Multimodal Reward Modeling for Alliance-Aware Mental Health LLMs

Ryan Zhao
Indiana University
Bloomington, IN, USA
rz25@iu.edu

Aijia Yuan
Indiana University
Bloomington, IN, USA
yuana@iu.edu

Yang Gao
Indiana University
Bloomington, IN, USA
gaoyang@iu.edu

Jianna Hur
Indiana University
Bloomington, IN, USA
jhur06@iu.edu

Edlin Garcia Colato
Indiana University
Bloomington, IN, USA
eggarcia@iu.edu

Sagar Samtani
Indiana University
Bloomington, IN, USA
ssamtani@iu.edu

Abstract

Therapeutic alliance, defined as the collaborative working relationship between therapist and patient, is among the most consistent predictors of psychotherapy outcome, yet existing mental health large language models (LLMs) are optimized primarily for surface helpfulness rather than the relational dynamics that shape clinical effectiveness. We propose **TARM** (Therapeutic Alliance Reward Model), a multimodal multi-task reward model that provides a turn-level alliance signal for downstream alliance-guided reinforcement learning (RL) of mental health LLMs. Grounded in the validated Rupture Resolution Rating System (3RS), TARM jointly predicts an overall Working Together alliance score from textual and vocal cues, with rupture and repair as auxiliary supervision tasks, integrating text and audio through gated residual cross-modal fusion with an auxiliary audio supervision head. On a curated 7,447-turn corpus of public educational and role-play psychotherapy videos, controlled ablations show that cross-modal fusion with a residual gated combination of fused and text representations gives the largest within-session alliance gain over text-only baselines, while auxiliary audio supervision additionally lifts pooled correlation. To our knowledge, TARM is the first turn-level alliance reward model grounded in established observer coding, providing the per-turn signal required for alliance-guided RL of mental health LLMs.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; *Speech recognition*; *Neural networks*; *Reinforcement learning*; • **Applied computing** → *Health informatics*.

Keywords

Therapeutic alliance, Multimodal reward modeling, Mental health LLMs, Reinforcement learning, Psychotherapy

1 Introduction

Mental health conditions impose substantial individual and societal costs, and psychotherapy remains a widely used treatment [33]. Decades of psychotherapy research identify the therapist–patient relationship as a central determinant of engagement and outcome [22, 30]. Within this broader relational construct, the *therapeutic alliance*, defined as agreement on goals, assignment of tasks, and development of bonds [2], has emerged as one of the most consistent predictors of clinical outcome and premature treatment dropout [8, 28]. Therapeutic alliance is inherently dynamic. During a session, the alliance may be disrupted by *ruptures*, defined as alliance deteriorations involving reduced task/goal collaboration or strains in the emotional bond, and expressed as patient withdrawal (moving away) or confrontation (moving against). Therapists address ruptures through *repair* strategies that restore collaboration and bond, such as exploring the rupture or refocusing on tasks and goals, or acknowledging one's contribution to it [6]. The *Rupture Resolution Rating System (3RS)* [7], one of the most widely used observer-based coding systems for alliance rupture and repair, operationalizes these dynamics by producing a Working Together rating for overall alliance and event-level codes for rupture and repair. However, 3RS coding is typically performed retrospectively by trained observers and therefore provides limited support for in-session alliance monitoring. This limitation is especially salient in online psychotherapy, where a reduced physical presence makes the assessment of real-time alliances more difficult.

Mental health large language models (LLMs) offer a promising basis for in-session decision support, but their utility for alliance support remains limited for two reasons. First, existing mental health LLMs primarily focus on mental-state assessment, interpretable diagnostic analysis, or response-level qualities such as empathy and fidelity [4, 13, 34, 36], but do not model therapeutic alliance or its rupture–repair dynamics. Second, current mental health LLMs are predominantly text-based and therefore miss alliance signals conveyed through vocal behavior. This departs from established alliance-coding practice, which incorporates paraverbal cues such as tone of voice, voice quality, and warmth [5–7]. By relying only on text, existing LLM-based systems may miss clinically meaningful alliance signals conveyed through speech.

