depression	Level classification	10,251	3	Not, Moderate, Severe
DepSeverity	Depression	Severity prediction	3,553	4	Minimum, Mild, Moderate, Severe
SDCNL	Suicide	Ideation detection	1,895	2	No, Yes
RSD	Suicide	Severity classification	1,265	4	Indicator, Ideation, Behavior, Attempt
LID	Loneliness	Intensity prediction	498	4	[1–2], [2–3], [3–4], [4–5]
FIG	Loneliness	State prediction	5,633	2	No, Yes

is performed for one epoch with a per-device batch size of 4 and gradient accumulation over 16 steps. We use 4 generations per prompt, with both the maximum prompt length and maximum completion length set to 512 tokens. All experiments are conducted in bf16 precision with DeepSpeed ZeRO-2 optimization. For all RL-based methods, we use Qwen3-8B (Yang et al., 2025) as the base model. The model is jointly trained and validated on the combined training and validation splits of all datasets, and then evaluated separately on each dataset's test split. All experiments are conducted on a Linux server equipped with eight NVIDIA RTX PRO 6000 GPUs. Our framework is implemented using PyTorch, Transformers, and TRL. Following (Yang et al., 2024; Zhai et al., 2025), we evaluate model performance with weighted F1-score. We report the mean and standard deviation over five independent inference.

We follow the standard train/validation/test splits provided in the original datasets. For datasets without predefined splits, we randomly divide the data into training, validation, and test sets with a ratio of 8:1:1. For datasets that only provide training and test sets, we further randomly split 10% of the training data as a validation set. As an exception, we align the partitions of DepSeverity with Dreaddit's splits due to their shared corpus.

C.1 Baseline and Ablation Configurations

The RL baselines, including GRPO and DAPO, use free-form reasoning with a standard `<think>`/`<answer>` format reward and an answer-correctness reward, without SER and cognitive-stage constraints.

The *w/o SER* variant retains the cognitive-stage instructions and reward design of CRPO, while removing the entropy regularization term. The *w/o CRS* variant removes the cognitive-stage instructions and the corresponding stage-specific format checks, while retaining the outer output structure and the answer reward used in full CRPO. Its reasoning tokens within `<think>` are divided into $S = 5$ contiguous segments of approximately equal length, and SER is applied to these segments using the same coefficient schedule as full CRPO, with $M = 0.06$ and $\tau = 3.5$.

D Parameter Study of SER

We analyze the influence of the magnitude parameter M and the transition parameter τ in SER. Specifically, M controls the overall strength of entropy modulation, while τ determines the stage at which the schedule shifts from exploration to certainty. We evaluate a grid with $M \in \{0.04, 0.06, 0.08\}$ and $\tau \in \{2.5, 3.5, 4.5\}$, and report the results in Figure 5. As shown, performance peaks at a moderate magnitude $M = 0.06$ and decreases only slightly when the modulation is either weaker or stronger, indicating that SER is relatively insensitive to M within a moderate range. In contrast, performance is consistently highest when $\tau = 3.5$, compared with earlier transitions at $\tau = 2.5$ and later transitions at $\tau = 4.5$. This suggests that shifting to certainty too early limits necessary exploration, while delaying the transition weakens decisiveness in later stages. The best-performing region therefore corresponds to a mid-

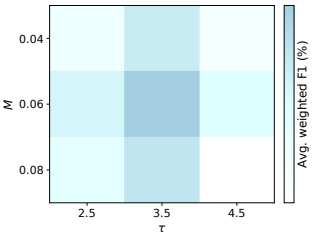

Figure 5: Heatmap of average weighted F1-score under different combinations of the magnitude M and transition parameter τ. Darker color indicates better performance.

stage transition after the second appraisal stage, allowing the model
to maintain exploration throughout the context-gathering and appraisal phases before pivoting to certainty for the final reaction and mental state prediction.

