

Does Rationale Quality Matter?

Enhancing Mental Disorder Detection via Selective Reasoning Distillation

Hoyun Song Huije Lee Jisu Shin Sukmin Cho
Changgeon Ko Jong C. Park*

School of Computing

Korea Advanced Institute of Science and Technology (KAIST)

{hysong, huijelee, jisu.shin, nellpic, pencaty, jongpark}@kaist.ac.kr

Abstract

The detection of mental health problems from social media and the interpretation of these results have been extensively explored. Research has shown that incorporating clinical symptom information into a model enhances domain expertise, improving its detection and interpretation performance. While large language models (LLMs) are shown to be effective for generating explanatory rationales in mental health detection, their substantially large parameter size and high computational cost limit their practicality. Reasoning distillation transfers this ability to smaller language models (SLMs), but inconsistencies in the relevance and domain alignment of LLM-generated rationales pose a challenge. This paper investigates how rationale quality impacts SLM performance in mental health detection and explanation generation. We hypothesize that ensuring high-quality and domain-relevant rationales enhances the distillation. To this end, we propose a framework that selects rationales based on their alignment with expert clinical reasoning. Experiments show that our quality-focused approach significantly enhances SLM performance in both mental disorder detection and rationale generation. This work highlights the importance of rationale quality and offers an insightful framework for knowledge transfer in mental health applications. The implementation code and dataset are publicly available¹.

1 Introduction

Detecting mental health issues at an early stage is crucial for initiating timely interventions that can significantly improve treatment outcomes. As online communities continue to grow, researchers in NLP have developed methods to screen user-generated content for signs of depression, anxiety, and other mental illnesses (Jiang et al., 2020, 2021;

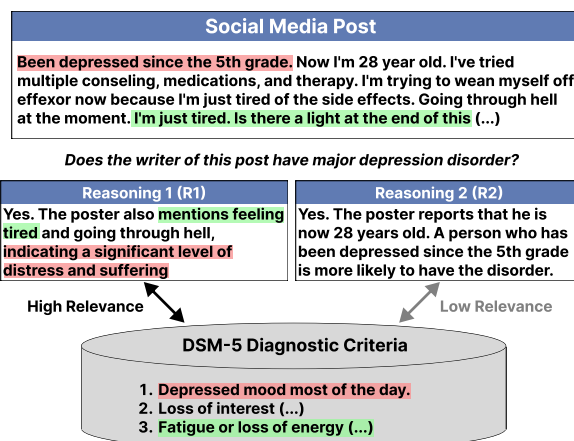


Figure 1: Illustration of varying rationale quality. R1 effectively connects the social media post to specific symptoms in the DSM-5 criteria for major depressive disorder, demonstrating high relevance. R2 lacks these connections, showing low relevance. These examples were generated by GPT-3.5.

Uban et al., 2022; Ji et al., 2022; Aragon et al., 2023). These studies aim to enhance detection accuracy by integrating domain-specific knowledge, such as symptom-related information (Nguyen et al., 2022; Zhang et al., 2022; Song et al., 2023; Zhang et al., 2023).

Some studies focus on interpreting detection results to improve diagnostic accuracy and assist mental health professionals in decision-making with clear reasoning (Song et al., 2018; Harrigan et al., 2020; Ji, 2022; Zanwar et al., 2023; Malhotra and Jindal, 2024). Large language models (LLMs), with their emerging ability to generate explanatory rationales, have demonstrated promising potential in this area through techniques such as Chain-of-Thought (CoT) prompting (Yang et al., 2023; Wang et al., 2024). However, this potential comes with a cost, as such reasoning abilities usually require a substantial amount of parameters, which limits their practicality in resource-constrained settings (Wei et al., 2022).

To address these limitations of smaller language models (SLMs) in conducting CoT reasoning, researchers are exploring distillation methods that

* Corresponding author

¹<https://github.com/HoyunS/acl25-selective-reasoning-distillation>

train these SLMs using teacher-generated rationales (Ho et al., 2023; Hsieh et al., 2023; Magister et al., 2023). Similarly, Yang et al. (2024) proposed a method for mental health condition classification that leverages the efficiency of SLMs. They leveraged LLMs to generate explanatory rationale data for fine-tuning an SLM, allowing it to perform similarly to larger models and provide human-like explanations. Therefore, inspired by these efforts, this study aims to effectively distill the mental health detection ability and rationale generation ability of the teacher model into the student model.

However, despite much advancement in distillation techniques, we found a critical challenge: the inconsistency in the quality of teacher-generated rationales. We observed that, even with identical inputs, LLMs may produce rationales that vary significantly in their relevance to the specific domain. For example, as shown in Figure 1, R1 explicitly references established clinical symptoms for major depressive disorder, reflecting how psychiatric professionals reason about diagnoses. R2, by contrast, is comparatively superficial, lacking references to symptom criteria and failing to incorporate domain-specific insights. Training on such low-quality rationales may hinder the student model’s development of accurate and reliable clinical reasoning, limiting its effectiveness in mental health detection. Consequently, we hypothesize that, by selectively focusing on high-quality rationales, those strongly aligned with domain-specific knowledge may turn out to be more effective for reasoning distillation in the specific domain.

To implement such a selective approach, we need to define what constitutes a high-quality rationale and establish a method for evaluating it. Previous research has employed various criteria to assess rationale quality, including fluency, consistency, reliability, and professionalism (Jeon et al., 2024; Yang et al., 2023, 2024). In this study, we focus on relevance with domain knowledge, a key aspect of professionalism. Specifically, this criterion refers to the extent to which a rationale is explained based on a sufficient understanding of domain knowledge, as shown in Figure 1. Our emphasis on domain relevance stems from its close alignment with the reasoning and diagnostic processes of mental health experts, who utilize established clinical criteria (American Psychiatric Association, 2013) to identify specific mental disorders. By prioritizing rationales that reflect this expert reasoning process,

we aim to enhance the student model’s ability to acquire and apply domain-specific knowledge.

In this paper, we aim to investigate whether selectively distilling high-quality rationales can improve the performance of student models in mental health detection and rationale generation, specifically for major depressive disorder (MDD). To this end, we propose a framework that includes a process for evaluating and selecting high-quality rationales generated by the teacher model, particularly those that show a strong understanding and integration of domain-relevant knowledge. This framework will allow us to assess the impact of prioritizing these rationales on the student model’s ability to learn and effectively apply clinical reasoning. Through the experiments, we demonstrated that our quality-evaluation method aligns well with expert reasoning processes, and that the selective distillation method effectively improves the student model’s performance in both mental health problem detection and rationale generation.

Our contributions are as follows:

- This is the first study to explicitly investigate the impact of rationale quality on student model performance within the critical context of mental health detection, highlighting the importance of incorporating domain-relevant knowledge into the rationale distillation process for improved mental health detection.
- We introduce a framework that includes a process for evaluating and selecting high-quality rationales based on domain-specific knowledge, thereby facilitating effective reasoning distillation for mental health detection.
- Experimental results demonstrate the efficacy of our quality-focused approach, significantly improving student model performance in both mental disorder detection and rationale generation.

2 Related Work

2.1 Reasoning Distillation from LLMs

Numerous studies have explored methods for distilling the CoT reasoning capabilities of advanced LLMs into SLMs (Fu et al., 2023; Ho et al., 2023; Hsieh et al., 2023; Magister et al., 2023; Wang et al., 2023; Dai et al., 2024). These methods typically involve extracting rationales from a teacher model and then fine-tuning a student model on these rationales. In the mental health domain, reasoning

distillation has also been employed, extracting explanatory rationales from advanced models to address the challenge of limited data for fine-tuning SLMs (Yang et al., 2024). This enables the student to elicit reasoning steps and knowledge from the teacher and achieve comparable performance with reduced size and computational cost.

