MIRAMIND: Benchmarking Reliable Clinical Mental Health Reasoning beyond Answer Accuracy

Anonymous ACL submission

Abstract

Mental-health reasoning with large language models (LLMs) requires more than producing a plausible answer: models must interpret limited and subjective evidence, calibrate risk, select appropriate interventions, synthesize clinical evidence, and verify mental-health claims while keeping their reasoning grounded and internally consistent. Existing mental-health benchmarks have advanced psychiatric diagnosis, counseling competence, crisis response, and ethical reasoning, but they usually target specific clinical roles or final-answer correctness, leaving the reliability of intermediate reasoning trajectories under-specified. To bridge this gap, we introduce MIRAMIND, a unified benchmark for reliable mental-health reasoning. MIRAMIND spans six task families and 13 datasets across five core reasoning aspects: appraisal, diagnosis, intervention, abstraction, and verification, and further provides a trajectory-level evaluation protocol that assesses usability, logical structure, and informational contribution beyond task accuracy. Our evaluation of 20 LLMs reveals a restraint gap: strong models often recognize familiar labels or generate plausible explanations, but struggle to keep conclusions proportional to limited mental-health evidence. As a benchmark-guided post-training case study, we train Mindora, an 8B model optimized for evidence-proportional reasoning. Despite its compact scale, Mindora achieves the best overall performance on MIRAMIND while producing more balanced reasoning trajectories. These results suggest that MIRAMIND can serve as both an evaluation benchmark and a research substrate for exploring more practically meaningful mental-health reasoning models.

1 Introduction

When people ask LLMs about mental health, they rarely ask a clean question with a checkable answer. They bring a story and ask the model to help decide what it means, whether a thought should worry them, whether a feeling suggests risk, what response would help, or whether a piece of advice should be trusted (Xiao et al., 2024; Singh et al., 2024; Yang et al., 2024; Qiu and Lan, 2024; Xiao et al., 2025). This is where today's reasoning LLMs face a problem that their gains on mathematics, coding, and scientific problem solving do not resolve. In those domains, post-training can often reward reasoning because the final answer is externally verifiable(Zelikman et al., 2022; Lightman et al., 2024; Guo et al., 2025).

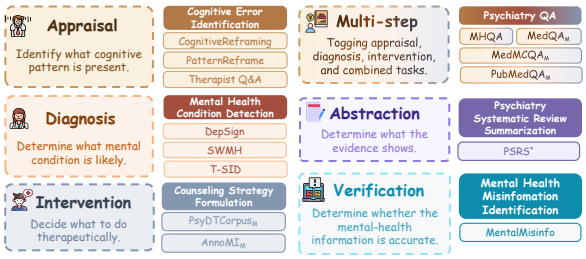

Figure 1: Tasks and datasets included in MIRAMIND.

In mental health, by contrast, the model's intermediate interpretation, what it treats as a symptom, a risk signal, an appropriate response, or credible evidence, may shape the user's self-understanding and subsequent action before any external verification is possible. Recent lawsuits and investigations have made these stakes visible[1]. Yet current evaluations rarely test this form of reasoning directly: whether LLMs can make cautious interpretations from limited narratives, avoid prematurely turning

[1]We refer to these cases as public allegations and investigations, not as established causal findings. Examples include public reports on lawsuits or investigations concerning chatbot interactions in adolescent suicide (https://time.com/7312484/chatgpt-openai-suicide-lawsuit/), delusional escalation before murder-suicide (https://apnews.com/article/97fd7da31c0fa08f3d3ea9efd6713151), and school or campus shooting planning (https://www.pbs.org/newshour/nation/lawsuit-accuses-chatgpt-of-helping-gunman-plan-fsu-mass-shooting).

Table 1: Comparison between MIRAMIND and existing mental-health reasoning benchmarks. MIRAMIND differs from prior work by jointly covering heterogeneous mental-health reasoning functions and evaluating reasoning trajectories beyond final-answer accuracy.

Benchmark	Main Focus	Appraisal	Diagnosis	Intervention	Abstraction	Verification / Safety	Multi-task Coverage	Trajectory Eval.
Psychiatric diagnosis and clinical decision-making benchmarks								
MentalBench (Song et al., 2026)	DSM-5 diagnostic decision-making	✗	✓	✗	✗	✗	✗	✗
MentalDx Bench (Sun et al., 2026)	Real-world psychiatric diagnosis from EHRs	✗	✓	✗	✗	✗	✗	✗
LingxiDiagBench (Xu et al., 2026)	Multi-turn Chinese psychiatric consultation and diagnosis	✗	✓	✓	✗	✗	✓	✗
PsychiatryBench (Fouda et al., 2026)	Psychiatry textbook/casebook QA	✗	✓	✓	✗	✗	✓	✗
MENTAT (Lamparth et al., 2025)	Real-world psychiatric decision-making	✗	✓	✓	✗	✓	✓	✗
Counseling and psychotherapy competence benchmarks								
CBT-Bench (Zhang et al., 2025)	CBT knowledge, cognitive model understanding, and response generation	✓	✗	✓	✗	✗	✓	✗
CounselingBench (Nguyen et al., 2025b)	Counseling competency examination	✓	✓	✓	✗	✓	✓	✗
CounselBench (Li et al., 2025)	Counseling response quality and adversarial safety	✗	✗	✓	✗	✓	✗	✗
MindEval (Pombal et al., 2025)	Multi-turn mental-health support interaction	✗	✗	✓	✗	✓	✗	✗
Safety, crisis, and ethical reasoning benchmarks								
EthicsMH (Kasu, 2025)	Ethical reasoning in mental-health AI	✗	✗	✗	✗	✓	✗	✗
Mental Health Crisis Benchmark (Arnaiz-Rodriguez et al., 2026)	Crisis detection and crisis response	✗	✓	✓	✗	✓	✗	✗
MIRAMIND (ours)	Reliable mental-health reasoning beyond answer accuracy	✓	✓	✓	✓	✓	✓	✓

distress into fixed labels, identify risk without exaggeration, select helpful responses, and explain their judgments in a way that remains grounded and useful.

Recent mental-health benchmarks have moved beyond early social-media classification, yet their coverage remains fragmented across clinical roles, task families, and risk scenarios. Psychiatric benchmarks evaluate DSM/ICD-grounded diagnosis, EHR-based disorder-level prediction, multi-turn psychiatric consultation, psychiatry QA, and real-world clinical decision-making (Song et al., 2026; Sun et al., 2026; Xu et al., 2026; Fouda et al., 2026; Lamparth et al., 2025); counseling-oriented benchmarks assess CBT assistance, counseling competencies, therapeutic response quality, adversarial counseling failures, and multi-turn mental-health support (Zhang et al., 2025; Nguyen et al., 2025b; Li et al., 2025; Pombal et al., 2025); and safety-focused resources examine crisis handling and ethical reasoning in high-risk scenarios (Arnaiz-Rodriguez et al., 2026; Kasu, 2025). While clinically valuable, these benchmarks usually test a specific role or endpoint rather than the full set of capabilities required for mental-health reasoning, as summarized in Table 1.

More importantly, they often emphasize final-answer correctness or response quality while leaving intermediate reasoning underspecified: a model may select the expected diagnosis, counseling strategy, or risk label without showing whether it followed a cautious, evidence-grounded path from the user's narrative to that conclusion. This is consequential because, in mental health, reasoning steps determine what the model treats as a symptom, risk signal, therapeutic need, or credible evidence. Current benchmarks therefore cannot fully distinguish clinically meaningful judgment from surface-pattern matching, motivating a benchmark that jointly evaluates heterogeneous mental-health reasoning capabilities and the reliability of the reasoning trajectories supporting final decisions.

To address this gap, we introduce MIRAMIND, a unified benchmark for evaluating whether LLMs can reason reliably in mental-health settings. Unlike prior evaluations that mainly test isolated knowledge, prediction, or final-answer selection, MIRAMIND is designed to examine how models progress from limited mental-health evidence to cautious, evidence-grounded judgments. We curate 13 datasets covering the main evidence forms encountered in mental-health use, including personal posts, counseling dialogues, psychiatry questions, systematic-review abstracts, and online mental-health claims. This contribution goes beyond aggregation: we refine existing resources, construct task-specific versions of medical and counseling datasets, and introduce new expert-annotated evidence summarization data. We organize these resources into six task families that make different parts of the judgment process testable, including interpreting problematic thought patterns, identifying possible mental-health conditions, selecting counseling responses, answering psychiatry questions, synthesizing clinical evidence, and verifying misleading claims. Finally, MIRAMIND evaluates not only whether a model's final answer matches a reference, but also whether its reasoning trajectory is usable, logically coherent, and informative. This allows MIRAMIND to distinguish models that reach the expected answer through surface-pattern matching

from those that justify it through grounded, reliable, and decision-relevant reasoning.

We evaluate 20 LLMs on MIRAMIND and first use the benchmark to characterize systematic failures in mental-health reasoning. The results reveal a restraint gap rather than a simple knowledge gap: strong models can often recognize familiar labels or produce plausible explanations, but they frequently overread weak cues, add generic clinical language, or give rationales that do not justify the strength of the conclusion. These failures become visible because MIRAMIND varies both the evidence source and the type of judgment required, and evaluates reasoning trajectories rather than only endpoints.

To test whether the failures identified by MIRAMIND can be reduced, we further train Mindora as a benchmark-guided post-training case study. Although Mindora is an 8B model, it obtains the highest overall score on MIRAMIND among all evaluated systems, surpassing strong baselines such as GPT-o4-mini, DeepSeek-R1, and Qwen-plus. Its more balanced reasoning trajectories further suggest that targeted optimization can improve evidence-proportional reasoning rather than merely increasing answer accuracy. Together, these results suggest that MIRAMIND may serve not only as an evaluation benchmark, but also as a useful research substrate for exploring more practically meaningful mental-health reasoning models.

Our contributions are summarized as follows: **(i)** We introduce MIRAMIND, a unified benchmark covering six task families and 13 datasets across appraisal, diagnosis, intervention, abstraction, and verification. **(ii)** We develop a reasoning-trajectory evaluation protocol that assesses usability, logical structure, and informational contribution beyond final-answer accuracy. **(iii)** We evaluate 20 LLMs and reveal a restraint gap in current models; as a benchmark-guided case study, Mindora shows that targeted post-training can improve evidence-proportional reasoning.

2 MIRAMIND Benchmark

MIRAMIND is designed to evaluate whether LLMs can perform reliable mental-health reasoning across the types of evidence, judgments, and explanations that arise in user-facing mental-health settings. Unlike benchmarks that focus on a single clinical role or final decision, MIRAMIND decomposes mental-health reasoning into five core aspects: appraisal, diagnosis, intervention, abstraction, and verification. These aspects correspond to distinct but complementary cognitive operations: interpreting problematic thought patterns, identifying possible mental-health conditions, selecting appropriate counseling actions, synthesizing clinical evidence, and verifying mental-health claims.

Our benchmark construction follows three principles. First, each task should correspond to a clinically meaningful reasoning operation rather than a generic NLP formulation. Second, each dataset should preserve realistic mental-health evidence whenever possible, including personal posts, counseling dialogues, psychiatry questions, systematic-review abstracts, and online mental-health claims. Third, evaluation should measure both task outcomes and the quality of the reasoning trajectory that supports the final decision. Table 2 summarizes the datasets, sources, annotation types, and metrics used in MIRAMIND. We provide details on dataset licenses, access conditions, intended use, personally identifying information, and sensitive-content handling in Appendix B.

2.1 Task Definition and Dataset Curation

Appraisal: Cognitive-Pattern Reasoning Appraisal tests whether a model can identify the maladaptive cognitive pattern expressed in a client statement. This ability is necessary because user-facing mental-health support often begins before formal diagnosis: the model must interpret how a person frames an event, whether the thought reflects overgeneralization, catastrophizing, personalization, mind-reading, or another cognitive distortion.

Task definition: Given a short situation, such as thought pair or a client utterance, the model selects the most appropriate cognitive-error label from the taxonomy defined by the source dataset. The task is formulated as multi-class classification and evaluated with Micro-F1.

Dataset curation: We instantiate this task using CognitiveReframing, PatternReframe, and Therapist Q&A. These datasets differ in source and taxonomy: CognitiveReframing combines simulated negative thoughts with Mental Health America self-reports annotated by trained mental-health professionals; PatternReframe contains crafted statements labeled by independent raters; Therapist Q&A provides naturally occurring distorted statements from therapist-client interactions. Because the taxonomies are not identical, we pre-

Table 2: Task and dataset summary of MIRAMIND. Datasets with $_M$ are processed in this work. *PSRS is newly curated and expert-annotated in this work. The *recall* metric for PSRS measures the coverage of expert-annotated scoring points.

