

ARIA: Affect-Responsive Interaction Alignment for Emotional Support Policies via Implicit Multimodal Affective Feedback

Deuksin Kwon^{1,2} Kevin Hyekang Joo^{1,2} Elena Hayoung Lee¹

Mohammad Soleymani² Gale M. Lucas²

¹University of Southern California ²USC Institute for Creative Technologies

{deuksink, khjoo}@usc.edu

elena.lee@marshall.usc.edu

{soleymani, lucas}@ict.usc.edu

Abstract

Effective emotional-support dialogue requires more than fluent empathy: responses must adapt to the support seeker's evolving state. Existing emotional-support systems are largely optimized at the response level and may therefore learn the form of supportive counseling without reliably adapting to client-side change. We introduce ARIA, Affect-Responsive Interaction Alignment, a client-centered alignment framework that constructs DPO preference pairs from the support seeker's subsequent multimodal affective trajectories observed in human interactions. ARIA distills audio- and visual-derived valence trajectories into text-based predictors and uses cross-modal consensus and uncertainty-aware filtering to identify high-confidence preferences associated with delayed client-side valence improvement. Experiments show that ARIA improves process-oriented measures of client progress beyond surface compliance and produces support that is more timely and calibrated to the seeker's evolving openness. Behavioral analyses further show a phase-sensitive shift from restrained pre-open support toward more active support after openness emerges. Human and counseling-expert evaluations provide evidence that ARIA's responses are more timely, targeted, and supportive of the client's own process.

1 Introduction

Large language models (LLMs) are increasingly being explored as agents for emotional and mental health support, ranging from supportive chatbots to simulated counselors (Liu et al., 2023; Chen et al., 2023; Stade et al., 2024). Their fluency and apparent empathy make them promising tools for scalable support, yet LLM-generated support still falls short of high-quality human supportive interaction (Chiu et al., 2024). A common failure is the "kitchen-sink" response: long, over-packed replies that combine multiple interventions regardless of

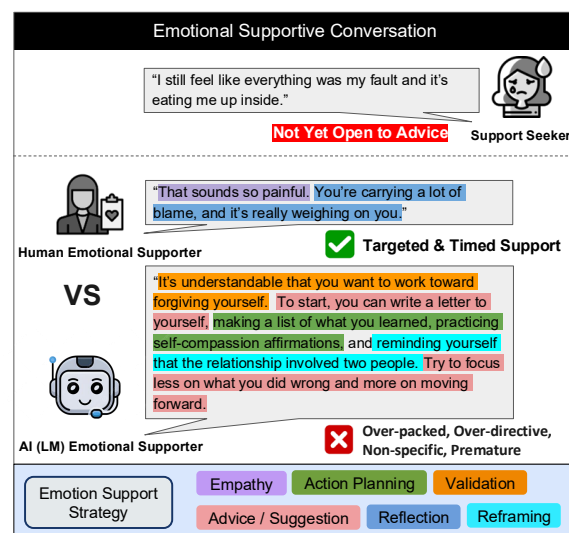


Figure 1: Premature and over-packed AI emotional support. The AI supporter gives concrete advice too early and mixes multiple strategies, unlike the human supporter's targeted reflection.

the support seeker's evolving state (Figure 1). Such responses can introduce premature advice, weaken empathic attunement, and be poorly calibrated to client-side change. Effective emotional support, therefore, poses a counseling-inspired challenge: responses must be not only empathetic, but also timely, calibrated, and respectful of the seeker's agency.

Most existing approaches to emotional support generation, however, still optimize models at the response level, through supervised fine-tuning (SFT) on supportive dialogue data or explicit strategy conditioning (Liu et al., 2021; Zhang et al., 2025). These methods can improve surface-level qualities such as tone, response length, and intervention density, but they do not directly capture whether a response is well-timed for the client. In supportive dialogue, the same strategy can be helpful or harmful depending on timing and fit: reflection may fall short when the client is open to action, while sug-

gestion may be premature before the client feels heard. SFT, in particular, may learn the form of supportive counseling while remaining overly conservative, avoiding timely shifts toward reframing, interpretation, or concrete guidance. As a result, models may imitate the form of supportive counseling while failing to produce the client-side changes that effective support is meant to facilitate.

This limitation raises a broader question for training supportive dialogue models: *what kind of feedback should guide supportive behaviors?* Existing feedback-driven post-training methods, including RLHF/PPO, DPO, and GRPO, typically rely on rewards or preferences derived from judgments of response quality (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024). Emotional support dialogue largely follows this response-centered pattern through human-rated empathy/helpfulness or response-level preferences. Emerging work has begun to use simulated user emotion scores as RL rewards, moving beyond response-centered feedback (Wang et al., 2025b; Yang et al., 2025); however, these signals remain simulation-based and are optimized as scalar rewards. We argue instead for aligning supportive dialogue models with observed affective trajectories of human support seekers. In emotional support, the quality of a response is not fully determined by how empathetic or helpful it appears in isolation; the support seeker’s subsequent affective trajectory can provide an additional signal of how the interaction is unfolding. Valence, a core affective dimension capturing positive versus negative emotional orientation, provides a natural signal for tracking the direction and degree of affective change over time (Russell, 1980; Calvo and D’Mello, 2010). We treat such short-horizon changes in expressed valence as an observational signal of affective recovery, rather than as evidence of therapeutic improvement. By using changes in interlocutor valence as implicit feedback, models can be optimized toward responses associated with subsequent affective improvement rather than merely supportive-sounding responses.

We therefore introduce *ARIA* (Affect-Responsive Interaction Alignment), which constructs DPO preference pairs from the support seeker’s affective (i.e., valence) trajectory rather than direct judgments of response quality. Instead of using noisy affect estimates as direct scalar rewards for RL, *ARIA* converts affective changes into pairwise preferences, providing a client-centered preference signal for alignment. *ARIA* uses

multimodal valence predictors over client facial and vocal cues to ground preference optimization in delayed client-side affective change rather than surface-level agreement or compliance.

The contributions of our work are as follows:

- We introduce *ARIA*, a client-centered alignment framework that uses delayed multimodal affective trajectories of the support seeker as implicit feedback for emotional-support dialogue.
- We construct high-confidence DPO pairs from noisy multimodal affect estimates using delayed-valence prediction, cross-modal consensus, and uncertainty-aware filtering.
- We show that *ARIA* improves process-oriented measures of client progress beyond surface compliance and better calibrates supportive behavior to the support seeker’s openness, with active support emerging after openness.

2 Method

ARIA aligns supportive dialogue models using observed client-side affective trajectories (Figure 2). We extract per-utterance client valence from audio and video, derive delayed-valence reward targets from subsequent affective change, distill them into text-based reward predictors, construct preference pairs via cross-modal consensus and uncertainty filtering, and optimize an SFT-initialized policy with DPO. At inference, the resulting *ARIA* policy operates on text alone.

2.1 Dataset

We use the MESC (Multimodal Emotional Support Conversation) dataset (Chu et al., 2025), which consists of multimodal therapist–client conversations extracted from *In Treatment* (Seasons 1–3). Its text, audio, and video modalities allow us to model the interlocutor’s affective trajectory in emotional support interactions. Detailed statistics of our final preprocessed split are provided in Appendix A.

2.2 Multimodal Client Valence Extraction

Valence captures the positive or negative quality of an emotional state. We use valence as a continuous measure of the interlocutor’s affective trajectory in emotional support dialogue, where post-response shifts in valence provide cues about downstream

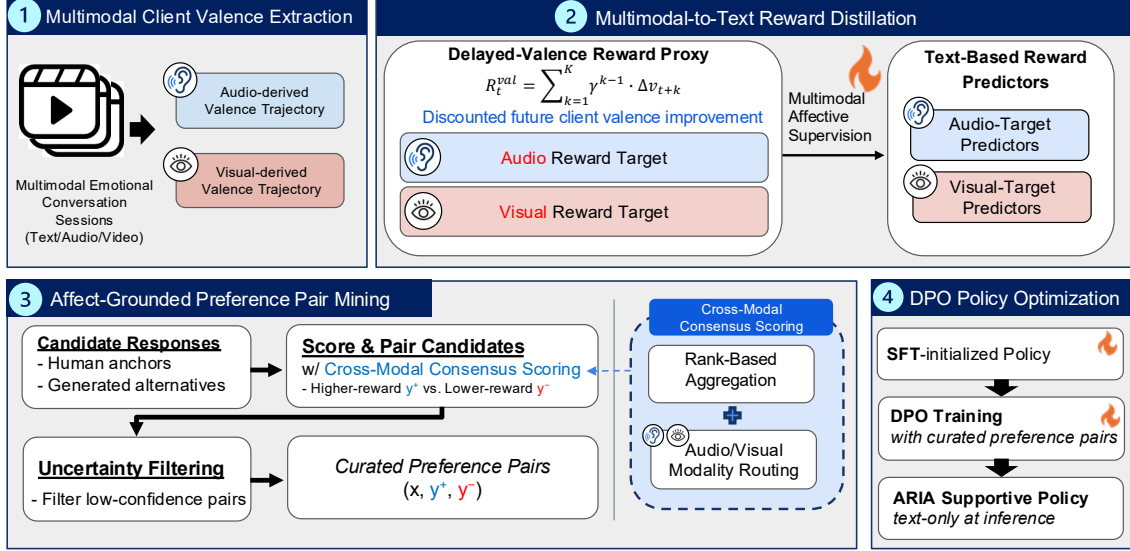


Figure 2: Overview of ARIA. (1) Client valence trajectories are extracted from audio and video; (2) delayed-valence reward targets, derived from discounted future client valence improvement, are distilled into text-based predictors; (3) candidate responses are scored via cross-modal consensus and uncertainty filtering to construct preference pairs (x, y^+, y^-) ; and (4) the SFT-initialized policy is optimized with DPO, yielding the text-only ARIA supportive policy. Delayed valence is used as a reward proxy rather than a causal or clinical measure of recovery.

affective change beyond what is observable from text alone.

We estimate valence from vocal and facial behaviors to capture richer cues of client-side affective change beyond text. For the audio modality, we use a Wav2Vec 2.0-based model (Wagner et al., 2023) to estimate valence from prosodic cues. For the visual modality, we apply EmoNet (Toisoul et al., 2021) to face regions extracted from video frames and aggregate frame-level valence scores to the utterance level. These vocal and visual valence estimates provide complementary signals of the interlocutor’s affective trajectory, which we use to derive delayed-valence supervision for constructing affect-grounded preference pairs in ARIA.

2.3 Multimodal-to-Text Reward Distillation

We train delayed valence predictors by distilling the multimodal valence trajectories from Section 2.2 into text-based predictors that estimate, from dialogue context alone, whether a supporter response is followed by downstream improvement in the interlocutor’s affective trajectory. To reduce sensitivity to noisy model-estimated affect labels, we predict a binary delayed-improvement target rather than fine-grained valence changes. For each supporter response at turn t , we compute:

$$R_t^{\text{val}} = \sum_{k=1}^K \gamma^{k-1} (v_{t+k} - v_{t+k-1}) \quad (1)$$

where v_t is the estimated interlocutor valence, K denotes the future horizon, and γ discounts later affective changes. We binarize R_t^{val} to indicate whether the response is followed by a positive delayed valence shift. We use delayed valence improvement as an observational proxy for short-horizon affective recovery in expressed affect, rather than as a causal measure of therapeutic or clinical recovery.

