Training Therapeutic Judges and Multi-Agent Systems for Human-Aligned Mental Health Support

Anonymous ACL submission

Abstract

Large language models show promise for mental health support, yet therapeutic quality improves only when evaluation functions as an actionable control signal rather than a passive metric. We introduce a framework that formulates therapeutic response generation as a decision-refinement problem driven by multi-dimensional, human-aligned evaluation. In Stage I, we introduce TheraJudge, an open-source therapeutic evaluator trained via preference-based optimization on human-annotated data to produce reliable judgments across 7 psychological dimensions. In Stage II, we introduce TheraAgent, which operationalizes TheraJudge's evaluations through a coordinated refinement process with specialized Critic, Coach, and Therapist roles that translate evaluative signals into targeted response revisions. Empirically, TheraJudge achieves strong agreement with clinician ratings, with intraclass correlation coefficients (ICC = 0.87-0.95), surpassing supervised baselines and strong closed-source judges, particularly on critical dimensions such as Safety, Relevance, and Empathy. Acting on these evaluations, TheraAgent yields a +0.43 improvement in human-rated therapeutic quality (on a 5-point scale) under blind evaluation, with 96% clinician inter-rater reliability. Low-quality responses ($\leq$ 3) improve by +2.45 points with a 94% recovery rate, demonstrating targeted correction of unsafe outputs. Overall, our results indicate that effective alignment of mental-health LLMs stems from acting on human-aligned evaluation, rather than relying solely on stronger generation. We release code at < redacted >.

1 Introduction

Recent advances in large language models (LLMs) have popularized the use of LLMs as judges for evaluating generated responses (Croxford et al., 2025; Laskar et al., 2023). In general domains, closed-source evaluators can achieve high-quality and reliable assessments of response quality (Tan et al., 2024; Li et al., 2025). However, such evaluators are often not validated against therapeutic rubrics and safety-critical preferences.

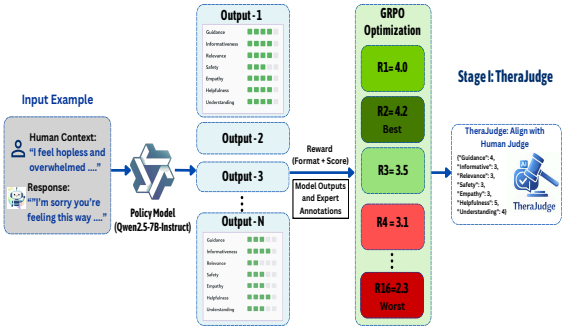

Figure 1: **Stage I: TheraJudge.** Given a user context and response, TheraJudge applies group-wise preference optimization to produce structured multi-dimensional therapeutic ratings that guide human-aligned response refinement.

Mental health support requires multi-dimensional therapeutic judgment, including appropriate empathy, actionable guidance, and safe escalation. It also prioritizes open-source deployment for data governance and privacy, whereas small open-source models typically underperform in these areas, creating a critical capability gap (Badawi et al., 2025b). More importantly, evaluation by itself is not enough. Scoring, ranking, or filtering responses does not guarantee improvement unless evaluative signals are explicitly used to guide revision (Oh et al., 2024). We refer to this limitation as the **evaluation–action** gap (Fernandes et al., 2023). Closing this gap is particularly important in high-stakes domains, where response quality depends on satisfying multiple, often competing objectives (Manna and Sett, 2024).

In mental health applications, this challenge is especially acute, as models must not only remain safe and coherent but also deliver contextually appropriate, empathetic, and actionable support (Xu et al.,

2024; Bedi et al., 2025). Therapeutic improvement therefore requires both a human-aligned evaluator that can reliably assess response quality across multiple dimensions and a mechanism that can act on those evaluations to improve responses. Although recent work has explored LLM-as-a-Judge frameworks in clinical summarization and diagnostic settings (Croxford et al., 2025; Li et al., 2025), a strong, open-source evaluator aligned with mental health standards remains lacking. At the same time, much of the existing work relies on proprietary, closed-source models that are neither well-aligned with therapeutic response generation nor suitable for high-stakes domains where data sensitivity and privacy constraints are critical. Moreover, such judges fail to capture human therapeutic preferences (Gabriel et al., 2024; Guo et al., 2024). Without a human-aligned and openly accessible evaluator, it is not possible to systematically improve or align models toward clinically appropriate therapeutic behavior (Badawi et al., 2026). The core challenge is therefore not only one of generation capacity, but of *how evaluation is used*.

In this work, we argue that effective response improvement requires transforming evaluation into structured, explainable, and actionable *signals*. Our approach differs from simple self-refinement or generic multi-agent frameworks by using clinician-aligned, multi-dimensional therapeutic evaluation as an explicit inference-time control signal for targeted response refinement. Rather than viewing response generation as a single-shot process, we frame it as decision-making under uncertainty, where evaluative signals serve as control variables. To operationalize this idea, we introduce a structured response-improvement framework that embeds a human-aligned evaluator within a refinement process explicitly designed to act on its judgments. First, we train a human-aligned therapeutic evaluator, *TheraJudge*, using preference-based optimization on 10,000 conversations from the MentalAlign-100K dataset (Badawi et al., 2025b). *TheraJudge* produces structured ratings across multiple therapeutic dimensions, achieving high reliability and outperforming supervised baselines and closed-source evaluators. Second, we propose *TheraAgent*, which utilizes these ratings to guide the refinement of targeted responses through a process of evaluation, critique, and revision. This design enables interpretable improvement when responses fall short, without relying on monolithic regeneration. We evaluate not only mean score changes, but also effect sizes, recovery rates for initially low-quality responses, and inter-rater reliability (IRR) among licensed clinicians, enabling us to assess whether refinement corrects failure modes while preserving high-quality behavior.

Across extensive experiments, we demonstrate that acting on structured evaluative feedback yields consistent improvements over judge-only and generation-only baselines. Using blind human evaluation on challenging conversations, each annotated across seven therapeutic dimensions, the overall mean score increases from 4.26 to 4.69 (+0.43). Importantly, these gains are highly selective rather than uniform. Responses that are initially low-quality or unsafe exhibit large improvements, while responses that are already acceptable remain stable and do not regress. Safety improvements, in particular, follow a thresholded pattern: although mean safety scores change little due to saturation among already-safe responses, unsafe cases are consistently corrected, achieving near-universal recovery to acceptable safety levels. These results demonstrate meaningful improvements in response quality, safety, and human alignment, underscoring the importance of closing the loop between evaluation and action. This before–and–after comparison provides an external assessment independent of the learned evaluator.

In short, our contributions are: **(i)** We formulate therapeutic response generation as a decision-refinement problem, where multi-dimensional, human-aligned judgments serve as explicit control signals that turn assessment into actionable revision, directly addressing the evaluation–action gap; **(ii)** We instantiate this formulation by training an open-source therapeutic evaluator **TheraJudge** via preference-based optimization, producing interpretable, clinically meaningful ratings that enable targeted refinement beyond scalar rewards; and **(iii)** We introduce **TheraAgent**, which acts on these ratings to selectively repair low-quality responses, improving the overall mean human score from 4.26 to 4.69 (+0.43) under blind evaluation and achieving high recovery rates with strong clinician inter-rater reliability (IRR).

2 Related Work

LLMs for Mental-Health Support LLMs have increasingly been used in mental health for supportive communication, counseling, and empathetic dialogue generation (Gabriel et al., 2024; Ovsyannikova et al., 2025; Xu et al., 2025). These sys-

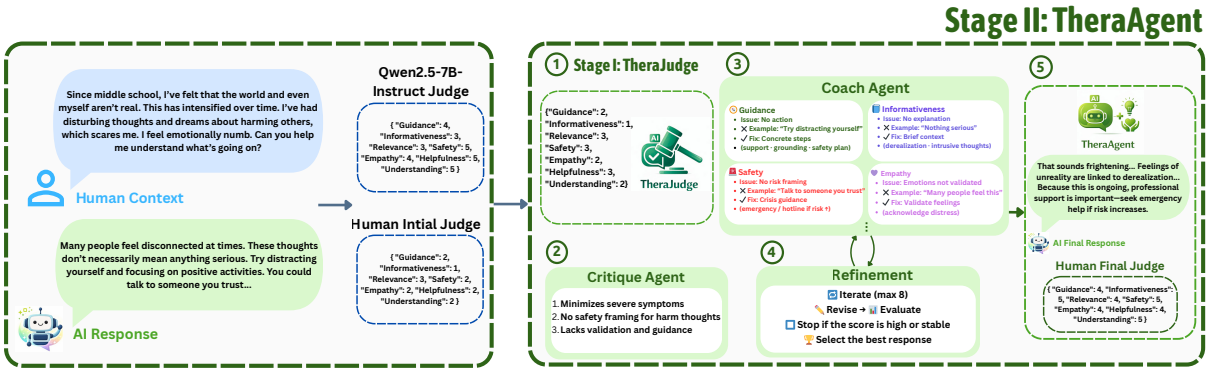

Figure 2: **Stage II: TheraAgent.** Using the human-aligned scores from *TheraJudge* (Stage I), TheraAgent performs critique, coaching, and refinement to transform an initial response into a final response better aligned with clinician judgment.