We propose **TARM** (Therapeutic Alliance Reward Model), a multimodal multi-task reward model for turn-level alliance assessment from textual and vocal cues. Grounded in the 3RS framework, TARM jointly predicts overall alliance through the Working Together score, with alliance rupture and repair as auxiliary binary tasks. Building on the alignment of LLM based on the reward-model [23], we further propose an alliance-guided reinforcement learning framework in which TARM provides the reward signal to fine-tune a domain-adapted mental health LLM to alliance-supportive responses. The framework is intended to support real-time, turn-level clinical decision making by helping therapists identify emerging disruptions and repair opportunities during psychotherapy sessions. This paper

reports on ongoing work on framework design and a preliminary evaluation of TARM; full RL training and policy evaluation will be reported in an extended version.

This paper makes three primary contributions. First, we formulate therapeutic alliance assessment as a theory-grounded multi-task reward modeling problem based on the 3RS framework, with overall alliance as the primary target and rupture and repair as auxiliary turn-level supervision. Second, to our knowledge, we provide the first controlled architectural ablation of multimodal reward models for the turn-level therapeutic alliance, showing that cross-modal text–audio fusion combined with residual gating gives the largest within-session alliance gain over text-only baselines, while auxiliary audio supervision further improves the pooled correlation. Third, we position TARM as a turn-level reward signal for alliance-guided reinforcement learning of mental health LLMs: because TARM is discriminative within sessions rather than only at the aggregate level, it can support future policy optimization toward alliance-supportive responses in context. As preliminary evidence, we train and evaluate TARM on a curated 7.4K-turn multimodal psychotherapy corpus and show that vocal cues improve within-session alliance prediction over text-only baselines.

2 Related Work

We situate TARM in three lines of recent work: the training objectives and approaches of mental health LLMs (§2.1), the utilization of vocal cues in mental health LLMs (§2.2), and reinforcement-learning fine-tuning of LLMs for conversational quality (§2.3).

2.1 Existing Mental Health LLMs

Mental health LLMs have been developed for mental-state classification, such as Mental-LLM [34], MentaLLaMA [36], and MACR [14]; counseling response quality and empathy, such as ChatCounselor [16] and SoulChat [4]; and fidelity to therapeutic frameworks such as cognitive behavioral therapy, as in CBT-LLM [20]. Most systems rely on prompt engineering or supervised fine-tuning on clinical or counseling data [12], with reinforcement-learning-based alignment explored only in limited settings [1, 29]. Despite the importance of therapeutic alliance for treatment outcomes, existing mental health LLMs do not explicitly optimize for therapist-patient relational dynamics. TARM addresses this limitation by providing a multimodal, turn-level reward model of therapeutic alliance grounded in an established observer rating system, enabling alliance-guided reinforcement learning for mental health LLMs.

2.2 Vocal Cues in Mental Health LLMs

Vocal cues provide clinically meaningful information: prosodic, energetic, and timing patterns vary systematically in depression and other affective disorders [17, 19], and alliance-coding systems such as the 3RS consider patient affect and paralinguistic expression when identifying alliance rupture [7]. However, mental health LLMs remain largely text-based. Existing audio-aware mental health systems are typically standalone classifiers for distress or depression detection [17] rather than conversational models, while recent voice-enabled mental health chatbots often rely on speech-to-text and text-to-speech modules that process transcripts but discard acoustic signals [21]. Speech-LLMs such as BLSP-Emo [31] show that paralinguistic emotion cues can improve empathetic response generation, but they target general emotional conversation rather than therapeutic dialog. TARM addresses the underutilization of vocal cues in mental health LLMs by integrating patient speech features with text for turn-level therapeutic alliance prediction.