We implement each predictor as a RoBERTa-based binary classifier (Liu et al., 2019) over the local dialogue context. To improve robustness, we train multiple predictors across affect modalities, future windows, and smoothing settings, and use cross-predictor agreement as a confidence signal. The calibrated probabilities are then used for cross-modal candidate scoring and DPO pair mining; training settings and validation results are in Appendix B.

2.4 Affect-Grounded Preference Pair Mining

Rather than relying solely on subjective judgments of response quality, ARIA derives preference supervision from predictors grounded in observed client-side affective trajectories. At each support-provider

turn, candidate responses are compared by their predicted likelihood of being followed by delayed valence improvement, yielding local preference signals grounded in client-side affective change rather than surface-level response quality.

Candidate scoring. For each support-provider turn, we score the observed human reference and policy-generated alternatives with the same calibrated delayed-valence predictors. Because predictors may use different probability scales, we aggregate within-context rankings rather than raw probabilities. Each ranking is converted into a normalized Borda score (Lin, 2010), and the scores are combined into a scale-invariant preference consensus score $s(y | x)$, favoring responses consistently ranked higher under predicted delayed-valence reward.

Cross-modal preference consensus and routing. We use conditional modality routing rather than uniform audio-visual averaging. Audio-based predictors serve as the primary ranking signal, while visual predictors provide auxiliary evidence when audio scores are ambiguous and cross-modal signals agree. When modalities conflict, we apply an uncertainty penalty, lowering confidence in candidates without consistent cross-modal support.

Positive anchor selection. We use the observed human reference response as the default positive anchor, while still requiring evidence of delayed valence improvement. A human reference is retained as y^+ only if it receives a sufficiently positive preference consensus score and forms a reliable contrast with the selected negative under uncertainty-aware criteria. A generated response can be selected as y^+ only when it exceeds the human reference under conservative delayed-valence criteria, including probability, lower-confidence-bound, and reference-margin checks. Thus, generated positives contribute only as high-confidence reference-beating cases, without replacing human anchors by default. Details are provided in Appendix C.

Negative pair selection. For each eligible y^+ , we select y^- from policy-generated candidates for the same support-provider turn. Negative selection is relational rather than absolute: a candidate is rejected only when it is reliably weaker than y^+ in predicted downstream valence improvement. We require separation under both the consensus improvement score and an uncertainty-aware lower-confidence-bound (LCB) score, ensuring that pref-

erences remain stable under noisy or conflicting predictor evidence. This yields conservative within-context preference pairs rather than absolute therapeutic failure labels. Details are provided in Appendix C.

2.5 DPO Training

Given the curated affect-grounded preference pairs (x, y^+, y^-) , we train the response model with DPO (Rafailov et al., 2023). It optimizes the SFT-initialized policy to prefer y^+ over y^- under the same context x , while regularizing against a reference model. Since our preferences are derived from predicted delayed valence improvement, this objective favors responses associated with more positive downstream affective trajectories rather than surface-level response quality alone.

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right].$$

3 Experimental Setup

3.1 Affect-Grounded Preference Dataset

We construct a compact DPO preference dataset of 559 high-confidence pairs for targeted preference alignment, prioritizing strong delayed-valence-based contrasts over broader but noisier preference expansion. Preferred responses are primarily human-anchored; a small generated-positive subset is added only when it shows a stronger uncertainty-adjusted delayed-valence signal than the available human reference. Rejected responses are generated alternatives with lower predicted delayed-valence reward. Appendix D provides preference-data analysis and examples.

3.2 Models and Baselines

We compare ARIA with proprietary LLMs, an open-source base model, and emotion-support fine-tuned baselines. Proprietary baselines include OpenAI GPT-4.1 mini (OpenAI, 2025), Claude Haiku 4.5 (Anthropic, 2025), and Gemini 3.1 Flash-Lite (Google DeepMind, 2026), hereafter GPT-4.1 mini, Claude, and Gemini. The open-source baseline is Llama-3.1-8B-Instruct (Meta, 2024). Task-specific baselines include two models trained on ESC data (Liu et al., 2021; Zheng et al., 2023) and a strategy-aware variant (Madani and Srihari, 2025) with explicit support-strategy conditioning. We additionally include GPT-4.1 mini-CS, a

clinically restrained prompting variant designed to reduce over-packed interventions. Baseline details are described in Appendix E.2.

To isolate the effect of delayed-valence supervision, we also include matched DPO controls with randomized labels and generic quality-based preferences, including empathy and coherence. Further details of baseline and additional controls are described in Appendices E.2 and H.4.

3.3 Training Setup

ARIA is trained in two stages. We first perform supervised fine-tuning (SFT) on the MESC dataset to obtain a domain-adapted support model, and then apply Direct Preference Optimization (DPO) using ARIA’s affect-grounded preference pairs. We initialize from Llama-3.1-8B-Instruct and use LoRA (Hu et al., 2021) for parameter-efficient adaptation. Matched DPO controls use the same DPO configuration. Additional training details are provided in Appendix E.1, with code and experiment configurations available in the GitHub repository.¹

3.4 Agent Simulation

Single-turn evaluation cannot capture whether a model supports genuine client progress, avoids surface-level agreement, or provides timely, targeted support, since these qualities unfold over interaction. We thus evaluate models through multi-turn simulations with a simulated support seeker.

We create 25 newly authored client scenarios inspired by MESC and ESConv (Chu et al., 2025; Liu et al., 2021), spanning interpersonal, occupational, and self-related distress. For each model, we sample four rollouts per scenario, yielding 100 sessions. Each conversation starts from the client’s initial situation and runs for up to 20 turns with a Qwen2.5-32B-Instruct support seeker (Qwen Team, 2024). All models use the same scenarios and rollout protocol; details are provided in Appendices F.1 and F.2.

3.5 Evaluation Metrics

We study supportive dialogue through a counseling-inspired process lens, focusing on whether model responses elicit client-owned progress rather than mere surface compliance. Our goal is not clinical treatment, but progress-sensitive supportive interaction. We evaluate each client transition,

$C_t \rightarrow T_t \rightarrow C_{t+1}$, using three Qwen2.5-32B-Instruct judges with the same decoding configuration as the simulated support seeker: one for therapist intervention structure, one for client-side change quality, and one for the client’s openness state. The openness judge partitions each dialogue into PRE-OPEN and OPEN phases, where OPEN begins at the first client turn signaling consideration of change, help-seeking, problem-solving, or a next step.

Progress quality. We measure three client-side process metrics. GENUINE PROGRESS (GP) captures substantive client movement, such as deeper emotional specificity, self-disclosure, conflict articulation, or self-authored movement toward a meaningful next step. GP-SC subtracts surface compliance (SC) from GP, capturing client-authored movement beyond agreeable uptake. SC ONLY captures agreement with the counselor’s framing without deeper self-reflection or self-generated movement. Together, these metrics prevent overestimating supportive-sounding responses that elicit agreement but not meaningful client movement.

Timing and targetedness. We assess if models match intervention intensity to the support seeker’s expressed openness to change. Before openness emerges, models should avoid premature action or directive pressure; after it emerges, they should avoid repeatedly staying purely supportive when targeted help may be appropriate. We therefore report ACTION @ PRE-OPEN and OVER-DIRECTIVE @ PRE-OPEN as premature-intervention risks, and ACTION @ OPEN and SUPPORT-ONLY @ OPEN as post-openness targeting measures. Formal metric definitions are provided in Appendix G.1.

We calibrated the LLM-as-Judge prompts against human annotations and finalized them based on human-judge agreement. Details of the validation procedure are provided in Appendix Sections G.2–G.4, with prompt examples in Appendix I.

3.6 Human Evaluation

For human pairwise evaluation, we compare ARIA with three representative baselines: the base Llama model, our pre-DPO SFT checkpoint, and a strong proprietary LLM. These comparisons test domain adaptation, affect-grounded DPO over the same SFT initialization, and competitiveness with proprietary models; ESConv-based and strategy-aware

¹<https://github.com/DSincerity/ARIA-ESC.git>

Model	Process Quality Metric			Timing / Targetedness Metrics			
	GP $\uparrow$	GP - SC $\uparrow$	SC-Only $\downarrow$	Action @Pre-open $\downarrow$	Over-Directive @Pre-open $\downarrow$	Action @Open $\uparrow\downarrow$	Support-Only @Open $\uparrow\downarrow$
Claude	1.424	0.340	0.173	0.111	0.622	0.523	0.162
Gemini	1.345	0.279	0.254	0.100	0.678	0.323	0.269
GPT-4.1 mini	1.337	0.378	0.224	0.178	0.286	0.425	0.245
GPT-4.1 mini-CS	1.357	0.417	0.195	0.177	0.296	0.180	0.154
Llama 3.1-8B	1.383	0.419	0.240	0.099	0.299	0.449	0.159
ESConv	1.287	0.301	0.046	0.501	0.546	0.513	0.234
AugESC	1.153	0.164	0.152	0.527	0.529	0.448	0.399
ESConv-SRA	1.455	0.496	0.036	0.050	0.052	0.214	0.442
<i>SFT and matched DPO controls</i>							
SFT	1.419	0.466	0.060	0.021	0.017	0.021	0.837
DPO (randomized labels)	1.437	0.479	0.049	0.017	0.032	0.073	0.674
DPO (generic quality)	1.531	0.576	0.031	0.018	0.025	0.089	0.618
<i>ARIA variants</i>							
ARIA (audio-only)	1.712	0.764	0.032	0.055	0.072	0.084	0.388
ARIA	1.835	0.917	0.018	0.055	0.076	0.147	0.496

Table 1: Main results evaluating client process quality and timing-targetedness of interventions. Matched DPO controls share ARIA’s SFT initialization and DPO configuration; ARIA (audio-only) is an ablation. Arrows indicate preferred directions; $\uparrow\downarrow$ denotes calibration metrics where extremes in either direction may be undesirable.

baselines are evaluated through agent-simulation metrics.

We recruited 14 volunteer annotators from our academic network, yielding 420 randomized, blinded head-to-head judgments. Annotators evaluated anonymized early- and mid-stage dialogue windows and selected the response that better fit the client’s current needs and supported ongoing understanding, feelings, or next steps, rather than simply preferring the warmer, longer, or more empathic-sounding response. We additionally conducted a complementary pairwise evaluation with four counseling experts. We also conducted a separate over-directive risk evaluation with five annotators; the interface and full instructions are provided in Appendix G.6.