Since reasoning distillation involves fine-tuning the student model, the quality of the data used for this fine-tuning is important. Recent studies have demonstrated the effectiveness of using small, high-quality datasets for fine-tuning models, further emphasizing the importance of data quality (Zhou et al., 2024; Xu et al., 2024b; Ye et al., 2025). Researchers also proposed a method to reduce the negative impact from imperfect teacher models, emphasizing the potential for these models to adversely affect students’ learning processes (Zhou and Ai, 2024). While these studies demonstrate the general importance of data quality, there remains a need to explicitly investigate the impact of selecting high-quality and domain-relevant data in the reasoning distillation process. This paper focuses on this critical aspect, investigating the impact of rationale quality on SLM performance.

2.2 Mental Health Detection from Social Media

Detecting mental disorders through social media involves analyzing a user’s posts to determine if they indicate any mental health issues (Yates et al., 2017; Tadesse et al., 2019; Kim et al., 2020; Murarka et al., 2021; Dinu and Moldovan, 2021; Chen et al., 2023). Researchers also investigated interpretable methods that analyze linguistic features, such as emotional expressions (Allen et al., 2019; Wang et al., 2021) or specific word choices (Tadesse et al., 2019; Jiang et al., 2020), to provide clinically useful insights and explain the model’s decisions. These methods primarily aimed to enhance detection accuracy by screening individuals who may be experiencing mental health issues more effectively.

Given the impressive achievements of LLMs in various domains, recent studies have examined their capability for detecting mental health issues and generating explanatory rationales (Yang et al., 2023, 2024; Wang et al., 2024; Xu et al., 2024a). Through CoT prompting, LLMs can include emotional cues (Yang et al., 2023) and symptom information (Wang et al., 2024) while generating detailed, domain-specific explanations. These studies

have also focused on the quality of rationales generated for interpretability, assessing factors such as consistency, reliability, and professionalism (Jeon et al., 2024; Yang et al., 2024). Inspired by these studies, this research aims to enhance both the performance of mental health problem detection and the rationale quality, thereby improving the clinical applicability of LLMs in mental health detection.

Recent studies have explored integrating standardized diagnostic tools into depression detection models, such as the PHQ-9 (Kocalevent et al., 2013) and the DSM-5 (American Psychiatric Association, 2013). These approaches aim to improve the accuracy and interpretability of these models by incorporating established clinical criteria and symptom-based assessments (Nguyen et al., 2022; Zhang et al., 2022; Song et al., 2023; Kang et al., 2024). The utilization of symptom-related information aligns with the diagnostic practices of mental health professionals in real-world clinical settings. This highlights the importance of incorporating such domain-specific knowledge to provide more clinically relevant and interpretable results. In line with this emphasis, we hypothesize that high-quality rationales are those that effectively incorporate clinical symptoms, and that fine-tuning a student model with these rationales will enhance both detection performance and rationale quality.

3 Method

This section introduces our proposed framework for improving the effectiveness of reasoning distillation by selectively focusing on high-quality rationales for major depressive disorder detection, as illustrated in Figure 2. This framework includes rationale generation, quality evaluation, and quality-based selection.

3.1 Problem Definition

We frame our explainable mental health problem detection task similarly to the previous studies (Yang et al., 2023, 2024). Suppose that we have a mental health detection dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x represents social media posts and y their detection labels. When a post x is given, we use a teacher model \mathcal{T} to predict the authors’ mental condition y' and generate rationales r using CoT prompting p , explaining why each x is detected as y' . This process is represented as $(x, y', r) \sim \mathcal{T}(y', r \mid x, p)$. We determine the detection accuracy by comparing the ground truth label y and

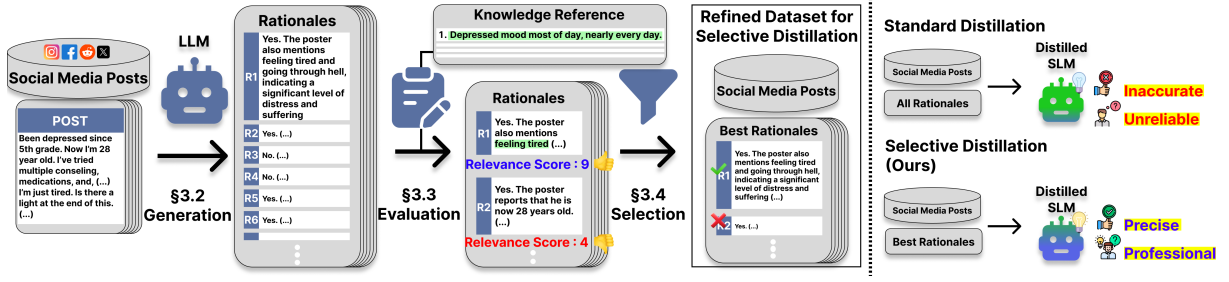


Figure 2: Overview of our proposed framework for selective reasoning distillation. Unlike standard reasoning distillation, our framework involves generating various rationales for each post, assessing their quality based on relevance to domain knowledge, and selecting the highest-quality rationale for distillation.

the teacher model’s prediction y' . Through this process, we can collect reasoning training samples: $\hat{D} = \{(x_i, y'_i, r_i)\}_{i=1}^N$. Similar to previous reasoning distillation studies (Magister et al., 2023; Hsieh et al., 2023; Ho et al., 2023), a student model \mathcal{S} is fine-tuned on \hat{D} . We evaluate the clinical interpretability by analyzing the rationales generated by the student models. Consequently, we aim to enhance SLM performance in both detection accuracy and clinical interpretability by distilling high-quality, domain-relevant rationales from teacher models.

3.2 Rationale Generation

First, we utilize a large teacher model to generate explanatory rationale. In typical reasoning distillation, when a depression post x is given, the teacher model is prompted to generate a detection result y' , and an explanatory rationale r . Unlike typical reasoning distillation, our framework prompts the teacher model to produce multiple diverse rationales for depression posts. This creates a pool of candidate rationales that are evaluated for quality, with only the highest-quality rationales selected to fine-tune the student model. While we utilize temperature adjustments in this study to generate a diverse set of rationales, we note that this is not the only method for achieving this goal. Details of the prompts for the rationale generation are in Appendix A.1.

3.3 Rationale Quality Evaluation

The evaluation of teacher-generated rationale quality, specifically regarding the incorporation of symptom-related information, can be approached through various methods. Prior work has explored methods such as symptom pattern matching (Nguyen et al., 2022), similarity score comparisons (Song et al., 2023), and the training of symptom classifiers (Zhang et al., 2022). Additionally, recent advancements in automated evalua-

tion, especially those utilizing LLMs, have demonstrated potential for achieving human-like assessment (Zheng et al., 2023). Therefore, this study leverages an LLM-based method, selected for its significant performance in previous studies. However, this does not exclude examining alternative approaches in future research.

We assessed rationale quality using an LLM-based evaluator, denoted by \mathcal{E} . For each post x_i and its corresponding rationale $r_{i,j}$, we compute a quality score $s_{i,j} = \mathcal{E}(x_i, r_{i,j})$. The evaluator is instructed to assess the alignment between generated rationales and the DSM-5 diagnostic criteria, a standard used by clinical experts in mental disorder diagnosis. This assessment is performed by GPT-4o (Achiam et al., 2023) and considers several key factors: (1) Domain Knowledge: incorporation of clinical symptom-relevant information (i.e., DSM-5 diagnostic criteria); (2) Symptom Recognition: accurate identification of relevant symptoms from the post; and (3) Symptom Relevancy: alignment of the rationale with identified symptom information. Details of the prompt for the LLM-evaluation are in Appendix A.2.