Dataset	Train/Valid/Test	Source	Annotation	Input / Output	Metric
Cognitive Error Identification					
CognitiveReframing (Sharma et al., 2023)	694/173/284	synthetic, MHA	Human	negative thought or self-report → cognitive-error label	Micro-F1
PatternReframe (Maddela et al., 2023)	1214/661/300	synthetic	Human	distorted statement → cognitive-error label	Micro-F1
Therapist Q&A (Shreevastava and Foltz, 2021)	1508/409/507	counseling dialogues (real)	Human	therapist–client utterance → cognitive-error label	Micro-F1
Mental Health Condition Detection					
DepSign (Poświata and Perełkiewicz, 2022)	1113/300/600	Reddit	weak supervision	Reddit post → depression severity	Micro-F1
SWMH (Ji et al., 2022)	1097/300/500	Reddit	weak supervision	Reddit post → mental-health condition	Micro-F1
T-SID (Ji et al., 2022)	1080/300/942	Twitter	weak supervision	Twitter post → disorder/risk category	Micro-F1
Counseling Strategy Formulation					
PsyDTCorpus$_M$	1105/300/-	counseling dialogues (syn)	Human	client case summary → counseling strategy set	Jaccard
AnnoMI$_M$	-/-/133	counseling dialogues (real)	Human	client case summary → counseling strategy set	Jaccard
Psychiatry QA					
MHQA (Racha et al., 2025)	1123/300/717	PubMed	Human/LLM	mental-health QA item → answer option	Micro-F1
MedQA$_M$	415/58/121	Medical exams	Human	psychiatry-related exam vignette → answer option	Micro-F1
MedMCQA$_M$	287/120/446	Medical exams	Human	psychiatry-related exam question → answer option	Micro-F1
PubMedQA$_M$	-/-/89	PubMed	Human	psychiatry-related research QA → yes/no/maybe label	Micro-F1
Psychiatry Systematic Review Summarization					
*PSRS	-/-/108	Cochrane Library	Human	systematic-review abstract → Main Results summary	Recall
Mental Health Misinformation Identification					
MentalMisinfo (Nguyen et al., 2025a)	123/31/130	YouTube, BitChute	Human	online mental-health claim or transcript → 0/1 label	Macro-F1

serve each dataset's original label set and prompt definitions rather than forcing a lossy unified taxonomy. This design evaluates whether models can adapt to related but non-identical appraisal schemas. Further details are in Appendix B.5.

Diagnosis: Mental-Condition Reasoning Diagnosis evaluates whether a model can infer possible mental-health conditions from short, subjective user narratives while avoiding over-pathologizing. This differs from psychiatric EHR diagnosis: inputs are informal posts or tweets, evidence is incomplete, and the correct behavior is calibrated condition detection rather than exhaustive clinical labeling.

Task definition: Given a user post, the model predicts the most appropriate condition or severity label according to the dataset-specific label space. DepSign evaluates depression severity, while SWMH and T-SID evaluate multiple mental-health condition categories. We evaluate classification performance with Micro-F1.

Dataset curation: We use DepSign, SWMH, and T-SID, covering Reddit and Twitter-style language. These datasets provide naturalistic mental-health expressions, including noisy, incomplete, and emotionally loaded narratives. We retain the original task labels to preserve dataset-specific clinical assumptions, while standardizing prompts and answer formats across datasets.

Intervention: Therapeutic-Action Reasoning Intervention tests whether a model can decide what therapeutic action is appropriate in a counseling context. This is distinct from generating an empathetic reply: the model must identify the counseling function needed by the client situation, such as clarification, reflection of feeling, summarizing, confrontation, encouragement, interpretation, or guidance.

Task definition: Given a concise client case derived from a counseling dialogue, the model selects one or more appropriate counseling strategies from 13 commonly taught intervention types. Because multiple strategies may be acceptable for a single counseling turn, we evaluate predictions with Jaccard similarity against the reference strategy set.

Dataset construction: We construct two processed datasets, PsyDTCorpus$_M$ and AnnoMI$_M$. Client utterances are compressed into concise case summaries, and counselor responses are annotated to extract strategy labels as reference answers. All constructed items are manually reviewed by an expert counselor with over ten years of clinical experience. To avoid evaluating on the same distri-

bution used for training, PsyDTCorpus$_M$ is used only for training because of its partially synthetic nature, whereas AnnoMI$_M$ is reserved for testing and is derived from real motivational-interviewing demonstrations. This source separation enables a cross-source test of whether intervention reasoning transfers beyond the training distribution. Construction details are in Appendix B.6.

Multi-step: Psychiatry Question Answering Multi-step reasoning evaluates whether models can integrate symptom interpretation, diagnostic inference, intervention selection, and evidence-based medical knowledge within a single case. This setting approximates clinician-side reasoning more closely than single-label appraisal or condition detection.

Task definition: Given a psychiatry-related question or clinical vignette with candidate answers, the model selects the correct answer and provides a reasoning trajectory. We evaluate final-answer correctness with Micro-F1.

Dataset curation: We use MHQA and construct psychiatry-focused subsets from MedQA, MedMCQA, and PubMedQA, denoted with the $_M$ suffix[2]. For the broad medical QA datasets, we manually extract items involving psychiatric disorders, psychopharmacology, therapeutic management, mental-health symptoms, or psychiatric research evidence. This processing isolates mental-health reasoning from general medical QA and ensures that the task evaluates psychiatry-relevant multi-step judgment rather than broad biomedical knowledge alone.

Abstraction: Evidence-Based Psychiatry Summarization Abstraction evaluates whether a model can synthesize psychiatric research evidence into clinically meaningful conclusions. This ability is important because mental-health users and practitioners often encounter claims grounded in studies, reviews, or guidelines, where the core challenge is not choosing a label but determining what the evidence actually supports.

Task definition: Given a structured abstract from a psychiatry-related Cochrane systematic review, the model generates the Main Results summary, focusing on the direction, magnitude, and certainty of the reported findings. Unlike extractive summarization, the task requires evidence abstraction: models must convert statistical and methodological details into a concise clinical interpretation.

Dataset construction: We introduce Psychiatry Systematic Review Summarization (PSRS), a newly curated and expert-annotated generation task. We collect 115 structured abstracts from the Cochrane Library in the mental-health domain, preserving review identifiers for traceability and using only publicly available abstract text. For annotation, LLM-generated Main Results summaries are evaluated by two psychiatrists using a 0–2 rubric: accurate and complete, partially correct or incomplete, and unsupported or substantially incorrect. A pilot calibration phase is used to align judgment thresholds, and a cross-check subset is used for quality assurance. The benchmark reports recall over expert-annotated scoring points, measuring whether a model covers the clinically relevant evidence conclusions. Construction and annotation details are in Appendix B.8.

Verification: Mental-Health Misinformation Detection Verification evaluates whether a model can judge the accuracy of mental-health information in noisy online environments. This aspect is necessary because many real user interactions involve asking whether advice, claims, or social-media content should be trusted.

Task definition: Given video-derived textual content, including titles, hashtags, transcripts, and contextual metadata when available, the model predicts whether the mental-health claim is accurate or misleading. We evaluate this binary classification task with Macro-F1 to account for class imbalance.

Dataset curation: We use MentalMisinfo, which contains mental-health-related video content from platforms such as YouTube Shorts and BitChute. The materials are filtered, transcribed, and manually annotated as accurate or misleading. This task extends MIRAMIND beyond clinical decision-making by testing evidence-based verification under informal and persuasive online discourse.

2.2 Reasoning Trajectory Evaluation

MIRAMIND is designed to evaluate reasoning rather than only endpoints and each model output includes a final answer and a reasoning trajectory. We score the final answer with task-specific metrics and separately assess trajectory quality along three dimensions. Usability measures whether the

[2]A brief introduction of original datasets are provided in Appendix B.7.

reasoning is clear and concise enough to be followed; logical structure measures whether reasoning units are coherently connected to the input and conclusion; informational contribution measures whether each step adds decision-relevant information rather than restating prior content. These dimensions correspond to common failure modes in mental-health reasoning: overuse of clinical terminology, unsupported inferential leaps, contradictions, irrelevant steps, and verbose but uninformative explanations.

Operationally, we evaluate trajectory quality with Terminology Parsimony, Step Adequacy, Logical Coherence, Irrelevance Rate, Contradiction Rate, and Informativeness. We use an LLM-as-judge protocol to segment reasoning into reasoning content units and assign relation labels between each unit, the input context, and prior units. To validate the automatic protocol, we compare judge scores with expert human annotations on sampled trajectories; details, rubrics, and agreement results are reported in Appendix C.

3 Mindora

To examine whether the reasoning failures exposed by MIRAMIND can motivate targeted model improvement, we train Mindora as a benchmark-guided post-training case study based on Qwen3-8B. The goal is not to introduce a standalone clinical system, but to test whether a compact model can be optimized toward the type of evidence-proportional and internally consistent reasoning measured by MIRAMIND.

Mindora follows a trajectory-oriented post-training pipeline with three main components. First, we construct supervision from hard cases where a base model fails under zero-shot prompting, since these instances are less likely to be solvable through surface cues and provide higher-signal reasoning examples. Second, we generate and rewrite reasoning trajectories into concise, non-backtracking explanations with explicit intermediate steps and a final conclusion aligned with the answer. Third, we train the model with a hybrid SFT–RL objective using a consistency-aware reward that combines format validity, length validity, internal consistency, and task-specific correctness. This design encourages the model not merely to reason more, but to keep its intermediate judgments grounded, concise, and consistent with the final decision.

The full training pipeline, including hard-case filtering, verifier-guided trajectory search, structured rewriting, and reward construction, is provided in Appendix D. We use Mindora in this paper as a case study showing that MIRAMIND can serve not only as an evaluation benchmark, but also as a research substrate for exploring more practically meaningful mental-health reasoning models.

4 Experiments

4.1 Experimental Setup

We evaluate 20 LLMs on MIRAMIND to examine three questions: (i) whether general-purpose proprietary and open-weight models can handle heterogeneous mental-health reasoning tasks; (ii) whether reasoning-oriented models improve reliability beyond final-answer accuracy; and (iii) whether mental-health-specific post-training leads to more evidence-proportional reasoning. Accordingly, the evaluated models cover four groups: proprietary general-purpose and reasoning models, large open-weight models, compact open-weight models, and mental-health-oriented baselines.

Models. We evaluate a broad set of LLMs, including proprietary systems, open-weight models across multiple scales, reasoning-oriented variants, and mental-health-specific baselines. This grouping allows us to compare the effects of model scale, general reasoning post-training, and domain-specific mental-health adaptation under the same benchmark. Model details are provided in Appendix E.1.

Evaluation protocol. All general-purpose models except Qwen3-8B are accessed through official APIs, while domain-specific models and Qwen3-8B are deployed on a single NVIDIA A800-SXM4-80GB GPU. We keep decoding parameters at their default values unless otherwise specified. All models are prompted to produce both a final answer and a structured reasoning trajectory, enabling evaluation of task accuracy and trajectory reliability under a unified protocol. To examine whether the structured prompting format itself explains the observed gains, we additionally conduct a few-shot ICL analysis in Appendix G.4.

4.2 Main Benchmark Results

Table 3 summarizes model performance across the six task families in MIRAMIND. Overall,

Table 3: Main benchmark results on MIRAMIND. Results are grouped by task family. Avg_1, Avg_2, and Avg_4 denote average scores for appraisal, diagnosis, and multi-step psychiatry QA, respectively; Avg_{all} is the average across all evaluated datasets. Higher scores indicate better performance. Mindora w/o S1/S2/S3 are ablation variants, where S1 denotes hard-case filtering, S2 denotes structured rewriting, and S3 denotes the consistency-aware reward.