2.3 Reinforcement Learning (RL) in Mental Health LLMs

RL is a standard approach for aligning LLMs with holistic conversational qualities that are difficult to optimize through token-level supervision alone. Following the RLHF paradigm of InstructGPT [23], alignment commonly uses policy-gradient methods such as PPO [26] and GRPO [27], or preference optimization methods such as DPO [24]. In emotionally supportive dialogue, this paradigm has been applied through PPO with verifiable emotion rewards [32], future-oriented reward modeling [38], and GRPO with rubric-derived empathy rewards [39]. Therapeutic alliance is a natural target for reward-based alignment because alliance-supportive responses require optimizing relational outcomes, such as sustaining collaboration and repairing ruptures, rather than matching token-level labels. We therefore adopt GRPO [27], with TARM providing the learned turn-level reward signal.

3 Problem Definition

We cast the development of alliance-aware mental health LLM as a two-stage learning problem: *multimodal alliance reward modeling* followed by *alliance-guided policy optimization*. The optimization target is the turn of the therapist. Specifically, the policy generates a text response for the therapist given the context of the multimodal patient, and the reward model evaluates the state of the alliance given the context of the response.

For a target therapist turn at index t, we define a turn-level segment as

$$s_t = (\mathbf{x}_t^{\text{text}}, \mathbf{x}_t^{\text{audio}}),$$

where $\mathbf{x}_t^{\text{text}}$ denotes a context-windowed textual representation of the dialogue up to the target therapist turn, including speaker tags, and $\mathbf{x}_t^{\text{audio}}$ denotes an acoustic representation of the immediately preceding patient turn. The model uses audio only from the patient side, while the policy output remains text.

Each segment is annotated with three labels derived from the 3RS framework: (1) a continuous *Working Together* score $y_t^{\text{a}} \in [1, 5]$ (computed as the mean of five 3RS WT item ratings), which serves as the primary alliance target; (2) a binary *Alliance Rupture* indicator $y_t^{\text{r}} \in \{0, 1\}$ capturing patient withdrawal or confrontation behaviors that strain the alliance; and (3) a binary *Alliance Repair* indicator $y_t^{\text{rep}} \in \{0, 1\}$ capturing attempts to restore the alliance following rupture. The reward model learns

$$f_\theta : (\mathbf{x}_t^{\text{text}}, \mathbf{x}_t^{\text{audio}}) \rightarrow (\hat{y}_t^{\text{a}}, \hat{y}_t^{\text{r}}, \hat{y}_t^{\text{rep}}),$$

where rupture and repair are treated as auxiliary supervisory tasks alongside the primary alliance prediction. Given the trained reward model, the second stage optimizes a domain-adapted mental health LLM π_ϕ using the primary alliance head as the reward:

$$R(c, y) = R_{\text{alliance}}(c, y), \tag{1}$$

where $R_{\text{alliance}}(c, y)$ is obtained from the alliance prediction of f_θ for a patient context c and candidate therapist response y. The policy is initialized from a reference policy π_{ref} and fine-tuned to maximize

$$\mathbb{E}_{c\sim\mathcal{D},\, y\sim\pi_\phi(\cdot|c)}\,[R(c, y)],$$

thereby encouraging therapist responses that strengthen therapeutic alliance.

4 Proposed Method

Our method follows the reward-model-based alignment paradigm [23]: we first train **TARM**, a multimodal multi-task reward model that predicts turn-level therapeutic alliance from dialogue text and patient vocal cues (§4.1), and then use its per-turn alliance score as the reward signal for GRPO-based fine-tuning of a mental health LLM policy (§4.2). As ongoing research, this paper focuses on TARM's design, training, and empirical evaluation.

4.1 Multi-Modal Therapeutic Alliance Reward Modeling

We instantiate the reward model f_θ from Section 3 as **TARM**, a multimodal multi-task network for alliance-guided RL (Figure 1). TARM follows two design constraints. First, reward inference is causal: the model conditions only on prior dialogue and the target therapist turn, while the following patient response is used only for supervision. Second, because the downstream policy generates text-only therapist responses, acoustic input is restricted to the preceding patient turn, preserving train–inference symmetry.