3.4 Quality-based Selection

In the previous steps, we generated a set of L rationales, denoted by $\{r_{i,j}\}_{j=1}^L$, for each post x_i using the teacher model. The best rationale $r_{i,\text{best}}$ is selected as:

$$r_{i,\text{best}} = r_{i,j^*} \quad \text{with } j^* = \arg \max_{j \in \{1, \dots, L\}} s_{i,j} \quad (1)$$

Using this process, we construct a refined dataset for selective distillation:

$$\hat{D}_{\text{SD}} = \{(x_i, y'_i, r_{i,\text{best}})\}_{i=1}^N \quad (2)$$

Here, only the highest-quality rationale for each post is paired with its corresponding prediction. This refined dataset is then used to fine-tune the

Teacher (\rightarrow)	X (Vanilla Models)						GPT-3.5-turbo					
Student (\rightarrow)	Llama-2		Llama-3.1		Mistral		Llama-2		Llama-3.1		Mistral	
Prompt (\downarrow)	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Std-Cot	59.24	67.24	75.82	69.36	70.51	70.88	84.38	78.15	80.49	70.75	83.01	75.69
Std-Cot + Selective	85.35	80.22	81.38	72.33	83.57	76.63	85.35	80.22	81.38	72.33	83.57	76.63
Step-by-Step	54.34	64.84	74.34	66.17	61.69	67.60	83.98	77.71	77.56	64.43	80.53	70.82
Step-by-Step + Selective	85.72	80.62	83.16	75.93	79.86	69.92	85.72	80.62	83.16	75.93	79.86	69.92
Emotion	65.61	70.69	73.33	63.07	79.93	75.85	82.12	74.25	79.53	68.85	82.94	75.43
Emotion + Selective	86.05	81.27	83.12	75.84	85.94	80.69	86.05	81.27	83.12	75.84	85.94	80.69

Teacher (\rightarrow)	GPT-4o						Llama-3-70B					
Student (\rightarrow)	Llama-2		Llama-3.1		Mistral		Llama-2		Llama-3.1		Mistral	
Prompt (\downarrow)	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Std-Cot	86.24	81.29	85.57	80.04	86.57	81.68	81.86	73.29	84.76	78.58	82.38	74.17
Std-Cot + Selective	88.95	85.66	86.83	82.19	91.02	88.63	84.24	77.74	86.87	82.39	89.21	85.96
Step-by-Step	83.53	76.58	81.64	73.10	82.27	74.19	88.91	85.62	86.65	81.89	90.62	88.36
Step-by-Step + Selective	89.17	85.92	85.27	79.42	90.95	88.64	89.32	86.19	86.80	82.02	90.69	88.32
Emotion	76.74	62.66	80.12	69.65	87.20	83.29	71.77	50.10	76.34	61.47	81.86	73.35
Emotion + Selective	85.76	80.49	81.90	73.25	89.35	86.39	76.93	62.80	85.05	79.15	85.98	80.95

Table 1: Experimental results on the performance of depression detection tasks. Each number represents the accuracy and F1 score of the corresponding student model trained with rationales generated by the corresponding teacher model and CoT prompting strategy. “+Selective” indicates that the model was trained using our proposed framework for selectively distilling high-quality rationales.

student model \mathcal{S} , thereby enhancing its ability to produce domain-relevant explanations and improving its performance in mental health detection.

3.5 Selective Reasoning Distillation

We fine-tune the student model \mathcal{S} on the selective distillation dataset $\hat{\mathcal{D}}_{SD}$. For each training sample, we concatenate the input post, predicted label, and the selected rationale into a single sequence $z = [x; y'; r_{best}]$. The objective loss function is:

$$\mathcal{L}_{SD} = - \sum_{t=1}^{|z|} \log P(z_t | z_{<t}; \theta), \quad (3)$$

where θ is the set of parameters of \mathcal{S} .

4 Experiments and Results

4.1 Experimental Setup

Dataset To validate the capability of smaller (distilled) models in both mental health problem detection and explanatory rationale generation, we utilize the Reddit_depression dataset (Song et al., 2023). This dataset comprises Reddit posts from two groups: mental disorder-related subreddits and random subreddits (clean text). Detailed statistics of this dataset are provided in Appendix B.1. For evaluating mental disorder detection, we use accuracy and F1 score as the primary metrics. Details of evaluation metrics are described in Appendix B.2

Models For the teacher models, we employed two closed-source LLMs, GPT-3.5 (Ouyang et al.,

2022) and GPT-4o (Achiam et al., 2023), and one open-source LLM, Llama-3-70B (Dubey et al., 2024). For the student models, we experiment with Llama-2-7B (Touvron et al., 2023), Llama-3.1-8B (Dubey et al., 2024), and Mistral-7B (Jiang et al., 2023). The details of the model versions are in Appendix B.3.

CoT Prompts We followed the prompt for generating rationales as suggested by Yang et al. (2023). To generate explanatory rationales, we conducted experiments with the following CoT prompts: (1) **Std-Cot** (Magister et al., 2023; Ho et al., 2023), the standard CoTs distillation method; (2) **Step-by-Step** (Hsieh et al., 2023), a multi-step reasoning approach; and (3) **Emotion-enhanced** (Yang et al., 2023), designed to elicit rationales that consider emotional aspects of the input text. We provide details in Appendix A.1.

Setup For rationale generation, we create a pool of 10 rationale candidates for each depression post. We set the temperature to 1.0 when using teacher models to generate these candidates. Other than that, we set the temperature to 0.0 for baseline experiments and during the test phase. Each request is attempted up to 5 times, with posts being excluded from the dataset if the generation is refused. The evaluator model scores rationales on a scale from 1 to 10. We employ LoRA (Hu et al., 2021) for fine-tuning the student SLMs. More details of the hyperparameters can be found in Appendix B.4.

Models	Human Evaluation Metrics (0–3†)		
	Consistency	Reliability	Professionality
Llama-2	1.88	1.62	1.56
+Distillation	2.41	2.25	1.96
+Selective	2.81	2.69	2.44
Llama-3.1	2.15	1.93	1.70
+Distillation	2.32	2.22	1.85
+Selective	2.54	2.62	2.55
Mistral	1.97	1.84	1.77
+Distillation	2.33	2.27	1.97
+Selective	2.79	2.66	2.73

Table 2: Human evaluation results in three metrics. The values that are **bold** mean the first outperforming groups, determined by Tukey’s HSD pairwise test at a significance level of $\alpha = 0.05$.

Human Evaluation Metrics We conducted human evaluations to assess the quality of generated rationale, following the guidelines from Yang et al. (2024). Three key metrics are employed: (1) *Consistency*, which measures the agreement between the model’s diagnosis and the information in the given post; (2) *Reliability*, which evaluates the credibility of the rationale, ensuring it is grounded in facts from the post; and (3) *Professionality*, which evaluates whether the rationale adheres to diagnostic standards. Further details of the evaluation scheme are provided in Appendix D.

4.2 Depression Detection Results

To comprehensively evaluate the impact of rationale quality on reasoning distillation for mental health detection, we conducted experiments with three distinct teacher models (GPT-3.5, GPT-4o, Llama-3-70B), three different student models (Llama-2, Llama-3.1, Mistral), and three CoT prompting strategies to generate rationales (Std-Cot, Step-by-Step, Emotion-enhanced). By testing our approach across these various settings, we aimed to demonstrate its robustness and broad applicability. Table 1 presents the depression detection performance across various combinations of these models and strategies.

The experimental results show that distilling rationales from the teacher models generally improved the performance of the student models, confirming that domain-specific knowledge can be transferred effectively by reasoning distillation. Moreover, applying our proposed framework (+*Selective*), which involves selectively distilling high-quality rationales, further enhanced performance across all combinations. This suggests that focusing on high-quality rationales leads to more effective knowledge transfer. The consistent improvement observed in the experiments provides compelling evidence that prioritizing rationale quality

is essential for maximizing the effectiveness of knowledge distillation in this domain.

4.3 Rationale Generation Results

We conducted human evaluations with two domain experts to assess the impact of our quality-focused approach on rationale generation. The experts hold degrees in psychology and were actively involved in related research. They are familiar with the DSM-5 criteria and other diagnostic tools. We randomly sampled 30 examples per model from a total of nine models: three vanilla student models, three models distilled using standard reasoning distillation (+*Distillation*), and three models distilled with selective reasoning distillation (+*Selective*). All models used the standard CoT prompting strategy with GPT-3.5 as the teacher model. Each expert independently assessed a total of 270 explanation samples, using three metrics from previous research (Yang et al., 2024): *Consistency*, *Reliability*, and *Professionality*. These metrics were rated on a scale of 0 to 3, and we averaged the scores from the two experts². The results are presented in Table 2.