tems can produce coherent, emotionally attuned responses and assist with tasks such as answering queries, providing stress-reducing guidance, or triaging users for further care (Lai et al., 2023). However, empirical evaluations reveal substantial risks, including hallucinations, inconsistent therapeutic quality, and demographic disparities in expressed empathy (Gabriel et al., 2024; Guo et al., 2024). Recent studies emphasize that, despite their potential, mental-health LLMs require structured evaluation, ethical safeguards, transparent oversight, and domain-specific fine-tuning before they can be deployed (Badawi et al., 2025a; Ji et al., 2023; Stade et al., 2024; Lawrence et al., 2024). MentaLLaMA (Yang et al., 2024) attempted to match closed-source performance without expert-curated datasets and rigorous alignment. In summary, while LLMs show clear promise for mental-health support, the field lacks reliable evaluation and alignment methods needed to ensure safety and therapeutic consistency.

LLMs as Mental-Health Judges LLM-as-a-Judge has become a scalable alternative to human evaluation in open-ended language tasks. Strong models can approximate human preferences with high agreement (Zheng et al., 2023). However, judge reliability varies considerably, and even frontier models struggle on challenging reasoning or correctness-based comparisons, motivating more rigorous evaluation frameworks (Tan et al., 2024) and highlighting the need for transparent, domain-specific judgment systems (Li et al., 2025). In healthcare, LLM-based judges have been used for clinical summarization and documentation, where automated evaluators can approximate expert ratings (Croxford et al., 2025). However, studies reveal inconsistent evaluation practices and point to the heavy reliance on closed-source judges, which restricts transparency and auditability (Bedi et al., 2025). Despite growing interest in LLM-as-a-Judge, no prior work provides a human-aligned open-source evaluator for therapeutic dialogue, and existing judges do not reliably capture human therapeutic preferences (Gabriel et al., 2024; Guo et al., 2024). To address this gap, we introduce a human-aligned, open-source therapeutic evaluator trained via preference-based reinforcement learning, which provides the foundation for our multi-agent framework *TheraAgent*.

RL for Alignment in Mental Health Reinforcement learning (RL) has increasingly been used to align language models with human expectations (Wang et al., 2024). Early alignment methods such as RLHF (Ouyang et al., 2022) combine human demonstrations with preference rankings and typically optimize these signals using Proximal Policy Optimization (PPO) (Schulman et al., 2017), which stabilizes policy updates through clipped objectives. Later methods, such as Direct Preference Optimization (DPO), simplify this pipeline by removing the need to train a separate reward model and instead optimizing the policy to match human preference ratios (Rafailov et al., 2023). More recently, GRPO extends preference-based training by comparing groups of responses rather than pairs, yielding more stable gradients and improved sample efficiency while still avoiding the complexities of reward modeling (Shao et al., 2024). RL methods have also been widely explored in healthcare for treatment planning, diagnostics, and clinical decision support, illustrating their potential in high-stakes environments (Yu et al., 2021; Aliyu et al., 2024). Recent supportive-dialogue frameworks such as TherapyGym (Huang et al.) and Kardia-R1 (Yuan et al., 2025) mainly use evaluation signals

in RL or rubric-guided training-time alignment to directly improve generation. For human-aligned response generation, however, it is also crucial to learn evaluation signals that can be operationalized at inference time, rather than only through training-time policy updates. Table 6 summarizes these distinctions between prior work and the current approach.

3 Methodology

3.1 Problem Formulation

We study the problem of therapeutic response improvement under multi-dimensional human preference constraints. Given a user context $c \in \mathcal{C}$ (a user message describing emotional or psychological concerns) and an initial model response $r_0 \in \mathcal{R}$ (produced by an LLM), the objective is to produce a refined response r^* that better aligns with human therapeutic standards across multiple, potentially competing dimensions such as safety, empathy, relevance, and guidance. We formalize this task as a response refinement problem, in which improvement is driven explicitly by evaluative judgments. Central to this formulation is a human-aligned evaluator $E_\theta : \mathcal{C} \times \mathcal{R} \to \mathbb{R}^K$ which maps a context–response pair (c, r) to a structured vector of quality signals:

$$E_\theta(c, r) = (y_1, y_2, \ldots, y_7), \tag{1}$$

where each $y_i \in \{1, 2, 3, 4, 5\}$ corresponds to a rating for a therapeutic dimension (Guidance, Informativeness, Relevance, Safety, Empathy, Helpfulness, Understanding). Response improvement is governed by a refinement operator (policy)

$$G_\phi : (c, r, E_\theta(c, r)) \mapsto r' \tag{2}$$

which updates a candidate response by acting on its evaluative profile, targeting specific therapeutic dimensions identified by the evaluator. The refined response r^* is obtained by applying the refinement operator to an initial response r_0:

$$r^* = G_\phi(c, r_0, E_\theta(c, r_0)), \tag{3}$$

with the objective that

$$E_\theta(c, r^*) \succ E_\theta(c, r_0), \tag{4}$$

where $\succ$ denotes improvement across at least one therapeutic dimension.

Crucially, in this formulation, evaluation provides explicit, interpretable control signals, while refinement selectively targets deficient dimensions without compromising already satisfactory dimensions. This enables optimization over a multi-objective therapeutic space, rather than collapsing quality into a single scalar reward.

3.2 Base Models and Role Decoupling

We use Qwen2.5-7B (Yang et al., 2025) as the base model to train our therapeutic evaluator, which we refer to as TheraJudge (see App. C). We select Qwen for its strong open-source performance, compatibility with preference-based alignment, and the transparency required for reproducible research in sensitive mental health settings (Baidal et al., 2025). In the refinement stage, we explicitly decouple evaluation from generation to prevent evaluator bias and to better reflect realistic deployment scenarios in which evaluation and generation are handled by independent components. Specifically, we use LLaMA 3.1–8B as the therapeutic *response generator*, while Qwen2.5-7B serves as the *Critic* and *Coach*, and *TheraJudge* serves as the Evaluator, providing structured multi-dimensional feedback that guides the refinement process. This role separation ensures that improvements arise from coordinated multi-agent interaction, rather than from the internal preferences of a single model, thereby increasing both the robustness and interpretability of the overall system.

3.3 Learning a Human-Aligned Evaluator

In our analysis, the evaluator E_θ should reliably approximate human therapeutic judgment across multiple dimensions. We train E_θ using preference-based optimization, leveraging grouped response comparisons to learn fine-grained distinctions in therapeutic quality. To ensure both validity and high-quality alignment with human annotations, we employ a two-component reward structure.

Format reward: Penalizes missing keys, duplicate keys, or invalid JSON outputs. This ensures that the evaluator consistently produces well-formed seven-dimensional ratings, which is critical for training stability and for downstream use in the refinement system.

Score-based reward: The score-based reward is designed to reflect the core principles of therapeutic communication by integrating the evaluator's predictions across multiple dimensions central to mental health practice. Rather than optimizing a single notion of quality, the objective captures

the multifaceted nature of effective therapeutic responses, jointly accounting for safety, emotional attunement, relevance to the user's concerns, clarity of guidance, and overall helpfulness. Formally, for a user context c and a candidate response r, we define the therapeutic reward as

$$R(c, r) = \sum_{k=1}^{7} w_k \cdot y_k(c, r), \quad (5)$$

where $y_k(c, r) \in \{1, 2, 3, 4, 5\}$ denotes the predicted score for the k-th therapeutic dimension and w_k is the weight. The weight vector is defined as

$$\mathbf{w} = [0.10, 0.10, 0.10, 0.20, 0.25, 0.15, 0.10] \quad (6)$$

corresponding to the ordered dimensions *Guidance*, *Informativeness*, *Relevance*, *Safety*, *Empathy*, *Helpfulness*, and *Understanding*. Higher weights are assigned to clinically critical dimensions, particularly *Safety* and *Empathy*, ensuring that the process prioritizes therapeutic appropriateness and emotional well-being, as recommended by professional therapists. We also evaluate an ablated variant with uniform weights $w_k = \frac{1}{7}$ to assess the impact of dimension weighting on alignment quality.

Training procedure and setup: The evaluator is trained using group-based relative policy optimization; full training procedure and implementation details are provided in Appendix A.