Frozen encoders with lightweight adaptation. We encode text with DeBERTa-v3-large [10] and audio with WavLM-large [3], freezing both encoders and adapting each modality with a Houlsby-style bottleneck adapter [11] with width 128. Text input contains a six-turn dialogue window, comprising five prior turns and the target therapist turn, encoded into a vector $\mathbf{h}^{\text{text}} \in \mathbb{R}^{1024}$. Audio input is the frame-level WavLM representation of the immediately preceding patient turn. We retain all 24 hidden layers and collapse them through a learnable softmax-weighted sum [37], then pass the resulting frame sequence through the audio adapter to obtain $\mathbf{H}^{\text{audio}} \in \mathbb{R}^{T\times 1024}$ as keys and values for cross-modal fusion.

Cross-modal fusion. To inject vocal information into alliance prediction, we apply a single pre-LN cross-modal attention block with $\mathbf{h}^{\text{text}}$ as the query and $\mathbf{H}^{\text{audio}}$ as keys and values, producing a fused turn representation $\mathbf{h}^{\text{fused}}$. The primary alliance head reads from a gated residual combination of fused and text representations:

$$\mathbf{h}^{\text{a}} = \alpha\,\text{LN}(\mathbf{h}^{\text{fused}}) + (1-\alpha)\,\text{LN}(\mathbf{h}^{\text{text}}), \qquad \alpha = \sigma(g), \tag{2}$$

where g is a learnable scalar initialized at zero. This design preserves the text pathway as a strong baseline while allowing audio to contribute additively rather than replace it.

Auxiliary audio supervision. Because fusion can be dominated by text and underutilize audio, we add an auxiliary alliance regression head over an attention-pooled audio summary [25], predicting the same alliance target as the primary head. This auxiliary objective injects direct gradient signal into the audio pathway during training, while remaining unused at inference.

Task heads. TARM predicts three outputs. The primary alliance head regresses the Working Together score from $\mathbf{h}^{\text{a}}$. The rupture and repair heads are binary classifiers that read only from $\mathbf{h}^{\text{text}}$. We restrict these auxiliary heads to the text pathway because their 3RS markers are primarily linguistic, and our audio diagnostic showed that audio improves alliance prediction but does not provide additional predictive value for rupture or repair beyond text. This routing focuses audio capacity on the primary alliance task.

Training objective. TARM is trained end-to-end via a multi-task supervised objective combining the primary alliance regression with the rupture, repair, and auxiliary alliance heads:

$$\mathcal{L}_{\text{TARM}} = w_{\text{a}}\,\mathcal{L}_{\text{a}}^{\text{MSE}} + w_{\text{r}}\,\mathcal{L}_{\text{r}}^{\text{focal}} + w_{\text{rep}}\,\mathcal{L}_{\text{rep}}^{\text{focal}} + w_{\text{aux}}\,\mathcal{L}_{\text{aux}}^{\text{MSE}}, \tag{3}$$

where $\mathcal{L}_{\text{a}}^{\text{MSE}}$ and $\mathcal{L}_{\text{aux}}^{\text{MSE}}$ are mean-squared errors on the primary and auxiliary alliance heads, and $\mathcal{L}_{\text{r}}^{\text{focal}}$, $\mathcal{L}_{\text{rep}}^{\text{focal}}$ are focal cross-entropy losses [15] on the rupture and repair classification heads, addressing their class imbalance in the training split (6.8% rupture, 3.6% repair). Loss weights $(w_{\text{a}}, w_{\text{r}}, w_{\text{rep}}, w_{\text{aux}}) = (1.0, 0.5, 0.1, 0.5)$ balance the alliance regression objective with auxiliary supervision.

4.2 Alliance-Guided Reinforcement Learning

We propose to fine-tune a mental health LLM policy π_ϕ with TARM's per-turn alliance prediction as the reward signal, using Group Relative Policy Optimization (GRPO) [27]. The policy is initialized from a reference policy π_{ref} obtained by supervised fine-tuning a general-purpose LLM on therapy session transcripts; π_{ref} both seeds clinically appropriate dialogue style and serves as the KL anchor that constrains policy drift during RL.

For each patient context c drawn from our therapy corpus, the policy samples a group of K candidate therapist responses; TARM scores each under identical patient-side context, group-relative advantages are computed from the resulting rewards, and π_ϕ is updated via a clipped surrogate objective with a KL penalty to π_{ref}.

