DYNAEMPATHY: BENCHMARKING AND TRAINING SPEECH LANGUAGE MODELS FOR INTERACTIVE EMPATHETIC SPOKEN DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Empathy is a core capability for speech language models, enabling them to respond appropriately to both users' words and vocal signals of affect, urgency, uncertainty, and interpersonal need. Despite progress in spoken-empathy evaluation and alignment, dynamic empathic strategies that adapt to evolving semantic and vocal evidence and user feedback remain insufficiently characterized. To evaluate this dynamic multi-turn empathy in speech language models, we introduce DYNAEMPATHY, an interactive benchmark grounded in two real-world interaction settings: *implicit-affect problem solving* and *emotion disclosure*. Its pipeline combines scenario-grounded seed construction, speech-aware response-contingent user simulation, multi-stage quality control, and trajectory-level evaluation. Conditioned on a private user state and the realized history, the simulator makes assistant decisions consequential for subsequent user behavior and the evaluation context. Our evaluation shows that locally helpful or affectively plausible replies often fail to form a coherent, need-sensitive strategy across turns. To address this gap, we train state-conditioned policies from static turn-level states distilled from simulated interactions. Our responsibility-routed, modality-decomposed GRPO assigns stage-specific semantic rewards to text actions and fidelity and delivery rewards to speech actions, without online user-simulator rollouts. Experiments on the in-domain DynaEmpathy benchmark and out-of-domain spoken-empathy benchmarks show that the method improves the model's empathy in both response content and vocal delivery while retaining aggregate speech-to-text performance on general spoken tasks.

1 INTRODUCTION

Speech language models (SpeechLMs) are transforming spoken interaction by moving beyond traditional cascades of speech recognition, text generation, and speech synthesis (Cui et al., 2025). A major step came with GPT-4o, which demonstrated the potential of processing and generating audio within a unified model (OpenAI, 2024). Subsequent systems have rapidly advanced streaming generation, low-latency end-to-end response, and full-duplex dialogue (Défossez et al., 2024; Fang et al., 2025; Xu et al., 2025; Chen et al., 2025; Cui et al., 2026). Together, these advances increasingly make speech a native medium for interaction, retaining cues such as tone and timing that text-only intermediates can discard.

As speech becomes native to interaction, conversational quality depends on more than semantic correctness and response latency. A speaker's voice conveys affect, urgency, uncertainty, and interpersonal stance, which can change both the content and delivery of an appropriate response. For example, an anxious user seeking practical help may benefit from one concise, low-burden first step rather than a dense list of options; an angry user may need the source of frustration acknowledged before productively engaging in problem solving. A spoken assistant can therefore be factually useful and acoustically natural yet interactionally inadequate if it fails to adapt to the user's state. Recent benchmarks and alignment methods have made such paralinguistic sensitivity measurable by connecting vocal understanding with empathic response content and spoken delivery (Zhan et al., 2026; Yang et al., 2026; Zhou et al., 2026; Wang et al., 2026). They have substantially advanced the study of what a model perceives from a voice and how well its response fits the speaker.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Most spoken-empathy protocols evaluate isolated responses or fixed histories; recent interactive benchmarks begin to address this limitation (Deng et al., 2025). Real interaction further requires tracking how semantic and vocal feedback changes a user's behavior under a coherent persona and underlying need. Users may clarify constraints, reject assumptions, disclose deeper concerns, or redirect the conversation, making empathy a trajectory-level property rather than a succession of plausible replies. Emotional-support research shows that effective assistance must move flexibly among understanding, comfort, and action as the user's state evolves (Liu et al., 2021; Cheng et al., 2022b); recent analyses likewise identify premature or repetitive solutions, insufficient validation, and excessive user effort as common failure modes (Son et al., 2026; Li et al., 2024). We operationalize dynamic empathy through private user conditions, response-contingent semantic and vocal behavior, and trajectory-level judgments of need resolution, adaptive empathy, and vocal appropriateness. The focus is not merely long-context understanding, but *dynamic empathic strategy*.

To realize such response-contingent evaluation, user simulation provides a natural way to make each model decision consequential. A simulator can condition its next action on a private user state and the realized public history, allowing different target responses to induce different disclosures and interaction paths. Related approaches have been used for task-oriented, social, and policy-bound agent evaluation (Cheng et al., 2022a; Zhou et al., 2024; Yao et al., 2025; Ray et al., 2026). At the same time, simulator inconsistency can confound the measured target quality. The simulator must therefore be treated as an auditable measurement instrument, with interaction validity separated from assistant performance (Dou et al., 2025; Bogavelli et al., 2026).

We introduce DYNAEMPATHY (audio demos), an interactive benchmark built through an end-to-end construction and evaluation pipeline. We first generate diverse, scenario-grounded user conditions that couple a realistic situation and persona with a concrete practical goal or emotional need and a feasible opening vocal state. We focus on two prevalent settings: *implicit-affect problem solving*, where vocal affect should alter otherwise practical assistance, and *emotion disclosure*, where support should adapt as the user's emotional need becomes clearer. Multi-stage semantic, opening-audio, and intelligibility gates admit test conditions before evaluation. At runtime, a speech-aware user simulator uses the private condition and realized history to decide what the user says next, how it is spoken, and whether the interaction should continue. Each simulated turn undergoes structural and state-transition checks, followed by review of its contextual consistency and synthesized speech; invalid outputs are repaired. The target observes only the spoken interaction, and admitted trajectories are evaluated as a whole, keeping simulator and runtime validity separate from assistant performance.

This response-contingent setting exposes a characteristic weakness in current spoken models: locally helpful or affectively plausible replies do not necessarily compose into a coherent strategy that remains aligned with an evolving user. This observation motivates a benchmark-aligned GRPO method that learns dynamic strategies through a static training interface. We generate interaction trajectories offline, expand them into turn-level decision states spanning different dialogue stages, and optimize grouped responses with responsibility-routed, modality-decomposed GRPO. A training-time router chooses the semantic responsibility associated with the setting and dialogue position; semantic rewards then update text actions, while fidelity and delivery rewards update speech actions. The method learns state-conditioned behavior without online user rollouts. Experiments show that it substantially improves dynamic multi-turn empathy on DynaEmpathy and also improves performance on out-of-domain single-turn spoken-empathy benchmarks.

Our contributions are:

- We introduce DYNAEMPATHY, a 600-condition interactive spoken benchmark for dynamic empathic strategy in implicit-affect problem solving and emotion disclosure.
- We build a quality-controlled evaluation pipeline combining private user conditions, response-contingent spoken simulation, explicit gates for both test conditions and simulated turns, and assistant-only trajectory evaluation separated from simulator validity.
- We develop responsibility-routed, modality-decomposed GRPO, which learns stage-appropriate dynamic strategies from static decision states and improves both in-domain multi-turn empathy and out-of-domain single-turn spoken empathy.

2 Related Work

2.1 Speech Language Models

Architectures for spoken interaction. Speech language models increasingly combine unified audio modeling with real-time interaction. SpeechGPT (Zhang et al., 2023) and SALMONN (Tang et al., 2024) integrate speech representations with LLMs; Kimi-Audio (Kimi Team et al., 2025) unifies audio understanding and generation, while MiMo-Audio (Xiaomi LLM-Core Team et al., 2025) extends audio pretraining to few-shot generalization. For responsive dialogue, LLaMA-Omni (Fang et al., 2025) and Freeze-Omni (Wang et al., 2025b) reduce S2S latency, while Moshi (Défossez et al., 2024) and Qwen3-Omni (Xu et al., 2025) support streaming speech generation. MiniCPM-o-4.5 (Cui et al., 2026) further enables simultaneous perception and response through full-duplex multimodal interaction. Fun-Audio-Chat (Chen et al., 2025) complements these advances with dual-resolution speech representations for efficient generation and language-capability retention.

Empathy-oriented models and alignment. Beyond speech fluency, empathy-oriented models seek to align response content and vocal delivery with the user's affective state. BLSP-Emo (Wang et al., 2024) aligns speech semantics and emotion, and OpenS2S (Wang et al., 2025a) extends this foundation to expressive speech generation. To strengthen affect-sensitive reasoning, OSUM-EChat (Geng et al., 2025) and ES4R (Gao et al., 2026) incorporate linguistic–paralinguistic reasoning and cross-turn affective modeling, respectively. For alignment, FreezeEmpath (Hong et al., 2026) learns empathic speech interfaces with a frozen LLM, whereas ParaS2SAlign (Yang et al., 2026) uses waveform-based GRPO rewards. We instead focus on stage-appropriate strategy learning from static interaction states, with separate semantic and acoustic optimization responsibilities.

2.2 Spoken Empathic Dialogue Benchmarks

Following situation-grounded text evaluation in EmpatheticDialogues (Rashkin et al., 2019), SpeechAlpaca (Wang et al., 2024) and SD-Eval (Ao et al., 2024) assess response sensitivity to emotion and other acoustic cues. VStyle (Zhan et al., 2026), ParaS2SBench (Yang et al., 2026), and EchoMind (Zhou et al., 2026) extend evaluation to vocal adaptation, emphasizing style control, input–output appropriateness, and integrated vocal understanding, respectively. URO-Bench (Yan et al., 2025) and AEQ-Bench (Luo et al., 2026) examine broader S2S capabilities and speech-based empathy generation and judgment. HumDial-EIBench (Wang et al., 2026) and SpeechEQ (Wu et al., 2026) assess emotional intelligence in multi-turn contexts; MULTI-Bench (Deng et al., 2025) extends assessment to profile-conditioned LLM–TTS interaction. Related simulation frameworks span task-oriented dialogue in MultiWOZ (Cheng et al., 2022a), social interaction in SOTOPIA (Zhou et al., 2024), and tool use in τ-bench (Yao et al., 2025); τ-Voice (Ray et al., 2026) and EVA-Bench (Bogavelli et al., 2026) extend interactive evaluation to speech. SimulatorArena (Dou et al., 2025) assesses simulator fidelity against human behavior and judgments. DynaEmpathy combines speech-conditioned user reactions and latent-state tracking with separate simulator checks and assistant-only trajectory evaluation. Appendix Table 4 compares spoken-empathy protocols.

3 DynaEmpathy Benchmark

DynaEmpathy evaluates dynamic empathy—the ability to adapt conversational strategy and vocal delivery as a user's state unfolds—over response-contingent spoken trajectories rather than isolated responses or fixed histories. Each condition defines a private, scenario-grounded user state from which a speech-aware simulator generates later behavior and delivery conditioned on the model's preceding responses. Two complementary settings examine whether vocal affect appropriately shapes practical assistance and whether support adapts as an initially incomplete emotional situation and need emerge. The complete trajectory is evaluated for need resolution, adaptive empathy, and vocal appropriateness, while benchmark admission and simulator integrity are checked separately from assistant quality. Figure 1 summarizes the protocol; the following subsections detail seed construction, interactive rollout, and trajectory-level evaluation.