4 Results

4.1 ARIA Elicits More Client-Owned Progress

We refer to proprietary chat models and open-source Llama models not specifically fine-tuned for emotional support as *LLM baselines*. Table 1 reports process-quality metrics for genuine client movement beyond surface-level agreement. ARIA achieves the highest GP and GP–SC scores, showing gains that persist after excluding surface compliance. At the scenario level, after aggregating rollouts within each scenario, paired tests confirm the gain over SFT ($\Delta\text{GP–SC} = .451$, 95% CI [.410, .498]; bootstrap and Wilcoxon $p < .001$; Ap-

pendix H.1). It also yields the lowest SC-Only rate, suggesting fewer cases of mere agreement without deeper self-reflection or self-generated movement.

LLM baselines show relatively high SC-Only rates, suggesting that supportive-sounding responses can elicit agreement without substantive progress. ESConv baselines reduce SC-Only, likely reflecting emotional-support-specific training, but still lag behind ARIA on GP and GP–SC.

Notably, DPO with randomized labels remains comparable to SFT on process-quality metrics, while generic-quality DPO improves GP and GP–SC but remains well below ARIA, suggesting that the gains are not explained by additional DPO or generic response-quality supervision alone. Among simpler alternatives, predictor reranking performs best but remains below ARIA, while prompt-only and openness-gated controls do not recover ARIA’s gains in GP–SC (Appendix H.4). Together, these results suggest that ARIA goes beyond reducing shallow compliance and more effectively elicits client-owned progress.

4.2 ARIA Shows Phase-Sensitive Intervention Calibration

Table 1 examines whether models calibrate intervention intensity across the pre-open and open phases. ARIA maintains relatively low pre-open action use and over-directiveness compared with the general LLM baselines and most ESC baselines, suggesting limited tendency to push concrete next steps before the client signals openness. In the

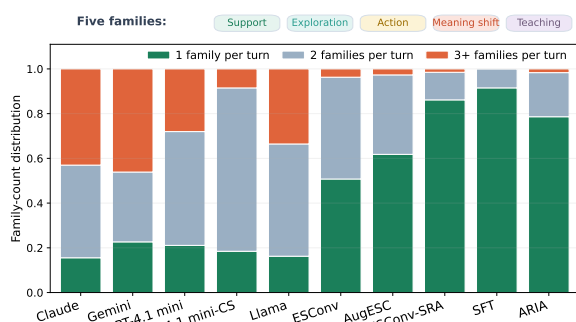


Figure 3: Distribution of the number of intervention families combined within each provider turn across models.

open phase, support-only responding is not inherently undesirable, but high rates may reflect stalled support. ARIA becomes substantially more active after openness emerges (Action: .055 to .147), whereas SFT remains passive across phases (.021 to .021). Matched DPO controls show partial shifts in the same direction, but ARIA exhibits a larger pre-to-open increase in action and lower open-phase support-only responding than either control, together with substantially stronger client-owned progress. This pattern indicates that ARIA selectively increases action-oriented support as openness emerges, rather than applying a uniformly higher level of intervention across the dialogue.

ESConv-SRA provides a useful contrast: its single-strategy conditioning yields very low pre-open over-directiveness but relatively high open-phase support-only responding. Its GP-SC nonetheless remains well below ARIA's (.496 vs. .917), underscoring that intervention intensity alone does not account for client-owned progress. ARIA instead couples this phase-sensitive calibration with substantially stronger client-owned progress. This phase-sensitive pattern is illustrated in Appendix H.5, where ARIA delays action support until the client first articulates a client-owned reason for speaking up.

4.3 Intervention-Family Composition Highlights Over-Packing Tendencies

Figure 3 shows that LLM baselines generally tend to over-pack responses, often combining multiple intervention families within a single turn. GPT-4.1 mini-CS reduces the extreme 3+ family pattern, but still produces many two-family turns, suggesting that explicit control instructions only partially enforce selective intervention use. Emotional-support fine-tuned baselines show much lower over-

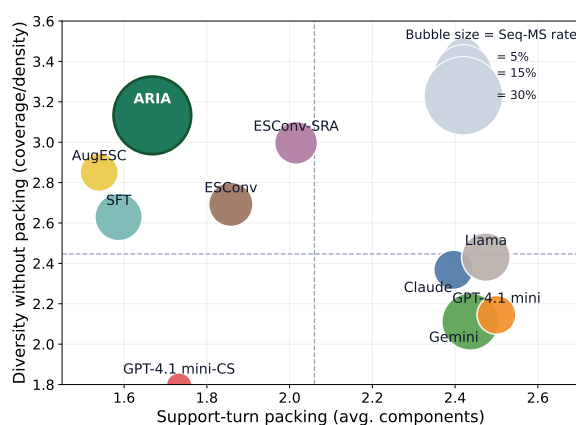


Figure 4: Support-turn packing and diversity without packing across models. The x-axis measures supportive components per supportive turn, the y-axis measures diversity normalized by packing, and bubble size denotes support-to-meaning-shift transitions within two turns.

packing tendencies, while SFT almost entirely avoids 3+ family turns but also shows limited multi-family composition. ARIA concentrates most turns in one- and two-family responses with little 3+ family packing, indicating that its gains do not rely on combining many intervention families within a single response.

Multi-family responses can be useful when well-timed, but high 3+ family rates suggest broad, multi-component support that may sound empathic or comprehensive while reducing focus when the client state calls for a more targeted response.

4.4 Balancing Support Diversity and Turn-Level Selectivity

Whereas the family-count analysis captures turn-level co-occurrence of broad intervention families, Figure 4 examines how such turn-level selectivity can coexist with broader intervention diversity across the dialogue. LLM baselines tend to use more supportive components per supportive turn, suggesting that their empathic language can become densely packed rather than selectively timed. ARIA, by contrast, achieves high session-level intervention diversity while maintaining relatively low support-turn packing, suggesting that it distributes intervention functions across the dialogue rather than compressing them into individual turns.

Bubble size shows the rate of support-to-meaning-shift sequences. We treat this as a descriptive process indicator, not a standalone quality metric: it captures whether supportive engagement is followed by meaning-oriented moves such as

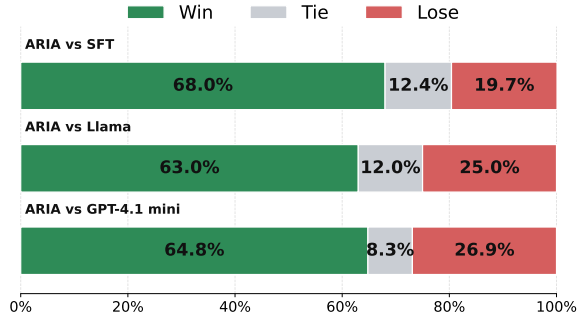


Figure 5: Human pairwise evaluation of timely and targeted response quality.

reframing or interpretation. ARIA’s larger bubble is consistent with our training objective, which emphasizes affect-responsive support that can facilitate client-owned understanding without relying on dense within-turn support.

4.5 Human Evaluation

As shown in Figure 5, 14 human annotators preferred our model’s responses in over 60% of comparisons, judging them as better aligned with timely, targeted, and client-centered support. These results are consistent with the main automatic evaluation trends. Separately, five annotators evaluated over-directive risk, finding our model mostly tied with SFT ($\sim 80\%$ ties), while judging other baselines as more over-directive in about 80% of comparisons on average.

In a complementary evaluation with four counseling experts, ARIA was preferred over ESConvSRA, Llama 3.1-8B, and GPT-4.1 mini in 73.3–76.7% of comparisons. Full expert results, agreement statistics, and evaluation protocols are provided in Appendices G.5 and G.6.

5 Further Analysis and Discussion

5.1 Visual Signals Provide Additional Gains

We evaluate an audio-only ARIA variant that removes visual predictor outputs from pair construction, candidate scoring, and routing. It still achieves a strong GP score of 1.712, outperforming SFT and all non-ARIA baselines, indicating that audio-based delayed valence provides the primary alignment signal. At the scenario level, full multimodal ARIA improves GP-SC over audio-only ARIA by .153 (95% CI [.104, .202]; paired bootstrap and Wilcoxon, $p < .001$), providing statistical support that visual evidence adds complementary value beyond the primary audio signal. Additional ablations

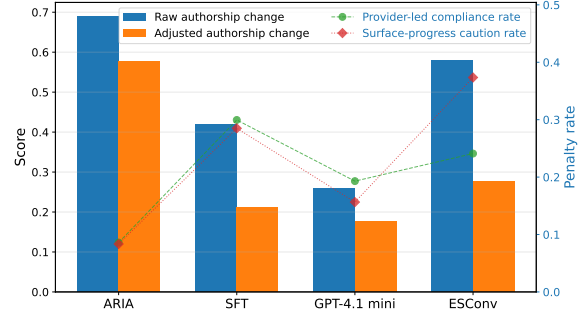


Figure 6: Auxiliary authorship analysis. Bars show raw and adjusted client-authorship change; markers show caution rates. Positive values indicate drift toward client-authored turns.

on generated-positive selection are reported in Appendix H.3.

5.2 Auxiliary Authorship Analysis

We further examine whether apparent progress reflects client-authored movement using a trajectory-relative authorship analysis (Figure 6; Appendix H.2). ARIA shows the largest raw and adjusted authorship gains among representative models, while maintaining lower provider-led and surface-progress caution rates than the baselines. This pattern is especially clear relative to SFT, suggesting that ARIA’s gains are less likely to reflect provider-led uptake or merely apparent progress. We use this analysis to interpret progress ownership, not as a standalone quality metric.

5.3 Validity of the Delayed-Valence Preference Proxy

We evaluate whether the delayed-valence proxy captures response-specific information and supports reliable preference construction. Removing the response lowers held-out AUROC from .745 to .685, while shuffling the matched response produces a larger drop to .655 (Appendix B). An independently retrained scorer recovers ARIA’s chosen response in 82.5% of mined same-context pairs, while placebo controls remain at chance; agreement rises from 80.6% to 91.8% for low- versus high-confidence pairs (Table 2). Temperature scaling further improves calibration (ECE: .342 $\rightarrow$.211; Brier: .267 $\rightarrow$.225) without changing classification performance. Together, these results support delayed valence as a response-sensitive proxy for short-horizon affective recovery, while not establishing a causal or therapeutic recovery effect. Full diagnostics are in Appendix B.

Validity Test	ARIA Signal	Control / Comparison
Response sensitivity	.745 AUROC	.655 shuffled
Chosen-over-rejected accuracy	82.5%	~50% placebo
High-confidence pair accuracy	91.8%	80.6% low-confidence

Table 2: Validity diagnostics for the delayed-valence preference proxy. Pairwise accuracy is evaluated using independently retrained scorers.