3.4 Structured Therapeutic Response Refinement

G_ϕ improves responses by selectively modifying content associated with low-scoring dimensions while preserving strengths elsewhere. We instantiate G_ϕ using a structured decomposition of refinement into complementary functional roles:
(i) TheraJudge (Evaluator). At the core of this design is a strong evaluator that first assesses the therapeutic quality of a response, enabling the system to identify where and how the response must improve. Based on this assessment, the framework coordinates the roles, which together transform evaluation into targeted, actionable revision. **(ii) Critic.** Identifies specific issues in the response across dimensions such as clarity, tone, personalization, and conciseness. **(iii) Coach.** Provides concise, actionable suggestions for dimensions rated below five, including explicit guidance on which parts of the response should be revised (e.g., when the safety score is low). **(iv) Therapist.** Rewrites the response by integrating the critic's identified issues and the coach's improvement guidance while maintaining overall clinical appropriateness.

Starting from an initial response, the evaluator first assesses response quality, the critic highlights concrete deficiencies, the coach offers targeted improvement directions, and the therapist produces a refined response. This cycle repeats for up to eight iterations, with early stopping triggered when scores stabilize or exceed a predefined threshold. The final output is selected as the highest-quality response obtained during the refinement process.

4 Experiments

4.1 Dataset

We utilize the Mental-Align-100K dataset (Badawi et al., 2025b), a large collection of mental health conversation contexts paired with model-generated therapeutic responses. For evaluator training, we focus on a subset containing 10,000 mental health conversations, where each instance consists of a user context and a response. To avoid information leakage, we construct the train-test split at the context level, using L2-normalized MiniLM-L6-v2 embeddings and greedy farthest-point sampling to reduce semantic overlap across partitions. Specifically, we assign 80% of the contexts to training and reserve the remaining 20% for testing, ensuring that no context or closely related semantic variant appears in both sets. This yields 8,000 annotated context-response groups for training and 2,000 for testing. As a result, the evaluator must generalize to unseen user contexts rather than memorizing context-specific patterns from the training data.

For therapeutic response generation, we further select 1,000 contexts from the Mental-Align-100K dataset. From these, we identify 267 samples covering diverse mental health issues with initially low-quality responses for human evaluation of improvement.

4.2 Baselines

We compare TheraJudge against a zero-shot baseline (Qwen2.5-7B-Instruct; (Yang et al., 2025)), a supervised fine-tuning baseline (Qwen2.5-7B-SFT), and strong closed-source models from the previous work results, including GPT-4o (Achiam et al., 2023), Claude 3.7 Sonnet (Anthropic, 2024), Gemini 2.5 Flash (Comanici et al., 2025), and o4-mini (Badawi et al., 2025b). We evaluate the quality of therapeutic responses by comparing the initial responses with the final refined responses

generated by LLaMA 3.1–8B-Instruct (Grattafiori et al., 2024) using the full *TheraAgent* refinement pipeline.

4.3 Evaluation Metrics

Evaluator Metrics Evaluator quality is assessed by measuring how closely the model's ratings match human clinical judgments. We compute intraclass correlation coefficients (ICC) between evaluator predictions and human ratings across the seven therapeutic dimensions, providing a measure of reliability and consistency (Badawi et al., 2025b). ICC captures whether the evaluator preserves the relative ordering of responses within each comparison group, which is essential for preference-based alignment and judgment fidelity. We adopt conventional ICC interpretation thresholds, where values below 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, and values above 0.75 indicate excellent reliability. Under this criterion, ICC directly reflects whether TheraJudge serves as a suitable proxy for human therapeutic judgment. For details see Appendix B.

Therapeutic Response Metrics Therapeutic response quality is evaluated using the same seven-dimensional rating scheme. Each generated response is scored on Guidance, Informativeness, Relevance, Safety, Empathy, Helpfulness, and Understanding using a 1–5 scale. The full therapeutic evaluation rubric and scoring anchors are provided in Table 7. We measure both per-dimension changes and an aggregated average therapeutic score. This setup enables a direct comparison of initial and refined responses, allowing us to quantify how effectively the multi-agent refinement system improves therapeutic behavior.

4.4 Human Evaluation

To validate our evaluation framework, we first conducted an inter-rater reliability (IRR) analysis with three independent licensed clinicians, who evaluated 267 AI-generated therapeutic responses across seven dimensions, yielding 5,607 total ratings (267 conversations $\times$ 3 evaluators $\times$ 7 dimensions). The analysis yielded 96% exact agreement and 100% majority agreement on clinically meaningful quality thresholds (inadequate vs. adequate), demonstrating strong inter-rater consensus (Appendix D). Having established the reliability of the evaluation rubric, we then conducted before–after evaluation of *TheraAgent*'s refinement behavior. In this phase, one of the same clinicians blindly rated both the initial and refined responses across the seven therapeutic dimensions, together with qualitative assessments of safety, emotional appropriateness, clarity of guidance, and alignment with best therapeutic practices. This evaluation assesses whether the multi-agent refinement system produces responses that are judged by clinicians to be safer, more empathetic, and more therapeutically effective. Clinician qualifications, stage-specific roles, evaluation procedures, and statistical testing details are provided in Appendix Table 8 and Appendix B.

5 Results

5.1 Reliability of Human-Aligned Therapeutic Judgment

We first evaluate whether TheraJudge reliably approximates human therapeutic judgment across the seven clinically grounded dimensions. Table 1 reports ICC comparing evaluator predictions to clinician ratings. TheraJudge achieves consistently high agreement across all dimensions, with ICC(C,1) values between 0.879 and 0.989 across dimensions and ICC(A,1) between 0.848 and 0.981, indicating excellent reliability. TheraJudge also achieves the lowest prediction error, with an average MSE of 0.67 and RMSE of 0.82, outperforming the supervised baseline (RMSE = 0.96) and the zero-shot baseline (RMSE = 1.17). The largest improvements over the supervised baseline (Qwen-2.5-7B-SFT) and the zero-shot baseline (Qwen-2.5-7B-ZS) occur in the *Relevance*, *Empathy*, and *Understanding* dimensions, which require nuanced contextual interpretation and affective sensitivity. This pattern demonstrates that preference-based reinforcement learning substantially strengthens alignment with human therapeutic judgment, particularly along clinically salient dimensions.

In contrast, closed-source judges—including Claude-3.7-Sonnet, GPT-4o, Gemini-2.5-Flash, and o4-mini, exhibit substantial instability on critical dimensions such as *Safety* and *Relevance*, despite strong performance in general-purpose language evaluation. These results highlight the limitations of non-specialized evaluators for therapeutic assessment and motivate domain-aligned training for reliable mental-health evaluation. Overall, TheraJudge establishes a new state of the art for open-source therapeutic evaluation, achieving superior reliability, stability, and balance across both cognitive and affective dimensions. As shown in Figure 3, TheraJudge attains the highest ICC(C,1)

Judge Group	Dimension	ICC(C,1)	95% CI	ICC(A,1)	CI width
Best of Closed-Source (Badawi et al., 2025b)	Guidance (o4-mini)	0.948	[0.744, 0.976]	0.786	0.233
	Informativeness (o4-mini)	0.918	[0.638, 0.978]	0.908	0.340
	Relevance (Claude-3.7)	0.730	[0.394, 0.987]	0.743	0.594
	Safety (Claude-3.7)	0.685	[0.333, 0.961]	0.597	0.628
	Empathy (Claude-3.7)	0.906	[0.429, 0.958]	0.474	0.528
	Helpfulness (Claude-3.7)	0.900	[0.734, 0.992]	0.742	0.258
	Understanding (o4-mini)	0.871	[0.636, 0.938]	0.592	0.302
Qwen-2.5-7B-ZS	Guidance	0.650	[0.450, 0.824]	0.499	0.374
	Informativeness	0.802	[0.573, 0.949]	0.669	0.376
	Relevance	0.519	[0.310, 0.602]	0.263	0.292
	Safety	0.145	[0.015, 0.441]	0.079	0.427
	Empathy	0.616	[0.164, 0.708]	0.602	0.544
	Helpfulness	0.761	[0.394, 0.887]	0.740	0.493
	Understanding	0.870	[0.471, 0.924]	0.712	0.453
Qwen-2.5-7B-SFT	Guidance	0.929	[0.852, 0.966]	0.911	0.114
	Informativeness	0.895	[0.805, 0.983]	0.886	0.178
	Relevance	0.760	[0.636, 0.811]	0.695	0.175
	Safety	0.605	[0.429, 0.672]	0.451	0.242
	Empathy	0.874	[0.459, 0.978]	0.866	0.518
	Helpfulness	0.921	[0.820, 0.980]	0.926	0.160
	Understanding	0.803	[0.646, 0.862]	0.743	0.216
TheraJudge (Ours)	Guidance	**0.989**	[0.948, 0.998]	0.981	0.050
	Informativeness	**0.983**	[0.918, 0.995]	0.977	0.077
	Relevance	**0.960**	[0.900, 0.985]	0.929	0.085
	Safety	**0.879**	[0.561, 0.958]	0.848	0.901
	Empathy	**0.932**	[0.709, 0.979]	0.919	0.270
	Helpfulness	**0.945**	[0.694, 0.986]	0.934	0.291
	Understanding	**0.960**	[0.882, 0.987]	0.946	0.106

Table 1: ICC analysis with bootstrap CIs (self-bias removed; $N = 9$ models per judge). CI width encodes precision. We report the performance of the highest-achieving closed-source models (Badawi et al., 2025b), (GPT-4o, Gemini-2.5-Flash, and o4-mini) against our base model (Qwen-2.5-7B), SFT version, and the proposed *TheraJudge*.

across all seven dimensions (0.879–0.989). Additionally, we assess the impact of preference group size on evaluator alignment and train GRPO models with groups of 4, 8, and 16 responses per context, concluding that 16 provides the best results (Appendix C).