5 Preliminary Results

We evaluate TARM and its ablation variants as turn-level alliance predictors on held-out test sessions; the downstream RL stage is planned as follow-up work.

5.1 Experimental Setup

Dataset and Label Generation. We curate a multimodal psychotherapy corpus of 274 publicly available therapy demonstration and role-play sessions from YouTube, yielding 7,447 turn-level segments. We use a session-level 80/10/10 train/dev/test split (219/28/27 sessions; 5,989/730/728 turns), stratified by session length and rupture/repair prevalence to preserve coverage of these rare events. In general, 6.8% of turns are rupture-positive and 3.6% are repair-positive. Turn-level 3RS labels are generated by Qwen3-30B-A3B-Thinking-2507 [35] as an LLM judge through a four-pass pipeline (withdrawal markers, confrontation markers, repair strategies, and Working Together items). Each turn receives three labels: (i) a continuous *alliance* regression target in $[1, 5]$, taken as the mean of five 1-5 Working Together item ratings; (ii) a binary *rupture* indicator, set to 1 if any withdrawal or confrontation marker is present; and (iii) a binary *repair* indicator, set to 1 if any repair strategy is present. The

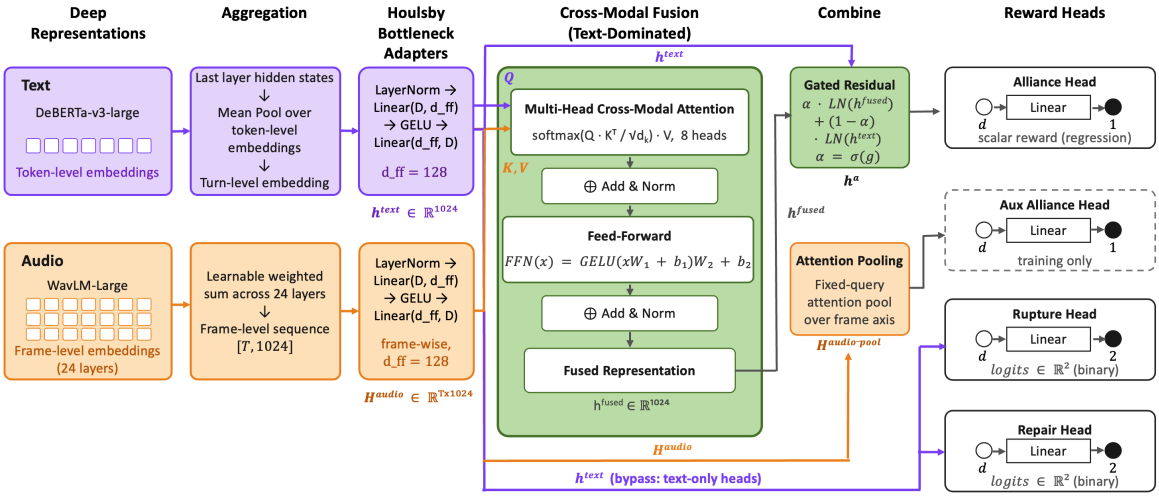

Figure 1: Full TARM architecture: frozen DeBERTa-v3-large (text) and WavLM-large (audio) encoders with Houlsby bottleneck adapters, cross-modal attention fusion, gated residual combination, auxiliary audio supervision head, and three task heads (alliance, rupture, repair).

full 3RS coding scheme used as the judge rubric is summarized in Appendix A.

Model configurations and Implementation Details. We ablate five variants against our full TARM model, RM-A3-RES, to isolate the contribution of each design choice. Two text-only baselines remove the audio modality entirely: RM-T1 predicts only the primary alliance score, whereas RM-T3 predicts alliance, rupture, and repair jointly. The remaining four multimodal models share the three-task setting and form a 2×2 design over fusion type and auxiliary audio supervision: RM-M3-TRI, RM-M3-RES, RM-A3-TRI, and RM-A3-RES. Here, RES denotes the residual fusion in Equation 2, TRI denotes a 3-way softmax fusion over fused, text, and pooled-audio branches, and A3 indicates that the auxiliary audio supervision head is enabled.