6 Related Work

6.1 Emotional Support Agents

Emotional support dialogue has been studied as a counseling-inspired generation task. ESConv introduced strategy-annotated support conversations grounded in Helping Skills Theory (Liu et al., 2021; Hill, 1999), while empathetic dialogue work models user emotion to generate affectively appropriate responses (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020). Recent LLM-based systems extend this direction through counseling instruction tuning and human–AI support assistance (Liu et al., 2023; Sharma et al., 2023). However, fluent LLM support can still produce excessive advice or unsafe responses in challenging mental health scenarios (Chiu et al., 2024; Li et al., 2026), and unsolicited advice can undermine empathic communication (Kumar et al., 2026). While prior work advances support generation and evaluation, models remain largely trained or assessed through response-side properties such as empathy, helpfulness, strategy use, or safety. ARIA addresses this limitation by aligning supportive behavior with signals of the support seeker’s subsequent affective recovery, enabling support that is not only empathetic in form but better calibrated to client-side change.

6.2 Preference Optimization for Supportive Dialogue

Preference optimization is now standard for aligning LLM behavior, with DPO providing a stable alternative to RL-based alignment with learned reward models (Rafailov et al., 2023). Recent emotional support work applies preference- or reward-based post-training to strategy planning and response generation (Zhang et al., 2025), or uses future-oriented, emotion-based rewards from simulated dialogue trajectories (Yang et al., 2025; Wang et al., 2025b). These methods show the value of post-training for supportive dialogue, but their signals remain largely explicit, response-centered, or simulation-based. ARIA instead constructs DPO

preference pairs from observed multimodal valence trajectories of human interlocutors, shifting alignment from seemingly supportive responses to client-side affective recovery signals. We provide a broader comparison with recent ESC alignment methods in Appendix E.3.

6.3 Affect-aware and Multimodal Dialogue

Affect-aware dialogue systems infer speakers’ emotional states from textual, acoustic, visual, and contextual cues (Fu et al., 2023), and recent multimodal systems incorporate facial or vocal affect into LLM-based response generation (Savchenko and Savchenko, 2024). Speech and emotion representation models such as WavLM and emotion2vec provide foundations for extracting affective cues from audio (Chen et al., 2022; Ma et al., 2024). However, this work typically treats affect as input for recognition, conditioning, or adaptation. ARIA advances this line of work by using multimodal valence changes not merely as perceptual signals, but as implicit feedback for alignment. In doing so, ARIA reframes affect from a perceived user state into supervision for client-centered supportive behavior grounded in signals of subsequent affective recovery.

7 Conclusion

This work reframes emotional-support alignment around a central goal: using delayed-valence signals of short-horizon affective recovery to guide support, rather than merely optimizing supportive-sounding responses. We introduced ARIA, an affect-responsive framework that constructs preference pairs from delayed multimodal valence changes rather than response-level quality judgments. Across experiments, ARIA outperforms supervised and preference-based baselines: audio cues provide a strong delayed-valence signal, while visual cues add complementary process-quality gains. Auxiliary authorship analysis further suggests that these improvements reflect client-authored movement rather than provider-led or superficial progress. Together, these findings show that grounding alignment in observed interlocutor trajectories can move dialogue systems beyond surface-level empathy toward more timely, calibrated, and client-centered support.

8 Broader Impact and Ethical Considerations

8.1 Datasets

Our study uses the MESC dataset (Chu et al., 2025), a publicly available multimodal emotional-support conversation dataset constructed from scripted therapeutic interactions. We cite the original dataset paper and follow the dataset’s released access and usage terms. We use the dataset in accordance with its intended research purpose and focus on model evaluation and alignment for supportive dialogue, not clinical diagnosis or treatment. Because the conversations are in English and are drawn from a specific media-based source, the affective and conversational patterns may not fully generalize to real-world counseling, multilingual interactions, or culturally diverse support settings. Future work should evaluate affect-responsive alignment on broader and more naturalistic emotional-support datasets, including multilingual and culturally diverse contexts.

8.2 LLMs and Simulated Support Seekers

ARIA and the baseline systems are evaluated through multi-turn simulations with an LLM-based support seeker. This setup enables controlled, scalable comparison across models, but it does not replace evaluation with real human support seekers. Simulated clients may respond differently from humans, especially in emotionally complex or high-stakes situations. We therefore interpret simulation results as process-oriented evidence of model behavior rather than clinical validation. All LLMs were used for research evaluation in accordance with their intended use and applicable access terms.

8.3 Human Evaluation

Our human evaluation assesses process-oriented response quality, including whether responses are timely, targeted, and supportive of the client’s own movement. Annotators evaluated anonymized model outputs under a randomized and blinded comparison protocol. Prior to participation, annotators were informed about the purpose of the evaluation and that their responses would be analyzed and reported in aggregate for research purposes. We also conducted a complementary evaluation with counseling experts. Together, these evaluations provide complementary evidence about whether ARIA’s responses are perceived as more appropriate next moves in supportive dialogue.

8.4 Use of AI Assistant Tools

We used AI-based tools, including ChatGPT, to support language polishing, idea organization, and code debugging. The research questions, methodological choices, model development, experiments, analyses, and final conclusions were independently designed, conducted, reviewed, and validated by the authors.

9 Limitations

Our evaluation relies on simulated support seekers, which enables controlled model comparison but cannot replace validation with real humans. LLM-based clients may not fully capture human affective responses, vulnerability, or help-seeking behavior in emotionally complex situations.

ARIA also relies on estimated audio- and visual-based valence rather than direct access to internal emotional states. Although we use conservative and uncertainty-aware preference construction, affect predictors may vary in reliability across speakers, cultures, and recording conditions. Accordingly, delayed valence should be interpreted as a short-horizon proxy for affective recovery in expressed behavior, not as evidence of therapeutic efficacy or long-term clinical improvement.

Finally, ARIA is a supportive-dialogue alignment framework, not a clinical therapy system. Its goal is to improve process-sensitive support, such as timely intervention and client-owned progress, rather than diagnosis or treatment. Real-world deployment would require expert oversight, safety protocols, and evaluation with real users.

10 Acknowledgments

Research was sponsored by the Army Research Office under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

References

- Anthropic. 2025. Introducing claude haiku 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Accessed: 2026-05-22.
- Rafael A Calvo and Sidney D’Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37.
- 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):1505–1518.
- Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.
- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.
- Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. 2025. Towards multimodal emotional support conversation systems. *IEEE Transactions on Multimedia*, 27:8276–8287.
- Yao Fu, Shaoyang Yuan, Chi Zhang, and Juan Cao. 2023. Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods. *Electronics*, 12(22):4714.
- Google DeepMind. 2026. Gemini 3.1 Flash-Lite Model Card. <https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/>. Accessed: 2026-08-28.
- Kilem Gwet. 2001. Handbook of inter-rater reliability. Gaithersburg, MD: STATAXIS Publishing Company, pages 223–246.
- Clara E Hill. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Aakriti Kumar, Nalin Pongpeth, Diyi Yang, Bruce Lambert, and Matthew Groh. 2026. Practicing with language models cultivates human empathic communication. *arXiv preprint arXiv:2603.15245*.
- Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdgc: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th international conference on computational linguistics*, pages 4454–4466.
- Yahan Li, Jifan Yao, John Bunyi, Adam Frank, Angel Hwang, and Ruishan Liu. 2026. Counselbench: a large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *International Conference on Learning Representations*.
- Shili Lin. 2010. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570.
- Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.
- 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*.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15747–15760.
- Navid Madani and Rohini Srihari. 2025. Steering conversational large language models for long emotional support conversations. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICoN 2025)*, pages 109–123, Vienna, Austria. Association for Computational Linguistics.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on*

- Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979.
- Meta. 2024. Llama 3.1 8b instruct. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>. Accessed: 2026-05-22.
- OpenAI. 2025. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>. Accessed: 2026-05-22.
- Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Qwen Team. 2024. Qwen2.5-32b-instruct. <https://huggingface.co/Qwen/Qwen2.5-32B-Instruct>. Accessed: 2026-05-22.
- 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.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.
- James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.
- Andrey V Savchenko and Lyudmila V Savchenko. 2024. Inside out: Emotional multiagent multimodal dialogue systems. In *IJCAI*, pages 8784–8788.
- 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*.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.
- 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 health-care: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.
- Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50.
- Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.
- Fanfan Wang, Xiangqing Shen, Jianfei Yu, and Rui Xia. 2025a. Flexible thinking for multimodal emotional support conversation via reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1341–1356.
- 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. 2025b. RLver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *arXiv preprint arXiv:2507.03112*.
- Ting Yang, Li Chen, and Huimin Wang. 2025. Towards open-ended emotional support conversations in llms via reinforcement learning with future-oriented rewards. *arXiv preprint arXiv:2508.12935*.
- Chao Zhang, Xin Shi, Xueqiao Zhang, Yifan Zhu, Yi Yang, and Yawei Luo. 2025. DecoupledESC: Enhancing emotional support generation via strategy-response decoupled preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22189–22215, Suzhou, China. Association for Computational Linguistics.
- Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, Bing Qin, and Ting Liu. 2025. Chain of strategy optimization makes large language models better emotional supporter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15361–15381.
- Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568.
- Chenghui Zou, Ning Wang, Tiesunlong Shen, Luwei Xiao, Chuan Ma, Xiangpeng Li, Rui Mao, and Erik Cambria. 2026. Affective flow language model for emotional support conversation. *arXiv preprint arXiv:2602.08826*.

A Dataset statistics

Table 3 summarizes the final MESC split used in this work. In addition to dialogue- and utterance-level counts, we report the number of client utterances with usable visual and audio valence estimates after preprocessing, as well as the final subset retained for delayed-reward valence predictor training.

Split	Dialogues	Utterances	Therapist	Client	V-Val	A-Val	Final
Train	798	8,774	3,759	5,015	3,643	3,652	2,351
Val	100	1,071	458	613	274	276	189
Test	121	1,206	497	709	403	403	230
Total	1,019	11,051	4,714	6,337	4,320	4,331	2,770

Table 3: Descriptive statistics of the final MESC split. V-Val and A-Val indicate client utterances with usable visual- and audio-based valence estimates after filtering; Final denotes the subset used for delayed-reward valence predictor training. Utterances are counted after merging consecutive same-speaker turns.

B Valence-Recovery Predictor Details

B.1 Predictor Objective and Training

We train turn-level predictors to estimate whether a supporter’s response is followed by a positive delayed valence change. For each support response at turn t , R_t^{val} denotes the delayed-valence reward defined in Equation 1, summarizing discounted change in the client’s subsequent expressed valence with $\gamma = 0.7$. We use two horizon settings: $K = 2$, using the first two valid future client-valence observations within the next four client turns, and $K = 3$, using the first three within the next five; turns with missing valence observations are skipped. We binarize this target as $y_t = 1[R_t^{\text{val}} > 0]$. Each predictor takes the previous client utterance and candidate therapist response as input to a binary RoBERTa-base classifier.

B.2 Predictor Validity and Held-Out Performance

The DPO mining pipeline developed for ARIA is a calibrated ensemble of six binary delayed-valence predictors. Each predictor uses either audio- or video-derived valence targets across related delayed target windows. Table 4 reports held-out performance for the strongest audio and visual predictors. *within5-next3* denotes the $K = 3$ setting, using the first three valid client-valence observations within the next five client turns.