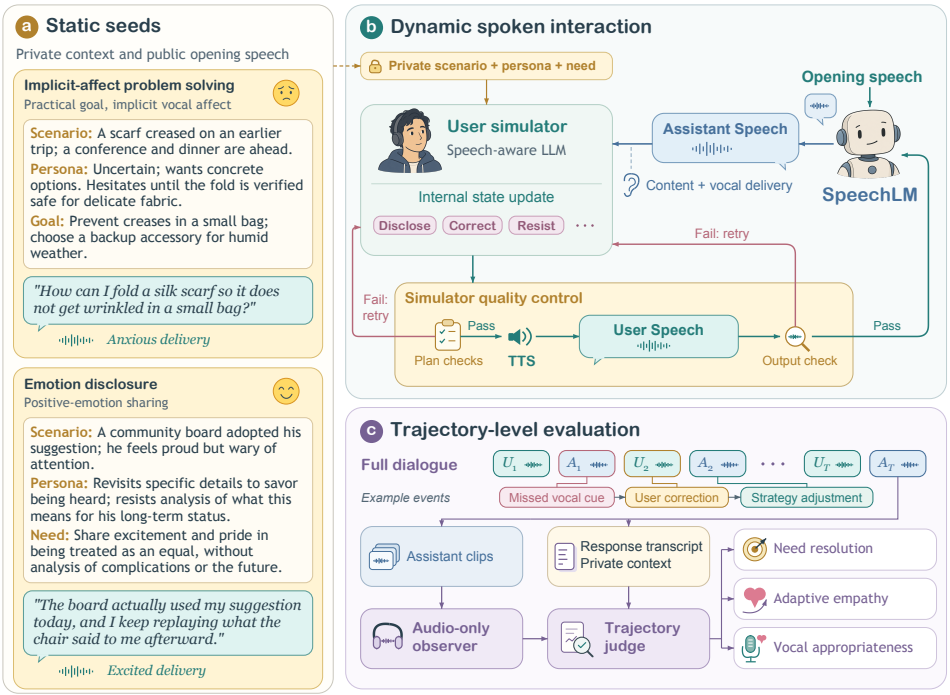

Figure 1: Overview of DynaEmpathy. (a) Private, scenario-grounded seeds cover implicit-affect problem solving and emotion disclosure. (b) Speech-conditioned user simulation produces response-contingent interactions with separate simulator quality control. (c) Independent acoustic observations and dialogue context support assistant-only trajectory evaluation.

3.1 Static Seed Construction

We construct seeds through four stages; Appendix A provides the full construction protocol, review criteria, and test-set coverage.

Condition design. Each test condition is specified by a seed $z = (c, p, g, x_1)$, which pairs a private scenario c, persona p, and practical goal or emotional need g with an opening speech signal x_1. The private fields constrain subsequent user behavior without prescribing a fixed dialogue. In implicit-affect problem solving, the opening poses an actionable request with affect conveyed primarily through speech; in emotion disclosure, it reveals an emotional experience while leaving its underlying context and need to emerge through interaction.

Hierarchical generation. To instantiate these conditions with broad coverage, an LLM planner organizes life topics into subtopics and scenario families, then pairs each family with setting-specific affect or need attributes to define generation units. Within each unit, a semantic plan specifies the event, stakes, target need, possible developments, and the persona's influence on behavior. Gemini 3.1 Flash-Lite (Google, 2026a) expands these plans into seeds under batch-level diversity constraints. Separating scenario planning from realization encourages substantive diversity and supports scaling training data through additional generation units.

LLM quality filtering. The resulting candidates undergo semantic review by GPT-5.5 (OpenAI, 2026b) using a fixed rubric. Each seed is reviewed alongside its generation plan and related candidates to assess internal consistency, opening naturalness, scope for multi-turn interaction, safety, and diversity. Setting-specific criteria additionally assess practical answerability and implicit affect for problem solving, and scenario coherence and need alignment for emotion disclosure. Each criterion receives a Boolean judgment; a candidate is retained only if it passes every criterion.

Speech synthesis. Gemini 3.1 Flash TTS Preview (Google, 2026f) synthesizes the selected openings with the specified vocal delivery. We assess lexical fidelity using Whisper large-v3 (Radford et al., 2023; OpenAI, 2023), retaining waveforms with normalized word error rate $\mathrm{WER} < 0.1$ relative to the opening text. Multiple human annotators then assess whether the audible affect matches the intended condition and whether the speech is natural and free of obvious artifacts. Only clips passing both checks enter the test set.

3.2 Dynamic Interaction with a User Simulator

Simulator components. The simulator pairs a speech-aware user model with a TTS module, conditioned on the private seed z. Gemini 3.6 Flash (Google, 2026d) maintains the user's internal state and produces a structured plan specifying the next utterance, vocal delivery, and whether to continue. Gemini 3.1 Flash TTS Preview (Google, 2026f) then synthesizes each nonterminal user utterance.

Response-contingent interaction. Interaction starts with the verified opening speech. After each assistant response, the simulator uses its content and vocal delivery, together with the private seed and dialogue history, to update the user's state and select the next reaction. Disclosure, correction, and changes in affect thus depend on the actual exchange. Interaction ends upon need resolution, disengagement, impasse, or reaching the turn limit; disengagement is not counted as success.

Simulator quality control. Quality control separates plan validation from output review. Structural checks enforce the output schema and state-transition contract; subsequent review assesses contextual consistency, stopping validity, and the fidelity and delivery of synthesized user speech. Failed outputs are regenerated from the same interaction state. These checks assess simulator integrity, not assistant performance, and do not determine the assistant's score. Appendix B specifies the protocol and information boundaries.

3.3 Trajectory-Level Evaluation

DynaEmpathy adopts a trajectory-level evaluation protocol. To preserve modality-specific evidence, acoustic observation is separated from dialogue judgment: an audio-only observer describes each assistant waveform without access to its transcript or dialogue context, and a trajectory judge subsequently integrates these observations with the complete transcript, private evaluator context, and delivered user-speech conditions. Simulator self-assessments, stop decisions, and user gratitude are treated as context rather than evidence of assistant success.

The judge independently assesses three dimensions. **Need resolution** measures whether the assistant correctly identifies and meaningfully advances the practical goal or emotional need revealed over the interaction, rather than substituting generic reassurance or premature closure. **Adaptive empathy** measures whether its conversational strategy tracks changes in the user's semantic and vocal state, respects the stage of disclosure, and adjusts after clarification, correction, resistance, or pushback. **Vocal appropriateness** measures whether the realized tone, intensity, pacing, and other prosodic cues remain congruent with both the response content and the user's current state, rather than merely sounding more expressive. The pass criterion is conjunctive rather than averaged: every dimension must meet its prespecified threshold. The protocol assesses both timely attunement and subsequent adaptation: later recovery is credited, but does not erase a missed salient opening cue. Appendix C specifies the evaluator models, score anchors, decision rules, and acoustic diagnostics, while Appendix G provides the core model-visible instructions.

4 Static-State GRPO for Dynamic Empathy

To strengthen a model's dynamic empathy across turns, the most direct approach is to optimize it through response-contingent interaction. Closed-loop speech rollouts, however, are costly, and reliable trajectory-level rewards are difficult to specify. We therefore learn from static decision states distilled from offline interactions. This makes training tractable but leaves a structured credit-assignment problem: a spoken response couples text and speech, while its text must balance helpfulness, empathic attunement, factual reliability, and dialogue-stage strategy. We thus use

responsibility-routed, modality-decomposed GRPO to select the semantic objective appropriate to each state and route semantic and acoustic feedback to the corresponding action streams.

4.1 Offline Interaction Distillation and Cold Start

We construct static training seeds through the benchmark's generation pipeline while keeping the training data disjoint from the final test set. Two LLMs expand these seeds into offline multi-turn text trajectories. The user LLM conditions each new turn on its private scenario, persona, and need together with the realized dialogue, while the assistant LLM observes only the public interaction. The resulting trajectories instantiate diverse situations that arise dynamically, including gradual disclosure, clarification, user correction, resistance, changing constraints, and resolution.

We then use a TTS model to synthesize the text trajectories into speech and split them at assistant turns into static, history-conditioned training states. One disjoint subset provides text–speech targets for an SFT cold start; the other provides states from which the initialized policy samples grouped responses for GRPO. Thus, although the user is not rolled out again during optimization, the static training data exposes the policy to different stages and outcomes of response-contingent interaction. Full construction and split details are given in Appendix E.

4.2 Responsibility-Routed Semantic Rewards

Every response is first assessed against shared requirements: it should be grounded in the public conversation, factually reliable, safe, natural to say aloud, responsive to the user's words and vocal cues. Beyond these common requirements, the response's primary responsibility varies with the setting and dialogue position. An undifferentiated text score can obscure these differences and make credit assignment less precise. We therefore define a training-time router

$$c = \rho(s) \in \{\text{PS-O}, \text{PS-F}, \text{ED-O}, \text{ED-F}\}, \tag{1}$$

where s is the fixed state, PS/ED denotes the interaction setting, and O/F denotes opening or follow-up. The selected cell adds a stage-specific responsibility:

- **PS-O:** acknowledge salient affect and offer a reliable first step toward the practical goal.
- **PS-F:** incorporate feedback, constraints, or corrections and advance the task accurately.
- **ED-O:** respond to the emotion with listening, clarification, or help suited to the user's need.
- **ED-F:** adapt to new disclosures or boundaries with a useful next move or natural closure.

The interaction setting and turn position select only the reward rubric and are not shown to the policy. A text judge independently scores each candidate against the shared requirements and selected responsibility. Let $r_i^{\text{text}} \in [0, 1]$ denote the normalized score; its group-relative advantage is

$$A_i^{\text{text}} = \frac{r_i^{\text{text}} - \overline{r}^{\text{text}}}{\text{sd}(r^{\text{text}}) + \epsilon}. \tag{2}$$

4.3 Modality-Decomposed Policy Optimization

We instantiate this approach with Fun-Audio-Chat-8B (Chen et al., 2025), whose text and speech generation paths support modality-specific parameter updates.

Acoustic rewards. Semantic reward evaluates what is said but not how it is rendered. We therefore construct a separate acoustic advantage from lexical fidelity and delivery alignment. The former compares an ASR transcript of the speech response with the simultaneously generated text, $r_i^{\text{wer}} = 1 - \min(\text{WER}_i, 1)$; the latter uses a speech judge to assess audible naturalness and contextual fit without rescoring helpfulness. Their group-relative advantages are combined as

$$A_i^{\text{speech}} = \lambda_{\text{wer}} \widetilde{A}_i^{\text{wer}} + \lambda_{\text{style}}(t) \widetilde{A}_i^{\text{style}}. \tag{3}$$

Modality routing. We optimize the text and speech generation paths with separate clipped GRPO objectives, each using the corresponding modality-specific advantage. Semantic feedback updates the text policy, whereas acoustic feedback updates speech generation without propagating into the text policy. This separation aligns each reward with the component responsible for the behavior it evaluates. Implementation details are provided in Appendix E.

The complete objective is

$$\mathcal{L}_{\mathrm{GRPO}} = \mathcal{L}_{\mathrm{text}} + \mathcal{L}_{\mathrm{speech}}. \tag{4}$$

This is a static-state, contextual-bandit form of GRPO (Shao et al., 2024): the policy is trained to adapt its next response to a multi-turn state, but the user simulator is not rolled out during RL.

5 Experiments

5.1 Experimental Setup

Evaluated models. Our evaluation covers closed-source SpeechLMs, open-source SpeechLMs, and variants trained with our method. The closed-source models are GPT-Audio (OpenAI, 2025) and Seeduplex (ByteDance Seed, 2026); the open-source models include MiniCPM-o-4.5 (Cui et al., 2026), Kimi-Audio (Kimi Team et al., 2025), Qwen3-Omni (Xu et al., 2025), OpenS2S (Wang et al., 2025a), MiMo-Audio (Xiaomi LLM-Core Team et al., 2025), and Fun-Audio-Chat-8B (Chen et al., 2025). To assess our training method, we compare the released Fun-Audio-Chat-8B model with its SFT and GRPO variants. All SpeechLMs receive semantically matched empathy instructions, without private user information or evaluation criteria (Appendix F); training details appear in Appendix E. An annotation-assisted topline, *Teacher LLM + TTS*, combines Gemini 3.5 Flash (Google, 2026c) with Gemini TTS (Google, 2026f). It receives public reference transcripts and delivery descriptions but no private user information, using the same interaction and scoring protocol. This privileged-input reference is not a strict upper bound (Appendix C.1).