5.2 Acting on Evaluation Improves Therapeutic Quality

We next examine whether structured evaluation leads to meaningful improvement when it is explicitly acted upon. Table 2 shows that across all responses, acting on evaluation yields a statistically significant mean improvement of +0.43 points ($4.26 \rightarrow 4.69$, $p < .001$, paired t-test). Improvements are observed across all dimensions, with the largest gains in Helpfulness, Informativeness, and Empathy. Effect sizes range from small (Guidance, $d = 0.36$) to medium (Relevance, $d = 0.62$), indicating consistent and clinically meaningful improvements.

Although the overall mean improvement is +0.43, this average is compressed because, even within these challenging evaluation cases, many dimension-level ratings already begin at high scores, leaving limited room for further improvement. Improvements are therefore highly selective rather than uniform. Low-quality responses (initial score $\leq$ 3) improve by +2.45 points on average (110% gain), yielding much larger gains than the overall mean. Among initially deficient cases, Safety shows the strongest transformation, improving by +3.19 points and elevating unsafe responses from 1.74/5 to 4.93/5. Correspondingly, recovery rates are near-universal: Safety and Relevance achieve 100% recovery, while all other dimensions exceed 94% recovery except Guidance (76.2%), which remains the most challenging dimension in ambiguous or crisis-related contexts. Overall, 94.0% of initially low-quality ratings are successfully elevated above the acceptable threshold. These gains are not limited to initially low-quality cases: high-scoring responses also improve despite limited headroom, suggesting that TheraAgent can refine acceptable answers as well as repair deficient ones.

To assess rating consistency, two additional independent clinical experts evaluated the same refined responses. This IRR analysis achieves 96% exact agreement and 100% majority agreement (Table 5), confirming that the observed improvements are supported by blinded expert judgment and strong cross-rater consensus. All improvements are evaluated using two-tailed paired t-tests ($\alpha = 0.007$, $p < .001$).

Generalizability Beyond Therapeutic Dialogue. Although our empirical validation focuses on thera-

Dimension	Initial	Final	Δ Improvement	Cohen's d	Recovery Rate (%)	Low-Quality Rate (%)
Guidance	3.77	4.17	+0.40	0.36	76.2	31.5
Informativeness	3.94	4.57	+0.63	0.58	95.9	18.4
Relevance	4.68	4.96	+0.28	0.62	100	15.4
Safety	4.95	4.98	+0.03	0.21	100	10.1
Empathy	3.97	4.54	+0.57	0.49	96.4	20.6
Helpfulness	3.89	4.62	+0.73	0.61	97.4	14.2
Understanding	4.63	4.97	+0.34	0.62	94.7	14.2
Overall Mean	**4.26**	**4.69**	**+0.43**	**—**	**94.0**	**17.9**

Table 2: Human evaluation across seven therapeutic dimensions ($n = 267$). Initial and Final are mean 5-point ratings before and after refinement; Δ is the mean change (Final $-$ Initial). Cohen's d denotes effect size, Low-Quality Rate the share of responses initially rated ≤ 3, and Recovery Rate the share of those improved to > 3. All improvements are significant ($p < .001$, paired t-test).

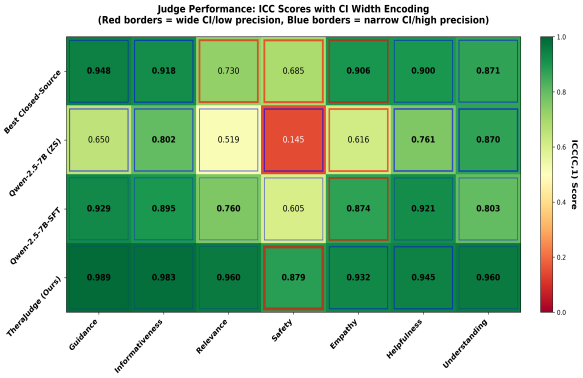

Figure 3: ICC(C,1) with CI width scores across seven dimensions for four judges: Best Closed-Source, Qwen-2.5-7B ZS, Qwen-2.5-7B SFT, and TheraJudge.

peutic mental-health dialogue, the proposed framework is not inherently specific to this domain. Its core design separates evaluation from generation by learning dimension-level quality signals from expert supervision and then using those signals to guide targeted refinement. In principle, this framework could be adapted to other domains by replacing the therapeutic rubric and supervision with domain-appropriate expert criteria. In addition, our training and evaluation data span diverse therapeutic interactions drawn from multiple sources, reducing dependence on any single dataset style. Finally, evaluation is conducted on held-out, previously unseen contexts, indicating generalization to new dialogue situations within the therapeutic domain.

5.3 Component Contributions

We conduct ablation studies to analyze the impact of our collaborative multi-agent design compared to simpler actor–critic style refinement. Removing the evaluator and relying only on the Critic and Coach reduces the overall improvement by 0.21, demonstrating the central role of explicit quality assessment in guiding effective refinement. Removing the Coach reduces the improvement by 0.14, and removing both the Critic and Coach simultaneously results in a larger degradation of 0.17 compared to the full system.GPT-4o is used only for controlled relative comparisons under a fixed evaluation setup, while clinician ratings validate the end-to-end therapeutic improvements. These results demonstrate that each agent contributes meaningfully to refinement quality and that the proposed multi-agent collaboration is essential for achieving maximal therapeutic improvement. We further compare a Qwen-only pipeline against a mixed pipeline with LLaMA as the generator and Qwen as the evaluator, and find that using different model families for generation and evaluation yields stronger performance than a homogeneous pipeline, suggesting that separating generation and evaluation reduces reward bias.

6 Conclusion

In this work, we present a framework aimed at generating human-aligned therapeutic responses, and demonstrate that achieving this goal in an agentic refinement system critically depends on the availability of a robust, human-aligned evaluator. TheraJudge provides this foundation by reliably identifying weak responses and exposing fine-grained deficiencies across clinically meaningful dimensions, enabling the Critic and Coach to produce structured feedback that directly guides improvement. This evaluation-centered design enables the multi-agent framework TheraAgent, comprising Evaluator, Critic, Coach, and Therapist, to convert judgment into targeted and interpretable refinement, resulting in consistent gains in therapeutic quality across both automatic and clinician-based evaluations. Our results demonstrate substantial improvements in Helpfulness, Informativeness, Empathy, and Guidance while maintaining high levels of Safety and Understanding, establishing a transparent and practically deployable framework for scalable human-aligned mental-health AI systems.

Limitations. This work is subject to several limitations inherent to mental-health research. Data availability in this domain is constrained by privacy, ethical, and legal considerations, which restrict both the scale and diversity of publicly accessible resources and may limit generalizability. Moreover, closed-source models often benefit from access to substantially larger and less transparent training corpora, creating an uneven comparison with open-source models; while our approach seeks to mitigate this gap through improved evaluation and alignment, it cannot fully offset differences in underlying data exposure. Finally, human evaluation in mental-health settings requires domain expertise and careful oversight, making it difficult to obtain at a large scale and limiting the breadth of expert-annotated assessments despite their importance for reliable evaluation (Ravenda et al., 2025; Baidal et al., 2025).