The results demonstrate that applying our selective distillation method (+*Selective*) consistently improves the quality of the generated rationales, as evidenced by higher scores across all three metrics. This improvement is particularly pronounced in the *Professionality* metric, indicating that our quality-focused approach effectively enhances the clinical relevance of the explanations. It suggests that by selectively distilling high-quality rationales, we not only enhance the accuracy of student model predictions but also improve the quality and clinical relevance of their explanations.

5 Analysis of Quality Evaluation

5.1 Evaluation Method Validation

To assess the validity of our LLM-based evaluation method, we conducted a comparative analysis with other established evaluation metrics. Our goal was to assess whether our LLM-based approach aligns more closely with human evaluations of rationale quality in comparison to these alternative methods. Table 3 presents the Spearman rank correlation coefficients between the scores generated by each evaluation method and the human evalua-

²Inter-annotator agreement (IAA) was assessed by converting scores to ranks and calculating Cronbach’s Alpha. The resulting score of $\alpha = 0.69$ suggests acceptable reliability.

Evaluation Method	Consistency	Reliability	Professionality
Pattern Matching	0.075	-0.001	0.223**
BLEU Score	0.322***	0.286***	0.257***
Cosine Similarity	0.255***	0.232***	0.180*
BERTScore	0.374***	0.328***	0.248***
LLM-Evaluation	0.431***	0.327***	0.565***

Table 3: Results of comparing different rationale evaluation methods, assessing their correlation with human judgments of *Consistency*, *Reliability*, and *Professionality*. A high correlation indicates a strong alignment with human judgment. (Spearman correlation, *: $p < .05$, **: $p < .01$, ***: $p < .001$)

Symptoms (Reference)	Relevance	Corr.w/Human	Corr.w/LLM
Vocal Nodule	X	0.172*	0.307***
Schizophrenia (DSM-5)	Δ	0.383***	0.513***
Anxiety (DSM-5)	Δ	0.413***	0.563***
Depression (PHQ-9)	\bigcirc	0.470***	0.716***
Depression (DSM-5)	\bigcirc	0.565***	1.000

Table 4: Correlation of evaluation scores with human (*Professionalism*) and LLM evaluations using varying knowledge sources (X: Irrelevant, Δ : Moderately Relevant, \bigcirc : Highly Relevant). Details for each symptom reference are in Appendix C. (Spearman correlation, *: $p < .05$, **: $p < .01$, ***: $p < .001$)

tion scores across three human evaluation metrics. The evaluation methods under comparison include symptom pattern matching (Nguyen et al., 2022), semantic similarity scores (Cosine Similarity and BERTScore) (Song et al., 2023), and BLEU score.

As shown in Table 3, our LLM-based evaluation method exhibits a strong correlation with *Professionalism* (0.565, $p < .001$), demonstrating its ability to assess the subtleties of domain-specific knowledge and clinical reasoning, similar to human evaluations. Significant correlations are also observed for *Consistency* (0.431, $p < .001$) and *Reliability* (0.327, $p < .001$). Importantly, these correlations are consistently higher than those achieved against the other automated metrics, suggesting that our LLM-based method aligns more closely with human judgments of rationale quality.

5.2 Knowledge Reference Comparison

We performed a cross-reference analysis to assess how the choice of knowledge source influences our LLM-based evaluation method for major depressive disorder detection. We used various sources, ranging from irrelevant (e.g., vocal nodule, a physical condition) to highly relevant (e.g., PHQ-9, another depression assessment tool). We also included references with moderate relevance to mental health, such as those related to other mental disorders (generalized anxiety disorder and schizophrenia from the DSM-5). Table 4 presents the Spearman rank correlations between the eval-

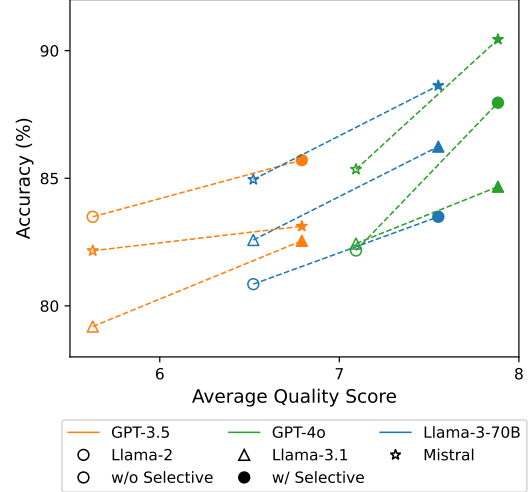


Figure 3: Correlation between the quality of teacher-generated rationales and the detection performance of student models. Lines connect the performance of the same student model with and without selective distillation. Markers indicate different student models, while colors indicate different teacher models.

uation scores obtained using different knowledge references and both human evaluations (*Professionalism*) and LLM evaluations.

The results show that, the more relevant the knowledge reference is to mental health, the stronger the correlation with both human and DSM-5 evaluations. This suggests that our proposed method accurately reflects domain knowledge when assessing rationale quality, highlighting the importance of utilizing a relevant knowledge reference like the DSM-5 for accurate and reliable evaluation.

5.3 Quality-Performance Correlation

To investigate how the quality of rationales relates to the performance of student models in detecting mental health issues, we analyzed the correlation between the average quality score of rationales and the accuracy of student models trained on those rationales. Figure 3 illustrates this correlation, plotting the average quality score against the corresponding accuracy. Each point represents a student model trained with rationales generated by a teacher model. The lines connect points representing the same student model trained with and without our selective distillation method.

Figure 3 shows a clear positive trend, indicating that higher rationale quality scores are generally associated with greater accuracy in student models. While there may be some variation depending on the specific teacher and student models used,

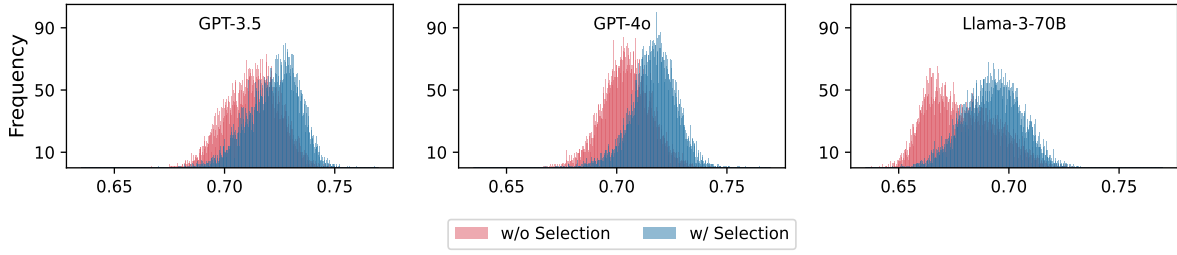


Figure 4: Distribution of semantic similarity scores between teacher-generated rationales and DSM-5 diagnostic criteria symptom descriptions for depression. We utilized BERTScore to measure the similarity scores and standard CoT prompts to generate rationales. The histograms in each panel, colored in red and blue, represent rationales generated without and with our proposed quality-based selection method, respectively.