Text inputs are capped at 512 tokens (left-truncated); audio is 16 kHz mono. The cross-modal fusion block uses 8 attention heads, FFN dimension 1024, and dropout 0.3; focal-loss $\gamma = 2.0$ for the rupture and repair heads. We train with AdamW [18], 500-step linear warmup, cosine decay, weight decay 0.01, batch size 32, and up to 30 epochs with early stopping (patience 5) and gradient clipping at 1.0. Learning rates are selected from {1e−4, 3e−4, 1e−3, 3e−3} based on pooled alliance Pearson. We use a single NVIDIA A100 40 GB GPU per run.

Evaluation metrics. We report Pearson correlation between predicted and gold Working Together scores (i.e., turn-level alliance ratings) under two aggregation schemes. *Pooled Pearson* is computed over all test turns and measures the aggregate prediction quality across the test set. *Session-mean Pearson* averages per-session Pearson correlations over the test sessions and measures within-session turn-level discrimination, the property most relevant for using TARM as a per-turn reward in downstream RL.

Table 1: Reward model results: alliance Pearson under pooled and session-mean aggregation, mean±std over 3 seeds. All multimodal models and RM-T3 use three tasks (alliance, rupture, repair); RM-T1 uses alliance only. RM-A3-RES is the full TARM. Bold = best per column.

Model	Audio	Aux	Fusion	Pooled r	Session r
RM-T1	–	–	–	$0.557_{\pm.002}$	$0.368_{\pm.006}$
RM-T3	–	–	–	$0.545_{\pm.011}$	$0.349_{\pm.009}$
RM-M3-TRI	✓	–	tri	$0.553_{\pm.006}$	$0.399_{\pm.015}$
RM-M3-RES	✓	–	res	$0.540_{\pm.016}$	$\mathbf{0.413}_{\pm.025}$
RM-A3-TRI	✓	✓	tri	$0.555_{\pm.009}$	$0.371_{\pm.018}$
RM-A3-RES	✓	✓	res	$\mathbf{0.562}_{\pm.008}$	$0.409_{\pm.025}$

5.2 Reward Model Evaluation

Table 1 reports alliance Pearson under both aggregation schemes across six configurations, with best-performing methods in bold.

The full TARM (RM-A3-RES) matches the strongest text-only baseline on pooled alliance Pearson (0.562 vs. 0.557 for RM-T1; $\Delta < 1\sigma$) and matches the best session-mean alliance Pearson within noise (0.409 vs. 0.413 for RM-M3-RES; $\Delta < 1\sigma$), lifting session-mean by +0.041 over the strongest text-only baseline (0.368, RM-T1) while preserving pooled performance.

The pooled result confirms TARM's predicted alliance closely tracks gold ratings at the aggregate level, supporting its use as an absolute alliance estimator. The session-mean result is the more consequential for downstream RL: TARM's elevated session-mean exposes within-session variation that GRPO can exploit to train a policy toward turn-level alliance-supportive behavior, whereas a low session-mean would collapse policy optimization into a global style mimic that issues near-constant rewards within a session.

The ablation suggests the two metrics are driven by different components. *Within-session gain* is driven by audio modality and residual gated fusion: removing audio at matched 3-task supervision

(RM-T3: 0.349) drops session-mean by 0.060 relative to RM-A3-RES, and the 1-task text-only baseline RM-T1 (0.368) is similarly outperformed, confirming text alone cannot capture within-session variation; switching residual fusion to tri-branch (RM-A3-TRI) drops session-mean to 0.371, indicating the gated residual design is what successfully integrates vocal information. *Pooled performance* is driven by auxiliary audio supervision: removing it (RM-M3-RES) drops pooled to 0.540, showing the audio pathway needs direct gradient signal to align with the primary head.