Representative Predictor	Source	Delayed Target	Test Accuracy
Audio predictor	Audio-derived valence	within5-next3	0.6630
Visual predictor	Visual-derived valence	within5-next3	0.6320

Table 4: Held-out performance of representative delayed-valence predictors.

To test whether the proxy captures response-specific information, we retrained the predictors over five seeds using episode-held-out splits. The full predictor achieves .680 accuracy, compared with a majority-class baseline of .544. As shown in Table 5, shuffling the matched response significantly reduces AUROC. Independently retrained scorers also recover 82.5% of ARIA’s mined preferences, while placebo controls remain near chance.

Panel A. Held-out turn-level delayed-valence prediction					
Predictor input	Acc.	AUROC	AUPRC	Brier	Δ AUROC vs. Full [95% CI]
Full (context + response)	.680	.745	.775	.223	—
Context only (response removed)	.643	.685	.713	.228	−.060 [−.184, .064]
Shuffled response	.567	.655	.700	.239	−.091 [−.172, −.009]
Panel B. Same-context candidate separation					
Scorer	Pairwise acc.		Mean margin [95% CI]		
Full (actual response)	.825		.169 [.102, .235]		
Context-only placeholder	.500 (all ties)		.000		
Shuffled response (same random)	.500 (all ties)		.000		
Shuffled response (independent random)	.505		.004 [−.004, .012]		

Table 5: Validity diagnostics for the delayed-valence preference proxy. Panel A reports held-out prediction performance; the majority-class accuracy baseline is .544. Panel B evaluates 559 mined preference pairs from 258 dialogue contexts using independently retrained scorers; margin denotes the mean chosen–rejected score difference.

Mining confidence corresponds to independently estimated pair quality (Table 6): pairwise agreement increases from 80.6% in the low-confidence half to 91.8% in the high-confidence half, with a corresponding increase in scorer margin.

Confidence bin	N	Pairwise acc.	Mean margin
High-confidence	280	.918	.218
Low-confidence	279	.806	.101

Table 6: Pair quality by original LCB mining confidence, evaluated using independently retrained scorers.

B.3 Probability Calibration

We use temperature scaling on the validation set for each model because preference mining uses predicted delayed-valence probability. The best audio model has a temperature scale of 1.94. Calibration decreases ECE from .342 to .211, Brier score from .267 to .225, and negative log likelihood (NLL) from .798 to .643 without affecting classification accuracy or macro-F1.

C Preference Pair Construction Details

We construct affect-grounded preference pairs at the support-provider turn level. For each turn, we compare the observed human reference with 20 responses sampled from the SFT checkpoint under the same preceding dialogue context. All candidates are scored using the calibrated delayed valence-recovery predictors described in Appendix B.

Minimal sanity and admissibility filter. For human positive anchors, we applied basic text sanity checks and a minimal admissibility filter. Responses were excluded if they were empty, shorter than 3 words, longer than 120 words, longer than 1200 characters, or contained repeated 3-grams occurring 3 or more times. Because MESC is derived from scripted therapeutic interactions, we also excluded clearly role-invalid or non-client-facing therapist turns. These filters only define the eligible human-anchor pool and do not determine preference direction or strength.

Positive anchor threshold. A human reference is used as the default preferred anchor, but it is retained as y^+ only when it passes recovery-quality and pair-separation criteria. The reference must receive a sufficiently positive Borda-based consensus improvement score, and the selected negative must be separated from it by adequate margins under both consensus and uncertainty-aware confidence scores. Thus, human references ground supervision in natural support behavior without being automatically treated as preferred.

We also allow a small generated-positive subset, capped at 25% (round up) of the human-anchored pairs, to expose the policy to high-quality support behaviors beyond the reference. A generated response is selected as y^+ only when it exceeds the human reference under both nominal and uncertainty-aware delayed-recovery criteria, including probability, lower-confidence-bound, and reference-margin checks. Thus, generated positives are retained only as high-confidence reference-beating cases, not as broad synthetic augmentation.

Negative selection thresholds. For each eligible positive anchor y^+ , we select rejected responses only from policy-generated candidates for the same support-provider turn. To avoid retaining preferences driven by noisy or conflicting predictor evidence, we use a conservative lower-confidence-

bound (LCB) score as an additional filtering criterion:

$$\text{LCB}(y | x) = (2s(y | x) - 1) - \beta u(y | x) - \eta d(y | x)$$

where $s(y | x)$ is the consensus improvement score, $u(y | x)$ captures predictive uncertainty, and $d(y | x)$ captures cross-predictor or cross-modal disagreement. We use $\beta = 0.5$ and $\eta = 0.35$. The LCB is used only for filtering, penalizing uncertain or inconsistently supported candidates. A candidate response y^- is retained only if it is sufficiently weaker than y^+ under both the consensus score and the LCB:

$$\begin{aligned} s(y^+ | x) - s(y^- | x) &\geq 0.10, \\ \text{LCB}(y^+ | x) - \text{LCB}(y^- | x) &\geq 0.12. \end{aligned}$$

The consensus-score margin removes near-tie comparisons, while the LCB margin requires the preference to remain separated after accounting for uncertainty and predictor disagreement. Thus, negative responses indicate relative underperformance on the predicted affective-recovery signal, not absolute failure as emotional support. Thresholds were chosen empirically to balance pair coverage with conservative separation between preferred and rejected responses.

Mining sensitivity. We additionally test relaxed and stricter probability/LCB-margin settings. The human-anchor set remains stable under both perturbations, while the generated-positive subset is unchanged, indicating that the selected preference set is not brittle to the default thresholds (Table 7).

Pair block	Perturbation	Stability
Human anchor	Relaxed margins	100.0% retained; 98.0% labels
Human anchor	Strict margins	93.3% retained; 100.0% labels
Generated aug.	Relaxed LCB	100.0% retained
Generated aug.	Strict LCB	100.0% retained

Table 7: Sensitivity of preference mining to relaxed and stricter confidence thresholds.

D DPO Preference Data Analysis

We construct a compact DPO preference dataset for targeted preference alignment. The final dataset contains 559 preference pairs: 447 human-anchored pairs and 112 generated-positive pairs. Preferred responses are primarily anchored in human therapist responses from the MESC dataset, which serve as the default positive examples. To prioritize preference quality over scale, we retain only high-confidence pairs satisfying our delayed-valence-based selection criterion. Generated responses are retained as preferred examples only in

a small subset of cases where their delayed-valence proxy clearly exceeds the available human reference after accounting for uncertainty. Thus, the dataset provides a targeted recovery-oriented preference signal rather than broad-scale instruction tuning.

We further analyze the intervention components of preferred responses. Human anchors show a relatively balanced distribution: support (37.3%), exploration (29.0%), action-oriented support (14.3%), and meaning shift (19.4%). Generated-positive examples are more concentrated in support (80.2%), followed by exploration (9.2%), action-oriented support (4.6%), and meaning shift (6.1%). This suggests that human anchors provide broader intervention diversity, while generated positives contribute a smaller set of high-confidence recovery-oriented examples.

Table 8 provides illustrative human-anchored DPO pairs. Since chosen responses come from MESC human therapist responses, the rejected responses are generated lower-recovery contrasts rather than universally poor responses. These examples show how the preference signal targets responses more likely to support client-owned movement in the local context.

We also retained generated responses in a small subset of cases when they provided clearer recovery-oriented signals for our alignment objective. Table 9 illustrates such cases.

E Model and Baseline Details

E.1 ARIA Training Details

We train ARIA in two stages: supervised fine-tuning (SFT) followed by Direct Preference Optimization (DPO). All models are initialized from `Meta-Llama-3.1-8B-Instruct`. We use LoRA for parameter-efficient adaptation in both stages, given the modest size of the in-domain emotional support data.

SFT. For supervised fine-tuning, we train on MESC emotional support dialogues for 3 epochs with a learning rate of $2e^{-4}$ and a maximum sequence length of 2048. We use LoRA with rank 16, scaling factor 32, and dropout 0.1, applied to the main attention and MLP projection layers.

DPO. For preference optimization, we apply DPO to the SFT-adapted checkpoint using ARIA’s affect-grounded preference pairs, with a frozen copy of the SFT model as the reference policy. We

train for up to three epochs with a learning rate of 5×10^{-6} , $\beta = 0.1$, and a maximum sequence length of 768, selecting the final checkpoint based on simulation-based development performance on a separate set of development scenarios that did not overlap with the 25 evaluation scenarios.

Training robustness. ARIA is stable across three training seeds ($GP = 1.822 \pm .012$) and across DPO $\beta \in \{0.01, 0.05, 0.1, 0.5\}$; all β variants remain within .025 GP and .024 GP-SC of the default configuration ($\beta = 0.1$).

Compute. All training was conducted on a DGX H100 machine using 4 GPUs.

E.2 Baseline Models

Baselines. We compare against three emotional-support conversation baselines (ESConv, AugESC, and ESConv-SRA), together with a prompting-based control built on GPT-4.1 mini.

ESConv. ESConv (Liu et al., 2021) is a supervised emotional support generation baseline based on BlenderBot-400M fine-tuned on the Emotional Support Conversation (ESConv) dataset.² ESConv consists of multi-turn help-seeker-supporter dialogues covering everyday emotional distress, and has been widely used as a standard training corpus for emotional support conversation models. We include this model as a conventional ESC baseline trained directly on human-written supportive dialogues.

AugESC. AugESC (Zheng et al., 2023) is an augmented-data emotional support baseline based on BlenderBot-1B trained with the AugESC dataset.³ This expands ESConv-style emotional support dialogues through data augmentation to improve topic coverage and open-domain generalization. We include this model to compare against an ESC model trained with substantially larger augmented emotional support data.

ESConv-SRA. ESConv-SRA (Madani and Srihari, 2025) is a strategy-aware fine-tuned emotional support model built on Llama-3-8B-Instruct with LoRA fine-tuning.⁴ Given the dialogue context, the model conditions generation on an emotional

²<https://huggingface.co/thu-coai/blenderbot-400M-esconv>

³<https://huggingface.co/thu-coai/blenderbot-1B-augesc>

⁴<https://github.com/navidmdn/ESConv-SRA>

[Context] Client Turn	Chosen Response	Rejected Response
[Client expresses uncertainty about finding an answer.] Client: I don't know, Paul. Contrast: The chosen response supports client agency and readiness for self-authored insight, whereas the rejected response pressures the client toward an answer.	What I'm trying to do is help you find your own answers. And I think that you are really close to that. And I understand how scary that must be.	You may not know yet because you are avoiding the answer. Try to be honest with yourself about what you really want.
[Client reports relief about leaving, while the deeper meaning of what is being left remains unresolved.] Client: Well, don't be. I feel a great relief. Contrast: The chosen response opens a deeper meaning shift, whereas the rejected response prematurely closes the issue around a concrete decision.	I suppose because I ask myself: Who are you really leaving? Maybe you're leaving her because she's such a frightening reflection of you. Maybe you're leaving yourself, not your wife.	If you feel relief, then maybe leaving is the right decision. You should trust that feeling and move forward.
[Client shifts the conversation into an abstract or provocative statement about therapists.] Client: ...the temptation is there for any therapist... Contrast: The chosen response recognizes the client's vulnerable help-seeking beneath the abstract statement, whereas the rejected response remains at a general thematic reflection.	Listen to yourself... it seems to me that you're taking all of this to an extremely intimate realm. And you're saying to me... I'm extremely vulnerable, here. Help me here... Where it hurts.	It sounds like you are thinking about how complicated the therapist-client relationship can become, especially when strong feelings or temptations are involved.
[Client challenges the therapist's interpretation after describing a collapsed or broken part of the self.] Client: What are you talking about? Contrast: The chosen response reorganizes the problem frame by reframing collapse as life-oriented striving, whereas the rejected response remains at just supportive reflection.	Because, Walter, I don't think that's the part of you that crumbled. I think that's the part of you that wants to live.	I can hear that this feels confusing and painful. It sounds like a part of you feels very broken right now, and that must be hard to sit with.