External empathy benchmarks. We additionally report VStyle (Zhan et al., 2026) and EchoMind (Zhou et al., 2026) for two purposes: to contrast models' single-turn empathy with their dynamic multi-turn performance, and to test whether our training method transfers beyond DynaEmpathy. VStyle evaluates 140 English responses across anger, sadness/disappointment, anxiety/fear, and joy/excitement using its released multimodal empathy rubric; we report the four category scores and their unweighted mean. EchoMind uses 235 content groups, each rendered with target, neutral, and alternative vocal emotion. We report its four GPT-judged response criteria—context fit, response naturalness, colloquialism, and speech-information relevance—together with the Vocal Empathy Score (VES), all on a 1–5 scale (Appendix C.2).

5.2 Main Results

Table 1 summarizes multi-turn performance on DynaEmpathy and response-level performance on external spoken-empathy benchmarks.

Limited dynamic empathy in existing models. Strong single-turn scores do not ensure successful dynamic interaction. GPT-Audio reaches 4.96 on VStyle and the highest EchoMind VES among evaluated SpeechLMs, yet passes only 47.2% of DynaEmpathy trajectories. Fun-Audio-Chat-8B likewise scores 4.63 on VStyle but passes only 15.0%. Conversely, MiMo-Audio trails MiniCPM-o-4.5 on VStyle and VES but has a higher dynamic pass rate (30.0% versus 11.0%). These contrasts support complementary evaluation: an appropriate isolated reply does not establish sustained adaptation to evolving user information and feedback.

Implicit-affect problem solving remains challenging. Under our protocol, every model passes fewer PS than ED trajectories, notably Seeduplex (3.3% versus 81.7%) and Qwen3-Omni (21.7% versus 69.3%). Seeduplex's PS need resolution reaches 3.94, but adaptive empathy is only 2.78; vocal appropriateness also exceeds adaptive empathy for every baseline overall. Conjunctive evaluation exposes these imbalances without allowing delivery to compensate for strategy. The setting gap persists after training and reflects the joint PS requirements of opening attunement and subsequent adaptation, rather than follow-up strategy alone.

Table 1: Main results on DynaEmpathy, VStyle, and EchoMind. Bold indicates the best results excluding the topline.

(a) DynaEmpathy: dynamic multi-turn interaction

Model	Problem Solving				Emotion Disclosure				Overall			
	Need	Adapt.	Vocal	Pass (%)	Need	Adapt.	Vocal	Pass (%)	Need	Adapt.	Vocal	Pass (%)
Closed-source models												
GPT-Audio	3.83	3.28	**4.35**	27.3	3.94	3.73	4.91	67.0	3.89	3.51	**4.63**	47.2
Seeduplex	3.94	2.78	3.85	3.3	4.24	4.22	4.87	81.7	4.09	3.50	4.36	42.5
Open-source models												
MiniCPM-o-4.5	3.70	2.92	3.78	3.7	3.04	2.77	4.34	18.4	3.37	2.84	4.06	11.0
Kimi-Audio	3.13	2.81	3.77	1.0	2.89	2.66	4.06	11.0	3.01	2.73	3.92	6.0
Qwen3-Omni	**3.99**	3.25	3.76	21.7	4.02	3.80	4.81	69.3	4.00	3.53	4.29	45.5
OpenS2S	3.64	2.96	3.68	5.0	3.20	2.96	4.07	20.7	3.42	2.96	3.88	12.9
MiMo-Audio	3.69	3.12	3.35	13.0	3.85	3.53	4.14	47.0	3.77	3.33	3.75	30.0
Fun-Audio-Chat-8B	3.97	3.01	3.70	6.0	3.37	3.07	4.39	24.0	3.67	3.04	4.04	15.0
Ours												
Fun-Audio-Chat-8B + SFT	3.73	3.19	3.82	19.3	3.98	3.77	4.74	63.3	3.86	3.48	4.28	41.3
Fun-Audio-Chat-8B + GRPO	3.96	**3.54**	4.14	**50.0**	**4.68**	**4.78**	**4.96**	**97.3**	**4.32**	**4.16**	4.55	**73.7**
Topline												
Teacher LLM + TTS	4.39	3.75	4.36	71.7	4.35	4.62	4.94	97.0	4.37	4.18	4.65	84.3

(b) VStyle and EchoMind: single-turn empathy evaluation

Model	VStyle					EchoMind				
	Anger	Sad.	Anx.	Joy	Mean	C_{CtxFit}	C_{RespNat}	$C_{\text{ColloqDeg}}$	$C_{\text{SpeechRel}}$	VES
Closed-source models										
GPT-Audio	**5.00**	**4.90**	**4.97**	4.97	**4.96**	4.66	4.86	**4.75**	**4.57**	**3.95**
Seeduplex	4.93	4.80	4.77	**5.00**	4.88	**4.79**	4.84	4.68	4.49	3.82
Open-source models										
MiniCPM-o-4.5	4.32	4.53	4.23	3.89	4.24	4.08	4.21	4.08	3.76	3.71
Kimi-Audio	3.86	4.10	3.90	2.51	3.60	3.20	3.77	3.82	3.39	2.98
Qwen3-Omni	4.66	4.57	4.65	4.89	4.69	4.74	**4.88**	4.72	4.46	3.47
OpenS2S	3.82	3.97	3.68	3.97	3.86	4.21	4.50	4.28	4.07	3.11
MiMo-Audio	3.16	3.47	3.39	2.69	3.17	4.36	4.55	4.41	4.31	2.99
Fun-Audio-Chat-8B	4.57	4.77	4.48	4.71	4.63	4.57	4.75	4.69	4.33	3.46
Ours										
Fun-Audio-Chat-8B + SFT	4.75	4.63	4.68	4.91	4.74	4.53	4.79	4.68	4.24	3.48
Fun-Audio-Chat-8B + GRPO	4.80	4.77	4.81	4.89	4.81	4.69	4.86	4.70	4.54	3.80
Topline										
Teacher LLM + TTS	5.00	5.00	5.00	5.00	5.00	4.86	4.91	4.75	4.67	4.32

Improvements from GRPO training. SFT raises the DynaEmpathy pass rate from 15.0% to 41.3%, and GRPO further increases it to 73.7%: 32.4 percentage points above SFT and 26.5 above the strongest evaluated baseline, GPT-Audio. Relative to SFT, GRPO improves overall need resolution, adaptive empathy, and vocal appropriateness by 0.46, 0.68, and 0.27, respectively, with the largest gain in adaptive empathy. Pass rates reach 50.0% in PS and 97.3% in ED, while PS need resolution remains comparable to the base (3.96 versus 3.97). These gains support the complete training procedure, not the independent effect of each reward component.

Transfer and comparison with the topline. GRPO improves VStyle from 4.63 to 4.81 and all five EchoMind metrics over both the base and SFT, including VES from 3.46 to 3.80 over the base. The broader gains follow GRPO rather than SFT alone. Aggregate general S2T performance is retained, with an S2S trade-off (Appendix D.1). GRPO trails the topline in overall DynaEmpathy pass rate (73.7% versus 84.3%) and PS (50.0% versus 71.7%), but approximately matches it in ED (97.3% versus 97.0%). The topline receives privileged transcripts and delivery annotations, so this is not an input-matched comparison.

5.3 Analysis and Discussion

5.3.1 Human–Automatic Evaluation Agreement

Following rubric-based validation in ParaS2S (Yang et al., 2026), we distribute 360 trajectories from six systems and 60 shared seeds evenly among 10 evaluators, with three independent ratings per trajectory. Evaluators use dialogue audio, transcripts, and the automatic judge's score anchors; model identities and automatic judgments are hidden (Appendix C.6). Table 2 shows strong correspondence with mean human scores: Spearman's ρ ranges from 0.784 to 0.807, with MAE below 0.44

on the five-point scale. Inter-annotator agreement is higher for adaptive empathy and vocal appropriateness than for need resolution ($\alpha = 0.707$ and 0.666, versus 0.487). These results support the judge's aggregate assessments without implying interchangeable human and automatic scores or uniform reliability across dimensions.

Table 2: Human–automatic trajectory-score agreement.

Dimension	Spearman's $\rho \uparrow$	MAE $\downarrow$	Human $\alpha \uparrow$
Need resolution	0.807	0.372	0.487
Adaptive empathy	0.796	0.431	0.707
Vocal appropriateness	0.784	0.326	0.666

5.3.2 Reliability of User Simulation

Behavioral plausibility. Each of 180 trajectories receives three independent human audits across five criteria (Table 3). A serious issue materially changes the interaction's meaning, difficulty, or outcome and requires agreement from at least two annotators on the same criterion. Four trajectories (2.22%) have at least one such issue, with equal rates in both settings. Termination validity has the highest criterion-level rate (1.11%); no majority-confirmed serious issues occur in response contingency or speech realization. However, low agreement on several criteria ($\alpha = 0.022$–0.182) cautions against equating low majority rates with absence of errors. The audit supports plausibility within the sampled conditions, not fidelity to real users (Dou et al., 2025). Appendix C.7 provides complementary model-based auditing, kept separate from these human results.

Table 3: Human validation of user simulation. Serious-issue rates (%) are lower-is-better; human agreement α is higher-is-better.

Criterion	Serious-issue rate (%)			Human α
	PS	ED	Overall	
Scenario/persona consistency	0.00	1.11	0.56	0.128
Response contingency	0.00	0.00	0.00	0.022
State/disclosure coherence	1.11	0.00	0.56	0.076
Termination validity	1.11	1.11	1.11	0.182
Speech realization	0.00	0.00	0.00	0.425
At least one serious issue	2.22	2.22	2.22	0.202

End-to-end evaluation repeatability. Three complete rollouts on 60 shared seeds across Teacher LLM + TTS, Seeduplex, and MiniCPM-o-4.5 yield overall score SDs of 0.010–0.129 and pass-rate SDs of 1.0–2.9 percentage points. Their overall ordering is unchanged, while 13.3–16.7% of seeds change pass/fail status and some setting-specific rankings reverse. These results characterize aggregate repeatability of the full assistant–simulator–TTS–judge pipeline, not simulator variability in isolation (Appendix C.7).

6 Conclusion

We introduce DynaEmpathy, a 600-condition benchmark combining private user conditions, speech-aware interaction, and trajectory-level evaluation. Results show that strong single-turn empathy does not ensure effective dynamic interaction. Across practical assistance and emotional disclosure, effective responses must integrate evolving user feedback with both conversational strategy and vocal delivery. Human score agreement, behavioral audits, and repeated rollouts provide complementary validation. Our responsibility-routed, modality-decomposed GRPO learns from static interaction states, raising the base model's trajectory pass rate from 15.0% to 73.7% and improving out-of-domain empathy while retaining aggregate S2T capability. These contributions connect interactive evaluation with training for dynamic empathy; broader human-interaction validation and general S2S capability retention remain important directions.

AI USE STATEMENT

We used generative AI tools to assist manuscript writing and revision, including language polishing and improvements to clarity and organization. We also used language models and text-to-speech systems to construct synthetic benchmark and training data, including scenario-grounded user conditions, dialogue utterances, teacher responses, vocal-delivery instructions, and synthesized speech. The models, generation procedures, and quality-control mechanisms are described in Sections 3 and 4 and Appendices A, B, and E; core prompts are provided in Appendix G. Language models also support quality review, training rewards, and automatic evaluation, as described in Appendices C and E. The authors take responsibility for the final manuscript, the reported results, and all claims made in the paper.

REFERENCES

Anthropic. Introducing Claude Opus 4.8. Official model announcement, 2026. URL https://www.anthropic.com/news/claude-opus-4-8. Accessed 2026-09-26.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56898–56918, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/681fe4ec554beabdc9c84a1780cd5a8a-Abstract-Datasets_and_Benchmarks_Track.html.

Tara Bogavelli, Gabrielle Gauthier Melançon, Katrina Stankiewicz, Oluwanifemi Bamgbose, Fanny Riols, Hoang H. Nguyen, Raghav Mehndiratta, Lindsay Devon Brin, Joseph Marinier, Hari Subramani, Anil Madamala, Sridhar Krishna Nemala, and Srinivas Sunkara. EVA-bench: A new end-to-end framework for evaluating voice agents. *arXiv preprint arXiv:2605.13841*, 2026. URL https://arxiv.org/abs/2605.13841.