6 Discussion and Future Work

Our results show that vocal cues improve within-session alliance prediction through residual gated cross-modal fusion, while auxiliary audio supervision further improves pooled correlation. These findings support TARM as a turn-level alliance reward for downstream RL. We are fine-tuning a domain-adapted Llama-3.1-8B-Instruct [9] with GRPO using TARM as the reward signal. The resulting policy will be evaluated against general-purpose and mental health LLM baselines using NLG and 3RS-based alliance metrics, modality ablations, and a scenario-based therapist study measuring perceived alliance, usefulness, and trust.

References

[1] Mehrab Beikzadeh, Yasaman Asadollah Salmanpour, Ashima Suvarna, Sriram Sankararaman, Matteo Malgaroli, Majid Sarrafzadeh, and Saadia Gabriel. 2026. Multi-Objective Alignment of Language Models for Personalized Psychotherapy. *arXiv preprint arXiv:2602.16053* (2026).

[2] Edward S. Bordin. 1979. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice* 16, 3 (1979), 252–260. doi:10.1037/h0085885

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. doi:10.1109/JSTSP.2022.3188113

[4] Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1170–1183. doi:10.18653/v1/2023.findings-emnlp.83

[5] Andrew Darchuk, Victor Wang, David Weibel, Jennifer Fende, Timothy Anderson, and Adam Horvath. 2000. Manual for the Working Alliance Inventory – Observer Form (WAI-O): Revision IV. (2000). Unpublished manuscript, Department of Psychology, Ohio University.

[6] Catherine F. Eubanks and J. Christopher Muran. 2022. Rupture Resolution Rating System (3RS): Manual Version 2022. (2022). Unpublished manuscript.

[7] Catherine F. Eubanks, J. Christopher Muran, and Jeremy D. Safran. 2019. Rupture Resolution Rating System (3RS): Reliability and Validity. *Psychotherapy Research* 29, 3 (2019), 306–319. doi:10.1080/10503307.2018.1552034

[8] Christoph Flückiger, A. C. Del Re, Bruce E. Wampold, and Adam O. Horvath. 2018. The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy* 55, 4 (2018), 316–340. doi:10.1037/pst0000172

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2111.09543

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 2790–2799. https://proceedings.mlr.press/v97/houlsby19a.html

[12] Yining Hua, Hongbin Na, Zehan Li, Fenglin Liu, Xiao Fang, David Clifton, and John Torous. 2025. A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine* 8, 1 (2025), 230. doi:10.1038/s41746-025-01611-4

[13] Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 14245–14274. doi:10.18653/v1/2024.findings-emnlp.832

[14] Jun Li, Xiangmeng Wang, Haoyang Li, Yifei Yan, Shijie Zhang, Hong Va Leong, Ling Feng, Nancy Xiaonan Yu, and Qing Li. 2026. Multi-Agent Causal Reasoning for Suicide Ideation Detection Through Online Conversations. In *Database Systems for Advanced Applications*, Vol. 16539. Springer Nature Singapore, 53–69. doi:10.1007/978-981-92-0375-8_4

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. doi:10.1109/ICCV.2017.324

[16] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. ChatCounselor: A Large Language Models for Mental Health Support. *arXiv preprint arXiv:2309.15461* (2023).

[17] Clara Lombardo, Giulia Esposito, Silvia Carbone, Salvatore Serrano, and Carmela Mento. 2025. Speech analysis and speech emotion recognition in mental disease: a scoping review. *Frontiers in Psychology* 16 (2025), 1645860. doi:10.3389/fpsyg.2025.1645860

[18] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Bkg6RiCqY7

[19] Felix Menne, Felix Dörr, Julia Schräder, Johannes Tröger, Ute Habel, Alexandra König, and Lisa Wagels. 2024. The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry* 24 (2024), 794. doi:10.1186/s12888-024-06253-6

[20] Hongbin Na. 2024. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. *arXiv preprint arXiv:2403.16008* (2024).

[21] Nhat Ngo and Akane Sano. 2025. Evaluating Voice-Enabled Generative AI for Mental Health: Real-Time Performance and Safety Analyses. *medRxiv preprint* (2025). doi:10.1101/2025.11.14.25340246

[22] John C. Norcross and Michael J. Lambert. 2018. Psychotherapy relationships that work III. *Psychotherapy* 55, 4 (2018), 303–315. doi:10.1037/pst0000193

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. 27730–27744.