Model	Cognitive Reframing	Pattern Reframe	Therapist Q&A	Avg_1	DepSign	SWMH	T-SID	Avg_2	AnnoMI_M	MHQA	MedQA_M	MedMCQA_M	PubMedQA_M	Avg_4	PSRS	Mental Misinfo	Avg_{all}
close-source models																	
GPT-o4-mini (OpenAI, 2025)	0.7002	0.6607	0.4468	0.6026	0.3607	0.7374	0.7515	0.6165	0.2086	0.5462	**0.8925**	0.8102	0.7118	0.7402	0.8650	**0.7781**	0.6515
GPT-4o (OpenAI, 2024)	0.6791	0.6667	0.4543	0.6000	0.4064	0.7500	0.7114	0.6226	0.3358	0.4531	0.6138	0.7936	0.6949	0.6389	0.9065	0.6894	0.6273
Deepseek-R1 (DeepSeek-AI, 2025a)	**0.7516**	**0.7069**	0.4472	0.6352	0.4085	0.7624	0.7218	0.6309	0.2020	0.4984	0.8608	0.8988	0.7256	0.7459	0.9037	0.5689	0.6505
Deepseek-V3 (DeepSeek-AI, 2025b)	0.6540	0.6755	0.4131	0.5809	0.4211	0.7715	0.7581	0.6502	0.2140	0.4768	0.8184	0.8696	0.6848	0.7124	0.9296	0.6156	0.6386
Qwen-plus (Yang et al., 2025)	0.6821	0.6395	0.4057	0.5758	0.4043	0.7624	0.7662	0.6443	0.1866	0.4795	0.8242	0.8895	0.7199	0.7283	0.9048	0.6018	0.6387
QwQ-plus (Qwen et al., 2025)	0.6698	0.6871	0.3698	0.5756	0.4085	0.7547	0.7330	0.6321	0.1848	0.4201	0.6405	0.8354	0.7226	0.6546	0.9335	0.4838	0.6034
70B+ open-source models																	
LLaMA-4 (AI, 2025)	0.6588	0.6486	0.4031	0.5702	0.3914	0.7310	0.7884	0.6369	0.2066	0.4684	0.7414	0.8900	0.7239	0.7059	0.7634	0.6158	0.6178
LLaMA-3.3-70B (Grattafiori et al., 2024)	0.6444	0.6301	0.3566	0.5437	0.4148	0.7163	0.7157	0.6156	0.2023	0.4148	0.7712	0.8904	0.7299	0.7016	0.6318	0.6894	0.6006
dsdistill-LLaMA-70B (DeepSeek-AI, 2025a)	0.6667	0.6667	0.3905	0.5746	0.4000	0.7516	0.6953	0.6156	0.2367	0.4404	0.7097	0.7957	0.7452	0.6727	0.8914	0.6814	0.6209
Qwen2.5-72B (Qwen et al., 2025)	0.6852	0.6207	0.3930	0.5663	0.3806	0.7639	0.8136	0.6527	0.2796	0.4022	0.7449	0.8402	0.7229	0.6775	**0.9555**	0.6091	0.6316
32B open-source models																	
dsdistill-Qwen32B (DeepSeek-AI, 2025a)	0.6540	0.6696	0.3539	0.5592	0.3979	0.7358	0.7373	0.6237	0.2178	0.4010	0.6697	0.7795	0.7229	0.6433	0.9225	0.6177	0.6061
Qwen3-32B (Yang et al., 2025)	0.6247	0.6395	0.4255	0.5632	0.3806	0.7685	0.8038	0.6510	0.2016	0.4355	0.7415	0.8206	0.7083	0.6765	**0.9333**	0.5025	0.6143
QwQ-32B (Qwen et al., 2025)	0.6791	0.6842	0.3905	0.5846	0.4085	0.7500	0.7296	0.6294	0.1749	0.4305	0.6831	0.8470	0.7375	0.6745	0.9296	0.4642	0.6084
14B open-source models																	
dsdistill-Qwen14B (DeepSeek-AI, 2025a)	0.6635	0.5981	0.3404	0.5340	0.4169	0.7421	0.7750	0.6447	0.2470	0.3823	0.6198	0.6660	0.7497	0.6045	0.8641	0.5795	0.5880
Qwen3-14B (Yang et al., 2025)	0.6540	0.6486	0.4820	0.5949	0.3871	0.7484	0.7540	0.6298	0.1680	0.4215	0.7047	0.8102	0.6882	0.6562	0.8155	0.6600	0.6109
7∼8B open-source models																	
LLaMA3.1-8B (Grattafiori et al., 2024)	0.5871	0.5714	0.1957	0.4514	0.4476	0.7358	0.7581	0.6472	0.2147	0.2771	0.5791	0.7925	0.7096	0.5896	0.6805	0.6703	0.5553
dsdistill-LLaMA-8B (DeepSeek-AI, 2025a)	0.5906	0.5507	0.2396	0.4603	0.4190	0.7294	0.7490	0.6325	0.2001	0.3321	0.4430	0.6070	0.7587	0.5352	0.8157	0.7617	0.5536
EmoLLM (Team, 2024)	0.6180	0.5472	0.3322	0.4991	0.4334	0.7229	0.7931	0.6498	0.1408	0.3373	0.4311	0.5749	0.6735	0.5042	0.7906	0.6346	0.5407
Psyche-R1 (Dai et al., 2025)	0.5013	0.5222	0.2021	0.4085	0.4169	0.6877	0.7962	0.6336	0.2424	0.5872	0.6989	0.8401	0.7194	0.7114	0.8164	0.6954	0.5943
Qwen3-8B (Yang et al., 2025)	0.5941	0.6395	0.4057	0.5464	0.4169	0.7358	0.7515	0.6347	0.1489	0.4009	0.5824	0.7927	0.6899	0.6165	0.8115	0.4773	0.5729
Mindora_{SFT}	0.5693	0.6207	0.4472	0.5457	0.4887	0.7437	0.7597	0.6640	0.4193	0.5830	0.7263	0.8535	0.7376	0.7251	0.7764	0.5512	0.6367
Mindora_{RL}	0.6112	0.6301	0.4279	0.5564	0.3849	0.7358	0.7140	0.6116	0.1456	0.5844	0.8020	**0.9001**	0.7808	0.7668	0.8015	0.4493	0.6129
Mindora_{SFT+RL}	0.5975	0.6207	0.4566	0.5583	**0.5240**	0.7715	**0.8140**	0.7032	0.3803	0.5996	0.7839	0.8593	0.8212	0.7660	0.7159	0.5681	0.6548
Mindora w/o S1	0.6912	0.6456	**0.5262**	0.6210	0.4252	0.7013	0.8106	0.6457	**0.4719**	0.5927	0.7716	0.8737	0.8133	0.7628	0.7982	0.6286	0.6731
Mindora w/o S2	0.6044	0.6270	0.5240	0.5851	0.4576	0.7730	0.8225	0.6844	0.4430	0.5486	0.7330	0.8416	0.6308	0.6885	0.7955	0.6038	0.6465
Mindora w/o S3	0.6729	0.6637	0.5240	0.6202	0.4868	0.7685	0.7806	0.6786	0.4400	0.6145	0.7778	0.8695	**0.8516**	**0.7784**	0.6948	0.6796	0.6788
Mindora	0.7293	0.6842	0.5088	**0.6408**	0.4655	**0.7760**	0.8030	0.6815	0.4016	**0.6317**	0.7590	0.8535	0.8442	0.7721	0.8379	0.7178	**0.6933**

Table 4: Multi-dimensional evaluation of reasoning-trajectory quality across models.

Model	Accuracy	Terminology Parsimony ↑	Step Adequacy ↑	Avg Steps↓	Logical Coherence ↑	Irrelevance Rate ↓	Contradiction Rate ↓	Informativeness ↑
GPT-o4-mini	0.4593	0.8750	**0.9042**	5.1681	0.8875	0.0064	0.0101	0.7500
Deepseek-R1	0.5102	0.7458	0.6750	7.7480	1.0542	0.0138	0.0410	0.9958
Qwen-plus	0.4555	0.6208	0.4500	9.2035	0.8708	0.0306	0.0263	0.9042
LLaMA-4	0.4882	**0.9375**	0.9333	5.0826	0.9292	0.0092	**0.0049**	0.7917
Psyche-R1	0.4396	0.7250	0.6458	5.6422	0.8708	0.0074	0.0233	0.7667
Qwen3-8B	0.4614	0.6792	0.4208	7.8696	0.8625	0.0286	0.0214	0.9083
Mindora w/o S1	0.4969	0.8458	0.6958	6.2377	0.9625	**0.0060**	0.0163	0.8875
Mindora w/o S2	0.5223	0.5542	0.1917	10.0391	0.9917	0.0160	0.0383	**1.0042**
Mindora w/o S3	0.5199	0.8750	0.8250	5.2248	1.0042	0.0102	0.0115	0.9250
Mindora	**0.5467**	0.9042	0.8625	**5.0444**	**1.0625**	0.0065	0.0154	0.9917

the results show that mental-health reasoning remains challenging even for strong proprietary and reasoning-oriented models. Performance varies substantially across task families: models that perform well on structured psychiatry QA or misinformation verification do not necessarily perform well on narrative-based appraisal, diagnosis, or counseling intervention. This confirms that MIRAMIND does not measure a single generic capability, but exposes distinct reasoning requirements across evidence interpretation, clinical judgment, response selection, evidence synthesis, and verification.

Within this benchmark, Mindora achieves the highest overall average score (0.6933), outperforming strong baselines including GPT-o4-mini (0.6515), DeepSeek-R1 (0.6505), and Qwen-plus (0.6387). Importantly, its advantage is not concentrated in a single task family. Compared with the Qwen3-8B backbone, Mindora improves from 0.5729 to 0.6933 overall, with gains across appraisal, diagnosis, intervention, multi-step reasoning, abstraction, and verification. This suggests that benchmark-guided post-training improves cross-task robustness rather than only fitting one dataset.

4.3 Key Findings of MIRAMIND

Finding 1: Scale alone does not solve mental-health reasoning. Across open-weight models, increasing parameter count does not lead to uniform gains across MIRAMIND. Models rang-

ing from 14B to 70B achieve similar average scores around 0.6, while several reasoning-oriented variants differ only marginally from their chat-oriented counterparts. This suggests that the benchmark is not simply measuring general language capacity or generic reasoning strength. Instead, mental-health reasoning requires calibrated interpretation of incomplete, subjective evidence.

Finding 2: Reasoning-oriented models remain vulnerable to restraint failures. Strong reasoning models often perform competitively on structured QA tasks, but they do not consistently dominate narrative-based diagnosis, appraisal, or intervention tasks. This supports our central observation that mental-health reasoning is not only a matter of producing longer or more explicit reasoning. Models must decide how much the available evidence actually supports, avoiding overreading weak cues or inserting generic clinical templates.

Finding 3: Different task families expose different model strengths. The group-wise rank analysis in Appendix E.3 further confirms that model performance is uneven across task families. Mindora achieves the best overall average rank (4.92) while remaining strong across appraisal, diagnosis, and multi-step reasoning, whereas models such as DeepSeek-R1 and GPT-o4-mini are competitive on selected task families but less balanced overall. This indicates that MIRAMIND captures heterogeneous mental-health reasoning requirements rather than rewarding dominance on one high-scoring dataset.

4.4 Mindora as a Benchmark-Guided Case Study

As a benchmark-guided case study, Mindora tests whether the failures exposed by MIRAMIND can be reduced through targeted post-training. Pure SFT (0.6367), pure RL (0.6129), and naive SFT→RL (0.6548) all underperform the full model (0.6933), showing that imitation and reward-based refinement are complementary. Removing S1, S2, and S3 reduces Avg_{all} to 0.6731, 0.6465, and 0.6788, respectively; the largest drop comes from removing structured rewriting, suggesting that learnable non-backtracking trajectories are especially important for transferring reasoning patterns into an 8B model. Detailed analyses are in Appendix G.

4.5 Reasoning-Trajectory Evaluation

We evaluate trajectory quality on a 5% sample using the protocol in Section 2.2; incorrect final answers receive zero trajectory scores. Table 4 shows that Mindora achieves the strongest overall balance: highest subset accuracy (0.5467), best logical coherence (1.0625), near-best informativeness (0.9917), and low irrelevance/contradiction rates (0.0065/0.0154). Compared with Qwen3-8B, it improves Terminology Parsimony from 0.6792 to 0.9042 and Step Adequacy from 0.4208 to 0.8625 while reducing average steps from 7.8696 to 5.0444. Compared with DeepSeek-R1, it keeps similar informativeness but much lower contradiction rate (0.0154 vs. 0.0410), confirming that trajectory-level evaluation captures reliability differences beyond final-answer accuracy.

5 Conclusion

We introduced MIRAMIND, a unified benchmark for evaluating reliable mental-health reasoning beyond final-answer accuracy. Covering six task families and 13 datasets across appraisal, diagnosis, intervention, abstraction, and verification, MIRAMIND tests whether LLMs can move from limited mental-health evidence to cautious, evidence-grounded judgments. Our evaluation of 20 LLMs reveals a restraint gap: strong models often produce plausible answers, but show uneven performance across task families and may overread weak cues, rely on generic clinical language, or provide rationales that do not justify the strength of the conclusion. To capture these failures, MIRAMIND further introduces a reasoning-trajectory evaluation protocol that measures usability, logical structure, and informational contribution, showing that final-answer correctness alone is insufficient for assessing mental-health reasoning reliability. As a benchmark-guided case study, we train Mindora, an 8B model that achieves the best overall performance on MIRAMIND and produces more balanced reasoning trajectories. These findings suggest that MIRAMIND can support both systematic evaluation and future research on more evidence-proportional, interpretable, and practically meaningful mental-health reasoning models, while real-world deployment remains outside the scope of this benchmark and requires clinical validation, safety monitoring, and human oversight.

Limitations

MIRAMIND provides a controlled and reproducible testbed for reliable mental-health reasoning, but it does not cover the full scope of mental-health practice. The current benchmark focuses on single-instance reasoning over posts, counseling dialogues, psychiatry questions, systematic-review abstracts, and online claims, while longitudinal therapy, multi-session risk monitoring, clinician–patient collaboration, culturally situated care, and multimodal behavioral signals remain future directions. In addition, MIRAMIND combines newly curated resources with task-adapted existing datasets, so some tasks inherit source-specific assumptions, such as weak supervision in social-media condition detection or strategy-level abstraction in counseling intervention. Our reasoning-trajectory protocol operationalizes usability, logical structure, and informational contribution and is validated against expert annotations, but broader expert review and cross-cultural validation would further strengthen the benchmark.

Ethical Statement

Mindora is used only as a benchmark-guided post-training case study. Strong benchmark performance does not imply clinical safety, treatment validity, deployment readiness, or regulatory compliance; real-world mental-health applications require clinical validation, human oversight, safety monitoring, and safeguards against unsupported advice, over-pathologizing, and crisis mismanagement.

References

Meta AI. 2025. Introducing llama 4: Advancing multimodal intelligence.

Adrian Arnaiz-Rodriguez, Miguel Baidal, Erik Derner, Jenn Layton Annable, Mark Ball, Mark Ince, Elvira Perez Vallejos, and Nuria Oliver. 2026. Between help and harm: An evaluation of mental health crisis handling by LLMs. *JMIR Mental Health*, forthcoming:88435. In press.

Franziska Burger, Mark A Neerincx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: extracting schemas from thought records. *PloS one*, 16(10):e0257832.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Xun Yang, and Meng Wang. 2025. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. *Preprint*, arXiv:2508.10848.

DeepSeek-AI. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Aya E. Fouda, Abdelrahman A. Hassan, Radwa J. Hanafy, and Mohammed E. Fouda. 2026. Psychiatrybench: A multi-task benchmark for llms in psychiatry. *npj Digital Medicine*, 9(320).

Aaron Grattafiori, Abhimanyu Dubey, and et al Abhinav Jauhri. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34(13):10309–10319.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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)*, pages 2567–2577.

Sai Kartheek Reddy Kasu. 2025. EthicsMH: A pilot benchmark for ethical reasoning in mental health AI. *arXiv preprint arXiv:2509.11648*.

Max Lamparth, Declan Grabb, Amy Franks, Scott Gershan, Kaitlyn N. Kunstman, Aaron Lulla, Monika Drummond Roots, Manu Sharma, Aryan Shrivastava, Nina Vasan, and Colleen Waickman. 2025. Moving beyond medical exam questions: A clinician-annotated dataset of real-world tasks and ambiguity in mental healthcare. *arXiv preprint arXiv:2502.16051*.

