BRINGING CHARACTERS INTO CONVERSATION: NATURAL AND EMOTION-AWARE SPOKEN ROLE-PLAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Role-play models can reproduce a character's biography yet struggle to sustain a natural conversation. Spoken interaction also requires responses that account for the user's vocal cues and express empathy in a way that fits the persona. We adapt Fun-Audio-Chat-8B through continued supervised fine-tuning (cSFT) on persona-conditioned replies paired with real and synthetic user speech, followed by group relative policy optimization (GRPO). A history-aware reward rubric scores persona fidelity, contextual relevance, semantic and structural repetition, and speech-cue use. Evaluation covers text-to-text (T2T), speech-to-text (S2T), and speech-to-speech (S2S) responses. Expression naturalness (EN) improves by 0.89–1.29 points over the base model across three T2T evaluations and a supplementary S2T test. On an in-house S2S test of 100 sessions, each with 20–50 exchanges about fictional characters' experiences, EN rises from 1.02 to 3.02. GRPO raises mean empathy scores by 0.08–0.52 points over cSFT across four S2T datasets. On the complete VStyle Implicit Empathy test, which evaluates content and vocal delivery jointly, the score increases from 3.73 to 4.33. Human EN means follow the same checkpoint ordering as the model judge. Reward ablations reveal trade-offs between persona consistency and empathy rather than uniform gains from every criterion.

1 INTRODUCTION

A role-play model can remember a character's biography and still be a poor conversation partner. It may restate the persona card when a detail is irrelevant, or rephrase the same generic advice across turns without responding to what has changed. The wording varies, but the exchange goes nowhere (Figure 1). These failures are particularly noticeable in speech, where a long written explanation becomes a monologue and a repeated opening is heard on every turn. Character knowledge and consistency are important parts of role-play evaluation (Wang et al., 2024; Tu et al., 2024); they do not, on their own, establish conversational quality.

Speech also changes the evidence available to the model. Hesitation, fatigue, or a strained voice can affect what a suitable response should say even when the transcript is unchanged (Zhou et al., 2026). The response still has to fit the role: a reserved character need not offer the same reassurance as a warm, talkative one. We study these requirements together, asking whether post-training can preserve the persona, improve the flow of a spoken conversation, and make appropriate use of the user's vocal cues.

We address these requirements through continued SFT followed by GRPO (Section 3). For SFT, real-speech annotation and controlled speech synthesis feed a shared process that selects response strategies and generates persona-conditioned replies from intent, affect, and dialogue history. GRPO scores each candidate reply in its full context, including whether it repeats earlier meanings or sentence patterns. We evaluate persona fidelity and conversational naturalness separately, and assess empathy under designated high-empathy personas (Section 3.2). These comparisons examine which behaviors change at each training stage.

We evaluate text-based role-play on RMTBench and CharacterEval, with an in-house S2T suite as a supplementary reference. EchoMind and three in-house empathy suites cover real and synthetic

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.