[24] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[25] Pooyan Safari, Miquel India, and Javier Hernando. 2020. Self-Attention Encoding and Pooling for Speaker Recognition. In *Proc. Interspeech 2020*. 941–945. doi:10.21437/Interspeech.2020-1446

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[27] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[28] Iris Sijercic, Rachel E. Liebman, Shannon Wiltsey Stirman, and Candice M. Monson. 2021. The effect of therapeutic alliance on dropout in cognitive processing therapy for posttraumatic stress disorder. *Journal of Traumatic Stress* 34, 4 (2021), 819–828. doi:10.1002/jts.22676

[29] 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 (2024), 12. doi:10.1038/s44184-024-00056-z

[30] Bruce E. Wampold and Zac E. Imel. 2015. *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work* (2nd ed.). Routledge.

[31] Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024. BLSP-Emo: Towards Empathetic Large Speech-Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 19186–19199.

[32] Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025. RLVER: Reinforcement Learning with Verifiable Emotion Rewards for Empathetic Agents. *arXiv preprint arXiv:2507.03112* (2025).

[33] World Health Organization. 2022. *World Mental Health Report: Transforming Mental Health for All.* Technical Report. World Health Organization, Geneva.

https://www.who.int/publications/i/item/9789240049338

[34] 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 (2024), 1–32. doi:10.1145/3643540

[35] 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

[36] 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 (WWW '24)*. 4489–4500. doi:10.1145/3589334.3648137

[37] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*. 1194–1198. doi:10.21437/Interspeech.2021-1775

[38] Ting Yang et al. 2025. Towards Open-Ended Emotional Support Conversations in LLMs via Reinforcement Learning with Future-Oriented Rewards. *arXiv preprint arXiv:2508.12935* (2025).

[39] Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. 2026. Kardia-R1: Unleashing LLMs to Reason toward Understanding and Empathy for Emotional Support via Rubric-as-Judge Reinforcement Learning. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*. doi:10.1145/3774904.3793022

A 3RS Coding Scheme

The Rupture Resolution Rating System [6] is an observer-based coding system for therapeutic alliance dynamics. Coders rate each segment on a 5-point salience scale (1 = not salient, 3 = somewhat salient, 5 = very salient). Our LLM judge applies the rubric in Table 2 to produce three turn-level labels used to train TARM.

Table 2: 3RS dimensions used as the LLM-judge rubric, and their mapping to TARM labels. Each item is rated on a 1–5 salience scale. The alliance regression target is the mean of the five Working Together ratings ($[1, 5]$); the binary rupture indicator is set to 1 if any withdrawal or confrontation marker is rated as salient; the binary repair indicator is set to 1 if any repair strategy is rated as salient.

3RS dimension	Items / behaviors
Working Together — 5 items; mean used as the alliance regression target	
WT1	Patient & Therapist are collaborating on the work of therapy.
WT2	Patient & Therapist have a bond of mutual trust.
WT3	Therapist accepts and validates the Patient.
WT4	Therapist is curious and engaged.
WT5	Patient is engaged in the work; authentic and open.
Rupture markers — union used as the binary rupture indicator	
Withdrawal	Patient/Therapist moves *away*: shuts down (avoidant denial, minimal response, giving up), avoids (abstract communication, avoidant storytelling, topic shift), or masks (deferential/appeasing, content/affect split).
Confrontation	Patient/Therapist moves *against*: complains/criticizes (the other, activities, parameters), pushes back (rejecting ideas, defending self, hostile denial), or controls/pressures (overly directive, pushing ideas).
Repair strategies — any salient strategy yields the binary repair indicator	
Focusing on task/goal	Discussing changing the task/goal; explaining or reframing the rationale for a task.
Exploring the rupture	Naming, observing, or inviting exploration of the rupture or related vulnerable feelings.
Acknowledging contribution	Patient or therapist acknowledges their own contribution to the rupture.
Linking to patterns	Connecting the rupture to broader interpersonal patterns or to the patient's life.