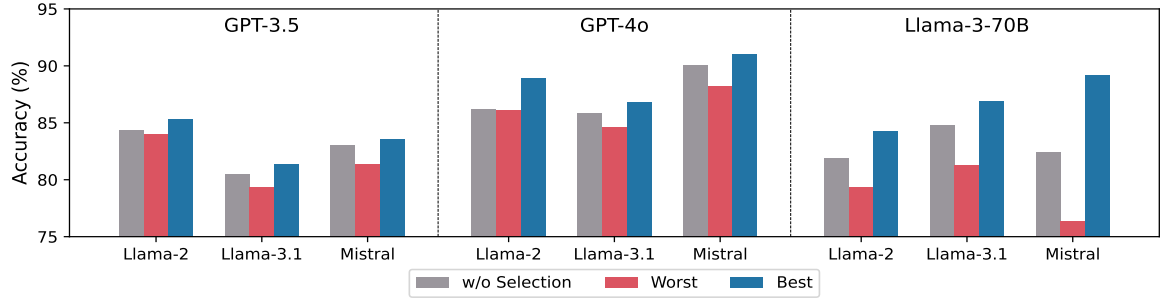


Figure 5: Ablation study on different selection criteria. We utilized standard CoT prompts for this experiment. Each bar represents the detection accuracy on the test dataset of the corresponding student model trained with the corresponding teacher.

the overall pattern suggests that learning with high-quality rationales leads to better performance. This observation highlights the importance of prioritizing high-quality rationales for effective knowledge distillation in mental health detection.

6 Analysis of Quality-Based Selection

6.1 Clinical Relevance Improvement

In order to investigate how our quality-based selection approach enhances the clinical relevance of rationales, we compared the distribution of semantic similarity scores between rationales generated by different teacher models and symptom descriptions from the DSM-5 diagnostic criteria for depression. We utilized BERTScore (Zhang et al., 2019) to compute the semantic similarity, making it suitable for assessing the alignment between rationales and clinical descriptions. We visualized these distributions in Figure 4. The histograms within each panel show the frequency of different BERTScore values for teacher-generated rationales with and without our quality-based selection.

In most cases, applying quality-based selection results in a noticeable shift of the semantic similarity score distribution towards higher values compared to the *w/o Selection* condition. This suggests that our method improves the clinical relevance of

the rationales by increasing their semantic similarity to the DSM-5 criteria, indicating that they convey more precise and relevant information about the associated symptoms. This improvement is essential because it ensures that the rationales used for reasoning distillation are clinically meaningful and informative, leading to enhanced student models through fine-tuning with them.

6.2 Ablation of Rationale Selection

We conducted an ablation study to analyze the impact of rationale selection by comparing the performance of models trained with different selection strategies: no quality-based selection (*w/o Selection*), selecting highest quality rationales (*Best*), and selecting lowest quality rationales (*Worst*). Figure 5 illustrates the impact of our quality-based selection method on the detection performance of student models.

As shown in Figure 5, across all teacher-student model combinations, the *Best* rationales consistently yield the highest accuracy, highlighting the effectiveness of our quality-based selection method. The *Worst* rationales often result in the lowest accuracy, emphasizing the adverse effect of incorporating low-quality rationales. These findings demonstrate the importance of prioritizing high-quality

rationales for effective knowledge transfer and improved mental health detection.

7 Conclusion

This paper empirically investigated how the quality of rationales, especially their relevance to domain knowledge, impacts the rationale distillation process. We proposed a quality-focused framework that fine-tunes smaller language models with rationales exhibiting high domain relevance, achieved through a process of evaluating and selecting rationales generated by teacher models. Our experiments demonstrated that our selective reasoning distillation significantly improves both detection accuracy and clinical interpretability in major depressive disorder detection. These results shed light on the importance of refining distillation data to enhance the model performance and interpretability, especially in mental health applications requiring high-quality domain knowledge. This work provides a promising direction for enhancing the performance and interpretability of smaller language models in mental health applications. Our selective distillation framework, which prioritizes high-quality, domain-specific knowledge, offers a valuable approach, especially in domains where accurate and reliable explanations are crucial.

Limitations

While this study offers valuable insights, it is important to acknowledge that there are some limitations beyond the primary scope of this research. Each paragraph below discusses these limitations and suggests meaningful directions for future work.

Efficient Modeling. Our framework incurs significant computational costs due to the necessity of repeated rationale generation and LLM-based quality evaluation. These processes, especially with large datasets or complex tasks, can become expensive due to API costs and computational resource requirements. Although this study prioritizes demonstrating the importance of refining high-quality rationales for enhancing SLM performance, future work should explore more efficient methods. This could involve developing computationally less expensive evaluation metrics or investigating alternative selection strategies that balance quality and efficiency.

Collaborative Approach with Experts. This study utilizes LLM-generated rationales for distillation, but future work could explore incorpo-

rating human-generated rationales to potentially enhance the quality and interpretability of the student models' explanations. However, obtaining such data presents practical challenges, including the costs associated with recruiting and compensating experts, as well as the logistical complexities of coordinating data collection and ensuring adherence to privacy and ethical guidelines. As demonstrated in § 5.1, our proposed method aligns with human judgment, suggesting its potential as an alternative to direct expert generation of rationales. Future research could investigate incorporating expert-generated ground truth data into the modeling process, potentially through collaborative approaches or by leveraging expert knowledge to refine evaluation metrics.

Clinical Usefulness Evaluation. This study incorporates human evaluation (§ 4.3) to assess the quality and interpretability of rationales generated by LLMs, utilizing a method previously established in mental healthcare research (Jeon et al., 2024; Yang et al., 2024). However, further research is needed to investigate how to measure both the quantitative and qualitative potential impacts of these rationales in clinical practice. Evaluating the effectiveness of LLM-generated rationales in real-world clinical settings is crucial to understanding their potential benefits and limitations (Won et al., 2025). Future research could investigate more deeply the human-AI interaction regarding how these rationales can assist mental health professionals in diagnosis, treatment planning, and patient communication by conducting user studies with clinicians.

Adapting to Different Domains. While this study primarily focuses on mental disorder detection, specifically for major depressive disorder, the proposed framework has the potential for broader applicability. As demonstrated in § 5.2, incorporating relevant knowledge references can enhance the effectiveness of our framework. Therefore, by utilizing appropriate knowledge sources, our framework could be adapted for specific domains that require different types of domain knowledge. Future research should investigate the feasibility and effectiveness of applying this framework to different domains by incorporating relevant knowledge sources and evaluating its performance on diverse tasks.

Maintaining Generalizability. Generalizability is also an important research topic. Research on how to maintain performance on general tasks

while performing specific tasks is also a very active field of research. Fine-tuning the student model on a domain-specific dataset may lead to some reduction in generalizability to other tasks. While this work focuses on optimizing performance for a specific domain, future research could explore strategies to enhance domain-specific performance without sacrificing generalizability.

Extending to a Broader Range of LLMs. We assessed the robustness and validity of our findings across different architectures by exploring a range of teacher and student models, including both recent and older versions, as well as open-source and closed-source models. Our experimental results (§ 4.2) demonstrate consistent trends across the models tested, but it is important to recognize that new models are continuously emerging. While this work could not encompass all available models, the core principle of training with high-quality, domain-relevant rationales, as demonstrated in this study, remains crucial for achieving optimal performance. Future research should explore the application of this principle to newly developed models, further investigating how to refine training data to enhance performance across diverse models and domains.

Ethical Statements

Our human evaluation was approved by the Institutional Review Board (IRB)³. All participants were fully informed about the procedure for the evaluation and acknowledged their consent to participate.

To ensure the privacy and anonymity of users, the examples presented in our paper have been paraphrased. Additionally, we adhered strictly to the ethical guidelines proposed for those datasets. The data were used solely for research purposes and were not shared with anyone outside the research team. In our data release for selective reasoning distillation, we will only provide the rationale generated by the model and the index of the original data, excluding any existing dataset content.

While our models showed strong performance in our experiments, it is important to note that these models are not intended to label or stigmatize individuals. Instead, they are designed to serve as tools to enhance public health and provide potential early warnings. We emphasize that any potential mental health-related symptoms and descriptions generated by the models should not be considered definitive diagnoses. For medical diagnosis, the

model’s output should only serve as supplementary indicators, and consultation with professional psychiatrists or clinical practitioners is essential.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208054).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Kristen Allen, Shrey Bagroy, Alex Davis, and Tamar Krishnamurti. 2019. ConvSent at CLPsych 2019 task a: Using post-level sentiment features for suicide risk prediction on Reddit. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 182–187.
- American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*. VA: American Psychiatric Association, Arlington.
- Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318. Association for Computational Linguistics.
- Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. 2023. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9071–9084. Association for Computational Linguistics.
- Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024. Improve student’s reasoning generalizability through cascading decomposed CoTs distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15643. Association for Computational Linguistics.
- Anca Dinu and Andreea-Codrina Moldovan. 2021. Automatic Detection and Classification of Mental Illnesses from General Social Media Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 358–366.