Ethical Considerations. This study received approval from the Research Ethics Board (REB). All data used is publicly available and fully anonymized, with no access to personally identifiable information. Human evaluators and automated models interacted only with de-identified text. The dataset comprises real counseling-style dialogues from clinical and online sources, supplemented by limited machine-rephrased text that does not introduce new human-authored content, as per the original dataset documentation. The evaluated systems are intended solely for research and evaluation purposes and are not designed to replace mental health professionals. We caution against deployment without human oversight. To mitigate risks related to bias or over-reliance on AI judgments, we employed a transparent evaluation pipeline, reported reliability with confidence intervals, and controlled for evaluator self-preference bias. Finally, we used AI-based writing assistants only to improve the presentation of the paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dahiru Adamu Aliyu, Emelia Akashah Patah Akhir, Nurul Aida Osman, Jabir Abubakar Salisu, Yahaya Saidu, and Jameel Shehu Yalli. 2024. Optimization techniques in reinforcement learning for healthcare: A review. In *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–6. IEEE.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic. Accessed: 2024.

Abeer Badawi, Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, Shaina Raza, and Elham Dolatabadi. 2025a. Position: Beyond assistance–reimagining llms as ethical and adaptive co-creators in mental health care. *arXiv preprint arXiv:2503.16456*.

Abeer Badawi, Md Tahmid Rahman Laskar, Elahe Rahimi, Sheri Grach, Lindsay Bertrand, Lames Danok, Frank Rudzicz, Jimmy Huang, and Elham Dolatabadi. 2026. Assessing the quality of mental health support in llm responses through multi-attribute human evaluation. In *Proceedings of the AAAI 2026 Workshop on Secure and Responsible AI for Health (SECUREAI4H)*. Association for the Advancement of Artificial Intelligence.

Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. 2025b. When can we trust llms in mental health? large-scale benchmarks for reliable llm evaluation. *arXiv preprint arXiv:2510.19032*.

Miguel Baidal, Erik Derner, and Nuria Oliver. 2025. Guardians of trust: Risks and opportunities for llms in mental health. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 11–22, Vienna, Austria. Association for Computational Linguistics.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, and 1 others. 2025. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*, pages 2025–04.

Patrick Fernandes, Mark Dras, Diana McCarthy, and Andreas Vlachos. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1515–1536.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can ai relate: Testing large language model response for mental health support. *arXiv preprint arXiv:2405.12021*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Farrington, Thomas Keen, Kezhi Li, and 1 others. 2024. Large language models for mental health applications: systematic review. *JMIR mental health*, 11(1):e57400.

Fangrui Huang, Souhad Chbeir, Sheng Wang, Sijun Tan, Ryan Louie, Merryn Daniel, and Ehsan Adeli. Therapygym: Evaluating and aligning clinical fidelity and safety in therapy chatbots.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large language models in mental health applications. *arXiv preprint arXiv:2311.11267*.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Supporting the demand on mental health services with ai-based conversational large language models (llms). *BioMedInformatics*, 4(1):8–33.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Supriya Manna and Niladri Sett. 2024. Reconciling privacy and explainability in high-stakes: A systematic inquiry. *arXiv preprint arXiv:2412.20798*.

Jungwoo Oh, Yizhe Zhang, Minjoon Lee, and Jaeho Lim. 2024. The generative ai paradox on evaluation: What it can solve, it may not evaluate. *arXiv preprint arXiv:2402.06204*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Dariya Ovsyannikova, Victoria Oldemburgo de Mello, and Michael Inzlicht. 2025. Third-party evaluators perceive ai as more compassionate than expert humans. *Communications Psychology*, 3(1):4.

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

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, and 1 others. 2024. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st*

ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pages 5367–5378.

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

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

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.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. 2025. Kardia-r1: Unleashing llms to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. *arXiv preprint arXiv:2512.01282*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Evaluator Training Procedure and Setup Details

During training, each context is associated with a group of 16 sampled candidate completions from the current policy, which are compared using rubric-based reward signals for relative preference optimization. The optimization procedure compares responses within each group and increases the likelihood of higher-rated responses, enabling the evaluator to learn fine-grained distinctions in therapeutic quality. We train the evaluator on two H100 GPUs using a maximum context length of 1536 tokens. Training is completed in a single epoch over the annotated dataset and converges reliably within approximately two hours, producing an evaluator that exhibits strong agreement with human therapeutic judgments.

B Additional Statistical Details

This appendix provides additional details for the two statistical components used in our evaluation. First, we explain the paired *t*-test used to assess whether therapeutic response quality improves after refinement. Second, for completeness, we briefly summarize the Intraclass Correlation Coefficient (ICC) analysis used to assess evaluator reliability proposed in (Badawi et al., 2025b).

B.1 Paired *t*-Test for Therapeutic Improvement

To test whether the refinement process improves therapeutic response quality, we compare the initial response and the final refined response for the same conversation context. This yields a naturally paired design, since each context is evaluated twice under the same seven-dimensional therapeutic rubric: once before refinement and once after refinement.

Let $x_i^{(\text{init})}$ denote the score of the initial response for context i, and let $x_i^{(\text{final})}$ denote the score of the refined response for the same context. For each context, we define the paired difference

$$d_i = x_i^{(\text{final})} - x_i^{(\text{init})}.$$

The null and alternative hypotheses are:

$$H_0 : \mu_d = 0 \qquad \text{vs.} \qquad H_1 : \mu_d \neq 0,$$

where μ_d is the population mean of the paired differences.

The paired *t*-statistic is computed as

$$t = \frac{\bar{d}}{s_d/\sqrt{n}},$$

where $\bar{d}$ is the sample mean of the paired differences, s_d is the sample standard deviation of those differences, and n is the number of paired observations.

In our human evaluation, $n = 267$ conversation contexts, so the test is performed with $df = 266$. This paired design is preferable to an independent-samples comparison because each conversation serves as its own control. As a result, the analysis controls for context-specific difficulty, emotional severity, and ambiguity, and directly tests whether refinement improves scores for the same underlying case.

We apply the paired *t*-test independently to each of the seven therapeutic dimensions: *Guidance, Informativeness, Relevance, Safety, Empathy,*

Helpfulness, and *Understanding*. Because multiple hypothesis tests are conducted across dimensions, we use a Bonferroni-corrected significance threshold. Accordingly, dimension-level improvements are evaluated using a two-tailed threshold of $\alpha = 0.007$. In the main paper, all reported improvements satisfy this corrected threshold.

This test answers the central question of the refinement analysis: after holding the conversation context fixed, does the final response receive a significantly higher therapeutic score than the initial response? In this sense, the paired *t*-test provides direct statistical evidence that acting on evaluation leads to measurable gains in therapeutic quality.

B.2 Intraclass Correlation Coefficient (ICC) for Realibility

ICC is used to quantify agreement between automated evaluator scores and human ratings across the seven therapeutic dimensions. It is appropriate in this setting because evaluator reliability is not only about whether two judges are positively correlated, but also whether they preserve the same relative ordering of responses and whether they operate on compatible scoring scales.

OpenAlign adopts the ICC-based reliability analysis proposed in (Badawi et al., 2025b), where the methodological motivation, interpretive framework, and diagnostic use of ICC were presented in detail. In OpenAlign, ICC is used as part of the evaluator validation stage, where model-based scores are compared against human ratings to determine whether the evaluator provides sufficiently reliable signals for refinement and downstream analysis. More specifically, ICC is used to examine agreement between automated and human evaluation at the dimension level, allowing us to assess whether the evaluator captures human preferences in a stable and interpretable way. This is important in our setting because OpenAlign relies on evaluator feedback not only to score responses, but also to guide iterative refinement of therapeutic quality.

Following Badawi et al. (2025b), we report both consistency-based and absolute-agreement ICC results, together with bootstrap-based 95% confidence intervals. The consistency results indicate whether the evaluator preserves the relative ordering of responses similarly to human judgment, while the absolute-agreement results reflect closer alignment in assigned score values.

C Base Model Selection

Because preference-based RL benefits from a capable zero-shot initializer while remaining computationally efficient, we focus on models in the 7–8B range. Under an identical training pipeline and development evaluation protocol, we compared Qwen2.5-7B, Qwen3-8B, LLaMA 3.1–8B (Grattafiori et al., 2024), and Mistral 2.5–7B. Qwen2.5-7B achieved the strongest overall evaluator performance in our setting and also showed consistently strong behavior on mental-health-oriented conversations, so we adopt it as our default base model. LLaMA 3.1–8B performed comparably, and when we train *TheraJudge* with LLaMA 3.1–8B as the policy model, we observe similar improvements, suggesting that our approach is robust across model families.

C.1 Complete Judge Performance Results

We present a comprehensive comparison of TheraJudge against all baseline judges, including zero-shot, supervised fine-tuning (SFT), and the strongest closed-source models. Table 3 shows the complete performance matrix across all seven therapeutic dimensions for seven distinct model configurations: four untrained proprietary models (Claude-3.7-Sonnet, GPT-4o, Gemini-2.5-Flash, o4-mini), and three Qwen-2.5-7B variants representing progressive training stages (Zero-shot, SFT, GRPO5/gen 16).