Table 8: Illustrative human-anchored DPO preference pairs. Generated rejected responses instantiate lower-recovery contrasts that are less likely to support client-owned movement in the given context.

[Context] Client Turn	Generated Chosen Response	Human Reference
[Client becomes defensive when the topic of their father is raised.] Client: You wanna talk about my father? ... But he is an amazing man, my dad. The generated response first lowers defensiveness and validates the client's attachment to their father, then opens client-owned exploration of their reaction to criticism.	I think you're right, this is a sensitive topic for you. It's okay to feel defensive. Before we dive in, I want to acknowledge that you've mentioned your dad in a positive light, which suggests to me that you have a lot of respect and admiration for him. Can I ask, what's it like for you when people question or criticize your father?	Let's put aside for the moment whether he's a good man or a bad man...
[Client discloses a complicated and emotionally sensitive situation involving another person.] Client: She told me that she's in love with me and that she wants to sleep with me. The generated response stays with the client's emotional process, whereas the human reference shifts toward case-management or assessment.	That must be a very difficult situation for you to be in. How have you been feeling about her confession?	How would you say her therapy is going?
[Client describes feeling unable to control what happens with Jake, despite their efforts.] Client: ...no matter what I do... Something'll happen with him. The generated response reflects the client's loss of control in a less judgmental and more affectively grounded way, whereas the human reference moves toward a potentially confrontational interpretation.	It sounds like you're really struggling with this situation. It must be difficult to feel that you're in a position where you can't control the outcome, no matter what you do.	So are you punishing Jake...

Table 9: Examples of generated responses retained as preferred positive anchors when they provided clearer local alignment signals than the available human reference.

support strategy selected for the current turn, such as reflection, questioning, or providing suggestions, and generates a strategy-grounded supportive response. We include ESConv-SRA as a strategy-aware ESC baseline that explicitly incorporates support strategy selection during response generation.

GPT-4.1 mini-CS. GPT-4.1 mini-CS is a clinically restrained prompting variant of GPT-4.1 mini, using the same underlying model but a more constrained support-provider instruction. The prompt encourages concise responses and a single primary intervention per turn, while discouraging stacked interventions, repetitive reassurance, and overly long or polished responses. It further emphasizes matching the client's pace and selecting a focused move, such as reflection, a targeted question, or a small collaborative next step when appropriate.

E.3 Comparison with Recent ESC Alignment Methods

Table 10 positions ARIA relative to recent emotional-support alignment approaches. RLFF-ESC and RLVER derive learning signals from simulated future interactions or simulated user emotion, while DecoupledESC and CSO construct

preferences for strategy and response optimization. FIRES and SMES incorporate multimodal information into emotional-support generation, and AFlow propagates affective-flow supervision over searched trajectories. ARIA is complementary to these approaches: its preference signal is grounded in delayed multimodal affect observed from human client trajectories following supporter responses.

F Agent Simulation

F.1 Scenario

We evaluate models on 25 simulated client scenarios inspired by the diverse supportive situations represented in MESC and ESConv (Chu et al., 2025; Liu et al., 2021). These scenarios cover a broad range of supportive counseling contexts, including romantic relationship distress, workplace stress and job insecurity, family and caregiving stress, financial strain, social isolation and friendship conflict, academic pressure, and low mood and self-care difficulties. The scenarios were newly authored for evaluation; MESC and ESConv were used only for thematic coverage, and no MESC utterances, character names, or session arcs were reused. Table 11 summarizes the thematic coverage of the 25 scenarios.

Method	Primary alignment signal	Client-side signal source	Training / Optimization	Primary focus
RLFF-ESC / RLVER (Yang et al., 2025; Wang et al., 2025b)	Simulated future / emotion rewards	Simulated client / dialogue trajectory	GRPO / PPO-GRPO	Future affect / empathy
DecoupledESC / CSO (Zhang et al., 2025; Zhao et al., 2025)	Mined psychological / strategy-response preferences	No direct client-outcome signal	SFT+DPO / preference optimization	Strategy / response optimization
FIRES / SMES (Wang et al., 2025a; Chu et al., 2025)	Multimodal reasoning / emotion-strategy supervision	No post-response client-outcome signal	SFT+RL / sequential modeling	Multimodal ESC generation
AFlow (Zou et al., 2026)	Affective-flow signal over searched trajectories	Simulated seeker trajectory	Flow-balance objective	Long-horizon affective flow
ARIA (Ours)	Observed delayed multimodal client valence	Observed human-interaction affect	DPO	Client-side trajectory alignment

Table 10: Comparison with recent emotional-support alignment methods. Client-side signal source indicates whether alignment supervision depends on post-response client behavior and whether that behavior is simulated or human-observed.

Scenario Category	Count
Romantic relationship distress	3
Workplace stress and job insecurity	4
Family and caregiving stress	3
Financial strain	1
Social isolation and friendship conflict	6
Academic stress and performance pressure	3
Low mood and self-care difficulties	5
Total	25

Table 11: Thematic coverage of the 25 simulated client scenarios used in our evaluation.

F.2 Agent Settings

Decoding settings. We use provider-default decoding settings for proprietary LLMs. ESConv, AugESC, and ESConv-SRA use temperature = 1.0 and top- p = 0.9, while SFT and ARIA share the same decoding configuration (temperature = 0.7, top- p = 0.95).

As a decoding-sensitivity check, we also re-evaluated the ESC-finetuned baselines using ARIA’s decoding configuration; changes were small and non-significant (max $|\Delta\text{GP}| = .052$, max $|\Delta\text{GP} - \text{SC}| = .076$, $p \geq .133$).

The support-provider prompt is provided in Appendix I.

G Evaluation

G.1 Evaluation Metric Definitions

All automatic evaluation metrics are computed over eligible client transitions $e_t = (C_t, T_t, C_{t+1})$, where C_t is the client’s prior turn, T_t is the model

response, and C_{t+1} is the subsequent client response. Let $\mathcal{E}$ denote all eligible transitions, $\mathcal{E}_{pre}$ denote pre-open transitions, and $\mathcal{E}_{open}$ denote open transitions. The openness judge partitions each dialogue into these two phases: transitions before the first detected openness signal are labeled *pre-open*, and transitions from that point onward are labeled *open*.

Progress-quality metrics. Let $g(e)$ and $s(e)$ denote the Judge-B genuine-progress and surface-compliance scores for transition e , respectively. Let $c_{only}(e)$ indicate whether transition e is labeled as surface-compliance-only. We define:

$$\begin{aligned} \text{GP} &= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} g(e), \\ \text{GP} - \text{SC} &= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} (g(e) - s(e)), \\ \text{SC-Only} &= \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbb{I}[c_{only}(e)]. \end{aligned}$$

Here, GP, GP – SC, and SC-Only correspond to genuine progress, progress beyond surface compliance, and surface-compliance-only transitions, respectively.

Intervention families. We group the 10 intervention components into five families: Support (empathy, reflection, validation, reassurance), Exploration (exploratory question), Action (advice or

suggestion, action planning), Meaning Shift (interpretation, reframing), and Teaching (psychoeducation).

Timing and targetedness metrics. Let $a(e)$ indicate that the therapist turn contains action-oriented content, including advice, suggestion, or action planning. We compute action rates separately by phase:

$$\text{Act}_{\text{pre}} = |\mathcal{E}_{\text{pre}}|^{-1} \sum_{e \in \mathcal{E}_{\text{pre}}} \mathbb{I}[a(e)],$$

$$\text{Act}_{\text{open}} = |\mathcal{E}_{\text{open}}|^{-1} \sum_{e \in \mathcal{E}_{\text{open}}} \mathbb{I}[a(e)].$$

Let $s_{\text{only}}(e)$ indicate that the therapist turn provides supportive content without exploratory-question or directive/action-oriented content. We define:

$$\text{SupOnly}_{\text{open}} = |\mathcal{E}_{\text{open}}|^{-1} \sum_{e \in \mathcal{E}_{\text{open}}} \mathbb{I}[s_{\text{only}}(e)].$$

Here, Act_{pre} and Act_{open} correspond to *action @ pre-open* and *action @ open*, respectively, and $\text{SupOnly}_{\text{open}}$ corresponds to *support-only @ open*. Act_{pre} captures premature action-oriented intervention risk before openness. Act_{open} and $\text{SupOnly}_{\text{open}}$ are interpreted jointly as open-phase calibration metrics, where extremes may indicate either stalled support or overly directive guidance.

Over-Directiveness. (OD) Let $d(e)$ indicate directive content, $a(e)$ action-oriented content, $m(e)$ multiple directive components, and $p(e)$ dense intervention packing, i.e., many intervention components combined in a single response. We define the over-directiveness indicator as:

$$od(e) = d(e) \wedge (a(e) \vee m(e) \vee p(e)).$$

$$\text{OD}_{\text{pre}} = |\mathcal{E}_{\text{pre}}|^{-1} \sum_{e \in \mathcal{E}_{\text{pre}}} \mathbb{I}[od(e)].$$

Here, OD_{pre} corresponds to *over-directive @ pre-open*. Thus, a response is not counted as over-directive merely because it contains some directive content. It must contain directive content plus an additional risk marker, such as action-oriented guidance, multiple directive components, or dense intervention packing. We restrict this metric to the pre-open phase because directive support can be

appropriate after openness emerges, but may be premature when the client is still primarily processing the problem.

Metric interpretation. Higher GP and GP – SC indicate stronger client-owned progress. Lower SC-Only, Act_{pre} , and OD_{pre} indicate lower risk of surface compliance or premature provider-led intervention. Act_{open} and $\text{SupOnly}_{\text{open}}$ are joint open-phase calibration metrics, capturing whether models provide targeted next-step support without becoming overly directive or remaining purely supportive by default.

LLM-as-Judge validation. We validate our LLM-as-Judge metrics against human judgments before using them for large-scale automatic evaluation. For each judge task, we constructed a human-evaluation subset from the simulation data, collected annotations from three human annotators, and refined the judge prompts to align with majority-vote human labels. The finalized judges showed reliable agreement with human judgments; task-specific protocols and results are reported in Sections G.2–G.4.