E SYSTEM PROMPT FOR MENTAL-R1

The prompt in last line will be replaced with the specific mental health question during usage.

"A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The Assistant must explicitly think through the reasoning process before giving the final answer.

The reasoning process should strictly follow the stages below:

1. <stimulus> Identify the key event, situation, or object described by the user without interpretation. </stimulus>

2. <primary_appraisal> Assess the personal relevance and potential impact of the stimulus. </primary_appraisal>

3. <secondary_appraisal> Evaluate available resources and options for coping with the situation. </secondary_appraisal>

4. <reaction> Describe the likely affective and behavioral responses. </reaction>

5. <mental_state> Conclude the probable mental health condition or state. </mental_state>

The reasoning process and answer must be enclosed within <think> </think> and <answer> </answer> tags respectively, i.e.:

<think>

<stimulus> stimulus here </stimulus>

<primary_appraisal> primary appraisal here </primary_appraisal>

<secondary_appraisal> secondary appraisal here </secondary_appraisal>

<reaction> reaction here </reaction>

<mental_state> mental state here </mental_state>

</think>

<answer> answer here </answer>

User: prompt. Assistant:"

F PROMPT TEMPLATES FOR DIFFERENT DATASETS

To ensure consistent evaluation across datasets with heterogeneous annotation schemes, we employ dataset-specific prompt templates that align with each task's original label space. All prompts follow a unified instruction–response format.

Binary Classification Tasks. For datasets formulated as binary classification, the model is instructed to respond with *Yes* or *No*.

- DATD (Anxiety/Depression Risk): *"Determine whether the individual who made the following statement is at risk of anxiety or depression. Respond with either 'Yes' or 'No'. Statement:"*
- Dreaddit (Psychological Stress): *"Determine whether the individual who made the following statement is experiencing psychological stress. Respond with either 'Yes' or 'No'. Statement:"*
- FIG (Loneliness Detection): *"Determine whether the individual who made the following statement is experiencing loneliness. Respond with either 'Yes' or 'No'. Statement:"*
- SDCNL (Suicide Risk Detection): *"Determine whether the individual who made the following statement is at risk of suicide. Respond with either 'Yes' or 'No'. Statement:"*

Multi-Class Severity Classification Tasks. For datasets requiring ordinal or categorical severity prediction, the prompt explicitly enumerates all valid label options.

- DepSeverity (Depression Severity): *"Determine the level of depression risk for the individual who made the following statement. Respond with one of the following options: 'Minimum', 'Mild', 'Moderate', 'Severe'. Statement:"*
- LT-EDI (Depression Risk Level): *"Determine the level of depression risk for the individual who made the following statement. Respond with one of the following options: 'Not depressed', 'Moderately depressed', 'Severely depressed'. Statement:"*
- RSD (Suicide Risk Severity): *"Determine the level of suicide-related content for the individual who made the following statement. Respond with one of the following options: 'Indicator', 'Ideation', 'Behavior', 'Attempt'. Statement:"*
- LID (Loneliness Intensity): *"Determine the intensity of loneliness for the individual who made the following statement. Respond with one of the following options: '[1–2]', '[2–3]', '[3–4]', '[4–5]'. Note that 1 indicates not lonely and 5 indicates the highest intensity. Statement:"*

G Prompt for Reasoning-Focused Subset Construction

You are curating items that genuinely REQUIRE reasoning to infer a mental-health label.

Dataset purpose: {*dataset_purpose*}
Label set: {*label_set*}
Text:
<<<{*text*}>>>

Score the degree to which CORRECTLY inferring the gold label requires nontrivial reasoning beyond explicit keywords or self-disclosure.

Guidelines:
- Penalize explicit self-report (e.g., "I have depression", "I'm suicidal") and single-word shortcuts.
- Reward implicit, multi-clue, temporal/causal, contrastive cues and so on (no need to enumerate axes).

Return ONLY JSON:
{
"notes": "reason",
"reasoning_need_score": 0-10,
}