Yahan Li, Jifan Yao, John Bosco S. Bunyi, Adam C. Frank, Angel Hwang, and Ruishan Liu. 2025. Counselbench: A large-scale expert evaluation and adversarial benchmark of large language models in mental health counseling. *arXiv preprint arXiv:2506.08584*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe.

2024. Let's verify step by step. In *International Conference on Learning Representations*.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13641–13660. Association for Computational Linguistics.

Viet Cuong Nguyen, Mini Jain, Abhijat Chauhan, Heather Jamie Soled, Santiago Alvarez Lesmes, Zihang Li, Michael L Birnbaum, Sunny X Tang, Srijan Kumar, and Munmun De Choudhury. 2025a. Supporters and skeptics: Llm-based analysis of engagement with mental health (mis) information content on video-sharing platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1329–1345.

Viet Cuong Nguyen, Mohammad Taher, Dongwan Hong, Vinicius Konkolics Possobom, Vibha Thirunellayi Gopalakrishnan, Ekta Raj, Zihang Li, Heather J. Soled, Michael L. Birnbaum, Srijan Kumar, and Munmun De Choudhury. 2025b. Do large language models align with core mental health counseling competencies? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7503–7526, Albuquerque, New Mexico. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert O. Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2024. Gpt-4o system card.

OpenAI. 2025. Openai o3 and o4-mini system card.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

José Pombal, Maya D'Eon, Nuno M. Guerreiro, Pedro Henrique Martins, António Farinhas, and Ricardo Rei. 2025. Mindeval: Benchmarking language models on multi-turn mental health support. *arXiv preprint arXiv:2511.18491*.

Rafał Poświata and Michał Perełkiewicz. 2022. OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282, Dublin, Ireland. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Suraj Racha, Prashant Joshi, Anshika Raman, Nikita Jangid, Mridul Sharma, Ganesh Ramakrishnan, and Nirmal Punjabi. 2025. Mhqa: A diverse, knowledge intensive mental health question answering challenge for language models. *arXiv preprint arXiv:2502.15418*.

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

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

Gopendra Vikram Singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal llm-based detection and reasoning framework. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 22546–22570.

Hoyun Song, Migyeong Kang, Jisu Shin, Jihyun Kim, Chanbi Park, Hangyeol Yoo, Jihyun An, Alice Oh, Jinyoung Han, and KyungTae Lim. 2026. Mentalbench: A benchmark for evaluating psychiatric diagnostic capability of large language models. *arXiv preprint arXiv:2602.12871*.

Xiao Sun, Yuming Yang, Junnan Zhu, Jiang Zhong, Xinyu Zhou, and Kaiwen Wei. 2026. Mentalseekdx: Towards progressive hypothetico-deductive reasoning for real-world psychiatric diagnosis. *arXiv preprint arXiv:2602.03340*.

EmoLLM Team. 2024. Emollm: Reinventing mental health support with large language models. `https://github.com/SmartFlowAI/EmoLLM`.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Mengxi Xiao, Ben Liu, He Li, Jimin Huang, Qianqian Xie, Xiaofen Zong, Mang Ye, and Min Peng. 2025. Moodangels: A retrieval-augmented multi-agent framework for psychiatry diagnosis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.

Shihao Xu, Tiancheng Zhou, Jiatong Ma, Yanli Ding, Yiming Yan, Ming Xiao, Guoyi Li, Haiyang Geng, Yunyun Han, Jianhua Chen, and Yafeng Deng. 2026. Lingxidiagbench: A multi-agent framework for benchmarking llms in chinese psychiatric consultation and diagnosis. *arXiv preprint arXiv:2602.09379*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. CBT-bench: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

A Related Work

A.1 Psychiatric Diagnosis and Clinical Decision-Making Benchmarks

Recent benchmarks have advanced the evaluation of LLMs from social-media mental-health classification toward clinically grounded psychiatric reasoning. MentalBench evaluates DSM-5-grounded diagnostic decision-making using a psychiatrist-built knowledge graph and controlled synthetic cases with varying information completeness and diagnostic complexity (Song et al., 2026). MentalDx Bench targets real-world disorder-level diagnosis from de-identified EHRs annotated under ICD-11 guidelines, while MentalSeek-Dx further models progressive hypothetico-deductive diagnostic reasoning (Sun et al., 2026). LingxiDiagBench extends this direction to Chinese psychiatric consultation by evaluating both static diagnostic inference and dynamic multi-turn diagnosis over EMR-aligned synthetic dialogues (Xu et al., 2026). Beyond diagnosis, PsychiatryBench covers multiple psychiatry-focused QA tasks, including diagnosis, treatment planning, longitudinal follow-up, management planning, and sequential case analysis (Fouda et al., 2026), while MENTAT evaluates real-world ambiguity across diagnosis, treatment, monitoring, triage, and documentation using clinician-annotated decision scenarios (Lamparth et al., 2025).

These benchmarks provide important clinical grounding, but they mainly evaluate LLMs as psychiatrists or mental-health practitioners making clinical decisions. In contrast, MIRAMIND treats mental-health reasoning as a broader set of cognitive operations, including appraisal, diagnosis, intervention, abstraction, and verification. This broader scope reflects user-facing mental-health interactions, where models must often interpret ambiguous narratives, identify risk without over-pathologizing distress, select helpful responses, summarize clinically relevant content, and assess the credibility of mental-health advice before any externally verifiable outcome is available.

A.2 Counseling and Psychotherapy Competence Benchmarks

Recent benchmarks also evaluate LLMs in counseling and psychotherapy-oriented settings. CBT-Bench assesses cognitive behavioral therapy assistance across basic CBT knowledge, cognitive model understanding, and therapeutic response generation (Zhang et al., 2025). CounselingBench evaluates 22 general-purpose and medical-finetuned LLMs on five core mental-health counseling competencies using NCMHCE-derived questions (Nguyen et al., 2025b). CounselBench provides a large-scale expert evaluation and adversarial benchmark for single-turn mental-health counseling, with 100 mental-health professionals rating model and human responses and adversarial questions designed to expose counseling-specific failures (Li et al., 2025). MindEval further evaluates LLMs in realistic multi-turn mental-health support conversations using simulated patients and automatic evaluation validated against licensed clinical psychologists (Pombal et al., 2025).

While these benchmarks are clinically valuable, they primarily evaluate counseling behavior or therapeutic response quality. MIRAMIND addresses a different evaluation target: whether LLMs can perform reliable mental-health reasoning across heterogeneous task families. Rather than focusing only on whether a model behaves like a counselor or therapist, MIRAMIND evaluates clinically aligned reasoning over appraisal, diagnosis, intervention, abstraction, and verification, together with reasoning qualities such as grounding, coherence, conciseness, hallucination avoidance, and internal consistency. In this sense, MIRAMIND complements psychotherapy competence benchmarks by testing the broader reasoning operations that underlie mental-health interpretation and assessment beyond therapeutic dialogue alone.

A.3 Safety, Crisis, and Ethical Reasoning Benchmarks

Safety-oriented benchmarks evaluate whether LLMs can handle high-risk mental-health scenarios. The Mental Health Crisis Benchmark evaluates crisis classification and response quality across categories such as suicidal ideation, self-harm, anxiety crisis, substance abuse or withdrawal, risk-taking behavior, and violent thoughts (Arnaiz-Rodriguez et al., 2026). EthicsMH targets ethical reasoning in therapeutic and psychiatric contexts, with scenarios involving confidentiality, autonomy, beneficence, fairness, and bias, together with structured decision options, expert-aligned reasoning, expected behavior, real-world impact, and stakeholder perspectives (Kasu, 2025). CounselBench-Adv simi-

larly stress-tests LLMs with expert-authored adversarial counseling questions designed to elicit counseling-specific failures (Li et al., 2025).

These benchmarks are indispensable for auditing high-risk failures, but they usually target specific failure domains such as crisis response, ethical dilemmas, or adversarial counseling. MIRAMIND complements this line of work by evaluating general mental-health reasoning reliability across both ordinary and high-stakes tasks. Rather than focusing only on whether a model responds safely to predefined risk scenarios, MIRAMIND evaluates whether it can reason coherently, cautiously, and consistently across appraisal, diagnosis, intervention, abstraction, and verification.

B MIRAMIND Benchmark Details

B.1 Dataset Licenses and Access Conditions

Table 5 summarizes the licenses and access conditions of the datasets used in MIRAMIND. For datasets without an explicit public license, we follow the original dataset access terms and cite the corresponding source. For restricted or platform-derived datasets, we do not redistribute raw texts and release only processed metadata, prompts, labels, or evaluation scripts where permitted.

B.2 Usage Declaration

MIRAMIND is intended for academic research on evaluating and analyzing mental-health reasoning in LLMs. We use existing datasets only for benchmark construction, model evaluation, and benchmark-guided post-training within research contexts. For datasets with explicit licenses or access conditions, we follow the original terms. For restricted, platform-derived, or license-unspecified resources, we do not assume permission to redistribute raw data and release only derived prompts, metadata, labels, evaluation scripts, or annotations where permitted. Processed datasets marked with $_M$ are intended for research evaluation and training under the access conditions of their original sources. PSRS is intended for evaluating evidence-grounded psychiatry summarization; for Cochrane-derived materials, we release review identifiers, prompts, expert annotations, and scoring metadata where permitted, but do not redistribute full copyrighted abstract text unless allowed by the original terms. Licenses of datasets are detailed in Appendix B.1.

B.3 Personally Identifying Information and Sensitive Content

MIRAMIND includes mental-health-related posts, counseling dialogues, online claims, and psychiatry questions, which may contain sensitive personal experiences or stigmatizing language. We therefore treat all user-generated and dialogue-based data as potentially sensitive. For existing datasets, we rely on the anonymization and access-control procedures of the original releases and follow their redistribution terms. For processed datasets marked with $_M$, we do not introduce new personal identifiers; client utterances are compressed into task-focused case summaries, and outputs are standardized into labels, prompts, or evaluation items. For restricted, platform-derived, or license-unspecified resources, we do not redistribute raw text unless explicitly permitted by the original source.

For newly curated PSRS, the source materials are publicly available systematic-review abstracts rather than patient-level records, so no individual patient identifiers are included. For all released artifacts, we will avoid publishing personally identifying information and will release only permitted metadata, prompts, labels, evaluation scripts, expert scores, and annotations where allowed. Because mental-health data can include distressing, stigmatizing, or crisis-related content, benchmark users should treat the data as sensitive research material and use it only under appropriate ethical and institutional safeguards.

B.4 Model Size and Computational Budget

For benchmark evaluation, most proprietary, hosted, and large open-weight models are accessed through official APIs, including GPT-o4-mini, GPT-4o, DeepSeek-R1, DeepSeek-V3, Qwen-plus, QwQ-plus, LLaMA-4, LLaMA-3.3-70B, DeepSeek-distilled 70B/32B/14B variants, Qwen2.5-72B, Qwen3-32B/14B, and QwQ-32B. Their exact provider-side compute budgets are not publicly available. The locally evaluated models include EmoLLM, Psyche-R1, and Qwen3-8B, all at the 7–8B scale. Local benchmark inference was conducted on 8 NVIDIA GPUs for approximately 2 hours in total, corresponding to about 16 GPU-hours.

Reasoning-trajectory evaluation is performed with gpt-5-mini-2025-08-07 as an API-based LLM-as-Judge. Combining answer generation through APIs and trajectory evaluation, the total monetary cost for running the benchmark experiments was approximately $20 under the API pricing used in our experiments. Mindora is initialized from Qwen3-8B and therefore has approximately 8B parameters; its benchmark inference is included in the local compute budget above.

B.5 Appraisal Dataset Details

Detailed descriptions of these datasets are shown below:

- CognitiveReframing (Sharma et al., 2023) combines simulated negative thoughts from the Thought Records Dataset (Burger et al., 2021) and self-reports from the Mental Health Amer-

Table 5: Dataset licenses and access conditions for MIRAMIND. "Not specified" indicates that we did not find an explicit dataset-specific license in the original release page or paper. For restricted or platform-derived datasets, we follow the original access terms and do not redistribute raw texts unless explicitly permitted.

Dataset	Original Source / Release	License / Access Condition	Redistribution Note
CognitiveReframing	Original GitHub release (Sharma et al., 2023)	CC BY-NC-ND 4.0	Non-commercial research use; no redistribution of modified data.
PatternReframe	ParlAI / original release (Maddela et al., 2023)	Released through ParlAI; dataset-specific license not clearly specified	Use follows original release terms; processed prompts/scripts only where permitted.
Therapist Q&A	Kaggle Therapist Q&A corpus (Shreevastava and Foltz, 2021)	Kaggle-hosted dataset; license follows original Kaggle page	Use follows original dataset terms; raw redistribution not assumed.
DepSign	LT-EDI shared-task dataset (Poświata and Perełkiewicz, 2022)	Shared-task access; explicit public license not specified	Use follows shared-task terms; raw redistribution not assumed.
SWMH	Original release / Zenodo (Ji et al., 2022)	Restricted research-only access under data access agreement	Redistribution not permitted; users should request access from source.
T-SID	Original Twitter-derived dataset (Ji et al., 2022)	Twitter-derived data; explicit dataset license not specified	Use follows original dataset and platform terms; raw tweet redistribution not assumed.
PsyDTCorpus_M	Processed from PsyDTCorpus (Xie et al., 2025)	Released by original authors; license follows original release	Used only for training; processed version follows original terms.
AnnoMI_M	Processed from AnnoMI (Wu et al., 2022)	Public dataset release; dataset-specific license not clearly specified	Use follows original access terms; processed labels/prompts released where permitted.
MHQA	Original MHQA release (Racha et al., 2025)	Dataset-specific license to be confirmed from original release	Use follows original release terms; raw redistribution not assumed.
MedQA_M	Processed from MedQA (Jin et al., 2020)	Official dataset license not clearly specified	Psychiatry subset extracted; redistribution follows original MedQA terms.
MedMCQA_M	Processed from MedMCQA (Pal et al., 2022)	Apache-2.0	Psychiatry subset extracted; redistribution permitted under Apache-2.0 terms.
PubMedQA_M	Processed from PubMedQA (Jin et al., 2019)	MIT License	Psychiatry subset extracted; redistribution permitted under MIT terms.
PSRS	Newly curated from Cochrane Library abstracts	Cochrane/Wiley terms; full abstract text may be copyright restricted	We release review identifiers, prompts, expert scores, and annotations where permitted; full abstract text is not redistributed unless allowed.
MentalMisinfo	Original MentalMisinfo release (Nguyen et al., 2025a)	Released by original authors; explicit dataset license not clearly specified	Use follows original dataset terms; raw video/transcript redistribution not assumed.

ica website[3], with distortions annotated by 15 trained mental-health professionals.