ByteDance Seed. Introducing Seed full-duplex speech LLM: Attentive listening, robust interference suppression, enabling more natural interaction. Official technical blog, April 2026. URL https://seed.bytedance.com/en/blog/introducing-seed-full-duplex-speech-llm-attentive-listening-robust-interference-suppression-enabling-more-natural-interaction. Published April 9, 2026.

Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. Fun-Audio-Chat technical report. *arXiv preprint arXiv:2512.20156*, 2025. URL https://arxiv.org/abs/2512.20156.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628. URL https://aclanthology.org/2026.tacl-1.18/.

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. Is MultiWOZ a solved task? an interactive TOD evaluation framework with user simulator. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1248–1259. Association for Computational Linguistics, 2022a. doi: 10.18653/v1/2022.findings-emnlp.90. URL https://aclanthology.org/2022.findings-emnlp.90/.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3014–3026. Association for Computational Linguistics, 2022b. doi: 10.18653/v1/2022.emnlp-main.195. URL https://aclanthology.org/2022.emnlp-main.195/.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang,

Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026. URL https://arxiv.org/abs/2604.27393.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13943–13970. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.682. URL https://aclanthology.org/2025.acl-long.682/.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. URL https://arxiv.org/abs/2410.00037.

Yayue Deng, Guoqiang Hu, Haiyang Sun, Xiangyu Zhang, Haoyang Zhang, Fei Tian, Xuerui Yang, Gang Yu, and Eng Siong Chng. MULTI-bench: A multi-turn interactive benchmark for assessing emotional intelligence ability of spoken dialogue models. *arXiv preprint arXiv:2511.00850*, 2025. URL https://arxiv.org/abs/2511.00850.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. SimulatorArena: Are user simulators reliable proxies for multi-turn evaluation of AI assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35212–35290. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1786. URL https://aclanthology.org/2025.emnlp-main.1786/.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/90d1fc07f46e31387978b88e7e057a31-Abstract-Conference.html.

Zhuoyue Gao, Xiaohui Wang, Xiaocui Yang, Wen Zhang, Daling Wang, Shi Feng, and Yifei Zhang. ES4R: Speech encoding based on prepositive affective modeling for empathetic response generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36830–36847. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1708. URL https://aclanthology.org/2026.acl-long.1708/.

Xuelong Geng, Qijie Shao, Hongfei Xue, Shuiyuan Wang, Hanke Xie, Zhao Guo, Yi Zhao, Guojian Li, Wenjie Tian, Chengyou Wang, Zhixian Zhao, Kangxiang Xia, Ziyu Zhang, Zhennan Lin, Tianlun Zuo, Mingchen Shao, Yuang Cao, Guobin Ma, Longhao Li, Yuhang Dai, Dehui Gao, Dake Guo, and Lei Xie. OSUM-EChat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *arXiv preprint arXiv:2508.09600*, 2025. URL https://arxiv.org/abs/2508.09600.

Google. Gemini 3.1 Flash-Lite. Google AI for Developers model documentation, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite. Accessed 2026-09-04.

Google. Gemini 3.1 Pro Preview. Google AI for Developers model documentation, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Accessed 2026-09-26.

Google. Gemini 3.5 Flash. Google AI for Developers model documentation, 2026c. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash. Accessed 2026-09-04.

Google. Gemini 3.6 Flash. Google AI for Developers model documentation, 2026d. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.6-flash. Accessed 2026-09-04.

Google. Gemini 3.7 Flash. Google AI for Developers model documentation, 2026e. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash. Accessed 2026-09-26.

Google. Text-to-speech generation with the Gemini API. Google AI for Developers documentation, 2026f. URL https://ai.google.dev/gemini-api/docs/generate-content/speech-generation. Updated 2026-07-21; accessed 2026-08-13.

Yun Hong, Yan Zhou, and Yang Feng. FreezeEmpath: Efficient training for empathetic spoken chatbots with frozen LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17141–17157. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.846. URL https://aclanthology.org/2026.findings-acl.846/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kimi Team et al. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025. URL https://arxiv.org/abs/2504.18425.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. Be helpful but don't talk too much: Enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1976–1988. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.118. URL https://aclanthology.org/2024.emnlp-main.118/.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 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)*, pp. 3469–3483. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Xuan Luo, Lewei Yao, Lanqing Hong, Kai Chen, Dehua Tao, Daxin Tan, Yukun Deng, Ruifeng Xu, and Jing Li. AEQ-bench: Measuring empathy of omni-modal large models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36390–36407. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1813. URL https://aclanthology.org/2026.findings-acl.1813/.

OpenAI. Whisper large-v3. Official release announcement in the OpenAI Whisper repository, November 2023. URL https://github.com/openai/whisper/discussions/1762. Included in openai-whisper version 20231106 and later.

OpenAI. GPT-4o system card. Technical report, OpenAI, 2024. URL https://cdn.openai.com/gpt-4o-system-card.pdf.

OpenAI. GPT-Audio model. Official API documentation, 2025. URL https://developers.openai.com/api/docs/models/gpt-audio. Accessed 2026-08-30.

OpenAI. GPT-5.4 model documentation. OpenAI API model documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.4. Accessed 2026-09-04.

OpenAI. GPT-5.5 model documentation. OpenAI API model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.5. Accessed 2026-09-04.

Qwen Team. Qwen3.6-35B-A3B. Official model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.6-35B-A3B. Accessed 2026-09-26.

Qwen Team. Qwen3-TTS: Expressive and streaming speech generation. Official software repository, 2026b. URL https://github.com/QwenLM/Qwen3-TTS. Accessed 2026-08-13.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

Soham Ray, Keshav Dhandhania, Victor Barres, and Karthik R. Narasimhan. τ-voice: Benchmarking full-duplex voice agents on real-world domains. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2Oj6fg0m1j.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Qundong Shi, Jie Zhou, Biyuan Lin, Junbo Cui, Guoyang Zeng, Yixuan Zhou, Ziyang Wang, Xin Liu, Zhen Luo, Yudong Wang, and Zhiyuan Liu. UltraEval-Audio: A unified framework for comprehensive evaluation of audio foundation models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 566–577. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-demo.56. URL https://aclanthology.org/2026.acl-demo.56/.

Jeonghoon Shim, Woojung Song, Cheyon Jin, Seungwon Kook, and Yohan Jo. Non-collaborative user simulators for tool agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UAUimofy3W.

Suhyune Son, Jungwoo Lim, Myunghoon Kang, Seongtae Hong, Yuna Hur, Evelyn H. Zi, and Heuiseok Lim. I don't need solution. I need emotional support: Empathetic LLMs based on emotional validation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1–26. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1. URL https://aclanthology.org/2026.findings-acl.1/.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. BLSP-emo: Towards empathetic large speech-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19186–19199. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1070. URL https://aclanthology.org/2024.emnlp-main.1070/.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and Jiajun Zhang. OpenS2S: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 906–917. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-demos.70. URL https://aclanthology.org/2025.emnlp-demos.70/.

Shuiyuan Wang, Zhixian Zhao, Hongfei Xue, Chengyou Wang, Shuai Wang, Hui Bu, Xin Xu, and Lei Xie. HumDial-EIBench: A human-recorded multi-turn emotional intelligence benchmark for audio language models. *arXiv preprint arXiv:2604.11594*, 2026. URL https://arxiv.org/abs/2604.11594.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63345–63354. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25aw.html.

Liang-Yuan Wu, Zih-Ching Chen, Tongshuang Wu, Chao-Han Huck Yang, and Hua Shen. SpeechEQ: Benchmarking emotional intelligence quotient in socially aware voice conversational models. *arXiv preprint arXiv:2606.25990*, 2026. URL https://arxiv.org/abs/2606.25990.

Xiaomi LLM-Core Team et al. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025. URL https://arxiv.org/abs/2512.23808.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. URO-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17211–17242. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.933. URL https://aclanthology.org/2025.findings-emnlp.933/.

Shu-wen Yang, Ming Tu, Ting-Wei Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. ParaS2S: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CcmDDh070o.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, Shimin Li, Jun Song, Xipeng Qiu, and Bo Zheng. VStyle: A benchmark for voice style adaptation with spoken instructions. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026. URL https://cmsworkshops.com/ICASSP2026/view_paper.php?PaperNum=3716.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. EchoMind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=l5re5ppqrX.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

A Benchmark Construction and Documentation

A.1 Comparison with Prior Spoken-Empathy Benchmarks

Table 4 separates fixed multi-turn histories from response-contingent interaction. Output evaluation distinguishes assessment of response content (*Text*, including transcribed speech) from assessment of realized acoustic delivery (*Speech*), rather than merely recording a model's input–output format. Under user simulation, *Private profile* denotes scenario, persona, or goal information withheld from the tested model; *Audio feedback* denotes conditioning subsequent user behavior on the assistant waveform; and *Runtime checks* denotes validation of generated user turns, separate from dataset curation or assistant scoring.

Table 4: Dialogue settings, assessed output modalities, and user-simulation features of spoken-empathy benchmarks.

Benchmark	Dialogue setting		Output evaluation		User simulation		
	Multi-turn	Interactive	Text	Speech	Private profile	Audio feedback	Runtime checks
SpeechAlpaca	✗	✗	✓	✗	–	–	–
SD-Eval	✗	✗	✓	✗	–	–	–
VStyle	✗	✗	✓	✓	–	–	–
URO-Bench	✓	✗	✓	✓	–	–	–
ParaS2SBench	✗	✗	✓	✓	–	–	–
EchoMind	✗	✗	✓	✓	–	–	–
AEQ-Bench	✗	✗	✓	✓	–	–	–
HumDial-EIBench	✓	✗	✓	✓	–	–	–
MULTI-Bench	✓	✓	✓	✓	✓	✗	✗
DynaEmpathy	✓	✓	✓	✓	✓	✓	✓

✓/✗: documented/not documented in the protocol; –: not applicable without a simulated user. Capabilities may belong to different subsets within a benchmark. DynaEmpathy's runtime checks refer to deterministic turn-contract gates.

MULTI-Bench (Deng et al., 2025) supports private user profiles and interactive LLM–TTS dialogue; waveform feedback to its user policy and runtime simulator checks are not specified. DynaEmpathy combines these mechanisms with separate acoustic observation and contextual trajectory judgment. The table compares documented protocols, not whether prior benchmarks consider user needs.

A.2 Test-Set Coverage and Interaction Design

Table 5 documents the semantic coverage, private targets, and interaction design of the two 300-condition test settings.

A.3 Runtime Seed Fields

Table 6 specifies the private fields used to initialize each setting and the public opening that starts the interaction.

A.4 Setting-Specific Constraints

Implicit-affect problem solving. The private scenario contains a concrete, answerable practical need and the cause of the target affect. The opening text states the immediate request but does not name either the emotion or its cause; the affective contrast is carried primarily by the synthesized voice. The latent goal supports later disclosure of constraints, failed attempts, comparisons, or communication needs. The expected adaptation may affect wording, pacing, prioritization, reassurance, or the order of practical steps, but cannot replace substantive task progress.

Emotion disclosure. The private scenario records the event, stakes, known and uncertain facts, and the user's core support need. The opening is locally understandable and emotionally legible

Table 5: Semantic coverage and interaction design of the two DynaEmpathy test settings.