³Approval number: KH2023-166

- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR.
- Keith Harrigan, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3774–3788.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882. Association for Computational Linguistics.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hyolim Jeon, Dongje Yoo, Daeun Lee, Sejung Son, Seungbae Kim, and Jinyoung Han. 2024. A dual-prompting for interpretable mental health language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 247–255. Association for Computational Linguistics.
- Shaoxiong Ji. 2022. Towards intention understanding in suicidal risk assessment with natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4028–4038. Association for Computational Linguistics.
- 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. European Language Resources Association.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th international workshop on health text mining and information analysis*, pages 147–156.
- Zhengping Jiang, Jonathan Zomick, Sarah Ita Levitan, Mark Serper, and Julia Hirschberg. 2021. Automatic detection and prediction of psychiatric hospitalizations from social media posts. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 116–121. Online. Association for Computational Linguistics.
- Migyeong Kang, Goun Choi, Hyolim Jeon, Ji Hyun An, Daejin Choi, and Jinyoung Han. 2024. CURE: Context- and uncertainty-aware mental disorder detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17924–17940. Association for Computational Linguistics.
- Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):11846.
- Rüya-Daniela Kocalevent, Andreas Hinz, and Elmar Brähler. 2013. Standardization of the depression screener patient health questionnaire (phq-9) in the general population. *General hospital psychiatry*, 35(5):551–555.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781. Association for Computational Linguistics.
- Anshu Malhotra and Rajni Jindal. 2024. Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *Cognitive Systems Research*, 84:101186.
- Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. [Classification of mental illnesses on social media using RoBERTa](#). In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 59–68, online. Association for Computational Linguistics.

- Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. [Improving the generalizability of depression detection by leveraging clinical questionnaires](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12. Association for Computational Linguistics.
- Hoyun Song, Jisu Shin, Huije Lee, and Jong Park. 2023. A simple and flexible modeling for mental disorder detection by learning from clinical questionnaires. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12190–12206. Association for Computational Linguistics.
- Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C Park. 2018. Feature attention network: interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia conference on language, information and computation*.
- Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Elsbeth Turcan and Kathy McKeown. 2019. Dreddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107. Association for Computational Linguistics.
- Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. 2022. Multi-aspect transfer learning for detecting low resource mental disorders on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3202–3219. European Language Resources Association.
- Ning Wang, Fan Luo, Yuvraj Shivtare, Varsha D Badal, KP Subbalakshmi, Rajarathnam Chandramouli, and Ellen Lee. 2021. Learning Models for Suicide Prediction from Social Media Posts. *arXiv preprint arXiv:2105.03315*.
- Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*.
- Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126. Association for Computational Linguistics.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Hyunseon Won, Migyeong Kang, Minji Kim, Daeun Lee, Hyein Choi, Yonghoon Kim, Daejin Choi, Min-sam Ko, and Jinyoung Han. 2025. "show your mind": Unveiling user experience on an ai-based mental health assessment system with symptom-based evidences. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–11.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024a. 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.
- Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.
- Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077. Association for Computational Linguistics.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mental-lama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.
- Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. [Depression and self-harm risk assessment in online forums](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,

pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Sourabh Zanwar, Xiaofei Li, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. What to fuse and how to fuse: Exploring emotion and personality fusion strategies for explainable mental disorder detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8926–8940.

Tianlin Zhang, Kailai Yang, Hassan Alhuzali, Boyang Liu, and Sophia Ananiadou. 2023. Phq-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing & Management*, 60(5):103417.

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

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. [Symptom identification for interpretable detection of multiple mental disorders on social media](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. *arXiv preprint arXiv:2406.05322*.

A Details of Instruction Prompts

A.1 Prompts for Rationale Generation

user:
Consider this post to answer the question: Is the poster likely to suffer from very severe Major Depressive Disorder? Only return Yes or No, then explain your reasoning.

Post: "{Post}".
Answer:

Table 5: The **Std-Cot** prompt used for rationale generation.

user:
Consider this post to answer the question: Is the poster likely to suffer from very severe Major Depressive Disorder? Only return Yes or No, then explain your reasoning **step by step**.

Post: "{Post}".
Answer:

Table 6: The **Step-by-Step** prompt used for rationale generation.

user:
Consider **the emotions expressed from** this post to answer the question: Is the poster likely to suffer from very severe Major Depressive Disorder? Only return Yes or No, then explain your reasoning **step by step**.

Post: "{Post}".
Answer:

Table 7: The **Emotion-enhanced** prompt used for rationale generation.

To evaluate the robustness of our proposed method across different Chain-of-Thought (CoT) prompting strategies, we prepared three distinct CoT prompts. Tables 5, 6, and 7 present the **Std-Cot**, **Step-by-Step**, and **Emotion-enhanced** prompts, respectively.

A.2 Prompt for Rationale Evaluation

To assess how well the rationale considers key factors, we employed GPT-4o as an evaluator using the prompt shown in Table 8. The prompt incorporates the diagnostic criteria of DSM-5 as a knowledge reference, as detailed in Table 9. Given a post and its corresponding rationale, the evaluator generates scores ranging from 1 to 10.

user:
You are a clinical psychologist with expertise in diagnosing Major Depressive Disorder (MDD). Your task is to score a reasoning text about potential MDD cases based on symptom identification and logical analysis. Return only a score without rationale.

Symptom Checklist
{*Knowledge Reference*}

Reasoning Quality Scale
1-2: SEVERELY INADEQUATE
No reference to checklist
Unclear or disconnected descriptions
Primarily subjective interpretations
3-4: INADEQUATE
Mentions symptoms without connecting to checklist
Some relevant points but significant gaps
Overlooks several key symptoms
5-6: BASIC
Basic references from checklist
Relevant but not insightful observations
Makes logical connections but lacks depth
7-8: PROFICIENT
Alignment with checklist
Relevant and accurate with minor omissions
Cover key symptoms
9-10: EXEMPLARY
Clear alignment with checklist
Highly relevant and comprehensive
Consider both key and minor symptoms with clear evidence

Input Format: ""
Post: [Original text]
Reasoning: [Analysis to evaluate]
""
Output Format: ""
Score: [1-10]
""

Post: {Post}
Reasoning: {Rationale}

Table 8: The prompt used for rationale quality evaluation.

B Details for Experimental Setup

B.1 Datasets

In our experiments, we utilized the Reddit_depression dataset (Song et al., 2023). This dataset comprises Reddit posts annotated for mental disorder detection, focusing on identifying major depressive disorder (MDD). It is divided into training, validation, and test sets with 17,678, 2,696, and 2,696 samples, respectively. Each post is labeled with one of two categories: “Yes” (indicating depression) or “No” (indicating non-depression).

B.2 Evaluation Metrics

To evaluate the performance of the models for mental health problem detection tasks, we use standard classification metrics, including Accuracy (Acc.) and F1-Score (F1). We frame the mental health problem detection task as an explainable mental health analysis, similar to the approach by Yang et al. (2024). Accordingly, the detection label y is included in the model’s generated output. If a model produces a label that is not part of the predefined detection labels for the task (yes or no), it is considered an unanswered and incorrect response.

B.3 Versions of Models

For the teacher model, we selected GPT-3.5 (*gpt-3.5-turbo-0125*) (Ouyang et al., 2022), GPT-4o (*gpt-4o-2024-08-06*) (Achiam et al., 2023), and Llama-3-70B (*meta-llama/Meta-Llama-3-70B-Instruct*) (Dubey et al., 2024) accessed through the DeepInfra API⁴. For the student model, we chose Llama-2-7B (*meta-llama/Llama-2-7b-chat-hf*) (Touvron et al., 2023), Llama-3.1-8B (*meta-llama/Llama-3.1-8B-Instruct*) (Dubey et al., 2024), and Mistral-7B (*mistralai/Mistral-7B-Instruct-v0.1*) (Jiang et al., 2023).