- PatternReframe (Maddela et al., 2023) is constructed from PERSONA-CHAT personas (Zhang et al., 2018) and contains statements crafted to manifest specific distortions and labeled by five independent raters.

- Therapist Q&A (Shreevastava and Foltz, 2021), derived from real therapistclient interactions in the Therapist Q&A corpus[4], provides naturally occurring distorted statements annotated by two clinical raters.

These three datasets adopt slightly different taxonomies of cognitive errors.

- Shared across all three datasets: All-or-Nothing Thinking; Overgeneralization; Labeling; Fortune Telling; Mind Reading; Should Statements; Personalization.

- Unique to CognitiveReframing: Emotional Reasoning; Comparing and Despairing; Blaming; Negative Feeling or Emotion; Catastrophizing; Discounting the Positive.

- Unique to PatternReframe: Mental Filtering; Catastrophizing; Discounting the Positive.

- Unique to Therapist Q&A: Emotional Reasoning; Mental Filtering; Magnification; No Distortion.

B.6 Intervention Dataset Details

We instantiate this dimension through the counseling-strategy formulation task, covering thirteen commonly taught intervention types. The strategies include Clarification, Paraphrasing, Reflection of Feeling, Summarizing, Questioning Skills, Immediacy, Use of Silence, Self-Disclosure, Confrontation, Encouragement, Repetition, Interpretation, and Guidance.

To construct evaluation data, we use two high-quality counseling dialogue corpora. Client utterances are compressed into concise case summaries using GPT-4o prompts, and counselor utterances

[3] https://screening.mhanational.org/
[4] https://www.kaggle.com/arnmaud/therapist-qa

are annotated to extract strategy labels as reference answers. All constructed items are manually reviewed and verified by an expert counselor with over ten years of clinical experience, resulting in the final $[\text{dataset}]_M$ versions. We deliberately separate the training and test sources for this task: PsyDTCorpus_M is used only for training, while AnnoMI_M is used for evaluation. This design tests whether intervention-planning ability transfers from the partially synthetic training corpus to real counseling dialogue demonstrations.

- PsyDTCorpus_M: Based on PsyDTCorpus (Xie et al., 2025), which contains 5,000 high-quality single-turn dialogues from SoulChatCorpus and 12 anonymized real counseling cases synthesized into multi-turn interactions. Due to its partially synthetic nature, this dataset is used to train Mindora but is excluded from benchmark evaluation.

- AnnoMI_M: Based on AnnoMI (Wu et al., 2022), constructed from authorized motivational interviewing (MI) demonstration videos sourced from YouTube and Vimeo, transcribed and curated into high-quality multi-turn counseling interactions.

B.7 Multi-step Dataset Details

We instantiate this dimension through a psychiatry QA task and evaluate it using four psychiatry-focused datasets.

- MHQA (Racha et al., 2025) provides knowledge-intensive question-answer pairs from 471k PubMed abstracts covering major mental-health conditions.

- MedQA_{en} (Jin et al., 2020) provides exam-style multiple-choice medical questions with clinical vignettes collected from professional medical licensing resources, covering broad disease knowledge and diagnostic decision-making.

- MedMCQA (Pal et al., 2022) provides large-scale multiple-choice questions curated from Indian medical entrance and licensing examinations, spanning diverse medical specialties and knowledge-intensive clinical reasoning.

- PubMedQA (Jin et al., 2019) provides research-focused question-answer pairs grounded in PubMed abstracts, requiring evidence-based inference over biomedical studies with yes/no/maybe-style conclusions.

B.8 PSRS Dataset Construction

In this section, we illustrate the construction pipeline of PSRS, a psychiatry-oriented systematic-review summarization resource derived from structured Cochrane abstracts and paired with expert evaluation of LLM-generated summaries. This dataset is introduced to fill a gap not covered by existing mental-health benchmarks: evaluating whether LLMs can abstract clinically relevant conclusions from psychiatric systematic-review evidence, rather than only answer diagnostic or counseling questions. The overall design follows a source-abstract $\rightarrow$ structured-summary generation $\rightarrow$ expert rubric-based assessment workflow.

B.8.1 Raw Data Collection

We curate 115 structured abstracts of systematic reviews from the Cochrane Library, restricted to the mental-health domain to ensure clinically coherent populations, interventions, and outcomes. Abstracts are sampled to preserve diversity in (i) target conditions, (ii) intervention families, and (iii) population characteristics. We retain review identifiers for traceability and use only publicly available abstract text, without any patient-level or trial-level records.

B.8.2 LLM Summary Generation

For each Cochrane abstract, we generate four structured summaries using GPT-4o, GPT-4o-mini, Claude Opus, and LLaMA-3.1-70B-Instruct, yielding 460 summaries in total. To minimize format variance, we enforce a fixed schema and focus evaluation on a single target field, Main Results, which summarizes the core outcome-level findings supported by the abstract evidence. We use low-temperature decoding (temperature ≤ 0.3) for determinism and provide no external content beyond the abstract.

B.8.3 Human Annotation

Annotators. Two clinical domain experts (psychiatrists) conduct the evaluation. Each generated *Main Results* statement is assessed under a structured rubric with mandatory rationales for non-maximal labels, enabling auditing and error analysis.

Pilot calibration. Before the main annotation, we run a pilot on 10 abstracts per expert to refine label definitions and align judgment thresh-

olds (e.g., borderline cases involving partial reporting vs. over-interpretation).

B.8.4 Evaluation Rubric and Labeling Scheme

Since our main paper evaluates only the Main Results field, we adopt a single-dimension scoring rubric.

Main Results (0–2). Annotators score the Main Results content on an ordinal scale:

- 2 = accurate and complete with respect to the Cochrane abstract;
- 1 = partially correct / incomplete / contains minor inaccuracies;
- 0 = missing, unsupported, or substantially incorrect.

Rationales are mandatory for scores in $\{0, 1\}$. In addition, annotators record a corrective action indicating the minimal edit needed to align the content with the abstract: add, remove, modify, rewrite, or none.

B.8.5 Inter-Annotator Agreement and Quality Assurance

Because full double-annotation is expensive with clinical experts, we adopt a targeted QA protocol.

Cross-check subset. After the main evaluation, we randomly select 30 reviews per model (i.e., 120 summaries in total) for cross-checking. In this phase, each expert reviews the other's prior Main Results labels and records agreement vs. disagreement.

Agreement reporting (IAA proxy). Rather than reporting chance-corrected coefficients (e.g., Cohen's κ), we report absolute disagreement counts on the QA subset. The IAA results are shown in Table 6. Notes that we report absolute disagreement counts due to the modest QA sample size. Disagreements mainly arise from borderline cases where summaries partially report effect directions/magnitudes or use stronger-than-supported wording, requiring subjective thresholds even among experts.

B.8.6 Output Artifacts and Release

For each generated summary, PSRS releases (i) the structured Main Results output, (ii) expert scores, (iii) rationales for non-maximal judgments, and (iv) QA cross-check agreement signals for the selected subset, enabling reproducible analysis of evidence-grounded summarization quality in psychiatry systematic reviews.

Table 6: Inter-annotator agreement (IAA) on the QA subset, reported as the number of disagreements between the two psychiatrists for the Main Results dimension ($n = 30$ reviews per model; 120 summaries in total).

Model	GPT-4o	Claude Opus	LLaMA-3.1-70B	GPT-4o-mini
Main Results	4	3	5	4

C Reasoning Trajectory Evaluation

Given an input context C, a reasoning chain $R = \{r_1, r_2, \ldots, r_n\}$, and a final answer A, our protocol scores R along three dimensions, each targeting one of the failure modes above. (1) **Usability:** is the chain clear and concise enough for its intended users to follow? (2) **Logical Structure:** Does the chain form a coherent structure that decomposes into explicit, connected steps? (3) **Informational Contribution:** Does each step add information relevant to the conclusion rather than restating it? A chain may satisfy any subset of these and violate the rest, for instance correct but tautological, informative but unsupported, or sound yet opaque, and the protocol scores them independently. Detailed scoring criteria are provided below.

C.1 Evaluation Settings

For reproducibility, we will release the complete judge outputs, including dimension-wise scores, relation annotations, and brief rationales. Since the judge prompts and released outputs are model-agnostic, future work can reproduce the protocol with alternative open-source judge models through few-shot prompting or judge-model fine-tuning.

C.2 LLM-as-Judge Settings.

To ensure scalable and consistent evaluation of reasoning-chain quality, we adopt an *LLM-as-Judge* paradigm to score all reasoning trajectories. Specifically, we use `gpt-5-mini-2025-08-07` as the evaluation model, prompted with structured rubrics corresponding to each reasoning dimension. The judge model is instructed to assess reasoning trajectories independently of model identity and to assign ordinal scores according to predefined criteria. To validate the reliability of automatic trajectory scoring, we conduct a human annotation study. We randomly sample 70 reasoning trajectories from the evaluation set and two domain expert coauthors with formal training in psychology to annotate trajectory quality. Each expert independently assigns ordinal scores to the same set of trajectories following the dimension-wise rubrics described below. Annotators are instructed to focus on the reasoning trajectory and to judge the extent to which each step is decision-relevant, structurally coherent, and informative. Disagreements are resolved via discussion to form a final consensus label, which is then used to quantify agreement between human annotations and the LLM-as-Judge scores reported in Table 7. Overall, the agreement and (weighted) κ scores in Table 7 indicate that our LLM-as-Judge ratings are highly aligned with expert human judgments across all reasoning dimensions, supporting their use as reliable automatic metrics for trajectory evaluation.

Table 7: Agreement between automatic annotations and human experts.

Dimension	Agreement	Cohen's κ	Weighted κ
Terminology Parsimony	0.743	0.426	0.489
Step Adequacy	0.686	0.479	0.593
Logical Structure	1.000	1.000	1.000
Informativeness	0.857	0.481	0.665

We also inspect cases where the automatic judge assigns relatively high scores while human experts assign lower scores. These disagreements vary in length, surface structure, and reasoning style, suggesting that the judge is not simply rewarding longer chains or more explicit formatting. Instead, most disagreements arise from borderline judgments about whether a step is genuinely decision-relevant or only rhetorically plausible. This analysis indicates that the judge captures more than structural compliance, although we acknowledge that the current human validation set remains limited.

C.3 Scoring Metrics

We adopt a three-level ordinal scale for each metric.

C.3.1 Usability

We assess whether the reasoning chain is presented in a clear and efficient manner by measuring *Terminology Parsimony* and *Step Adequacy*, which together reflect unnecessary terminology usage and avoidable verbosity.

Terminology Parsimony.

- Low (2 points): Only psychological concepts strictly required for explaining or justifying the decision are used.
- Medium (1 point): Additional background or contextual terminology is introduced but does not hinder comprehension.
- High (0 points): Terminology substantially exceeds decision needs and mainly serves as authority decoration.

Step Adequacy.

- Efficient (2 points): Every reasoning unit is essential and cannot be removed without weakening the decision rationale.
- Acceptable (1 point): Some content is compressible or could be merged, but the overall reasoning remains understandable and decision-relevant.
- Inefficient (0 points): A substantial portion of sentences can be removed with little or no impact on the decision, indicating verbosity that does not contribute to the conclusion.

C.3.2 Logical Structure

We evaluate whether the reasoning forms a coherent and well-connected structure using *Logical Coherence* and a set of *structure-derived submetrics*, which quantify unsupported leaps, irrelevant steps, and contradictions based on relation annotations.

Logical Coherence.

- Coherent (2 points): RCUs are well-formed and most units are meaningfully connected through clear support relations to the context and/or earlier RCUs.
- Partially Coherent (1 point): RCUs are mostly identifiable, but the chain contains some weakly connected units, minor leaps, or under-specified links.
- Incoherent (0 points): RCUs are difficult to identify or many units are disconnected, contradictory, or rely on unsupported leaps that undermine the reasoning structure.

Structure-derived Submetrics. In addition to the three-level coherence score, we compute quantitative submetrics directly from the judge-generated relation annotations in JSON. These submetrics provide fine-grained signals about common structural failure modes and enable scalable benchmark analysis.

Let $\mathcal{U} = \{R_1, \ldots, R_m\}$ denote the set of RCUs. For each R_i, the judge outputs a relation label:

$$\ell_i \in \left\{ \begin{array}{l} \text{ENTAILS, SUPPORTS, REFINES, CAUSES,} \\ \text{CONTRADICTS, IRRELEVANT} \end{array} \right\}. \quad (1)$$

We define the following rates:

$$\begin{aligned} \text{IrrelevanceRate} &= \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \mathbb{I}[\ell_i = \text{IRRELEVANT}]\,, \\ \text{ContradictionRate} &= \tfrac{1}{m} \textstyle\sum_{i=1}^{m} \mathbb{I}[\ell_i = \text{CONTRADICTS}]\,. \end{aligned} \quad (2)$$

Intuitively, IrrelevanceRate captures the fraction of steps that do not contribute to the decision, and ContradictionRate captures explicit conflicts within the reasoning structure. These submetrics complement the coarse coherence score by exposing which failure modes dominate a model's reasoning behavior.