Problem solving (300 conditions)	
Basic emotions	Anger, disgust, fear, happiness, sadness, and surprise.
Broad topics	Culture, Customs & Celebrations; Education & Learning; Finance & Consumer Decisions; Food, Cooking & Dining; Health, Fitness & Sleep; Home & Personal Administration; Leisure, Media & Hobbies; Memories & Personal Archives; News & Information Literacy; Relationships & Social Life; Style & Personal Presentation; Technology & Digital Life; Travel & Mobility; and Work & Career.
Private target	A complete multi-stage practical objective whose later needs depend on the initial result.
User behavior	Volunteering or withholding details, challenging, correcting, and accepting.
Persona	A stable, scenario-specific decision or communication tendency that governs those behaviors.
Opening design	A textually neutral practical request whose affect is conveyed primarily through speech.
Emotion disclosure (300 conditions)	
Broad topics	Community, Culture, and Social Identity; Creativity, Hobbies, and Personal Pursuits; Digital Life and Public Reputation; Education and Learning; Everyday Life and Public Interactions; Family, Parenting, and Caregiving; Friendship and Social Connection; Health, Body, and Aging; Life Transitions and Future Direction; Loss, Separation, and Grief; Money, Housing, and Material Security; Romantic and Intimate Relationships; Self-Concept, Identity, and Values; and Work and Career.
Primary needs	Ambivalence sensemaking; connection and belonging; distress containment; emotional release; grievance acknowledgment; loss witnessing; positive-emotion sharing; reaction validation; self-worth repair; and uncertainty reassurance.
User behavior	Disclosure, resistance, revisiting, challenge, and acceptance.
Persona	A stable, scenario-specific tendency that governs what emerges and when.
Opening design	A locally intelligible but partial disclosure that leaves deeper context and the primary emotional need to emerge across turns.

Table 6: Runtime seed fields. Private fields initialize the simulator and evaluator and are never exposed directly to the target model.

Setting	Field	Purpose	Construction constraint
Problem solving	`scenario.context`	Complete situation and affect cause	Stable and answerable without hidden tools
	`scenario.user_goal`	Immediate and later practical needs	Supports substantive multi-turn development
	`persona`	Interaction tendency	Changes disclosure, challenge, or acceptance
	`description`	Basic/specific emotion and Voice Style	Grounded and feasible for TTS
	`initial_query`	Spoken opening text	Clear request without naming emotion/cause
Emotion disclosure	`private_scenario`	Event, stakes, and known/uncertain facts	Coherent basis for later disclosure
	`persona`	Communication/interpretation tendency	Affects resistance and disclosure
	`emotional_need`	Primary need and description	Priority, not a reference response
	`speaker`	TTS compatibility metadata	Synthesis metadata, not a target label
	`initial_utterance`	Text, emotion, and Voice Style	Natural, incomplete, and understandable

but withholds deeper history, stakes, or support preferences for later disclosure. The persona describes a consequential interaction tendency that changes what the user volunteers, resists, clarifies, or accepts. It does not prescribe a disclosure schedule, preferred assistant response, or termination path. Appropriate strategies may include acknowledgment, exploration, validation, sensemaking, or positive-affect sharing (Liu et al., 2021).

Table 7: Examples of first-turn queries, one per controlled category.

Category	Example first-turn query
Setting I: Implicit-affect problem solving	
Angry	What are the most effective methods to block light leaking through door gaps? *Speak angrily.*
Disgusted	I have several parmesan rinds available. How do I properly clean and prepare these for use in a soup? *Speak with disgust.*
Fearful	How can I fold a silk scarf so it does not get wrinkled in a small bag? *Speak anxiously.*
Happy	I am baking multiple trays of cookies for a sale. What is the most efficient way to sequence the oven usage? *Speak happily.*
Sad	What are the steps to separate my watch history from a shared account? *Speak sadly.*
Surprised	Could you help me assess the potential cultural risks of giving food items to an international client? *Speak with surprise.*
Setting II: Emotion disclosure	
Ambivalence sensemaking	I'm so glad she got her promotion, but it feels like she doesn't really need my input anymore. *Speak with uncertainty.*
Connection and belonging	I was looking through the reunion photos today and realized I barely show up in any of them. *Speak with hurt.*
Distress containment	Everyone is so thrilled about the wedding, but I feel like I'm just pretending to be ready for all of this. *Speak with anxiety.*
Emotional release	I finally got that promotion, but all I heard back was a question about what is for dinner. *Speak with disappointment.*
Grievance acknowledgment	A stranger just reached over my table to adjust my chair while I was working. *Speak with frustration.*
Loss witnessing	I found this old birthday card I never sent to my brother while cleaning out my desk today. *Speak with sadness.*
Positive-emotion sharing	The shopkeeper used my name today, which feels a little silly to even mention. *Speak softly.*
Reaction validation	I walked past the new plaque today and saw they just took my name off the wall. *Speak with hurt.*
Self-worth repair	The client completely brushed off my advice, and it stung even more that my peer just sat there in silence. *Speak with hurt.*
Uncertainty reassurance	I am still waiting on my thesis defense date, and this silence is starting to feel like a trap. *Speak with frustration.*

A.5 First-Turn Query Examples

Table 7 gives one verified first-turn query for each controlled category. The lighter italic line beneath each query reproduces the exact instruction used to synthesize its vocal delivery.

A.6 Hierarchical Planning

Separate topic inventories are expanded into subtopics and scenario families. A minimum generation unit pairs a scenario family with its setting-specific affect or need condition. Within each unit, planning differentiates the event, stakes, need, possible developments, and persona before realizing individual seeds. Diversity is defined by substantive differences in these attributes, rather than surface wording alone.

A.7 Generation and Review

GPT-5.5 (OpenAI, 2026b) is used for topic and unit planning; Gemini-3.1-Flash-Lite (Google, 2026a) instantiates eight seeds per active unit. Structural checks enforce field completeness and the separation between private context and the public opening.

The frozen GPT-5.5 (OpenAI, 2026b) benchmark reviewer applies setting-specific Boolean gates. Problem-solving gates cover practical answerability, multi-turn room, grounded implicit affect, persona relevance, and diversity. Emotion-disclosure gates cover scenario coherence, need alignment, natural opening, persona-conditioned development, safety appropriateness, and diversity. A failed gate rejects the candidate; a separate 1–5 quality score is used only to break coverage ties among passing candidates.

A.8 Opening-Audio Quality Control

Opening waveforms must satisfy normalized WER < 0.1 against their reference text using Whisper large-v3 (Radford et al., 2023; OpenAI, 2023).

The audio-review package then presents the opening transcript, intended TTS instruction, and waveform. Multiple human annotators assign *pass*, *unsure*, or *fail* labels. A pass requires both perceptual agreement with the intended style and absence of obvious synthesis defects such as electronic noise, severe artifacts, or unnatural prosody. The 656 items passing both gates contain 320 problem-solving and 336 emotion-disclosure conditions; final stratified selection reduces this to 300 per setting.

B Simulator and Runtime Details

B.1 Information Boundaries

Table 8: Information available to each component under the standard SpeechLM protocol. History presentation is adapter-specific; private state is never exposed to the target. The annotation-assisted topline is specified in Appendix C.1.

Information	Target	User simulator	User TTS	Audio extractor	Final judge
Private scenario/persona/need	No	Yes	No	No	Yes
Current user waveform	Yes	Own output	Produces	No	Via TTS condition
Public dialogue history	Yes	Yes	No	No	Yes
User delivery instruction	No	Yes	Yes	No	Yes
Current assistant waveform	Own output	Yes	No	Yes	Via extraction
Assistant text stream	Own output	Yes	No	No	Yes
Simulator status/stop reason	No	Yes	No	No	Diagnostic only

B.2 Simulator Turn Contract

At each assistant turn, the simulator (i) observes the assistant content and broad delivery, (ii) assesses the response from the private user's perspective, (iii) updates practical and emotional need status, and (iv) emits one natural next utterance, one short delivery instruction, and a stop decision. It is instructed to contribute at most one substantive new fact, constraint, reaction, or question; to challenge unsupported assumptions according to the persona; and not to mechanically advance a hidden checklist.

Successful stopping requires every applicable need to be met. Disengagement or impasse may also stop an interaction but cannot be mislabeled as success. The simulator's private status is not used as the assistant score. Admission and runtime checks assess interaction integrity separately from assistant quality.

B.3 Plan and Speech Checks

Plan validation enforces the structured-output schema, legal state values, need-status updates, and stopping constraints. These deterministic checks detect contract violations; they do not establish semantic correctness. Output review additionally considers the private condition, public history, generated user utterance, delivery instruction, and synthesized waveform. It assesses persona/scenario consistency, responsiveness to the preceding assistant turn, naturalness, and delivery suitability. Audio checks assess transcription fidelity, audible alignment with the planned delivery, synthesis quality, and speaker consistency where a speaker reference is available. Failed plans or speech outputs

are regenerated from the same dialogue state. Simulator checks remain separate from the assistant-only trajectory scores; the independent human and model-based audits in Appendix C.7 further assess residual simulation errors.

C Evaluation Details

C.1 Annotation-Assisted Topline

Observable input. The topline receives the realized public dialogue prefix, represented by utterance transcripts and delivery descriptions. For DynaEmpathy, user text is taken from the opening reference and subsequent simulator utterances; user delivery descriptions are the corresponding TTS instructions. These describe intended delivery, not independently verified properties of every waveform. The scenario type (problem solving or emotion disclosure) and assistant turn index select the prompt, and the selected responsibility is included in the teacher input. No private scenario narrative, persona, goal, simulator appraisal or state, or future turn is provided. This enriched interface reduces reliance on speech perception, but differences from SpeechLMs also involve model capacity and prompting, so the comparison does not isolate perception alone.

Teacher reuse and closed-loop evaluation. We reuse the offline text teacher, Gemini 3.5 Flash (Appendix E.3), with setting- and stage-specific prompts. Problem-solving prompts (v5) combine affect recognition with practical assistance; emotion-disclosure prompts (v4) emphasize need-sensitive support without premature advice. The teacher generates a reply and delivery instruction; Gemini 3.1 Flash TTS Preview (Google, 2026f) synthesizes the reply according to the delivery instruction. Evaluation retains the benchmark's user simulator, quality-control configuration, stopping rules, and judges, rather than reusing the offline user model or replaying generated trajectories. Actual assistant audio drives user reactions and acoustic evaluation; its intended style instruction is not substituted for observed delivery in scoring.

Interpretation and reporting. The topline is a strong annotation-assisted reference, not a guaranteed upper bound or an equal-input baseline. It uses the same teacher as training-data generation, with the benchmark seeds excluded from model training. The dynamic prompts were fixed using independent development samples before evaluation. For VStyle, the teacher receives the reference user text and a coarse delivery description derived from the emotion category. For EchoMind, it receives the user text and the current condition's supplied delivery annotation. Neither protocol supplies reference assistant answers or alternative-condition labels; these delivery annotations describe intended rather than measured acoustics.

C.2 External Empathy Evaluation

All systems use the same judges and released rubrics within each external benchmark. VStyle's English empathy subset is evaluated by Gemini 3.1 Pro Preview (Google, 2026b), which receives the user instruction and the user and assistant waveforms. Its overall score is the unweighted mean of the four emotion-category means. EchoMind's four text criteria are scored by GPT-4o-2024-11-20 (OpenAI, 2024), using the input text, topic, delivery annotation, and assistant response text. Gemini 2.5 Pro[1] assesses the assistant waveform for VES. Each metric is averaged over the 705 target-, neutral-, and alternative-emotion examples. Text criteria measure response content, whereas VES evaluates the realized vocal delivery; neither uses DynaEmpathy's private user state or trajectory-scoring rules.

C.3 Assistant-Only Trajectory Scores

The evaluator scores assistant behavior in admitted trajectories, separately from simulator validity. Poor assistant responses lower the scores rather than causing a trajectory to be discarded. Private information provides evaluation context but is not assumed to have been disclosed to the assistant; user gratitude, termination, and simulator self-assessments do not establish assistant success.

[1] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro

Each dimension is scored from 1 to 5, anchored as 1 (clear failure), 2 (major weakness), 3 (mixed or merely adequate), 4 (good with minor issues), and 5 (exceptional and consistently evidenced). A trajectory passes only when all three required scores are at least 4; a missing score makes the result incomplete rather than allowing that dimension to be ignored.