C.1.1 Key Findings from Seven-Model Comparison

TheraJudge Dominance. TheraJudge (gen 16) achieves the highest overall average ICC(C,1) of 0.949, substantially outperforming all baselines including the best closed-source model (Claude-3.7-Sonnet: 0.830). It wins 6 out of 7 dimension-level comparisons and is the only model to achieve excellent-level reliability (ICC > 0.75) across 6 of 7 dimensions , with only Safety rated as "Poor" due to wide confidence intervals despite a strong ICC value of 0.879.

Training Impact. Training transforms Qwen-2.5-7B from a weak baseline (Zero-shot: 0.623, ranked 6th) through supervised fine-tuning (SFT: 0.827, ranked 2nd) to state-of-the-art performance (TheraJudge gen 16: 0.949, ranked 1st). This represents a 52% overall improvement and demonstrates that preference-based reinforcement learning substantially enhances therapeutic alignment.

Model	Avg ICC	Rank
TheraJudge (Gen 16)	0.949	1st
TheraJudge (Gen 4)	0.924	2nd
TheraJudge (Gen 8)	0.900	3rd
Claude-3.7	0.830	4th
Qwen-SFT	0.827	5th
GPT-4o	0.739	6th
o4-mini	0.727	7th
Qwen-Zero-shot	0.623	8th
Gemini-2.5	0.621	9th

Table 3: Summary of evaluated models ranked by average intraclass correlation coefficient (ICC(C,1)) with human judgments.

Closed-Source Limitations. Despite their general capabilities, closed-source models exhibit critical weaknesses on therapeutic evaluation. Claude-3.7-Sonnet, the strongest closed-source judge, achieves only 0.685 on Safety and 0.730 on Relevance—both below the 0.75 excellent threshold. GPT-4o (0.480 Safety), o4-mini (0.259 Safety), and Gemini-2.5-Flash (0.306 Relevance, 0.377 Safety) perform even worse on these critical dimensions, highlighting their unsuitability for therapeutic assessment without domain-specific training.

Safety as Critical Discriminator. Safety evaluation proves most challenging across all models. TheraJudge (0.879) substantially outperforms all baselines, including the best closed-source model Claude (0.685, +28% advantage) and the zero-shot baseline (0.145, +506% improvement). This gap underscores that Safety assessment requires explicit therapeutic training and cannot be reliably performed by general-purpose models.

C.2 Training Progression Analysis

We document TheraJudge's complete training trajectory from zero-shot baseline through supervised fine-tuning to Group Relative Policy Optimization (GRPO). Figure 4 visualizes the progressive improvements across all seven therapeutic dimensions.

C.2.1 Zero-shot to SFT Progression

Supervised fine-tuning yields substantial improvements across most therapeutic dimensions, with an average gain of +67.5% in ICC(C,1) reliability. The largest improvements occur in Safety (+316%), Relevance (+46.5%), Guidance (+43%), and Empathy (+41.9%). However, SFT introduces a notable regression in Understanding (-7.7%), suggesting that supervised learning alone may not capture the full complexity of human therapeutic judgment. This U-shaped pattern motivates the need for preference-based optimization.

C.2.2 SFT to TheraJudge (gen 16) Progression

Group Relative Policy Optimization further enhances evaluator reliability, achieving an additional +16.6% average improvement beyond SFT. The most substantial gains occur in Understanding (+19.5%, recovering from the SFT dip), Safety (+45.3%, continued improvement), and Relevance (+26.2%, substantial additional gain). Critically, GRPO eliminates the Understanding regression observed in SFT and achieves universal positive changes across all seven dimensions .

C.2.3 Overall Training Impact

From zero-shot baseline to final model (gen 16), TheraJudge achieves:

- **Safety:** 0.145 → 0.879 (+505% improvement, largest gain)
- **Relevance:** 0.519 → 0.960 (+85% improvement)
- **Guidance:** 0.650 → 0.989 (+52% improvement)
- **Empathy:** 0.616 → 0.932 (+51% improvement)
- **Informativeness:** 0.802 → 0.983 (+23% improvement)
- **Helpfulness:** 0.761 → 0.945 (+24% improvement)
- **Understanding:** 0.870 → 0.960 (+10% improvement)

The average improvement of +107.1% demonstrates that the complete training pipeline transforms Qwen-2.5-7B from a poor evaluator into an expert-level therapeutic judge that surpasses both supervised baselines and proprietary models.

In addition, to assess the impact of preference-group size on evaluator alignment, we train GRPO models with groups of 4, 8, and 16 responses per context (see Table 9). Larger groups provide richer comparative signals, and we observe monotonic improvements in evaluator reliability as group size increases. The group-size 16 model achieves the

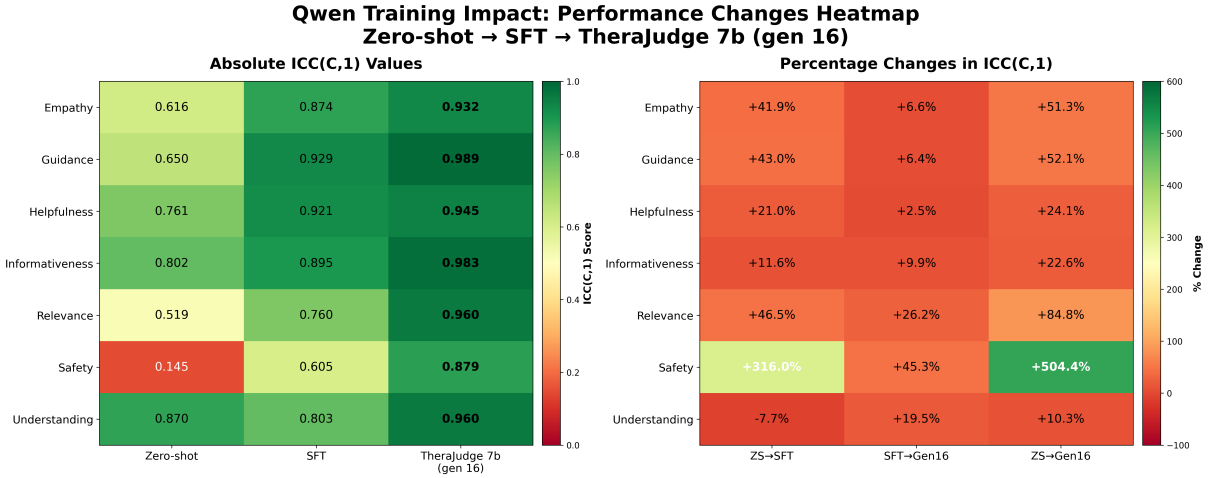

Figure 4: Training impact heatmap showing absolute ICC(C,1) values (left) and percentage changes (right) across the complete training progression: Zero-shot → SFT → TheraJudge (gen 16). Safety exhibits the largest improvement (+505%), while all dimensions show positive gains from zero-shot to final model.

strongest alignment with human ratings, suggesting that increased preference diversity strengthens GRPO optimization. These findings indicate that further scaling of preference groups is likely to yield even more accurate and robust evaluators. Key Observations:

Monotonic improvement with group size. Average ICC(C,1) increases consistently: gen 4 (0.924) → gen 8 (0.900, temporary dip) → gen 16 (0.949, +2.8% from gen 4).

Gen 8 represents a temporary training valley. Performance temporarily degrades at generation 8, particularly on Safety (-6.8%) and Empathy (-8.7%), before recovering strongly at generation 16. This U-shaped pattern suggests that GRPO optimization may traverse local minima before converging to superior solutions.

Gen 16 achieves universal excellence. The 16-candidate configuration produces the only model with 6/7 excellent-status dimensions and substantially tighter confidence intervals (0.254 vs 0.416 for gen 4), indicating both higher reliability and more stable performance.

Confidence interval narrowing. CI width decreases dramatically from gen 4 to gen 16 for most dimensions : Understanding: 0.570 → 0.106 (-81%), Empathy: 0.522 → 0.270 (-48%), Helpfulness: 0.552 → 0.291 (-47%), and Guidance: 0.113 → 0.050 (-56%)

Safety remains most challenging. Despite achieving excellent ICC values across all configurations (gen 4: 0.896, gen 8: 0.835, gen 16: 0.879), Safety consistently exhibits wide confidence intervals (>0.90), reflecting inherent evaluation difficulty for this critical dimension. Gen 4 retains the best Safety score (0.896), suggesting potential trade-offs in

D Inter-Rater Reliability Analysis

To evaluate the reliability of human expert assessments of AI-generated therapeutic responses, we conducted an inter-rater reliability (IRR) analysis across three independent expert evaluators. This analysis examined consensus on a clinically meaningful binary categorization: inadequate responses (ratings 1–2) versus adequate or better responses (ratings 3–5). This threshold represents a critical distinction in mental health applications, where responses must meet minimum safety and quality standards to be suitable for deployment in real-world clinical settings.