C.3.3 Informational Contribution

We assess how effectively the reasoning advances toward the final conclusion using *Step-level Informativeness*, which captures the usefulness of individual steps.

Informativeness.

- High (2 points): Most RCUs are informative and information accumulates toward the answer.
- Medium (1 point): Some RCUs are informative but redundancy is present.
- Low (0 points): many RCUs are redundant or irrelevant.

D Mindora Training Details

D.1 Overview

Mindora is a post-trained LLM built on Qwen3-8B and used as a benchmark-guided case study for improving reliable mental-health reasoning. The training pipeline contains two stages. The first stage, Reasoning Trajectory Generation (RTG), constructs concise and clinically interpretable chain-of-thought supervision from genuinely difficult cases. The second stage applies hybrid SFT–RL post-training with a consistency-aware reward, encouraging the model to produce reasoning that is task-relevant, internally consistent, and aligned with the final answer.

D.2 Reasoning Trajectory Generation

Effective supervision for mental-health reasoning must be both informative and learnable. Using all available training instances can dilute the signal with easy cases that are solvable through surface cues, while raw search traces can be too verbose or unstable for small-model learning. RTG therefore concentrates supervision on difficult cases and then converts successful trajectories into readable, non-backtracking rationales.

Hard-case filtering. We first run zero-shot question answering on each training split using LLaMA-3-8B-Instruct and retain only the instances answered incorrectly. These retained samples are treated as high-signal cases because they are less likely to be solved by superficial pattern matching and more likely to require explicit reasoning. This filtering step aims to improve the density of useful supervision rather than simply increasing training volume.

Verifier-guided search. For each retained instance, we generate candidate reasoning trajectories through iterative refinement with a verifier. The search process allows exploration and backtracking to recover a valid solution. We do not use the raw search dynamics as training targets; instead, we retain only the successful trajectory content that supports the correct final answer. This follows standard verifier-guided trajectory construction while keeping the downstream supervision focused on the final usable rationale.

D.3 Structured Rewriting for Non-backtracking Trajectories

Raw search traces often contain self-corrections, speculative branches, and backtracking artifacts. While such traces may help a strong teacher model explore the solution space, they can reduce readability and confuse smaller student models during post-training. We therefore rewrite the selected trajectory into a standardized, non-backtracking rationale.

Each rewritten example contains two phases. The reasoning phase is wrapped in `<think>` and segmented by domain-relevant subtitles, such as *Symptom Analysis*, *Differential Diagnosis*, or *Evidence Assessment*. It ends with a mandatory *Final Conclusion* that explicitly maps the reasoning to the decision. The answer phase is wrapped in `<answer>` and ends with an unambiguous `Answer: [option/result]` consistent with the final conclusion.

This rewriting stage enforces three properties. First, it improves linear coherence by presenting a single best path instead of visible exploration. Second, it strengthens reason–answer alignment by requiring the final conclusion and answer field to match. Third, it improves interpretability through clinically meaningful modularization. This design is particularly important in mental-health settings, where visible self-correction, excessive hedging, or loosely connected clinical language can undermine the clarity of the judgment.

D.4 Hybrid SFT–RL Post-training

Mindora is trained with a hybrid SFT–RL scheme that balances supervised imitation and reward-driven exploration. The SFT component encourages the model to imitate the structured reasoning trajectories generated by RTG, while the RL component further refines model behavior according to task-specific correctness and reasoning-quality constraints.

Our emphasis is the consistency-aware reward tailored to mental-health reasoning. The reward combines four components: format validity, length validity, internal consistency, and task-specific correctness. Format validity checks whether the output follows the required `<think>` and `<answer>` structure. Length validity discourages degenerate outputs that are either too short to justify the decision or unnecessarily verbose. Internal consistency is evaluated by an auxiliary

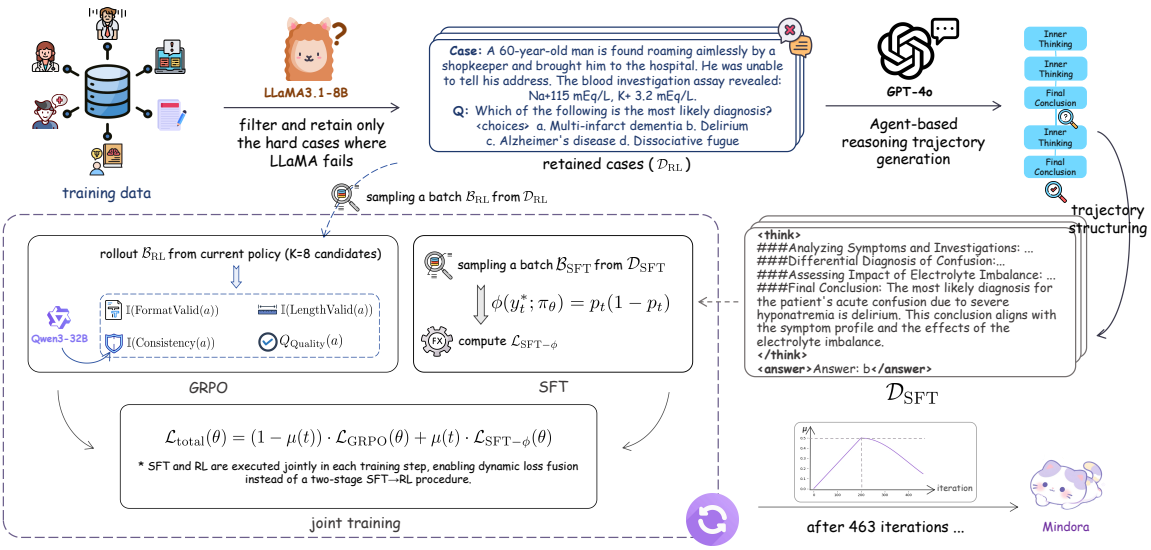

Figure 2: Mindora training pipeline: hard-case filtering, verifier-guided reasoning trajectory generation, structured rewriting into non-backtracking trajectories, and hybrid SFT–RL post-training with consistency-aware rewards.

model, Qwen3-32B, which checks whether the reasoning phase contains factual or logical inconsistencies. Task-specific correctness rewards final answers that match the reference labels or scoring criteria.

The consistency term is designed to penalize reasoning that appears fluent but contains unsupported or self-conflicting intermediate claims. In contrast to RL objectives that primarily optimize final-answer correctness or generic response quality, this reward shapes intermediate rationales toward concise, grounded, and internally consistent trajectories. This makes the training objective aligned with the evaluation target of MIRAMIND, where reliable mental-health reasoning depends not only on the endpoint but also on whether the path to that endpoint remains evidence-proportional and decision-relevant.

D.5 Role of Mindora in This Work

Mindora is used as a benchmark-guided post-training case study rather than as a deployable mental-health system. Its purpose is to test whether the failures surfaced by MIRAMIND can motivate targeted optimization toward more reliable reasoning behavior. The model results therefore support the utility of MIRAMIND as both an evaluation tool and a research substrate, while deployment-oriented safety, treatment validity, and clinical efficacy remain outside the scope of this work.

E Experimental Details

E.1 Evaluated LLMs

To assess whether model scale, distilled reasoning variants, and the gap between reasoning-oriented and chat-oriented LLMs affect performance on mental health reasoning tasks, we evaluated a broad range of GPT, DeepSeek, Qwen, and LLaMA models. For closed-source systems, we included the reasoning models GPT-o (o4-mini), DeepSeek-R (DeepSeek-R1), and Qwens QwQ series (QwQ-plus), as well as the leading chat models GPT-4o, DeepSeek-V3, and Qwen-plus[5]. For open-source models, we tested Qwen, LLaMA, and DeepSeek-distilled variants across multiple scales, with the full list provided in Table 3. For psychology-focused LLMs, we evaluated EmoLLM[6] , finetuned from Qwen2-7B-Instruct as a well-known mental-health chat model, and Psyche-R1[7] , an SFT+GRPO-trained Qwen2.5-7B-Instruct model representing the latest mental-health-reasoning LLMs.

E.2 Model Size and Computational Budget

For benchmark evaluation, most proprietary, hosted, and large open-weight models are accessed through official APIs, including GPT-o4-mini, GPT-4o, DeepSeek-R1, DeepSeek-V3, Qwen-plus, QwQ-plus, LLaMA-4, LLaMA-3.3-70B, DeepSeek-distilled 70B/32B/14B variants, Qwen2.5-72B, Qwen3-32B/14B, and QwQ-32B. Their exact provider-side compute budgets are not publicly available. The locally evaluated models include EmoLLM, Psyche-R1, and Qwen3-8B, all at the 7–8B scale. Local benchmark inference was conducted on 8 NVIDIA GPUs for approximately 2 hours in total, corresponding to about 16 GPU-hours.

Reasoning-trajectory evaluation is performed with gpt-5-mini-2025-08-07 as an API-based LLM-as-Judge. Combining answer generation through APIs and trajectory evaluation, the total monetary cost for running the benchmark experiments was approximately $20 under the API pricing used in our experiments. Mindora is initialized from Qwen3-8B and therefore has approximately 8B parameters; its benchmark inference is included in the local compute budget above.

E.3 Group-wise Performance Analysis

To complement the absolute performance scores in Table 3, we further conduct a group-wise rank analysis in Table 8. This analysis is intended to reduce the influence of any single high-scoring dataset and to examine whether a model performs consistently across heterogeneous mental-health reasoning tasks. For each dataset, models are ranked according to task performance, with lower ranks indicating better performance. We then compute average ranks within each task group, including appraisal-oriented cognitive error identification (Avg_1), mental-condition detection (Avg_2), multi-step psychiatry QA (Avg_4), as well as the overall average rank across all evaluated datasets.

The results show that Mindora achieves the best overall average rank (Avg_{all} Rank = 4.92), outperforming both strong proprietary models and domain-specialized baselines. This indicates that Mindora's advantage is not driven by a single dataset or task family, but reflects more stable performance across diverse reasoning settings. In comparison, DeepSeek-R1 obtains strong ranks on appraisal tasks (Avg_1 Rank = 3.33) and multi-step psychiatry QA (Avg_4 Rank = 6.50), but performs less consistently on mental-condition detection and misinformation identification, leading to a weaker overall rank (Avg_{all} Rank = 9.92). GPT-o4-mini ranks first on MedQA$_M$ and MentalMisinfo, but its relatively weak ranks on diagnosis-oriented datasets result in a higher overall average rank (Avg_{all} Rank = 12.15).

The group-wise view also reveals that different model families exhibit distinct strengths. Large proprietary and reasoning-oriented models remain competitive on structured QA and verification tasks, whereas their ranks fluctuate more substantially on narrative-based appraisal and diagnosis tasks. By contrast, Mindora maintains consistently strong ranks across appraisal (Avg_1 Rank = 3.00), diagnosis (Avg_2 Rank = 3.67), and multi-step reasoning (Avg_4 Rank = 5.75), while also ranking near the top on MentalMisinfo. These results support the conclusion that targeted post-training improves not only pooled task accuracy, but also cross-task robustness over heterogeneous mental-health reasoning scenarios.

The ablation variants further clarify the contribution of individual components. Mindora w/o S3 remains competitive overall (Avg_{all} Rank = 7.08),

[5]Qwen-plus is Qwens versatile chat model, also trained with reasoning capabilities.

[6]https://www.modelscope.cn/models/aJupyter/EmoLLM_Qwen2-7B-Instruct_lora/.

[7]https://huggingface.co/MindIntLab/Psyche-R1

Table 8: Group-wise rank comparison on MIRAMIND. Lower rank indicates better performance.