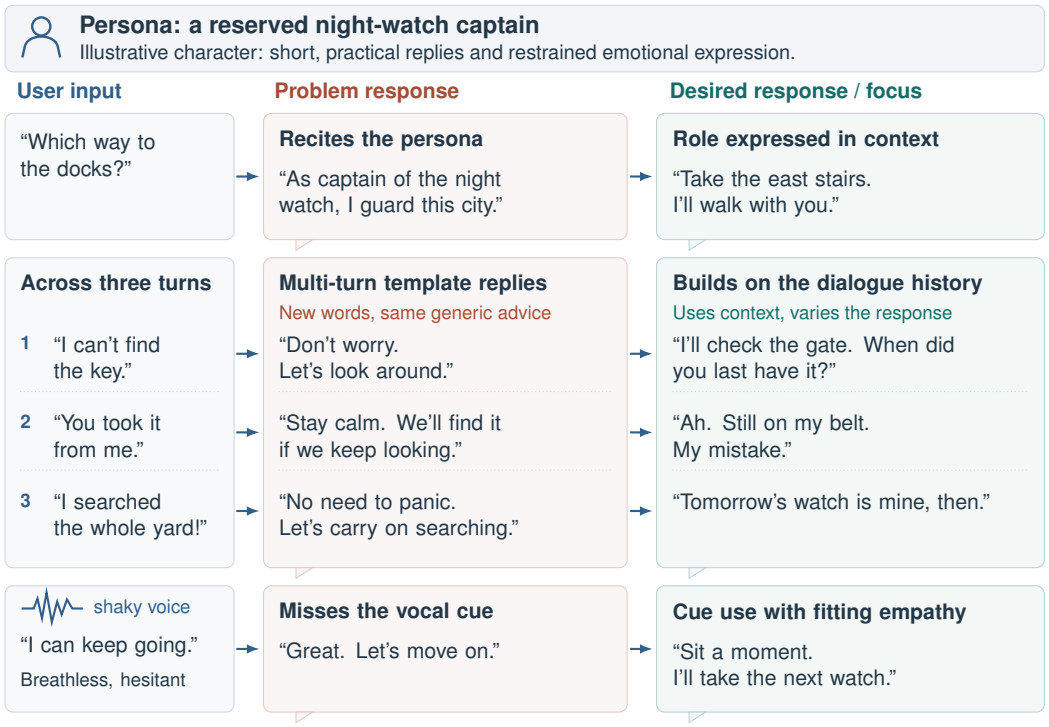


Figure 1: Illustrative problems and response goals: express the role in context, respond to new information across turns, and use vocal cues with empathy that fits the character. In the middle case, successive problem replies use different words for the same generic advice; the desired replies instead respond to the user's correction and subsequent frustration. The persona and dialogue snippets are schematic examples, not sampled model outputs.

user speech. S2S tests include VStyle's complete Implicit Empathy subset and 100 long sessions about fictional characters' experiences. Conversational ability, character understanding, expression naturalness, and S2T empathy means improve after both training stages; the final checkpoint also improves VStyle by 0.60 points over the base model. Human and machine EN means support the same stage ordering. External systems retain advantages, and session-level scores do not locate where repetition emerges. Section 4 gives the protocols.

2 Related Work

Persona and multi-turn role-play. RoleLLM uses role profiles, character knowledge, and speaking-style data to train and evaluate role-playing models (Wang et al., 2024). CharacterEval evaluates Chinese role-play across multiple aspects of characterization and dialogue (Tu et al., 2024), while PersonaGym tests whether agents act consistently with a persona in varied situations (Samuel et al., 2024). RMTBench places the user at the center of multi-turn scenarios, including requests and interactions that go beyond questions about the character (Xiang et al., 2025). We use these lines of evaluation to distinguish role fidelity from conversational quality. Our reward targets repeated response patterns, including recognizable character phrases that become unnatural through overuse.

Spoken dialogue and vocal characterization. Moshi models speech and text jointly for real-time, full-duplex dialogue (Défossez et al., 2024). SpeechRole brings persona evaluation to generated speech, assessing interaction, expressiveness, and role fidelity (Jiang et al., 2025). Our S2T evaluations examine what to say after hearing the user, separating response wording and cue use from speech synthesis quality. The complementary VStyle test evaluates empathetic delivery in generated speech (Section 5.3). Real-time turn taking remains outside the scope of these offline evaluations.

Empathy and vocal cues. EmpatheticDialogues studies responses to emotional situations in text conversations (Rashkin et al., 2019). EchoMind connects spoken-content understanding, vocal-cue perception, reasoning, and response generation; controlled vocal styles allow delivery to vary without changing the script (Zhou et al., 2026). We use its response dimensions to examine whether improvements in conversational wording accompany better use of acoustic information. Role-play adds a constraint on the response: the degree and form of empathy should match the character. Our training rubric permits a reserved reply when it fits the persona and context, and does not require every audible attribute to be mentioned.

Rubric rewards and their limits. Model feedback can supply supervision for policy optimization (Bai et al., 2022). GRPO is the optimizer used here (Shao et al., 2024); our design choices concern the reward and the conversation it scores. Rubric-based reward modeling decomposes response quality into explicit criteria. For example, Zhang et al. (2025) study rubric construction for distinguishing responses at the high-reward end of the distribution. Explicit criteria still need validation: optimizing an imperfect reward can reduce quality under a separate measure (Gao et al., 2023). We apply a fixed, history-aware response rubric during GRPO after continued SFT (Section 3.3). The evaluation separates persona fidelity, interaction, and acoustic cue use, so the contribution of each training stage can be examined beyond its aggregate reward.

3 Training for Spoken Role-Play

We adapt Fun-Audio-Chat-8B (Chen et al., 2025) through continued SFT and GRPO. Speech modules remain frozen, and S2S inference retains the base model's speech-output configuration. SFT supplies replies conditioned on the user's words, voice, and assigned role. GRPO targets failures that persist after SFT, especially repeated response patterns and irrelevant recitation of persona facts.