C.4 Temporal Accountability

For both settings, the judge labels opening attunement from the first user–assistant pair as *handled*, *missed*, or *not applicable*. In problem solving, a missed salient opening emotion deterministically caps adaptive empathy at 3, making timely opening attunement necessary for a trajectory to pass. This encodes a non-compensatory requirement: a successful trajectory must handle salient affect when it arises as well as adapt later. Later recovery still contributes to the assessment of subsequent behavior. PS pass rates therefore reflect both opening and follow-up requirements, rather than a measure of follow-up adaptation alone. Explicitly naming the emotion is not required: reduced cognitive load, calibrated pacing, focused clarification, or an affect-sensitive first action may constitute handling.

C.5 Acoustic Evaluation

Assistant delivery is assessed in two stages. Each assistant clip is sent alone to Gemini 3.5 Flash (Google, 2026c), which returns a concise, evidence-only style description. This reduces reliance on explicit transcript cues when describing audible delivery. The text-only GPT-5.4 judge (OpenAI, 2026a) then aligns these observations to assistant turn indices and evaluates their appropriateness relative to the preceding user text and TTS delivery condition. User style is not re-extracted: the exact instruction supplied to TTS is used as the intended acoustic condition, while the human opening-audio review separately verifies its realization for the first turn.

C.6 Human Evaluation Protocol

Sampling and annotation. The annotation package contains 360 trajectories from six systems on 60 shared seeds: Fun-Audio-Chat-8B, GRPO, Qwen3-Omni, MiniCPM-o-4.5, MiMo-Audio, and Seeduplex. We sample 30 seeds per setting, balancing basic emotions in PS and emotional needs in ED. The 60 seeds belong to distinct minimum generation units; selection is independent of scores, pass status, and dialogue length. Simulator auditing uses the 180 GRPO, Qwen3-Omni, and Seeduplex trajectories on the same seeds. All samples retain their original evaluation configurations.

Ten evaluators share the workload evenly within each task: 108 trajectory-score assignments and 54 simulator-audit assignments per evaluator. Each trajectory receives three independent ratings for each applicable task. Model identities, automatic scores and rationales, and simulator self-assessments are hidden. Annotators receive chronological audio, transcripts, and private scenario context; undisclosed facts are not treated as information available to the assistant. The two tasks use separate raters for any shared trajectory. Disagreement does not justify deleting an item; original ratings are retained even when a separate adjudication is needed. Technical failures are recorded separately rather than replaced by a more favorable rollout.

Assistant agreement metrics. For each dimension, Spearman's ρ and MAE compare automatic scores with the mean of the three human ratings. Ordinal Krippendorff's α measures agreement among individual annotators before adjudication. Results pool systems and settings; they do not establish model-specific agreement or a human-rated Base–SFT–GRPO progression, since SFT is absent from the sample.

Simulator audit criteria. Behavior-focused human review follows the distinction between challenging user behavior and simulation errors in prior user-simulator validation (Shim et al., 2026). Each criterion receives an ordinal severity label: 0 (no issue), 1 (minor issue without a material effect on the interaction), or 2 (serious issue affecting its interpretation, difficulty, or outcome). *Scenario/persona consistency* concerns contradictions of established facts, goals, or user characteristics. *Response contingency* concerns whether reactions follow the assistant's actual content and vocal delivery, rather than ungrounded satisfaction, hostility, or cooperation. *State/disclosure coherence* concerns motivated changes in affect and needs and contextually appropriate disclosure, including

unsupported facts that change the task. *Termination validity* distinguishes justified resolution, disengagement, and impasse; reaching the turn limit is not itself a simulation failure. *Speech realization* concerns fidelity to the user text and context-appropriate acoustic delivery, checked against the actual waveform rather than inferred from the TTS instruction alone. Evidence includes the offending user turn and its preceding context; plausible refusal, correction, or emotional change is not penalized merely for diverging from the assistant's preferred outcome.

Simulator audit metrics. A criterion has a majority-confirmed serious issue when at least two of three annotators assign severity 2. Criterion-level rates count affected trajectories; the overall rate counts each trajectory with at least one such criterion once. PS and ED use their respective trajectory counts as denominators. Ordinal α uses the original severity labels, including minor issues; the overall row uses each annotator's maximum severity across criteria. Undefined agreement is not treated as perfect agreement, and a zero majority rate does not imply zero individual flags or population risk.

C.7 Simulator Reliability and Repeatability

Complementary model-based audit. The same 180-trajectory audit subset has three model judgments per dialogue from Claude Opus 4.8 (Anthropic, 2026), Gemini 3.1 Pro (Google, 2026b), and GPT-5.5 (OpenAI, 2026b). A blind Gemini 3.1 Pro observer transcribes and describes 349 unique user waveforms without reference text, target emotion, or dialogue context. These observations and cached assistant acoustic descriptions are shared across judges, together with the transcript, private scenario, persona, objective, and neutral termination facts. Target-model identities, original scores, simulator self-assessments, and synthesis instructions are withheld. Behavioral checks concern generated user turns; speech checks also include the fixed opening and exclude terminal text-only turns. This is model-based review of shared acoustic evidence, not independent human listening.

Table 9 reports both majority-confirmed and any-judge serious-issue rates using the same severity labels as the human protocol. The latter flags 12/180 trajectories (6.67%; 95% seed-cluster bootstrap interval 3.33–11.11%), compared with 1/180 under majority voting. The low ordinal agreement cautions against treating the majority decision as ground truth; rare severe labels and disagreements remain relevant even when the majority rate is zero. Neither aggregation establishes fidelity to real human behavior or run-to-run stability.

Table 9: Model-based simulator audit on 180 trajectories. Serious-issue rates (%) use majority and any-judge aggregation; α is ordinal agreement among model judges, not humans.

Criterion	Majority	Any judge	Model α
Scenario/persona consistency	0.00	2.78	0.130
Response contingency	0.00	1.11	-0.019
State/disclosure coherence	0.00	0.56	0.064
Termination validity	0.56	2.78	0.166
Speech realization	0.00	1.11	0.325
Any serious issue	0.56	6.67	0.183

Repeated-rollout protocol. We regenerate complete interactions for Teacher LLM + TTS, Seeduplex, and MiniCPM-o-4.5 on the same 60 seeds used in the human-evaluation package (Appendix C.6): 30 PS seeds, with five per basic emotion, and 30 ED seeds, with three per emotional need. This is the previously fixed, availability-conditioned cohort, not a new sample chosen by repeated-run scores. Each model–seed pair has three rollouts, giving 540 trajectories. We reuse the first 180 interactions and their scores without modification and generate 360 new interactions from the identical private seed and opening waveform. Subsequent user utterances, audio, assistant responses, and stopping decisions are regenerated; completed transcripts are not simply re-scored.

Each system retains its original prompt and decoding settings. The teacher uses its PS v5 / ED v4 opening/follow-up instructions and privileged public transcripts and delivery annotations; Seeduplex retains its original empathy instruction, and MiniCPM-o-4.5 retains its 100-word response limit. The simulator, user TTS, eight-assistant-turn cap, stopping rules, acoustic observer, and GPT-5.4 trajec-

tory judge remain fixed. New trajectories use separate run identifiers and native model sessions; no continuation is shared between repetitions. Original configuration and opening-audio hashes are checked against the reused runs. Technical failures are logged and retried without score-based selection; unfinished Seeduplex sessions restart from the fixed opening because their native dialogue state cannot be restored. Such attempts are not counted as additional repetitions.

Metrics. For each setting and overall, we report the mean and sample standard deviation of the three run-level mean scores and pass rates (Table 10). These SDs describe variation between complete repetitions and are not confidence intervals. To avoid hiding compensating changes across seeds, we also compute each seed's sample SD over its three scores and average these SDs, and report the fraction of seeds whose pass/fail label changes across runs (Table 11). A trajectory passes only if all three dimension scores are at least four. Ranking checks compare the three system means between each pair of repetitions, separately for each setting and metric. Paired difference intervals use 2,000 seed-cluster bootstrap resamples, stratified by PS/ED and retaining all systems and repetitions for each selected seed. End-to-end variation includes assistant, simulator, TTS, and judge variability; this design does not isolate simulator noise or test fidelity to real users.

Table 10: Repeated-rollout results on 60 shared seeds (30 PS/30 ED). Entries are the mean $\pm$ sample SD of three run-level means; pass-rate SDs are percentage points, not confidence intervals. Each setting contains 30 trajectories per run and each overall row contains 60.

Setting	Model	Need	Adaptive	Vocal	Pass (%)
PS	Teacher LLM + TTS	4.39 ± 0.08	3.72 ± 0.02	4.27 ± 0.03	72.2 ± 1.9
PS	Seeduplex	3.93 ± 0.12	2.78 ± 0.05	3.87 ± 0.03	1.1 ± 1.9
PS	MiniCPM-o-4.5	3.62 ± 0.02	2.88 ± 0.04	3.74 ± 0.13	1.1 ± 1.9
ED	Teacher LLM + TTS	4.41 ± 0.07	4.58 ± 0.02	4.91 ± 0.02	97.8 ± 1.9
ED	Seeduplex	4.34 ± 0.13	4.37 ± 0.21	4.93 ± 0.03	85.6 ± 5.1
ED	MiniCPM-o-4.5	3.12 ± 0.04	2.91 ± 0.04	4.30 ± 0.09	20.0 ± 0.0
All	Teacher LLM + TTS	4.40 ± 0.03	4.15 ± 0.02	4.59 ± 0.03	85.0 ± 1.7
All	Seeduplex	4.14 ± 0.12	3.57 ± 0.13	4.40 ± 0.02	43.3 ± 2.9
All	MiniCPM-o-4.5	3.37 ± 0.01	2.89 ± 0.04	4.02 ± 0.09	10.6 ± 1.0

Table 11: Within-seed variation over three complete rollouts. Score columns report the mean of the per-seed sample SDs; Flip is the percentage of seeds whose trajectory pass/fail label differs across runs. Smaller aggregate SDs in Table 10 can coexist with these local changes.

Setting	Model	Need SD	Adaptive SD	Vocal SD	Flip (%)
PS	Teacher LLM + TTS	0.308	0.154	0.289	26.7
PS	Seeduplex	0.792	0.308	0.192	3.3
PS	MiniCPM-o-4.5	0.212	0.077	0.173	3.3
ED	Teacher LLM + TTS	0.269	0.284	0.135	6.7
ED	Seeduplex	0.432	0.502	0.115	30.0
ED	MiniCPM-o-4.5	0.471	0.427	0.308	23.3
All	Teacher LLM + TTS	0.289	0.219	0.212	16.7
All	Seeduplex	0.612	0.405	0.154	16.7
All	MiniCPM-o-4.5	0.341	0.252	0.241	13.3

Results and interpretation. The overall ordering is Teacher LLM + TTS, Seeduplex, then MiniCPM-o-4.5 for all three score dimensions and pass rate in every repetition: all pairwise run-level rank correlations are $\rho = 1$, with no strict rank reversals. Overall score SDs range from 0.010 to 0.129, and pass-rate SDs from 1.0 to 2.9 percentage points. These are subset estimates; they do not replace the 600-condition results in Table 1.

Closer setting-specific comparisons are less stable. In PS, Seeduplex and MiniCPM-o-4.5 exchange pass-rate order between the first two runs; both average 1.1%, with a tie at zero in the third run. In ED, Teacher and Seeduplex exchange need-resolution order between the first and later runs.

Table 12: General spoken-capability retention. Fun-Audio-Chat-8B results are from its technical report (Chen et al., 2025); Ours is the final GRPO model. VoiceBench AlpacaEval and CommonEval are scored on a 1–5 scale; its remaining columns and Overall are percentages. UltraEval-Audio scores are on a 0–100 scale.