B.4 Hyperparameters for Training and Inference

The student models were fine-tuned using LoRA (Hu et al., 2021). We used the following parameters: max epoch of 1, batch sizes of 2, gradient accumulation steps of 32, and learning rates of $2e - 4$. We used the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.01 and a linear scheduler starting with 50 warmup steps. For the initialization of LoRA weights, we used $r = 16$, $\alpha = 32$, and a dropout rate of 0.05. All models were fine-tuned on one NVIDIA A100 cluster.

During inference, we set top-p to 0.95, maximum sequence length to 300 tokens, and temperature to 0.0. We utilized vLLM (Kwon et al., 2023) to accelerate inference.

C Details of Knowledge References

We present the knowledge references used in §5.2. Tables 9, 11, and 12 present the DSM-5 diagnostic criteria for major depressive disorder (MDD), generalized anxiety disorder (GAD), and schizophrenia, respectively (American Psychiatric

⁴<https://deepinfra.com/meta-llama/Meta-Llama-3-70B-Instruct>

DSM-5 diagnostic criteria for MDD

Depressed mood most of the day, nearly every day
Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day
Insomnia or hypersomnia nearly every day
Significant weight loss when not dieting or weight gain, or decrease or increase in appetite nearly every day
Fatigue or loss of energy nearly every day
Feeling worthless or excessive or inappropriate guilt nearly every day
Diminished ability to think or concentrate, or indecisiveness, nearly every day
A slowing down of thought and a reduction of physical movement
Recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide

Table 9: DSM-5 diagnostic criteria for major depressive disorder (MDD).

PHQ-9

Feeling down, depressed, or hopeless.
Little interest or pleasure in doing things.
Trouble falling or staying asleep, or sleeping too much.
Poor appetite or overeating.
Feeling tired or having little energy.
Feeling bad about yourself - or that you are a failure or have let yourself or your family down.
Trouble concentrating on things, such as reading the newspaper or watching television.
Moving or speaking so slowly that other people could have noticed.
Thoughts that you would be better off dead, or of hurting yourself.

Table 10: PHQ-9 items assessing depressive symptoms.

DSM-5 diagnostic criteria for GAD

Excessive anxiety and worry, occurring more days than not for at least 6 months, about a number of events or activities.
The individual finds it difficult to control the worry.
The anxiety and worry are associated with irritability.
The anxiety and worry are associated with being easily fatigued.
The anxiety and worry are associated with sleep disturbance (difficulty falling or staying asleep, or restless, unsatisfying sleep).
The anxiety and worry are associated with difficulty concentrating or mind going blank.
The anxiety and worry are associated with muscle tension.

Table 11: DSM-5 diagnostic criteria for generalized anxiety disorder (GAD).

Association, 2013). Table 10 presents the PHQ-9 items (Kocalevent et al., 2013), a self-administered screening tool used to diagnose and assess a severity of depression. Table 13 describes the diagnostic features of vocal nodules from Wikipedia⁵, which

⁵https://en.wikipedia.org/wiki/Vocal_cord_nodule

DSM-5 diagnostic criteria for schizophrenia
The presence of one (or more) delusions with a duration of 1 month or longer.
Criterion A for schizophrenia has never been met. Note: Hallucinations, if present, are not prominent and are related to the delusional theme (e.g., the sensation of being infested with insects associated with delusions of infestation).
Apart from the impact of the delusions, or its ramifications, functioning is not markedly impaired, and behavior is not obviously bizarre or odd.
If manic or major depressive episodes have occurred, these have been brief relative to the duration of the delusional periods.
The disturbance is not attributable to the physiological effects of a substance or another medical condition and is not better explained by another mental disorder, such as body dysmorphic disorder or obsessive-compulsive disorder.

Table 12: DSM-5 diagnostic criteria for schizophrenia.

Diagnostic features of vocal nodules
Experiences persistent hoarseness or a raspy voice, especially after speaking for long periods.
Notices frequent vocal fatigue or difficulty projecting the voice.
Feels a sensation of strain or pain in the throat when speaking or singing.
Has a reduced vocal range, particularly in higher or lower pitches.
Experiences frequent throat clearing or the feeling of something stuck in the throat.
Has a history of overusing the voice, such as yelling, shouting, or excessive speaking.
Feels dryness or irritation in the throat despite staying hydrated.

Table 13: Diagnostic features of vocal nodules.

serves as an irrelevant source of knowledge compared to MDD diagnosis.

D Human Evaluation Scheme

We recruited two domain experts specializing in clinical psychology, providing each with a compensation of \$100 for their evaluation. The evaluation criteria proposed by Yang et al. (2024) were as follows:

Consistency: Evaluates whether the rationale is consistent with the detection result determined by the teacher model for the given post and if the rationale sufficiently supports the detection decision.

- 0: The detection result and the explanation do not match.
- 1: The detection result and the explanation match, but the explanation is difficult to read and contains serious errors.
- 2: The detection result and the explanation

match. The explanation is mostly consistent and readable, with a few minor errors.

- 3: The detection result and the explanation match perfectly. The explanation is natural, consistent, and error-free.

Reliability: Assesses the trustworthiness of the generated rationale, ensuring that it is fact-based and reliable.

- 0: Completely untrustworthy and contains false information (e.g., non-existent symptoms).
- 1: Partially trustworthy but includes explanations not based on facts.
- 2: Mostly trustworthy but contains minor misinformation or incorrect explanations.
- 3: Completely trustworthy.

Professionalism: Evaluates if the rationale adheres to diagnostic standards from clinical descriptions. Annotators used information from PHQ-9 and DSM-5 to determine symptoms and sorted them based on their own knowledge.

- 0: The explanation lacks any evidence or omits critical symptoms.
- 1: The explanation provides a few supportive pieces of evidence but omits more critical symptoms.
- 2: The explanation includes several supportive pieces of evidence but omits some minor symptoms.
- 3: The explanation includes all related supportive pieces of evidence in the post.

E Case Study

This section presents a case study to illustrate the practical application of our framework for selective reasoning distillation. We analyze the performance of student models trained with (+*Selective*) and without (+*Distillation*) our quality-based selection framework, focusing on their ability to detect major depressive disorder (MDD) and generate clinically relevant rationales. Tables 14 and 17 present the example responses generated by distilled Llama-2 and Mistral student models for the corresponding social media posts, respectively, after being trained

Post	
No one seems to care about me. My crush, who dumped me, is texting me again but mentions other guys. My family problems make me want to be alone for the rest of my life and just cry. I had high hopes for 2021, but I'm still not expecting anything good to happen. Normally my parents are disappointed in me; I was the dumbest kid in the class. I never expect to accomplish anything in life; I fully expect to die alone. I have zero energy for anything; I just lie in bed and want to cry nonstop. I still haven't forgotten her. I already know everything sucks for me, and that I'll never find happiness. I just hope the world ends soon.	
Method	Responses
+Distillation	Yes. 1. △Self-Perception and Emotional Confusion: The poster describes feeling disconnected from their emotions and struggling to understand their thoughts and feelings. This level of self-awareness and emotional confusion is often associated with Major Depressive Disorder (MDD). 2. △Doubt and Self-Doubt: The poster expresses doubt about their ability to talk, reason, and empathize with others, which can be indicative of low self-esteem and self-doubt, common in depression. 3. △Anxiety and Fear: The mention of a "sinking feeling of dreadful awfulness and revolting grossness" suggests a high level of anxiety and fear, which are common symptoms of depression. 4. ✓Lack of Clarity and Confusion: The poster expresses a lack of understanding about their emotions and feelings, which can be a sign of depression. 5. ✗Seeking Help and Uncertainty: The poster is seeking help and acknowledges that they may be experiencing something wrong, indicating a level of self-awareness and willingness to seek help, which are positive signs.
+Selective	Yes. 1. ✓Expression of Hopelessness: The poster expresses a deep sense of hopelessness and a belief that nothing good will happen in their life, which is a key symptom of Major Depressive Disorder (MDD). 2. ✓Lack of Energy and Motivation: The poster mentions having "0 energy for anything" and wanting to cry nonstop, indicating a significant lack of motivation and energy, which are common symptoms of depression. 3. △Social Isolation: The poster feels like no one cares about them and describes a lack of support from family and friends, which can exacerbate depressive symptoms. 4. ✓Negative Self-Perception: The poster describes themselves as "the dumbest kid in the class" and expresses a belief that they will never accomplish anything in life, which reflects a negative self-image and low self-esteem. 5. ✓Suicidal Ideation: The poster mentions hoping for the world to end faster so they could be "put down," which is a serious indicator of suicidal ideation and a critical