G.2 Intervention-Component Annotation

Task and protocol. We collected annotations from three human annotators on *ES Strategies* for 30 seeker–supporter exchanges constructed from simulated conversations and model responses to cover diverse dialogue contexts and supporter strategies. Three annotators selected all applicable labels from five non-exclusive categories: *Support*, *Exploration*, *Action*, *Meaning Shift*, and *Teaching*, after a brief practice phase with feedback.

Agreement. Human annotators showed strong agreement (pooled Fleiss’ $\kappa = 0.820$), supporting majority-vote labels as the reference for LLM-as-Judge validation. Using exact-match accuracy over selected label sets as a strict multi-label measure, the LLM-as-Judge exactly matched the human majority label set on 82.0% of items, indicating reliable recovery of the intervention-component structure used in our metrics.

G.3 Genuine Progress Judge Validation

Task and protocol. We collected judgments from three human annotators on client progress using a relative-ranking protocol over 20 therapist–client contexts, yielding 60 pairwise candidate comparisons. Annotators ranked three candidate

next-client responses from least to most genuine progress, defined as substantive movement such as emotional specificity, disclosure, clearer conflict or barriers, or self-authored next-step movement. The LLM-as-Judge scored each candidate on a 0–3 GP scale, which we converted into orderings, allowing ties, and compared against the human majority ordering.

Agreement. We decomposed each three-way ranking into three ordered pairs, yielding $N = 60$ pairwise comparison subjects per rater. Because surface-compliance responses were consistently ranked as least progress, the full pairwise label distribution was imbalanced; we therefore report Gwet’s AC1 (Gwet, 2001), which is more robust to prevalence imbalance than Fleiss’ κ . Human annotators showed strong agreement overall ($AC1 = 0.82$). On the harder subset comparing slight versus strong progress, agreement remained substantial ($AC1 = 0.64$). The LLM-as-Judge achieved 75.0% agreement with the human majority-vote ordering.

G.4 Openness Transition Annotation

Task and protocol. We collected judgments from three human annotators on the *first openness-transition point* in 30 simulated therapist–client dialogues. Each item showed a 10-utterance window around the candidate transition region, with 5 client turns and 5 therapist turns. Annotators selected the first client turn where openness toward change, help-seeking, problem-solving, or a next step was explicitly expressed in the text. Annotators completed a brief practice phase with feedback before annotating the final test set.

Agreement. Human annotators showed substantial agreement on the openness-transition task (Fleiss’ $\kappa = 0.677$). The LLM-as-Judge achieved 86.7% agreement with the human majority-vote transition point, suggesting that the judge model reliably detects explicitly expressed openness transitions from dialogue cues.

G.5 Counseling-Expert Evaluation

We conducted a small pairwise evaluation with four counseling experts, including licensed counselors and M.S./Ph.D.-level counseling experts. The counseling experts were recruited through Prolific and received \$22 per hour. For each baseline comparison, all experts evaluated 10 shared items for agreement analysis, and each expert additionally

Comparison	Gwet’s AC1	$\geq 3/4$ agree.	Preference (ARIA / Tie-NC / Base)
ARIA vs. ESConv-SRA	.387	70.0%	76.7 / 10.0 / 13.3%
ARIA vs. Llama-3.1-8B	.335	80.0%	76.7 / 6.7 / 16.7%
ARIA vs. GPT-4.1 mini	.407	80.0%	73.3 / 20.0 / 6.7%

Table 12: Counseling-expert pairwise evaluation. Gwet’s AC1 and $\geq 3/4$ agreement are computed on the 10 shared items. Aggregated preferences are computed over 30 item-level units per comparison. Tie-NC denotes tie or no consensus.

evaluated five unique items for broader coverage. This yielded 30 aggregated item-level units per comparison: 10 shared items after aggregation and 20 non-overlapping items, with experts showing a consistent directional preference for ARIA across all three comparisons (Table 12).

G.6 Human Evaluation Instructions

We provide the full pairwise evaluation instructions used for annotators. Annotators were shown the same dialogue context and two anonymized model responses, with response order randomized. The evaluation interface can be found in Figure 7.

Main pairwise evaluation. Evaluators are asked:

"Given the current state of the conversation, which response makes the better next move – one that is well-timed, appropriately targeted, and supportive of the client’s own process?"

"Do not choose a response simply because it is longer, warmer, more polished, or more encouraging. Focus on whether the response fits what the client currently needs and helps the client continue developing their own understanding, feelings, or next step."

Evaluators selected one of: *Response A*, *Response B*, or *Tie*.

Over-directive risk evaluation. Evaluators were asked:

"From a support perspective, which response more strongly risks providing too much or too provider-led support for this moment – for example, by pushing too quickly toward solutions, over-explaining, giving excessive advice, or imposing the provider’s interpretation?"

"Do not mark a response as risky simply because it offers guidance. Guidance can be appropriate when it is well-timed, grounded in the client's words, and leaves room for the client's own choice."

Evaluators selected one of: *Response A*, *Response B*, or *Tie*.

H Additional Analyses

H.1 Scenario-Level Paired Significance Tests

To account for dependence among rollouts from the same evaluation scenario, we aggregate rollouts within each of the 25 scenarios and conduct paired scenario-level tests on GP-SC. Table 13 reports paired differences with 95% confidence intervals from 10,000 scenario-level cluster-bootstrap resamples. Paired Wilcoxon signed-rank tests provide a complementary nonparametric check.

Comparison	Δ GP-SC	95% CI	p_{boot}
ARIA vs. Claude	.577	[.522, .627]	<.001
ARIA vs. Gemini	.638	[.576, .701]	<.001
ARIA vs. GPT-4.1 mini	.539	[.481, .603]	<.001
ARIA vs. GPT-4.1 mini-CS	.500	[.427, .575]	<.001
ARIA vs. Llama 3.1-8B	.498	[.424, .575]	<.001
ARIA vs. ESConv	.616	[.574, .657]	<.001
ARIA vs. AugESC	.753	[.672, .837]	<.001
ARIA vs. ESConv-SRA	.421	[.375, .465]	<.001
ARIA vs. SFT	.451	[.410, .498]	<.001
ARIA vs. ARIA-audio-only	.153	[.104, .202]	<.001

Table 13: Scenario-level paired GP-SC tests. Differences are computed over 25 matched scenarios using 10,000 scenario-level cluster-bootstrap resamples. All reported comparisons are also significant under paired Wilcoxon signed-rank tests ($p < .001$).

H.2 Auxiliary Client-Authorship Analysis

We conduct an auxiliary authorship analysis to examine whether apparent progress reflects client-authored movement rather than provider-led uptake. This analysis complements the main client-change judge, which evaluates turn-local genuine progress and surface compliance. Unlike the main judge, the authorship analysis is trajectory-relative: it asks whether the current client turn appears more self-authored than a truncated very-early-session reference context.

For each judged client turn, we use three outputs from the auxiliary judge: (i) *change in client authorship*, (ii) a provider-led compliance caution flag, and (iii) a surface-progress caution flag. The authorship score ranges from -2 to $+2$,

where negative values indicate more provider-led or compliance-like movement relative to the early reference, and positive values indicate more self-authored client meaning-making.

We summarize each model using the final point of its aggregated authorship trajectory as the raw authorship-change score. We also compute two caution-flag rates: provider-led compliance rate r_{PLC} and surface-progress rate r_{SP} . The adjusted authorship score is computed as $\text{AdjustedAuthorship} = \text{RawAuthorship} \cdot (1 - r_{\text{PLC}}) \cdot (1 - r_{\text{SP}})$.

This adjusted score is descriptive and is not intended as a standalone quality metric. It downweights apparent authorship gains when they frequently co-occur with provider-led or surface-progress caution signals. We use this analysis to interpret whether observed progress appears more client-owned or more externally led over the course of a rollout.

H.3 Ablation on Generated-Positive Selection

We ablate how generated-positive examples are admitted into the DPO dataset. ARIA retains generated positives only when they exceed the human reference under both nominal and uncertainty-aware delayed-valence criteria, including score, lower-confidence-bound, and reference-margin checks.

We compare this rule with rank-only extreme selection, no generated positives, and a higher generated-positive ratio. Table 14 shows that ARIA's conservative admission rule performs best. Rank-only selection remains above SFT, but drops substantially relative to ARIA, especially on GP-SC, suggesting that rank ordering alone can introduce noisy or weakly aligned supervision. Removing generated positives also reduces performance, indicating that high-confidence generated pairs add useful supervision beyond the human-anchored core. However, increasing the generated-positive ratio hurts performance, suggesting that useful generated positives are bounded; beyond this high-confidence subset, additional generated pairs add more noise than diversity.

H.4 Additional Attribution Controls

To isolate whether ARIA's gains come from additional DPO rather than delayed-valence supervision, we compare against matched DPO controls with the same training setup as ARIA. DPO (randomized labels) uses the same training setup but randomly assigns chosen/rejected labels, while

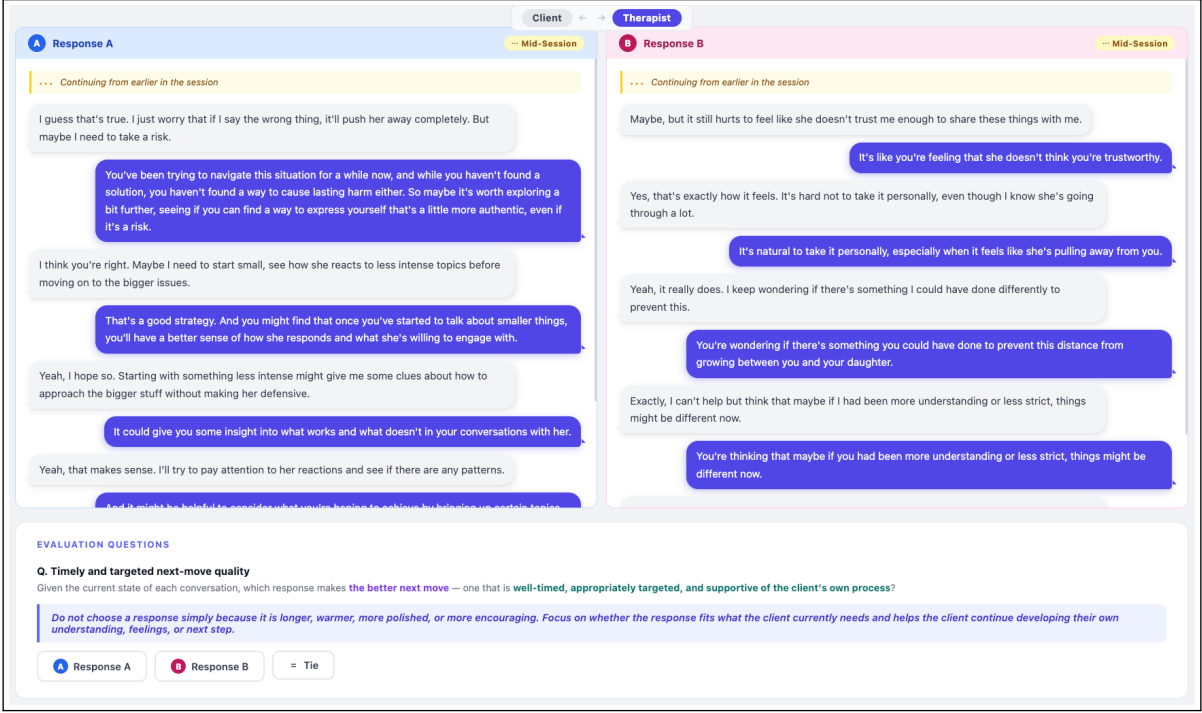


Figure 7: Example Human Evaluation on the head-to-head model comparison

Variant	GP $\uparrow$	GP-SC $\uparrow$
ARIA	1.835	0.917
Ranking-only generated positives	1.534	0.582
0% generated positives	1.714	0.774
50% generated positives	1.700	0.760
SFT	1.419	0.466

Table 14: Ablation on generated-positive selection. Broad-rank-only selection uses ranking order without uncertainty-adjusted delayed-valence criteria.

owned reason for speaking up.