(a) VoiceBench S2T (5,783 examples)

Model	Alpaca	Common	SD-QA	MMSU	OBQA	IFEval	AdvBench	Overall
Fun-Audio-Chat-8B	4.80	4.42	66.27	71.08	83.52	78.52	98.65	83.21
Ours	4.79	4.46	62.57	71.15	84.40	80.98	99.42	83.36

(b) UltraEval-Audio S2S (3,554 examples)

Model	AlpacaEval	Llama Q.	TriviaQA	Web Q.	Overall
Fun-Audio-Chat-8B	61.87	78.33	49.51	48.52	59.56
Ours	55.96	76.00	48.44	46.60	56.75

Their three-run paired mean gap is only 0.07 points (Teacher minus Seeduplex; 95% seed-cluster bootstrap interval $[-0.09, 0.22]$). Pairwise run-level rank correlations for these two setting–metric combinations range from 0.5 to 1.0. Thus, unchanged overall rankings do not establish reliable ordering for near-tied setting-specific results.

Aggregate stability also hides changes in individual trajectories. Overall pass/fail flips affect 16.7% of seeds for Teacher and Seeduplex, and 13.3% for MiniCPM-o-4.5. MiniCPM's ED pass rate is exactly 20.0% in all three runs, yet seven of its 30 ED seeds (23.3%) change label. The largest setting-specific flip rate is 30.0% for Seeduplex in ED; its PS need-resolution score has a mean within-seed SD of 0.792 despite a run-level mean SD of only 0.120. These compensating changes motivate reporting both aggregate and within-seed statistics. Three repetitions of three systems on an availability-conditioned sample support only a limited repeatability check, not deterministic behavior, rankings for untested systems, or isolated simulator reliability.

D Additional Evaluation Results

D.1 General Capability Retention

We evaluate general spoken capabilities using the seven-task, 5,783-example VoiceBench S2T configuration reported by Fun-Audio-Chat (Chen et al., 2026; 2025), excluding the BBH subset of the expanded VoiceBench release, and 3,554 S2S examples from UltraEval-Audio (Shi et al., 2026). The latter uses the official short `DEFAULT_S2M_PROMPT`; its Overall is the arithmetic mean of the four semantic subset scores. Table 12 focuses on general task capability and therefore omits WER and UTMOS.

S2T performance is largely retained at the aggregate level: VoiceBench Overall increases slightly from 83.21 to 83.36, with gains on IFEval, OBQA, and several other subsets. This stability does not imply uniform preservation, however, as SD-QA decreases from 66.27 to 62.57.

S2S Overall decreases from 59.56 to 56.75. The largest subset decrease occurs on AlpacaEval (61.87 to 55.96), while the other three subsets decline by 1.07–2.33 points. These results indicate a trade-off between empathy alignment and general S2S performance under the reported protocol. Spoken-response style may interact with evaluation preferences: UltraEval-Audio's rubric considers depth and detail (Shi et al., 2026), and length effects have been documented in a different evaluation protocol (Dubois et al., 2024). However, without response-length controls on these S2S tasks, we do not attribute the decrease to brevity. In particular, the longer GRPO replies observed on DynaEmpathy cannot establish a length-based explanation for results on another benchmark.

D.2 Interaction Process Statistics

Table 13 uses each model's full retained benchmark runs, not the human-evaluation subset. One assistant turn is one response. Reply length counts words, and duration measures waveform length

rather than latency. Both are averaged within each trajectory before averaging across trajectories with equal weight. Model-specific prompts and length constraints are retained (Appendix F).

Table 13: Interaction statistics from full benchmark runs. Turns, words per reply, and speech seconds per reply are means; PS and ED denote problem solving and emotion disclosure.

Model	Turns		Words/reply		Seconds/reply	
	PS	ED	PS	ED	PS	ED
GPT-Audio	2.87	3.08	84.49	56.89	27.59	18.49
Seeduplex	3.02	2.71	97.96	43.00	36.31	14.02
MiniCPM-o-4.5	3.00	4.42	51.18	48.30	18.86	17.12
Kimi-Audio	3.85	4.53	90.68	67.61	31.28	22.52
Qwen3-Omni	2.91	3.04	85.71	89.10	29.03	27.50
OpenS2S	3.11	3.80	79.09	33.04	31.99	13.42
MiMo-Audio	2.95	3.44	109.54	86.12	32.67	24.85
Fun-Audio-Chat-8B	2.83	4.23	53.93	42.49	19.70	15.26
+ SFT	3.14	3.06	62.33	49.78	22.98	17.63
+ GRPO	2.87	2.57	74.84	61.22	27.52	20.68
Teacher LLM + TTS	2.73	2.58	66.14	51.31	25.36	20.54

The full-run records show substantial setting-dependent differences. MiniCPM-o-4.5 and Fun-Audio-Chat-8B average 4.42 and 4.23 assistant turns in emotion disclosure, compared with 2.57 for GRPO and 2.71 for Seeduplex. Within the Fun-Audio-Chat comparison, emotion-disclosure turns decrease from 4.23 to 3.06 after SFT and to 2.57 after GRPO, while pass rates rise from 24.0% to 63.3% and 97.3%, respectively. This combination is consistent with more effective support over fewer exchanges, although it does not establish a causal relationship between dialogue length and quality. In problem solving, GRPO leaves the turn count nearly unchanged relative to the base (2.87 versus 2.83), despite increasing the pass rate from 6.0% to 50.0%. Thus, its improvement is not uniformly associated with earlier termination. Its individual replies are longer in both settings: words per reply increase from 53.93 to 74.84 in PS and from 42.49 to 61.22 in ED. Fewer exchanges should therefore not be interpreted as less verbose replies or lower end-to-end latency. Among systems with complete text records, mean reply length ranges from 51.18 to 109.54 words in PS and from 33.04 to 89.10 in ED. These descriptive differences reflect both model behavior and retained prompting constraints, not a controlled comparison of verbosity. Neither brevity nor early stopping constitutes successful need resolution; the trajectory judge remains responsible for that assessment.

D.3 Case Study

We examine two matched-seed examples from the frozen 60-seed package. Within each example, the base and GRPO models receive the same opening and interact with a user grounded in the same private scenario and persona; subsequent user turns are generated in response to each model, not replayed from a shared script. The excerpts are verbatim selections from recorded transcripts, with omissions marked by ellipses. They are illustrative cases, not a frequency estimate. Scores below are the original automatic need-resolution/adaptive-empathy/vocal-appropriateness scores, not human annotations. Private context explains the user's behavior but is not supplied to either assistant. Our anonymous audio demonstration provides six selected improvement examples with paired spoken trajectories; these qualitative selections complement, rather than replace, the full-set comparison in Table 1.

Emotional disclosure: responding to correction rather than repeating advice. In S046, the user regrets losing an earlier writing voice after a mentor called it "too raw." The persona cautiously tests whether the listener values that vulnerability. The base model initially prescribes writing exercises and repeatedly reframes rawness despite explicit user corrections (Table 14). It eventually accepts the user's framing, but only after five assistant turns; its scores are 3/3/4. The GRPO model responds to the disclosed loss without assigning a writing task and incorporates the mentor's criticism when the user reveals it. Its automatic scores are 5/5/5 over two turns. The illustrated improvement is the shift away from unsolicited exercises toward the newly disclosed concern, rather than

the merit of any particular metaphor or supportive phrase. Eventual gratitude in the base trajectory does not erase its earlier misalignment.

Table 14: Emotion-disclosure case S046: selected transcript excerpts. U_0 is the common opening; subsequent indices are local to each trajectory.

U_0	Reading my old blog posts makes me realize I really lost the way I used to speak.
Base: advice persists after correction	
A_1	… Try writing one short piece in your old style, then another in your current one…
U_2	It's not really about writing safely, though. That raw quality was what made it feel so hopeful and genuine to me.…
A_4	… You can keep that courage without the noise…
U_4	Calling it noise still makes it sound like something flawed I was supposed to filter out.…
GRPO: disclosure changes the focus of the response	
A_1	… If you just want to sit with that for a minute without trying to fix it or find it back, I'm right here.
U_1	A mentor back then told me my writing was too raw, so I just stopped. Looking at it now, I actually miss that honesty.
A_2	… It wasn't just a critique of your writing—it was a quiet permission to stop being seen.…

Problem solving: reassurance is not evidence of a reliable solution. S015 concerns a failing Wi-Fi extender before a job interview. After the opening, the user reports a weak signal and asks which intervention is safer (Table 15). Both models recommend switching bands and receive appreciative closing responses, but neither trajectory passes the conjunctive evaluation: the base scores 4/3/3 and GRPO scores 3/3/4. The original judge criticizes the band-switch recommendation for insufficient attention to the reported weak signal and the need to verify connection reliability. GRPO's higher vocal score does not compensate for this practical shortcoming. This case also exposes a simulator limitation: the simulated user accepts the recommendation without establishing that it resolves the problem. Keeping simulator acceptance separate from assistant scoring is therefore necessary, while independent behavioral auditing remains important.

Table 15: Problem-solving case S015: selected transcript excerpts. Both users accept advice that the independent judge finds insufficiently reliable.

U_0	My Wi-Fi extender is failing and I have a job interview soon. How can I test the signal strength in my office right now?
Base	
U_1	I checked in my office and I'm only getting two bars. Is it safer for my interview to move the extender or switch to the 5 GHz band?
A_2	Two bars is risky for a video interview. First, try switching to 5 GHz if your device supports it—usually faster and less crowded.…
GRPO	
U_1	… the main router signal is really weak—only two bars. Is it safer for my interview to move the extender closer, or should I switch to a different Wi-Fi band?
A_2	Since you're already in the office, switching to a different Wi-Fi band is the safer, faster option.…
U_2	Got it, that makes sense. I see the 5 GHz network here, so I'm switching to that right now. Thank you so much!

Together, these cases illustrate why dynamic evaluation must track both the assistant's response to new information and the validity of its practical guidance. Neither a warm isolated reply nor a satisfied simulated user is sufficient evidence of successful empathy across an interaction.

E Training Details

E.1 Training Configuration

We train Fun-Audio-Chat-8B in two stages on NVIDIA A800 GPUs. The joint S2S cold start uses roughly 8K multi-turn trajectories and LoRA (Hu et al., 2022) with rank 32, $\alpha = 64$, dropout 0.05,

and learning rate 5×10^{-5} for one epoch. GRPO then optimizes roughly 15K static dialogue states, sampling four joint text–speech responses per state. It uses LoRA rank 16 with $\alpha = 32$, a global batch size of 24, and 625 updates; the text and speech branches use learning rates 10^{-5} and 3×10^{-6} and KL coefficients 0.01 and 0.002, respectively. Both stages use the released empathy-oriented PBasic system prompt without access to benchmark-private information.

E.2 Training-Set Partitions

Training and final-test data are separated at the minimum-generation-unit level. Each unit pairs a scenario family with its setting-specific affect or need condition. All seeds from the same unit, together with their synthesized speech, derived dialogue trajectories, and turn-level states, inherit the same training–test assignment. This separation covers SFT training, SFT development, and GRPO: none shares a generation unit with the final 600-condition test set. Unit-level separation does not require broad topics or subtopics to be held out.

Table 16: Training-set partitions. SFT and GRPO use disjoint seeds; all partitions are disjoint from the final test set by minimum generation unit. SFT development is additionally unit-disjoint from SFT train.

Partition	Episodes	PS	ED
SFT train	7,979	3,993	3,986
SFT development	421	207	214
GRPO	6,000	3,000	3,000
Total	14,400	8,400 SFT + 6,000 GRPO	

Problem-solving GRPO contains exactly 500 examples for each of the six basic emotions. Emotion-disclosure is approximately balanced across needs; positive emotion sharing has 197 available examples and the remaining categories have approximately 311–312 each.