Model	Cognitive Reframing	Pattern Reframe	Therapist Q&A	Avg_1 Rank	DepSign	SWMH	T-SID	Avg_2 Rank	$AnnoMI_M$	MHQA	$MedQA_M$	$MedMCQA_M$	$PubMedQA_M$	Avg_4 Rank	PSRS	Mental Misinfo	Avg_{all} Rank
close-source models																	
GPT-o4-mini	3	10	10	7.67	27	17	17	20.33	15	9	**1**	18	19	11.75	11	**1**	12.15
GPT-4o	7	7	7	7.00	18	12	26	18.67	7	14	23	21	22	20.00	7	5	13.54
Deepseek-R1	**1**	**1**	8	3.33	15	8	23	15.33	18	10	2	2	12	6.50	9	20	9.92
Deepseek-V3	14	5	13	10.67	9	3	14	8.67	14	12	4	7	25	12.00	4	15	10.69
Qwen-plus	6	14	14	11.33	19	8	12	13.00	21	11	3	5	17	9.00	8	18	12.00
QwQ-plus	10	2	20	10.67	15	10	21	15.33	22	19	21	16	16	18.00	2	24	15.23
70B+ open-source models																	
LLaMA-4	13	11	16	13.33	22	22	9	17.67	16	13	13	4	13	10.75	23	14	14.54
LLaMA-3.3-70B	17	17	21	18.33	14	25	24	21.00	17	20	9	3	11	10.75	27	5	16.15
dsdistill-LLaMA-70B	11	7	18	12.00	20	11	27	19.33	11	15	16	20	8	14.75	10	7	13.92
Qwen2.5-72B	5	20	17	14.00	25	7	3	11.67	8	21	11	14	14	15.00	**1**	16	12.46
32B open-source models																	
dsdistill-Qwen32B	14	6	22	14.00	21	18	20	19.67	12	22	20	24	14	20.00	6	13	16.31
Qwen3-32B	18	14	12	14.67	25	5	5	11.67	19	16	12	17	21	16.50	3	23	14.62
QwQ-32B	7	3	18	9.33	15	12	22	16.33	23	17	19	12	10	14.50	4	26	14.46
14B open-source models																	
dsdistill-Qwen14B	12	23	23	19.33	11	16	11	12.67	9	24	22	25	7	19.50	12	19	16.46
Qwen3-14B	14	11	5	10.00	23	14	16	17.67	24	18	17	18	24	19.25	16	10	16.15
7∼8B open-source models																	
LLaMA3.1-8B	25	24	27	25.33	6	18	14	12.67	13	27	25	23	20	23.75	26	9	19.77
dsdistill-LLaMA-8B	24	25	25	24.67	10	23	19	17.33	20	26	26	26	6	21.00	15	2	19.00
EmoLLM	19	26	24	23.00	7	24	8	13.00	27	25	27	27	26	26.25	21	11	20.92
Psyche-R1	27	27	26	26.67	11	27	7	15.00	10	5	18	15	18	14.00	14	4	16.08
Qwen3-8B	23	14	14	17.00	11	18	17	15.33	25	23	24	22	23	23.00	17	25	19.69
$Mindora_{SFT}$	26	20	8	18.00	2	15	13	10.00	4	7	15	10	9	10.25	22	22	13.31
$Mindora_{RL}$	20	17	11	16.00	24	18	25	22.33	26	6	5	**1**	5	4.25	18	27	15.62
$Mindora_{SFT+RL}$	22	20	6	16.00	**1**	3	2	2.00	6	3	6	9	3	5.25	24	21	9.69
Mindora w/o S1	4	13	**1**	6.00	8	26	4	12.67	**1**	4	8	6	4	5.50	19	12	8.46
Mindora w/o S2	21	19	2	14.00	5	2	**1**	2.67	2	8	14	13	27	15.50	20	17	11.62
Mindora w/o S3	9	9	2	6.67	3	5	10	6.00	3	2	7	8	**1**	4.50	25	8	7.08
Mindora	2	3	4	**3.00**	4	**1**	6	3.67	5	**1**	10	10	2	5.75	13	3	**4.92**

but its weaker ranks on PSRS and MentalMisinfo suggest that the consistency reward is particularly important for tasks requiring stable evidence integration and verification. Mindora w/o S2 performs well on mental-condition detection (Avg_2 Rank = 2.67), but its weaker multi-step rank (Avg_4 Rank = 15.50) indicates that structured rewriting is important for transferring reasoning patterns across more complex clinical QA settings. Overall, the rank analysis is consistent with the absolute-score results and reinforces that Mindora provides the most balanced performance across the benchmark.

CognitiveReframing Case Content:

Question: Please choose the most proper cognitive error of this client from the following list: ...
Context:
- situation: Someone pushed me and made me bleed.
- thought: I feel like people out there try to take advantage of me and also betray me.

Response of Mindora:

###Analyzing Thought Pattern:
The client's thought pattern suggests a belief that people generally try to take advantage of and betray them, based on the experience of being pushed and bleeding.
###Evaluating Cognitive Errors:
The thought pattern reflects a belief that people are inherently trying to harm or betray them. This suggests a cognitive error that involves making broad assumptions about others based on specific experiences.
###Matching Cognitive Error:
The cognitive error that involves making broad assumptions about others based on specific experiences is 'Overgeneralization'. The client is generalizing the experience of being pushed to a belief that people generally try to take advantage of them.
###Final Conclusion:
The most proper cognitive error for this client is 'Overgeneralization', as they are generalizing their experience to a broader belief about others.

Figure 3: Reasoning trajectory comparison between Mindora and baseline models (1).

F Case Study

Figures 3–5 present a representative CognitiveReframing instance where the correct label (*Overgeneralization*) hinges on a minimal discriminative mapping: a *single negative incident* is extrapolated into a *global belief* about others' intentions. While most models can recover the label, their reasoning trajectories exhibit different trade-offs between readability, structure, and decision relevance.

Mindora provides the most balanced trace: it remains anchored to the clients stated *thought*, isolates the core cue with minimal redundancy, and performs an explicit evidence-to-label mapping under a stable multi-step template, yielding a trajectory that is simultaneously interpretable to users and well-posed as supervision for smaller models. In contrast, several baselines tend to over-index on a single axis: some favor brevity via generic "key analysis" templates that can become under-specified in harder cases, while others adopt exhaustive elimination or option-scanning that increases length and cognitive load without improving the discriminative justification. More generally, these trajectories either introduce compressible content (verbosity without additional decision signal) or omit critical linking steps (insufficient mapping from evidence to category), whereas Mindora better preserves a consistent structure that supports both human readability and reliable distillation.

Response of GPT-o4-mini:

1. Key Analysis:
The client experienced a single aggressive incident (being pushed and bleeding) and now believes broadly that all people intend to take advantage of or betray them.
2. Step Reasoning:
One negative event broad conclusion about everyone
Generalizing from a single experience
3. Final Answer:
Answer: Overgeneralization

Response of Deepseek-R1:

Based on the provided situation and thought, the client's cognitive error is Overgeneralization.
- Situation: A specific event where one person pushed them, causing injury.
- Thought: I feel like people out there try to take advantage of me and also betray me.
- Error: The client is taking the negative action of one specific person in one specific situation (being pushed) and generalizing it to all people (people out there). They conclude that this is a universal pattern (people trying to take advantage and betray them) based on this single incident. This fits the definition of Overgeneralization: generalize a particular event leading to a certain outcome to all similar events (or in this case, to all people).
Why other options are less fitting:
- Catastrophizing: Focuses on worst outcomes, not generalizing people's intent.
- Discounting the Positive: No positive event mentioned to discount.
- Personalization: Client isn't blaming themselves; they blame others.
- All-or-Nothing Thinking: Thought isn't about extremes of good/bad regarding a specific thing or person.
- Mental Filtering: Only one event is described; no other aspects are mentioned to filter out.
- Mind Reading: Client isn't inferring the specific pusher's thoughts; they are generalizing about all people's intentions.
- Fortune Telling: Client isn't predicting a future outcome; they are making a present generalization about people's character/intent.
- Should Statements: No demands for behavior are present in the thought.
- Labeling and Mislabeling: While related, labeling typically attributes the action to the character of the specific person involved (e.g., That person who pushed me is a monster). Here, the client is labeling all people (people out there) based on the action of one, which is the core of overgeneralization.
Answer: [Overgeneralization]

Response of Qwen-plus:

1. Key Analysis: The client experienced a single negative event (being pushed) and generalized it to believe all people are harmful and untrustworthy.
2. Step Reasoning:
- The client extrapolates one negative experience to an overarching belief about others.
- This aligns with drawing broad conclusions from isolated incidents (overgeneralization).
3. Final Answer: Answer: [Overgeneralization]

Figure 4: Reasoning trajectory comparison between Mindora and baseline models (2).

Response of LLaMA-4:

Key Analysis: The client feels that people are trying to take advantage of and betray them after being pushed and hurt. This suggests a pattern of negative thinking and attributing malicious intent to others based on a single incident.

Step Reasoning:

- The client's experience is a single incident of being pushed and hurt.
- The client generalizes that people try to take advantage of and betray them.
- This generalization seems to stem from jumping to conclusions based on one experience.

Final Answer: Answer: [Overgeneralization]

Response of Psyche-R1:

</think>

###Comparing with Given Choices:

- All-or-Nothing Thinking: The client does not seem to be thinking in extreme categories.
- Overgeneralization: The client is drawing a broad conclusion from a single event.
- Labeling: The client does not appear to be defining himself based on one action or characteristic.
- Fortune Telling: The client does not seem to be predicting the future.
- Mind Reading: The client does not assume he knows what others are thinking.
- Emotional Reasoning: The clients thoughts are not directly treated as facts.
- Should Statements: The client does not set unrealistic expectations for himself.
- Personalization: The client does not take the event personally.
- Discounting the Positive: The client does not ignore positive events.
- Catastrophizing: The client does not focus on the worst— Case scenario.
- Comparing and Despairing: The client does not compare his worst to someone else's best.
- Blaming: The client does not give away his own power to others.
- Negative Feeling or Emotion: The client is stuck on a distressing thought, but this is more a symptom than a cognitive error.

###Final Conclusion: The most proper cognitive error of this client is Overgeneralization.

</think>

Answer: [Overgeneralization]

Response of Qwen3-8B:

Key Analysis: The client generalizes a single negative event (being pushed) to a broader belief about others' intentions.

Step Reasoning:

- The incident (being pushed) is a single experience.
- The client concludes people "try to take advantage" and "betray," implying a universal pattern.
- This matches jumping to a broad conclusion from one event.

Final Answer: [Overgeneralization]

Figure 5: Reasoning trajectory comparison between Mindora and baseline models (3).

G Ablation Details

We conduct three ablations to isolate the contributions of S1 (hard-sample filtering), S2 (structured rewriting), and S3 (consistency reward). Removing any component consistently degrades average MIRAMIND performance, confirming that Mindoras gains do not come from a single training trick but from complementary design choices.

S1 primarily improves robustness on challenging, decision-sensitive instances: replacing hard-case filtering with random sampling reduces overall scores and yields less efficient trajectories, with lower step adequacy and informativeness despite often reaching similar final answers. S2 contributes the largest drop when removed, indicating that rewriting into a standardized, modular format is critical for transferring clinically aligned reasoning into an 8B-scale student; without S2, trajectory quality sharply deteriorates and inconsistency (irrelevance/contradiction) increases. S3 further stabilizes reasoning reliability, with the clearest impact on verification/abstraction tasks that require consistent evidence integration (e.g., PSRS and MentalMisinfo). Without S3, logical coherence and informativeness decline and irrelevant steps increase, suggesting the reward is effective at discouraging internally inconsistent or weakly grounded intermediate rationales. Detailed analyses and case studies are provided below.

G.1 Hard-sample Filtering (S1)

We examine the effect of hard-case filtering (S1) by comparing the full Mindora model with Mindora w/o S1, where training samples are randomly selected instead of being filtered by zero-shot failures. Removing S1 leads to a consistent degradation in overall performance, with the average score across MIRAMIND dropping (0.0202). The performance gap is more pronounced on diagnosis tasks (0.0358) and multi-step clinical reasoning tasks (0.0093), indicating weakened generalization to challenging and decision-sensitive cases.

Reasoning-chain evaluation further reveals a systematic quality gap. Compared with Mindora, Mindora w/o S1 shows substantially lower step adequacy (0.1667) and informativeness (0.1042), alongside reduced terminology economy (0.0584). Although both models often reach the same final answer, the variant without S1 tends to produce longer and less efficient reasoning chains, with additional steps that contribute limited new information.

Overall, these results indicate that S1 improves not only task accuracy but also the efficiency and structure of reasoning trajectories. By concentrating training on genuinely difficult cases, hard-case filtering encourages Mindora to learn concise, discriminative, and decision-oriented reasoning patterns rather than verbose or weakly informative explanations. A detailed case study is provided below.

Case Study of Hard-case Filtering (S1). To further illustrate the effect of hard-case filtering, we present a representative example from the mental health condition detection task, where both models predict the correct label (none), yet differ markedly in reasoning style. The full responses are shown in Figure 6.

The final Mindora model produces a compact and well-structured reasoning chain. Each step introduces a distinct decision-relevant check: (1) clarifying the task context, (2) examining the presence of emotional indicators, and (3) explicitly ruling out diagnostic criteria for depression, suicide, and PTSD. The reasoning terminates once sufficient evidence has been accumulated to justify the none classification. Notably, no speculative detours or meta-level uncertainty checks are introduced. This trajectory exemplifies high step adequacy and information efficiency, where every reasoning unit contributes directly to narrowing the decision space.

In contrast, the model trained without hard-case filtering is factually correct but noticeably more verbose. Its trajectory repeatedly restates symptom absence and adds a meta-level step (Considering Potential Misinterpretations) without introducing new evidence, resulting in redundant reasoning units and lower compressibility. This case suggests that hard-case filtering mainly changes how the model reasons rather than what it predicts: by focusing on difficult samples, Mindora learns to stop once the decision is sufficiently supported, avoiding unnecessary elaboration and speculative padding. This is important in mental-health settings, where over-explanation can obscure the core judgment and reduce interpretability.

T-SID Case Content:

Question: Please choose the most proper mental disorder of this post from: suicide, depression, ptsd, or none. Where 'none' means the poster shows no sign of mental disorder.
Context: @m_jxzFVNeQM I would like this specialist centre up (Chy Sawel) & running and there is interest from professionals willing to get involved.

Response of Mindora:

<think>
###Understanding the Context:
The context involves a desire to establish a specialist center and interest from professionals. This suggests a focus on a specific goal or project rather than personal emotional issues.
###Evaluating Emotional Indicators:
The statement lacks emotional indicators such as sadness, hopelessness, or trauma, which are typical signs of depression, suicide, or PTSD.
###Assessing for Signs of Mental Disorders:
There are no explicit mentions of symptoms related to depression, suicide, or PTSD. The focus is on professional involvement and establishing a center.
###Final Conclusion:
Based on the context and lack of emotional indicators, the most proper classification for the poster is 'none', indicating no sign of mental disorder.
</think>
<answer>Answer: none</answer>

Response of Mindora w/o S1 (i.e. the random sampling version):

<think>
###Understanding Context:
The context provided is a tweet expressing a desire for a specialist center to be established and mentions interest from professionals willing to get involved. This suggests a focus on professional and community involvement rather than personal mental health issues.
###Analyzing for Signs of Mental Disorders:
To determine if the poster shows signs of a mental disorder, we need to look for indicators of suicide, depression, or PTSD. The tweet does not mention any personal struggles, emotional distress, or traumatic experiences that would suggest these disorders.
###Evaluating for 'None':
The tweet is focused on professional and community initiatives, indicating no personal mental health concerns. Therefore, 'none' might be appropriate if the poster shows no signs of mental disorder.
###Considering Potential Misinterpretations:
While the tweet is about a professional initiative, it is important to ensure that there is no underlying personal mental health issue that is indirectly indicated. The tweet does not provide any such clues.
###Final Conclusion:
The poster does not show any signs of mental disorders such as suicide, depression, or PTSD. The tweet is focused on professional and community involvement, suggesting 'none' is the most appropriate classification.
</think>
<answer>Answer: none</answer>

Figure 6: Reasoning trajectory comparison between Mindora and Mindora w/o S1.