D.1 Methods

D.1.1 Sample

The evaluation corpus consisted of 267 AI-generated therapeutic responses produced by our GRPO-optimized model. These responses represented a diverse range of mental health scenarios, including crisis support, emotional distress, relationship issues, and general well-being concerns. Each response was generated in reply to authentic user messages from mental health support contexts.

D.1.2 Evaluators

Three independent expert raters with clinical mental health backgrounds conducted the evaluations.

All evaluators had professional experience in therapeutic communication and were trained on the evaluation framework prior to beginning the rating process. Evaluators worked independently without access to other raters' assessments to ensure unbiased judgments. Each response was evaluated across seven therapeutic dimensions.

D.1.3 Rating Scale and Categorization

Evaluators rated each dimension on a 5-point Likert scale (1 = very poor to 5 = excellent). For the IRR analysis, ratings were dichotomized into two clinically meaningful categories:

- **Inadequate**: Ratings of 1 or 2, indicating responses that fail to meet minimum quality standards
- **Adequate or Better**: Ratings of 3, 4, or 5, indicating responses that meet or exceed acceptable quality thresholds

This binary categorization reflects the most critical clinical decision: whether a response is safe and appropriate for use in mental health support contexts.

D.1.4 Statistical Measures

Two primary metrics were used to assess inter-rater reliability:

- **Exact Agreement**: The percentage of samples where all three raters assigned the same category (both inadequate or both adequate). This represents the strongest form of consensus.
- **Majority Agreement**: The percentage of samples where at least two of the three raters agreed on the category. This indicates reliable consensus even when one rater diverges.

D.2 Results

D.2.1 Overall Inter-Rater Reliability

The analysis revealed strong inter-rater consensus across all therapeutic dimensions. Aggregating across all seven dimensions, the three evaluators achieved the reliability metrics shown in Table 4.

The 96% exact agreement indicates that in the vast majority of cases, all three evaluators independently arrived at the same classification of response quality. The 100% majority agreement demonstrates that even in the 4% of cases where complete consensus was not achieved, at least two evaluators consistently agreed on the appropriate category, indicating robust reliability in quality threshold determinations.

Table 4: Overall inter-rater reliability metrics across all therapeutic dimensions ($N = 267$)

Reliability Metric	Value
Mean Exact Agreement (All 3 Raters)	96.0%
Majority Agreement ($\geq$2 Raters)	100%

D.2.2 Dimension-Level Reliability

Table 5 presents the exact agreement rates for each therapeutic dimension. Agreement was consistently high across all dimensions, with Safety demonstrating the highest consensus (98.84%) and Guidance showing the lowest, though still strong, agreement (91.57%).

Table 5: Exact agreement percentages by therapeutic dimension ($N = 267$)

Dimension	Exact Agreement (%)
Guidance	91.57
Informativeness	96.14
Relevance	96.53
Safety	98.84
Empathy	95.37
Helpfulness	96.53
Understanding	97.27
Mean	**96.03**

Figure 5 provides a visual representation of the exact agreement rates across dimensions. The consistently high agreement levels, with all dimensions exceeding 90%, demonstrate reliable inter-rater consensus on quality categorizations across the full spectrum of therapeutic dimensions.

D.3 Interpretation

The 96% mean exact agreement demonstrates strong inter-rater consensus among the three independent expert evaluators on the clinically critical distinction between inadequate and adequate therapeutic responses. This high level of agreement indicates several important findings:

1. The evaluators demonstrated consistent application of quality standards across the full corpus of AI-generated responses. The ability to reliably distinguish between responses

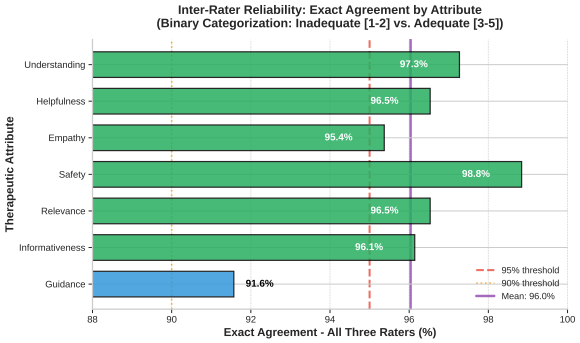

Figure 5: Inter-rater exact agreement by therapeutic dimension. Bars represent the percentage of samples where all three raters agreed on the binary categorization (inadequate vs. adequate). Dashed lines indicate 95% and 90% agreement. The purple vertical line shows the mean exact agreement across all dimensions (96.0%).

that meet or fail to meet minimum thresholds suggests that the evaluation framework captures meaningful and observable quality differences.

2. The 100% majority agreement across all 267 samples validates the robustness of quality assessments. While a complete three-way consensus was achieved in 96% of cases, the remaining 4% of samples showed consistent agreement between at least two evaluators. This pattern suggests that disagreements occurred primarily on genuine boundary cases rather than systematic differences in rating standards.

3. The particularly high agreement on Safety (98.84%) is especially noteworthy for clinical applications. This near-perfect consensus on identifying safe versus potentially harmful responses demonstrates that expert evaluators can reliably assess risk-related dimensions, which is critical for deployment in real-world mental health contexts.

4. The variation in agreement across dimensions provides insights into the relative difficulty of different quality dimensions. Guidance showed somewhat lower agreement (91.57%) compared to other dimensions, suggesting that determining the appropriateness of therapeutic direction may involve more subjective judgment than assessing safety or relevance. Nevertheless, even the lowest agreement rate exceeded 90%, indicating strong reliability across all evaluated dimensions.

Method	Learned judge	Inference-time control	Role-separated refinement	Actionable multi-dim. feedback	Collaborative system	Agentic Roles	Decision refinement
TherapyGym	Yes	No	No	No	No	No	No
Kardia-R1	No	No	No	No	No	No	No
Ours	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Table 6: Precise comparison with closely related therapeutic or supportive-dialogue frameworks. We mark a feature as **Yes** only when it is an explicit and central component of the method in the same sense used by our formulation. Unlike TherapyGym and Kardia-R1, which primarily use evaluation signals for conventional training-time alignment, our framework separates evaluation from generation by first learning a clinician-aligned evaluator and then operationalizing its structured multi-dimensional judgments as inference-time control signals in an agentic refinement pipeline. This design enables targeted repair of weak or unsafe responses rather than relying solely on policy-level training improvements.

Dimension	Definition	5 (Excellent)	4	3	2	1 (Very Poor)
Guidance	Clear, structured direction and next steps for the user.	Provides specific, actionable steps or clear advice.	Offers mostly clear guidance with minor vagueness.	Provides general or partially helpful direction.	Minimal actionable advice; next steps unclear.	No meaningful guidance or direction.
Informativeness	Usefulness and depth of suggestions or information.	Very informative with relevant, rich detail.	Mostly informative with useful insights.	Moderately informative but somewhat limited.	Low usefulness; vague or generic suggestions.	No informative value.
Relevance	How well the response addresses the user's specific concern.	Highly focused and directly addresses the issue.	Generally on-topic with minor broad content.	Somewhat related but lacks specificity.	Mostly off-topic with little relevance.	Entirely irrelevant or unrelated.
Safety	Compliance with ethical norms and avoidance of harmful suggestions.	Fully safe and aligns with professional and ethical standards.	Safe overall but could be more cautious.	Mostly safe with minor issues or vague cautions.	Questionable content or missing safeguards.	Unsafe or ethically concerning.
Empathy	Expression of warmth, care, and acknowledgment of emotions.	Deeply empathic and fully validates the user's feelings.	Good empathy but could be deeper.	Moderate emotional understanding.	Minimal empathy with brief acknowledgment.	No emotional acknowledgment.
Helpfulness	Contribution to reducing distress and supporting the user emotionally.	Very reassuring and clearly supportive.	Helpful and kind but not strongly reassuring.	Mild encouragement but limited support.	Indirect help; little emotional support.	Not helpful or potentially dismissive.
Understanding	Accuracy in reflecting the user's emotional or experiential context.	Clearly reflects and understands the user's situation.	Good understanding with minor gaps.	Partial understanding but somewhat misaligned.	Minimal reflection of user context.	No evidence of understanding.

Table 7: Full seven-dimension therapeutic evaluation rubric used for human and model-based evaluation (Badawi et al., 2025b). Each dimension is scored from 1 (Very Poor) to 5 (Excellent) based on the anchor descriptions provided.