Table 14: Example rationales generated by distilled student models assessing a social media post for detecting major depressive disorder and generating a rationale. For these examples, we utilized GPT-4o as a teacher model and Llama-2 as a student model. The notation ✓ indicates the high relevance with clinical symptoms, referring to proper symptoms in the post. △ indicates a sign of depression but not related to diagnostic criteria. ✗ indicates not proper reasoning. To ensure the privacy and anonymity of users, the example post presented in this table have been paraphrased.

on rationales produced by the GPT-4o and Llama-3-70B teacher models, respectively.

In both tables, the +*Selective* models generate rationales that are more focused, relevant, and clinically informative than the +*Distillation* models, demonstrating the effectiveness of our framework in enhancing the quality and interpretability of rationales generated by distilled student models. While the specific writing styles of the rationales vary depending on the student and teacher models used, applying our quality-based selection framework generally enhances interpretability. These case studies highlight the practical benefits of our framework for selective reasoning distillation, showing that by selectively distilling high-quality rationales, we can enhance the ability of student models to generate clinically relevant and interpretable explanations for mental health conditions.

F Performance Comparison with SOTA Models⁶

This section presents a performance comparison of our proposed method with existing state-of-the-art (SOTA) approaches for mental disorder detec-

Category	Model	Acc.	F1
Teacher models	GPT-3.5	80.71	74.14
	GPT-4o	86.92	83.94
	Llama-3-70B	86.86	83.79
Latest models	MentalLLaMA-13B	74.60	70.70
	MentalLLaMA-7B	79.57	78.57
Vanilla student	Llama-2-7B	59.24	67.24
	Llama-3.1-8B	75.82	69.36
	Mistral-7B	70.51	70.88
Ours	Llama-2-7B + ours	88.95	85.66
	Llama-3.1-8B + ours	86.83	82.19
	Mistral-7B + ours	91.02	88.63

Table 15: Results of the detection performance comparing our method with recent LLMs.

tion. Table 15 presents the depression detection performance, comparing our selective distillation framework, applied to student models (Llama-2-7B, Llama-3.1-8B, Mistral-7B), against other competitive SOTA models. Among these, we particularly highlight the MentalLLaMA models (Yang et al., 2024), which are specifically fine-tuned for diagnosing mental illness.

⁶We include these sections as a response to reviewers' comments.

Category	Model	DR		Dreddit		Reddit_anxiety	
		Acc.	F1	Acc.	F1	Acc.	F1
Vanilla model	Mistral-7B	51.70	53.24	56.45	56.15	71.34	70.77
	Llama-2-7B	72.43	83.38	61.43	64.06	62.35	69.06
Distilled model	Mistral-7B	65.94	69.56	67.40	67.77	83.19	84.64
	Llama-2-7B	62.08	68.24	73.98	71.15	85.42	83.45
Ours	Mistral-7B	69.19	73.49	71.43	75.08	86.60	84.19
	Llama-2-7B	82.16	87.78	74.38	77.25	87.92	86.25

Table 16: Results on three mental health problem detection tasks.

As demonstrated in Table 15, our selective distillation method consistently shows strong performance across various student models, significantly surpassing the MentalLLaMA baselines in both accuracy and F1 score. These results highlight the effectiveness of our quality-focused approach in enhancing the performance of smaller language models for mental health detection tasks.

G Test on Different Domains⁶

As highlighted in the Conclusion and Limitations sections, our framework, while primarily applied to major depressive disorder detection, possesses inherent versatility across diverse domains. In this section, we present preliminary experiments demonstrating its applicability and performance improvement across various domains, thereby guiding future work directions. For this exploration, we utilized three distinct datasets: (1) The Depression Reddit (DR) dataset (Pirina and Çöltekin, 2018), serving as another major depressive disorder de-

tection task; (2) the Dreddit dataset (Turcan and McKeown, 2019), for stress detection; and (3) the Reddit_anxiety dataset (Song et al., 2023), addressing generalized anxiety disorder detection. For the stress and anxiety detection tasks, symptom information from the DSM-5 was specifically utilized as the symptom reference.

The experimental results across these three datasets are detailed in Table 16. As shown, our method consistently improves both accuracy and F1 score when compared to vanilla and standard distilled models across all datasets. This consistent enhancement, observed even on out-of-domain datasets like Dreddit and Reddit_anxiety, provides strong evidence for the adaptability and broad applicability of our approach. These findings support our discussion in the Limitations section, indicating that our framework can be effectively adapted to specific domains by leveraging appropriate knowledge sources.

Post	
I deal with depression every day. Some days are a complete fog, and I feel like a helpless child. I struggle to hold a job, make enough money, and do things that would make me happy. My progress feels minimal, and I always end up giving up because my depression depletes nearly all of my energy. I don't know what I'm doing with my life. I feel so low and beside myself. I have no savings, no plan, no job, and often no will to get up and do things in the morning. My depression makes me feel like a shell of myself. I wouldn't wish this on my worst enemy. I'd give up everything just to not feel this way. I'm still trying, but I'm frustrated and impatient.	
Method	Responses
+Distillation	<p>Yes.</p> <p>I believe the poster is likely to suffer from very severe Major Depressive Disorder (MDD) for several reasons:</p> <ol style="list-style-type: none"> 1. △Intensity of negative emotions: The poster describes feeling “low”, “beside myself”, and experiencing a “fog” of depression that makes them feel like a “helpless child”. These phrases suggest a high level of emotional distress. 2. ✓Impact on daily life: The poster reports struggling to hold a job, make enough money, and engage in activities that would bring them joy, indicating a significant impact on their daily life. 3. ✓Lack of motivation and interest: The poster expresses a lack of interest in activities they might enjoy, such as acting or directing, and feels like they have no will to get up and do things in the morning. 4. △Feelings of hopelessness: The poster states that they wouldn't wish their current state on their worst enemy and would give up all their possessions just to not be depressed, indicating a sense of hopelessness and despair.
+Selective	<p>Yes.</p> <p>Here's my step-by-step reasoning:</p> <ol style="list-style-type: none"> 1. ✓Intensity of symptoms: The poster describes feeling like a “shell of myself” and having no energy, which suggests a high level of depressive symptoms. 2. ✓Loss of motivation and energy: The poster mentions that their depression “depletes nearly all of my energy”, which makes it difficult for them to engage in activities they enjoy or need to do. This is a common symptom of severe depression. 3. △Feelings of hopelessness: The poster states that they “don't know what I'm doing with my life” and feel like they're “not making any progress”. This sense of hopelessness is a hallmark of severe depression. 4. ✓Lack of interest in activities: The poster mentions that they “don't want to do anything” and feel like they're “just sometimes feel like a shell of myself”. This lack of interest in activities.

Table 17: Example rationales generated by distilled student models assessing a social media post for detecting major depressive disorder and generating a rationale. For these examples, we utilized Llama-3-70B as a teacher model and Mistral as a student model. The notation ✓ indicates the high relevance with clinical symptoms, referring to proper symptoms in the post. △ indicates a sign of depression but not related to diagnostic criteria. To ensure the privacy and anonymity of users, the example post presented in this table have been paraphrased.