G.2 Structured Rewriting (S2)

We next analyze S2 (structured rewriting), which is primarily designed to improve the readability and pedagogical quality of training trajectories. Compared with the full model, removing S2 yields a substantial drop in overall performance (0.0468). The degradation is even more pronounced in trajectory-level evaluation: Step Adequacy sharply declines (0.6708), and Terminology Economy decreases (0.3500). In addition, structural reliability deteriorates, as reflected by higher Irrelevance Rate (+0.0095) and Contradiction Rate (+0.0229). These results indicate that training with non-rewritten, unstructured reasoning traces tends to introduce compressible or weakly connected content, increasing cognitive burden without strengthening decision support.

Importantly, S2 also affects the teacher quality of supervision. While strong reasoning models may still succeed with unstructured trajectories, such traces can be suboptimal for distillation into smaller models: they often contain backtracking artifacts and loosely organized explanations, which provide weaker step boundaries and less stable alignment between intermediate rationale and the final decision. By enforcing a standardized, modular format with an explicit final conclusion, S2 converts exploratory traces into consistent and learnable supervision signals, which are critical for transferring clinically aligned reasoning patterns to an 8B-scale student model. Notably, although Informativeness is slightly higher without S2 (+0.0125), this gain coincides with large drops in adequacy and increased inconsistency, suggesting that additional content is not necessarily more useful when it is not organized into decision-relevant reasoning steps. A detailed case study is provided below.

Case Study of Structured Rewriting (S2). To illustrate why S2 is critical for trajectory readability and distillation quality, we compare the full Mindora output with the unstructured variant on a representative cognitive error identification instance, as is shown in Figure 7. Both models predict the same label (personalization), but the reasoning traces exhibit substantially different instructional properties.

The final Mindora model produces a concise, modular explanation that explicitly isolates the target thought (I am responsible for her mistakes), directly maps it to the definition of personalization (misattributing external responsibility to oneself), and terminates with a brief conclusion that tightly links evidence to the predicted label. This organization yields clear step boundaries and minimal compressible content, making the trajectory straightforward to imitate and evaluate.

In contrast, the unstructured trace contains substantial exploratory artifacts (e.g., hedging, self-questioning, and backtracking). The model introduces alternative hypotheses such as should statements, then revises its decision after several paragraphs of reflection. Although the final answer is correct, much of the intermediate text is pedagogically ambiguous: it mixes tentative reasoning with narrative fillers (Hmm, But wait), and delays the decisive criterion for personalization until late in the chain. For distillation into smaller models, such traces can be suboptimal supervision signals, since the student may learn stylistic verbosity or heuristic option-surfing rather than a stable decision template grounded in the cognitive-error definition.

This example highlights the functional role of S2: it converts exploratory, non-linear reasoning traces into standardized, decision-aligned trajectories with explicit modular structure and a clear evidence-to-label mapping. As a result, S2 improves not only human readability but also the quality of supervision available for training compact models under limited capacity.

CognitiveReframing Case Content:

Question: Please choose the most proper cognitive error of this client from the following list: ...

Context:

- situation: I got a parking ticket because my sister forgot to register for overnight parking in front of my house.
- thought: I am responsible for her mistakes.

Response of Mindora:

###Understanding the Thought:

The thought expressed is *I am responsible for her mistakes.* This suggests that the individual feels a sense of accountability for the actions of their sister.

###Identifying Cognitive Error:

The thought process involves taking responsibility for the sister's mistake, which aligns with the cognitive error of *Personalization*. Personalization is when one takes things personally or makes them about oneself, even when the responsibility does not lie with them.

###Final Conclusion:

The cognitive error that best fits the thought *I am responsible for her mistakes* is *Personalization*, as it involves taking personal responsibility for the actions of another person.

Response of Mindora w/o S2 (i.e. the unstructured version):

<think>

Okay, so the person is feeling responsible for their sister's mistake. That's interesting because they're taking on this blame for something that wasn't really their fault. It's like they're saying, *I should have known better*, or *It's my fault she didn't register*.

Hmm, this sounds like they're taking things personally, right? They're letting this one mistake by their sister affect how they see themselves. It's almost like they're saying, *I'm not a good sibling because I didn't stop her from making this mistake*.

But wait, let's think about this a bit more. They're not just blaming themselves; they're also saying they're responsible for the mistake. That's a bit different. It's not just about feeling bad about themselves; it's about thinking they should have done something to prevent the mistake.

Oh, this could be *Should statements*. They might be thinking, *I should have made sure she registered*, or *I should have known she'd forget*. It's like they're setting a high bar for themselves and feeling guilty when they don't meet it.

But hold on, there's more to it. They're not just setting expectations; they're also feeling responsible for the outcome. That's a bit more about taking ownership of the situation, even though they didn't actually cause the mistake.

Hmm, maybe this is *Personalization*. They're making the mistake their own problem, even though it was their sister's action. It's like they're internalizing the blame, thinking it's their fault even though it's not.

Alright, so after thinking it through, it seems like *Personalization* fits best. They're taking the mistake and making it about themselves, even though they didn't do anything wrong. It's about how they're letting this one event define their sense of responsibility.

</think>

<answer>Answer: personalization</answer>

Figure 7: Reasoning trajectory comparison between Mindora and Mindora w/o S2.

DepSign Case Content:

Question: Please choose the most proper depression degree of this poster from: 0,1,2. Where 0 for not depressed, 1 for moderately depressed, and 2 for severely depressed.

Context:

I don't see any point being here. I'm not going to kill myself, so don't worry about that. It's just that I don't see any point in me existing. Whenever I look at my peers, they're always happy, have an abundance of friends with them and have this aura of confidence and charisma that attracts anyone. For me, it's the direct opposite; I have a limited amount of friends (or acquaintances more likely), I don't feel happy most of the time and have 0 charisma. Granted, I'm watching videos on how to gain charisma and be more likeable but the progress of it is very gradual. And added with my family and the dynamics of it being fucked up, slogging through college for a degree that I both don't want and not good at (even if I eventually cared about it) and not feeling happy, I shouldn't be here. I honestly feel that I'm taking someone's space in college and they would be better privileged to have it rather than me. It's something that I had on my mind for a while and wanted to put it out there somewhere. Even if it sounds bad grammatically or something like that, I just wanted to get it off my chest.

Figure 8: Reasoning trajectory comparison between Mindora and Mindora w/o S3 (1).

G.3 Consistency Reward (S3).

We analyze S3 (consistency reward) by comparing the full Mindora model with Mindora w/o S3. Removing S3 leads to a measurable degradation in overall performance: $\mathrm{Avg}_{\mathrm{all}}$ decreases from 0.6933 to 0.6788 (-0.0145). The effect is most salient on verification- and abstraction-style tasks that require stable evidence integration, where PSRS drops from 0.8379 to 0.6948 (-0.1431) and MentalMisinfo decreases from 0.7178 to 0.6796 (-0.0382).

Trajectory-level evaluation further confirms that S3 improves reasoning reliability. Without S3, **Logical Coherence** decreases (-0.0583) and **Informativeness** decreases (-0.0667). Usability-related metrics also decline, including **Step Adequacy** (-0.0375) and **Terminology Economy** (-0.0292). In addition, the **Irrelevance Rate** increases ($+0.0037$), indicating a higher fraction of weakly contributive steps.

Overall, these results suggest that S3 contributes beyond marginal accuracy gains: it promotes more coherent and informative reasoning trajectories while reducing irrelevant content, and is particularly beneficial for tasks where internal inconsistency or ungrounded statements can directly compromise reliability.

Case Study of Consistency Reward (S3) This case shown in Figure 8 involves severity grading for depression (label 1 = moderate). As is shown in Figure 9, both Mindora and Mindora w/o S3 reach the same final answer, but their reasoning traces differ in how consistently they connect evidence to the severity decision.

The trajectory of Mindora (with S3) exhibits a clearer evidence-to-decision alignment by explicitly separating risk exclusion (not going to kill themselves) from severity evidence (lack of purpose, taking up space), and grounding the severity level in a combined rationale: moderate depressive features are present, while an acute self-harm signal is absent. The final conclusion explicitly composes these two premises into a severity judgment, making the chain internally well-anchored and decision-oriented.

The ablated model is similar but less well calibrated: it introduces an unsupported criterion (not accompanied by extreme despair that would suggest severe depression), shifting from text-grounded justification to a generic severity template. Although both outputs remain coherent and reach the correct label, this case illustrates the role of the consistency reward: discouraging loosely grounded templates and encouraging conclusions that stay tightly coupled to observed evidence. This supports the aggregate finding that S3 improves trajectory reliability even when final-answer accuracy is unchanged.

Response of Mindora:

<think>
###Evaluating Suicidal Ideation:
The poster explicitly states they are not going to kill themselves, indicating a lack of immediate suicidal intent. This suggests they are not in a state of severe depression where self-harm is considered.
###Analyzing Feelings of Hopelessness:
The poster expresses a lack of purpose and feeling like they are taking up space in college. These feelings indicate a level of hopelessness and self-doubt, which are common in moderate depression.
###Assessing Social and Academic Stress:
The poster mentions social isolation and academic struggles, which contribute to feelings of inadequacy. These factors can lead to moderate depression, as they are significant stressors but not necessarily severe.
###Final Conclusion:
The poster's feelings of hopelessness and social inadequacy, combined with their lack of suicidal intent, suggest a moderate level of depression. Therefore, the most appropriate degree of depression for this poster is 1, indicating moderately depressed.
</think>
<answer>Answer: 1</answer>

Response of Mindora w/o S3 (i.e. without consistency reward):

<think>
###Assessing Suicidal Ideation:
The poster explicitly states they are not going to kill themselves, indicating a lack of immediate suicidal intent. This suggests they are not in a state of severe depression where suicide is a consideration.
###Evaluating Emotional State:
The poster expresses feelings of hopelessness and inadequacy, which are indicative of depression. However, these feelings are not accompanied by extreme despair or hopelessness that would suggest severe depression.
###Analyzing Social and Academic Stress: The poster mentions social isolation and academic stress, which contribute to their feelings of inadequacy. These are common stressors that can lead to moderate depression but do not necessarily indicate severe depression.
###Final Conclusion:
The poster's expression of hopelessness and inadequacy, combined with their lack of suicidal intent and common stressors, suggests a moderate level of depression. Therefore, the most appropriate depression degree for this poster is 1, indicating moderately depressed.
</think>
<answer>Answer: 1</answer>

Figure 9: Reasoning trajectory comparison between Mindora and Mindora w/o S3 (2).

Table 9: Few-shot in-context learning results on representative MIRAMIND datasets. CR = CognitiveReframing; PR = PatternReframe.

Model	Shot	CR	PR	TherapistQA	DepSign	SWMH	T-SID	MHQA	MedQA	MedMCQA	Misinfo
Qwen3-8B	0-shot	0.5941	0.6395	0.4057	0.4169	0.7358	0.7515	0.4009	0.5824	0.7927	0.4773
	1-shot	0.7434	0.6111	0.4896	0.4000	0.7541	0.8247	0.6105	0.8020	0.8651	0.6997
	3-shot	0.7461	0.6364	0.4667	0.4106	0.7679	0.8305	0.6091	0.8195	0.8808	0.7048
GPT-o4-mini	0-shot	0.7002	0.6607	0.4468	0.3607	0.7374	0.7515	0.5462	0.8925	0.8102	0.7781
	1-shot	0.8906	0.6842	0.5473	0.3673	0.7923	0.8114	0.7127	0.9524	0.9431	0.5167
	3-shot	0.8571	0.6696	0.5738	0.3827	0.7805	0.8218	0.7219	0.9386	0.9543	0.4982

G.4 Few-shot In-Context Learning Analysis

To further examine whether prompt-level adaptation can substitute for post-training, we conduct a small few-shot in-context learning (ICL) analysis following the example-selection paradigm of Medprompt (Nori et al., 2023). We evaluate GPT-o4-mini and Qwen3-8B under 0-shot, 1-shot, and 3-shot settings on representative MIRAMIND datasets. Table 9 reports the results.

The results show that ICL can improve some task scores, especially on structured classification and QA datasets. For example, Qwen3-8B improves substantially on CognitiveReframing, MHQA, MedQA, MedMCQA, and Misinfo under few-shot prompting, while GPT-o4-mini also benefits on several classification and QA tasks. However, the gains are not consistent across datasets. For instance, Qwen3-8B decreases on PatternReframe and DepSign under 1-shot prompting, and GPT-o4-mini shows a large drop on Misinfo under both 1-shot and 3-shot settings. These mixed results suggest that few-shot prompting is useful as a prompt-level adaptation strategy, but it is insufficient as a general solution for robust mental-health reasoning. In contrast, Mindora obtains more consistent gains through targeted post-training, supporting the necessity of our SFT–RL optimization and reasoning-trajectory design.