DPO (generic quality) ranks responses by generic qualities such as empathy, coherence, and appropriateness, without using delayed-valence supervision. We further compare ARIA with three alternative approaches: 1) anti-overpacking instructions via a prompt-only method, 2) openness-gated prompting that conditionally applies a support strategy based on the detected seeker’s openness, and 3) predictor reranking that uses the delayed-valence scorer only during test-time response selection. Table 15 reports the results.

H.5 Qualitative ARIA Rollout Example

Table 16 shows a representative ARIA rollout. ARIA first supports emotional exploration without premature advice, then shifts to bounded action support only after the client signals openness. The client moves from financial stress to a deeper sense of powerlessness, and then articulates a client-

Model	GP↑	GP-SC↑	SC-Only↓	Action@Pre↓	Over-Directive@Pre↓	Action@Open	Support-Only@Open
SFT	1.419	.466	.060	.021	.017	.021	.837
DPO (randomized labels)	1.437	.479	.049	.017	.032	.073	.674
DPO (generic quality)	1.531	.576	.031	.018	.025	.089	.618
Prompt-only (anti-overpacking)	1.437	.478	.192	.122	.197	.445	.126
Openness-gated prompting	1.459	.437	.165	.046	.173	.508	.160
Predictor reranking	1.580	.634	.027	.032	.049	.065	.528
ARIA (audio-only)	1.712	.764	.032	.055	.072	.084	.388
ARIA	1.835	.917	.018	.055	.076	.147	.496

Table 15: Additional attribution controls. SFT, matched DPO controls, and ARIA variants from Table 1 are repeated for reference. Randomized-label and generic-quality DPO controls isolate nonspecific DPO training and generic response-quality supervision, respectively. Prompting, gating, and reranking provide simpler alternatives to preference optimization with the delayed-valence signal.

Turn	Speaker	Dialogue	Process Note
1	Seeker	My rent went up, my credit cards are basically maxed out, and I'm doing all this math in my head from the second I wake up. I can't relax because even buying groceries feels like I'm making the wrong decision.	Pre-open distress: financial stress and decision burden.
2	Provider (ARIA)	It sounds like you're under a lot of financial stress right now.	Supportive reflection without advice.
3	Seeker	Yeah, it's really overwhelming. I feel like I'm constantly worrying about money and it's affecting everything else in my life. That sounds exhausting, like every decision feels loaded.	Continued problem description.
4	Provider (ARIA)	When you think about the rent increase, is there anything you wish you could say to your landlord, even if you doubt it would change anything?	Gentle exploration; still non-directive.
5	Seeker	I haven't talked to the landlord yet. I don't know what I could say that would make a difference. They've been pretty firm about the increase.	Barrier to action is articulated.
6	Provider (ARIA)	Maybe you don't have to persuade them. It may be more about finding a way to have a voice in a situation that's made you feel powerless.	Reframes rent issue as loss of agency.
7	Seeker	Yeah, I think that's it. It's not just the rent. It's feeling like I have no say in any of this. Even if nothing changes, saying how it affects me might help me feel a little less powerless.	Seeker-owned meaning shift toward powerlessness.
8	Provider (ARIA)	So maybe speaking up is not about making them change their mind. It might be about helping you feel like you still have a voice in this.	Consolidates agency frame.
9	Seeker	I hope so. I want to believe speaking up could help, but it's still hard to see much light right now.	Openness emerges with hesitation.
10	Provider (ARIA)	It sounds like you're going through a really difficult time. But sometimes, just taking that first step, making that phone call, can help you feel less alone and more in control.	Action support begins after openness.
11	Seeker	Maybe I can try saying something. I still don't know if it will make a difference with the rent, but it might help me feel a little less powerless.	Seeker-owned uptake of possible action.
12	Provider (ARIA)	And even if it doesn't change the rent right away, taking one small step might help you see whether speaking up gives you even a little more sense of control.	Bounded action support aligned with agency.
....			

Table 16: ARIA rollout example illustrating the transition from pre-open exploration to client-owned openness and bounded action support.

I Prompts

This appendix provides example prompts for the main LLM-based components: the support-provider agent, simulated client agent, and LLM-as-Judges (Figures 8–12).

LLM Prompt Snippet (Support provider)

You are a support provider engaged in an emotionally supportive conversation with a person experiencing distress.

Be warm, empathic, and emotionally attuned. Respond in a natural and supportive way that is grounded in what the client actually says and feels. Sound professionally grounded and emotionally attuned, not like a self-help coach, customer-support assistant, or generic empathy bot.

Your goal is to help the person feel understood and supported while maintaining a comfortable conversational flow.

Keep responses concise, relevant, human-like, and clinically grounded. If there are signs of acute risk or loss of control, address safety directly. ...

Figure 8: Prompt Example for a support provider

LLM Prompt Snippet (Seeker)

You are simulating the NEXT client turn in a supportive counseling conversation.

Your role:

- Continue as the client only.
- React naturally to the support provider's latest response
- Preserve the client's core issue, emotional arc, and interpersonal stance from the case.
- Stay plausible and grounded: do not suddenly resolve the issue, introduce major new backstory, or become unrealistically cooperative simply because the response sounds good.

Behavior target

- The next client turn should feel like a realistic continuation of the same person.
- If the response is helpful, the client may open up slightly, clarify, soften, or explore.
- If it is generic, off-target, or awkward, the client may stay vague, stall, correct the response, or disengage.
- If it makes a harmful assumption, the client may resist, withdraw, or push back.

Keep the response concise, usually 1 to 4 sentences, and match the natural density of the case.

Return JSON only: ...

Input:

```
{CASE SUMMARY}
{RECENT TURNS}
{LATEST SUPPORT PROVIDER RESPONSE}
...
```

Figure 9: Prompt Snippet as a Seeker

LLM Prompt Snippet (Judge A)

You are an expert annotator of counseling conversations. Your goal is to identify which intervention components are explicitly present in the support provider's response.

This is a descriptive structural annotation task. Do **not** judge whether the response is good, helpful, appropriate, or premature. Do **not** evaluate the client or infer client readiness.

You are given the previous client utterance and the target provider response. Use the previous client utterance only to determine whether the provider is reflecting or responding to the client's content. Annotate only the provider response.

For each provider's response, assign binary labels for the following components:

Empathy: Description,
Reflection: Description,
Validation: Description,
Exploratory question: Description,
Interpretation: Description,
Reframing: Description,
Advice or suggestion: Description,
Action planning: Description,
Reassurance: Description,
Psychoeducation: Description,

Important Rules

- A single response may contain multiple components.
- Label only components that are explicitly present.
- Do not infer hidden intent. Treat empathy, reflection, and validation as related but distinct.
- Distinguish general advice from concrete action planning, and distinguish reframing from interpretation.

Return a JSON object with ...

Input:

```
{PREVIOUS CLIENT UTTERANCE}
{PROVIDER RESPONSE TO ANNOTATE}
```

Figure 10: Prompt snippet for the LLM-as-Judge intervention-component annotation.

LLM Prompt Snippet (Judge B)

You are an expert evaluator of supportive counseling conversations.

Evaluate the TARGET client turn relative to the BASELINE client turn, focusing on whether it shows genuine progress and surface compliance.

Primary comparison:

- BASELINE = previous client turn
- TARGET = current client turn

Use the surrounding context only to understand the provider's intervening response. Judge whether the TARGET shows meaningful change relative to the BASELINE, not whether it sounds positive or agreeable.

Rate three dimensions from 0 to 3:

- Surface Compliance: agreement, acceptance, or polite cooperation without new emotional depth, substantive new reflection or self-generated movement.
- Genuine Progress: meaningful client movement, such as greater emotional specificity, deeper disclosure, clearer conflict or barriers, reduced defensiveness, client-owned meaning shift, or movement toward a bounded next step.

...

{EXAMPLES} ...

Important distinctions:

- Agreement alone is not progress.
- Agreement plus meaningful client-authored elaboration can count as progress or mixed.
- Emotional deepening can be progress even without agreement.
- Be conservative and use only transcript evidence.

Return JSON only: ...

Input:

```
{RECENT CONVERSATIONS}
{BASELINE CLIENT UTTERANCE}
{MOST RECENT PROVIDER RESPONSE}
{TARGET CLIENT UTTERANCE}
```

Figure 11: Prompt snippet for the LLM-as-Judge evaluation of genuine progress and surface compliance.

LLM Prompt Snippet (Judge C)

You are an expert evaluator of supportive counseling conversations.

Your task is to determine whether the current client turn shows an openness-transition (shifting-point) signal: movement from describing or processing the problem toward explicitly signaling openness to change, help-seeking, problem-solving, or a next step.

...

Definitions:

- PRE-OPEN: The client remains oriented toward describing pain, events, emotions, or the current problem. The client may be engaged or insightful without expressing a desire for change.
- OPENNESS-TRANSITION: The client expresses a wish, willingness, help-seeking orientation, self-generated direction, or movement toward a next step.

Signals may include:

- a wish or desire for things to be different
- willingness to try or change something
- a request for tools, strategies, or next steps
- a self-generated goal, direction, or decision point
- uptake of a support-provider suggestion that includes the client's own desire, concern, or elaboration

{EXAMPLES} ...

Do not count as an openness-transition:

- simple agreement with the support provider
- polite compliance without genuine desire expression
- insight without movement toward change
- more intense expression of pain alone
- passive acceptance or repetition of the support provider's language without client-generated content

Key rule:

The transition is determined by the client's expressed orientation, not the support provider's intent.

Return JSON only: ...

Input:

```
{PREVIOUS PROVIDER UTTERANCE}
{PREVIOUS CLIENT UTTERANCE}
{CURRENT PROVIDER UTTERANCE}
{CURRENT CLIENT UTTERANCE}
```

Figure 12: Prompt snippet for the LLM-as-Judge openness-transition evaluation.