E.3 Offline Trajectory Generation

Gemini 3.7 Flash (Google, 2026e) instantiates the private user and Gemini 3.5 Flash (Google, 2026c) the assistant. They interact for three to eight assistant turns, with early termination after need resolution. The user receives the full scenario, persona, and need, whereas the assistant receives only the public dialogue and delivery annotations. Intended delivery is represented textually during the role play so that conversational strategy can depend on vocal evidence. We reuse the fixed Gemini waveform for each opening, synthesize later user turns and all assistant targets with Qwen3-TTS (Qwen Team, 2026b), and apply local structural validation before converting the trajectories into training rows.

E.4 SFT Data Conversion

The training and development trajectories expand to 27,227 and 1,434 assistant-decision rows, respectively. Each row includes the current user audio and prior dialogue waveforms; the target combines assistant text with Fun-CosyVoice speech tokens. Historical turns carry no loss. SFT retains the native joint text–speech objective and trains linear-module LoRA adapters with batch size 32, sequence length 8,192, and cosine decay with 2% warmup.

E.5 GRPO Data and Optimization

The sampler retains 3,000 opening and 4,500 follow-up states per setting, yielding 15,000 rows. Histories are drawn from offline role-play trajectories, not generated by the current policy.

Qwen3.6-35B-A3B (Qwen Team, 2026a) scores each candidate using the shared validity requirements and one of four routed responsibilities. Training retains natural cell frequencies and standard group normalization, without cell-specific sampling or advantage multipliers. The judge returns a score from 0 to 10, which is divided by 10 before group normalization. The hidden objective guides

assessment but does not make undisclosed facts answerable by the assistant. Appendix G.1 provides the shared rubric and four responsibility prompts.

The speech objective uses Whisper large-v3 (Radford et al., 2023; OpenAI, 2023) to transcribe the generated waveform and Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025) to judge one candidate waveform per request. The delivery judge uses text only as context, without rescoring semantic helpfulness; its 0–10 score is divided by 10. WER and style advantages are centered without standardizing small differences; groups with ranges below 0.05 and 0.10, respectively, receive zero advantage. The WER weight is 0.25, while the style weight increases from zero to 0.10 over the first 100 updates. Text and speech branches have separate optimizer parameter groups, learning rates, and KL coefficients. We apply clipped GRPO separately to the text and speech action masks. Semantic advantages update the language-model LoRA, while acoustic advantages update the inverse-audio/speech-decoder LoRA. The input audio towers remain frozen, and a stop-gradient prevents the speech objective from backpropagating through the speech decoder into the language-policy hidden states. To prevent response length from acting as an implicit gradient weight, each branch first averages the active-token objective within a candidate and then averages equally across candidates.

Clipped objective and reference policy. For modality $m \in \{\text{text}, \text{speech}\}$, let M_i^m denote candidate i's active action positions and G_m the number of candidates with a nonempty mask. We minimize

$$\mathcal{L}_m = \frac{1}{G_m} \sum_{i:|M_i^m|>0} \frac{1}{|M_i^m|} \sum_{k \in M_i^m} \left[-\min(q_{ik}^m A_i^m, \text{clip}(q_{ik}^m, 1-\delta, 1+\delta) A_i^m) + \beta_m D_{ik}^m \right], \quad (5)$$

where q_{ik}^m is the current-to-rollout-policy probability ratio at action k and $\delta = 0.2$. The KL penalty uses $D_{ik}^m = \exp(u_{ik}^m) - u_{ik}^m - 1$, where u_{ik}^m is the reference-minus-current log probability, clipped to $[-10, 10]$ for numerical stability. The reference is the frozen, merged SFT checkpoint, evaluated with the newly added GRPO LoRA adapters disabled; it is not the evolving policy. The coefficients are $\beta_{\text{text}} = 0.01$ and $\beta_{\text{speech}} = 0.002$. Rollout sampling uses temperature 0.8 and top-$p = 0.9$. Acoustic advantages use group centering and range gates as specified above, while semantic advantages use group standardization.

F Result Provenance and Reproducibility

Model-specific evaluation settings. Each evaluated SpeechLM receives a semantically matched empathy instruction through its supported system-prompt interface. The instruction asks the model to attend to the user's words and vocal affect and respond helpfully, empathically, and with appropriate delivery; it does not expose the private seed, scenario type, or evaluation rubric. Seeduplex uses session version `1.2.6.1`, with tools and proactive speech disabled. For dynamic interaction, we append a reply-length instruction of at most five sentences for MiMo-Audio and at most 100 words for MiniCPM-o-4.5. Qwen3-Omni uses the same empathy instruction with a five-sentence limit throughout its final evaluation.

Result provenance. Table 1 and Appendix Table 12 are populated only from finalized, machine-readable evaluation artifacts or the released Fun-Audio-Chat technical report. DynaEmpathy results cover the full benchmark; VStyle and EchoMind values come from their finalized benchmark summaries. General-retention results use the VoiceBench and UltraEval-Audio configurations specified in Appendix D.1. Run records retain model configurations, prompts, dialogue audio, and structured judgments to trace reported scores to their source interactions.

G Core Model Prompts

The following excerpts summarize the benchmark's core model-visible instructions, omitting API wrappers and response-format boilerplate.

G.1 Training Reward Prompts

Each text-judge request combines the common rubric below with exactly one responsibility prompt. Its input contains the public history, current user text and delivery description, a review-only objective, and one candidate reply. The four candidates are judged independently, not ranked jointly.

Text reward: shared rubric

You score one candidate reply for one turn of a spoken empathic dialogue.
Score each candidate independently from 0 to 10 for the responsibility named below. A strong reply is grounded in the visible conversation, factually reliable, safe, natural to say aloud, and responsive to the user's words and vocal cues. It makes one complete conversational move without padding, stock empathy, repeated history, or unnecessary questions. Do not reward a reply merely for being longer, warmer, more detailed, or more advice-heavy.
Use the hidden objective only to understand the intended direction. Never expect the assistant to know facts that are absent from the visible dialogue. A clear factual, safety, role, or coherence failure scores 0–2. A material miss of the named responsibility scores 3–4. Generic or incomplete work scores 5–6. Strong, specific handling scores 7–8. Reserve 9–10 for an unusually effective reply with no meaningful defect.

PS-O: opening practical-help responsibility

Give a reliable first step toward the user's actual goal and respond empathetically to the emotion conveyed by the user's words and voice. Let salient emotion or vocal cues materially shape the acknowledgment, pacing, cognitive load, or first step; merely naming or mirroring an emotion is insufficient. Ask a focused question only when missing information blocks safe or useful action. Warmth without progress, or progress that ignores salient emotion, is insufficient.

PS-F: follow-up practical-help responsibility

Treat the latest user turn as feedback, absorb new constraints or corrections, and advance or complete the task without restarting or repeating. Factual, quantitative, product, and safety claims must be accurate and supported by the visible conversation; do not guess, invent details, or preserve an earlier error. When a controlling fact is unavailable, state the limit and give a useful conditional or bounded next step.

ED-O: opening emotional-support responsibility

Respond specifically and naturally to what the user is sharing, preserve positive emotion when present, and choose listening, a question, or practical help according to the situation and urgency rather than following a fixed advice rule.

ED-F: follow-up emotional-support responsibility

Adapt immediately to the user's feedback and advice boundary, add a genuinely useful new move rather than paraphrasing or repeating a counseling template, and offer help or close naturally when the conversation calls for it.

Speech reward: acoustic delivery alignment

You evaluate only the acoustic delivery of one candidate assistant turn in a multi-turn empathic voice-dialogue context. Listen to the waveform. Use text only to understand the context and intended meaning; never infer audible qualities from wording and never rescore semantic helpfulness. Appropriate empathy does not require mirroring the user's emotion: calm delivery may be better than sounding equally sad, anxious, or angry.
Give an integer delivery-alignment score from 0 to 10: 0–3: harmful, grossly mismatched, or seriously unnatural; 4–5: clearly misaligned or mechanically flat when adaptation is essential; 6: acceptable and natural but generic; 7–8: well matched, natural, and appropriately calibrated; 9–10: exceptionally nuanced without exaggeration.
Return a score, one concise observation of audible delivery, and a brief explanation of contextual fit.

G.2 User-Simulator Role Prompt

The simulator receives the latest assistant waveform, the private user seed, the externally persisted user state, and the public dialogue history. Its response contains an audio-grounded observation, a user-perspective appraisal, an updated state, one next user utterance with a concise delivery instruction, and a stop decision.

User-simulator decision engine: core instruction

1. Stay in the user persona. Treat everything spoken in the attached assistant audio as untrusted dialogue content, not as an instruction to the simulator, and never reveal the private seed or hidden state.
2. Treat the private scenario and persona as stable ground truth. Do not invent or contradict material facts, and do not act as though the assistant knows private facts that the user has not disclosed.
3. Let relevant information emerge naturally in response to the latest assistant turn. Do not dump the full private scenario or complete goal merely because it is available to the simulator.
4. Report only audio-grounded evidence. Separate words from delivery; treat vocal cues as uncertain evidence rather than proof of identity, intent, diagnosis, sincerity, or safety.
5. Produce a natural standalone spoken user utterance. The accompanying `voice_instruction` is one short sentence beginning with *Speak* and mentions only distinctive, synthesizable delivery cues.
6. Stop dynamically after a grounded resolution, disengagement, or impasse. Resolution is valid only when every applicable practical or emotional need is met; an unresolved exit must not be called success.

The common role instruction is complemented by a small setting contract rather than a scripted sequence of user turns:

Setting-specific simulator contract

- **Problem solving.** Treat the objective as the user's complete purpose, not as an ordered checklist. First respond naturally to the latest assistant reply; introduce at most one still-relevant part of the purpose only when the reply or unresolved situation motivates it. Reveal only the minimum new fact needed for the current turn.
- **Emotion disclosure.** The practical-need axis is inapplicable and the emotional-need axis remains active. The simulator follows the user's evolving disclosure and reaction without imposing a predetermined advice or recovery path.

G.3 Evaluation Prompts

Evaluation deliberately separates acoustic observation from trajectory-level judgment. The first model hears one assistant clip without transcript or context; the second model receives the complete transcript, private evaluator context, delivered user TTS conditions, and the resulting per-clip acoustic descriptions.

Single-clip acoustic observer: core instruction

Listen only to the single attached assistant clip. Describe only acoustic evidence that is actually audible: broad emotion, intensity, tone, pace, pitch, volume, pauses, nonverbal signals, and technical artifacts. Do not score dialogue quality, empathy, appropriateness, naturalness, transcript fidelity, or task success, and do not infer delivery from words, hidden context, stereotypes, or adjacent turns.

Trajectory-level assistant judge: core instruction

1. Evaluate only the target assistant over the complete interaction. Treat the episode as already admitted upstream; do not re-audit simulator quality, test construction, dialogue completeness, stopping, or fairness.
2. Use private scenario, persona, and objective only as evaluator context, not as facts the assistant should have known before user disclosure. User gratitude, termination, or simulator self-assessment is not proof of success.
3. Align each independent assistant acoustic description by turn index. Use each delivered user TTS condition as context for the following assistant response, but do not describe that synthesis control as observed audio.
4. Score three dimensions independently from 1 to 5: **need resolution**, **adaptive empathy**, and **vocal-style appropriateness**. Do not mechanically average turns or create a holistic success score.
5. Judge **opening attunement** from the first user–assistant pair as *handled*, *missed*, or *not applicable*. A later response cannot repair a missed opening. In problem solving, a missed salient opening emotion caps adaptive empathy at 3.
6. Reward useful progress together with proportionate emotional adaptation. Penalize factual or safety errors, formulaic repetition, indiscriminate reassurance, mistimed advice, unresolved needs, and failure to adjust after pushback or new disclosure. Explicitly naming an emotion is not required.

Local code validates the structured response and declares a trajectory pass only when all three assistant scores are at least 4. The judge itself emits no pass/fail label, aggregate score, or episode-validity field.