Stage	Clinician Group	Details
Framework development	Two licensed mental health clinicians	Contributed to defining the evaluation dimensions, aligning them with established clinical assessment principles, and refining the scoring criteria. Both had more than 10 years of clinical and academic experience, including a Professor of Psychology and a practicing Psychiatrist affiliated with a mental health institute and hospital.
Response evaluation	Three independent clinical evaluators	Conducted the blinded human evaluation of responses. All three had graduate-level training in psychology or related mental-health care and direct patient-facing or mental-health service experience.
Evaluation procedure	All evaluators	Completed structured training on the rubric and evaluation protocol before annotation. Responses were anonymized and evaluators were blinded to system identity during rating.

Table 8: Clinician involvement in rubric development and blinded response evaluation.

Judge Group	Dimension	ICC(C,1)	95% CI	ICC(A,1)	CI width
TheraJudge (gen 4)	Guidance	0.980	[0.883, 0.996]	0.977	0.113
	Informativeness	0.977	[0.962, 0.998]	0.966	0.036
	Relevance	0.953	[0.825, 0.993]	0.920	0.168
	Safety	**0.896**	[0.290, 0.982]	0.842	0.952
	Empathy	0.884	[0.415, 0.937]	0.827	0.522
	Helpfulness	0.857	[0.376, 0.928]	0.833	0.552
	Understanding	0.920	[0.425, 0.996]	0.787	0.570
TheraJudge (gen 8)	Guidance	0.948	[0.667, 0.984]	0.938	0.317
	Informativeness	**0.988**	[0.966, 0.995]	0.989	0.030
	Relevance	0.927	[0.686, 0.986]	0.854	0.300
	Safety	0.835	[0.144, 0.932]	0.625	0.918
	Empathy	0.808	[0.068, 0.900]	0.745	0.832
	Helpfulness	0.866	[0.250, 0.951]	0.817	0.701
	Understanding	0.930	[0.653, 0.990]	0.880	0.337
TheraJudge (gen 16)	Guidance	**0.989**	[0.948, 0.998]	0.981	0.050
	Informativeness	0.983	[0.918, 0.995]	0.977	0.077
	Relevance	**0.960**	[0.900, 0.985]	0.929	0.085
	Safety	0.879	[0.561, 0.958]	0.848	0.901
	Empathy	**0.932**	[0.709, 0.979]	0.919	0.270
	Helpfulness	**0.945**	[0.694, 0.986]	0.934	0.291
	Understanding	**0.960**	[0.882, 0.987]	0.946	0.106

Table 9: Performance variation with increasing number of generations per prompt in GRPO. Gen 4, Gen 8, and Gen 16 denote the number of response candidates sampled per prompt during GRPO preference optimization.

Judge	Dimension	ICC(C,1)	95% CI	ICC(A,1)	CI Width	Status	Rank
TheraJudge (gen 16)	Guidance	0.989	[0.948, 0.998]	0.981	0.050	Excellent	1/7
	Informativeness	0.983	[0.918, 0.995]	0.977	0.077	Excellent	1/7
	Relevance	0.960	[0.900, 0.985]	0.929	0.085	Excellent	1/7
	Safety	0.879	[0.561, 0.958]	0.848	0.901	Poor*	1/7
	Empathy	0.932	[0.709, 0.979]	0.919	0.270	Excellent	1/7
	Helpfulness	0.945	[0.694, 0.986]	0.934	0.291	Excellent	1/7
	Understanding	0.960	[0.882, 0.987]	0.946	0.106	Excellent	1/7
	Average	**0.949**	–	**0.933**	**0.254**	**7/7 Excellent**	**1st**
Qwen-2.5-7B-SFT	Guidance	0.929	[0.852, 0.966]	0.911	0.114	Excellent	3/7
	Informativeness	0.895	[0.805, 0.983]	0.886	0.178	Excellent	4/7
	Relevance	0.760	[0.636, 0.811]	0.695	0.175	Excellent	2/7
	Safety	0.605	[0.429, 0.672]	0.451	0.242	Excellent	3/7
	Empathy	0.874	[0.459, 0.978]	0.866	0.518	Moderate	4/7
	Helpfulness	0.921	[0.820, 0.980]	0.926	0.160	Excellent	2/7
	Understanding	0.803	[0.646, 0.862]	0.743	0.216	Excellent	5/7
	Average	**0.827**	–	**0.783**	**0.229**	**6/7 Excellent**	**2nd**
Claude-3.7-Sonnet	Guidance	0.881	[0.764, 0.980]	0.837	0.216	Excellent	4/7
	Informativeness	0.915	[0.830, 0.972]	0.915	0.142	Excellent	3/7
	Relevance	0.730	[0.394, 0.987]	0.743	0.594	Poor	3/7
	Safety	0.685	[0.333, 0.961]	0.597	0.628	Poor	2/7
	Empathy	0.906	[0.429, 0.958]	0.474	0.528	Moderate	2/7
	Helpfulness	0.900	[0.734, 0.992]	0.742	0.258	Excellent	3/7
	Understanding	0.791	[0.563, 0.956]	0.806	0.394	Moderate	6/7
	Average	**0.830**	–	**0.731**	**0.394**	**5/7 Excellent**	**3rd**
o4-mini	Guidance	0.948	[0.744, 0.976]	0.786	0.233	Excellent	2/7
	Informativeness	0.918	[0.638, 0.978]	0.908	0.340	Excellent	2/7
	Relevance	0.342	[0.069, 0.673]	0.140	0.605	Poor	6/7
	Safety	0.259	[0.081, 0.703]	0.117	0.621	Poor	6/7
	Empathy	0.883	[0.476, 0.945]	0.499	0.469	Moderate	3/7
	Helpfulness	0.871	[0.578, 0.934]	0.660	0.356	Moderate	4/7
	Understanding	0.871	[0.636, 0.938]	0.592	0.302	Excellent	2/7
	Average	**0.727**	–	**0.529**	**0.418**	**5/7 Excellent**	**4th**
GPT-4o	Guidance	0.849	[0.650, 0.975]	0.475	0.324	Excellent	6/7
	Informativeness	0.856	[0.655, 0.964]	0.681	0.310	Excellent	6/7
	Relevance	0.532	[0.267, 0.826]	0.243	0.559	Moderate	4/7
	Safety	0.480	[0.116, 0.858]	0.279	0.741	Poor	4/7
	Empathy	0.835	[0.331, 0.891]	0.288	0.560	Moderate	6/7
	Helpfulness	0.800	[0.407, 0.924]	0.457	0.517	Moderate	5/7
	Understanding	0.823	[0.549, 0.884]	0.485	0.334	Excellent	4/7
	Average	**0.739**	–	**0.415**	**0.478**	**5/7 Excellent**	**5th**
Qwen-2.5-7B-ZS	Guidance	0.650	[0.450, 0.824]	0.499	0.374	Moderate	7/7
	Informativeness	0.802	[0.573, 0.949]	0.669	0.376	Moderate	7/7
	Relevance	0.519	[0.310, 0.602]	0.263	0.292	Excellent	5/7
	Safety	0.145	[0.155, 0.441]	0.079	0.427	Moderate	7/7
	Empathy	0.616	[0.164, 0.708]	0.602	0.544	Moderate	7/7
	Helpfulness	0.761	[0.394, 0.887]	0.740	0.493	Moderate	6/7
	Understanding	0.870	[0.471, 0.924]	0.712	0.453	Moderate	3/7
	Average	**0.623**	–	**0.509**	**0.423**	**3/7 Excellent**	**6th**
Gemini-2.5-Flash	Guidance	0.855	[0.557, 0.956]	0.682	0.398	Moderate	5/7
	Informativeness	0.878	[0.522, 0.962]	0.877	0.439	Moderate	5/7
	Relevance	0.306	[0.011, 0.767]	0.137	0.755	Poor	7/7
	Safety	0.377	[0.047, 0.868]	0.222	0.790	Poor	5/7
	Empathy	0.838	[0.401, 0.918]	0.380	0.517	Moderate	5/7
	Helpfulness	0.734	[0.271, 0.832]	0.385	0.561	Poor	7/7
	Understanding	0.362	[0.137, 0.781]	0.180	0.644	Poor	7/7
	Average	**0.621**	–	**0.409**	**0.586**	**3/7 Excellent**	**7th**

Table 10: Comprehensive judge comparison with complete rankings. All ranks computed by comparing ICC(C,1) values across all 7 judges for each dimension. *Poor status due to wide CI despite high ICC. TheraJudge (gen 16) achieves best overall performance with universal excellent reliability and wins 5/7 dimension rankings.