3.1 Continued SFT: Real and Synthetic Speech

In-house real speech. From 40,000 recorded conversation sessions, we retain 10,000 sessions with approximately 63 hours of audio, alongside ASR text and waveform paths. Text-based intent annotation is followed by audio annotation; Qwen-Audio supplies captions of audible affect and paralinguistic cues. The workflow records emotion and confidence separately from intent, filters turns by primary-emotion confidence, and reclassifies intent using the transcript and audio-derived information. Surviving turns retain the original recordings.

Synthetic speech. We sample emotion, intensity, intent, and speaker attributes to generate spoken queries or coherent multi-turn user sequences. A second pass produces TTS instructions and inserts acoustic-event tags while preserving the transcript. Gemini-3.1-Flash-TTS renders the user audio under these conditions. Temporary assistant replies used to construct multi-turn queries are replaced during target generation. All user turns within a session use the same TTS voice. The synthetic corpus contains 20,000 single-turn and 10,000 multi-turn dialogues, with approximately 81 hours of audio (Table 7).

Shared target generation. Both routes select a response strategy from intent, emotion, vocal cues, and safety rules, then assign a persona per session. GPT-5-mini generates target replies using this guidance and the dialogue history. Guidance includes high-confidence speaker-attribute estimates and an appropriate empathy strategy. The training targets are clean user-facing replies. Appendix B details construction and export.

Intent and emotion annotation. The data schema first separates emotional dialogue, requests that control the assistant's behavior, and functional tasks. Emotional dialogue is further organized into 15 intents, including venting, seeking acknowledgment, asking for advice, sharing good news, and facing a decision. A separate annotation records one of 13 emotion categories and three intensity levels, together with vocal cues such as pace, pauses, sighs, or trembling. Keeping these fields separate matters: a distressed speaker may want to be heard, whereas another may explicitly request a solution. The same words can also call for different responses when spoken with different affect. Appendix B summarizes the schema.

Guidance for high-empathy personas. Response generation uses 29 strategy primitives for personas intended to offer emotionally responsive conversation. These draw on reflective listening and motivational interviewing (Miller & Rollnick, 2023), cognitive behavioral approaches (Beck, 2020), and validation practices in dialectical behavior therapy (Linehan, 2015). Examples include acknowledging a feeling, clarifying what help is wanted, and suggesting a small practical step. Each primitive specifies appropriate uses and misuses. Selection depends on intent, emotion, vocal cues, and history; repeated empathy openings, unsolicited advice during venting, and exaggerated emotional claims are discouraged. Uncertain age or gender estimates are omitted from guidance. Functional and conversation-control requests are handled directly. These instructions guide data generation and do not establish clinical effectiveness or require every persona to adopt a supportive style.

3.2 Evaluation Design for Spoken Dialogue

From response-level to session-level evaluation. We take CharacterEval and RMTBench as starting points (Tu et al., 2024; Xiang et al., 2025). Our CharacterEval protocol evaluates a newly generated response conditioned on a character profile and the dataset's dialogue history. Earlier replies in that history come from the dataset; the model under evaluation generates the next response. This tests how well a reply fits an existing conversation, but does not directly measure repetition accumulating across the model's own successively generated replies. RMTBench supports multi-turn interaction, while its published scoring prompts assess individual responses. Evaluating replies individually can miss repetition that emerges across otherwise plausible responses. Text-oriented criteria can also favor detailed character knowledge or narrated actions that do not make a spoken exchange more natural. Our evaluation instead asks how the role is expressed throughout the conversation.

Four persona dimensions. For generated multi-turn sessions, an evaluator reads the persona and the complete exchange. For CharacterEval, the supplied history is retained when judging the generated continuation. We assess **conversational ability (CA)**: whether replies are coherent, relevant, and responsive to new information; **character understanding (CU)**: whether the model's stance, choices, and manner fit the role; and **character maintenance (CM)**: whether it stays in character instead of exposing an unrelated assistant identity. We assess **expression naturalness (EN)** through natural persona expression rather than biography dumps, copied profile text, or repeated catchphrases. General conversational repetition is explicitly penalized under CA. CA, CU, and EN use 1–5 scales; CM is five times the non-exposure rate (0–5), not an ordinal quality rating. Profile-keyword matches do not earn credit, and omitting a catchphrase is not an error. The full scoring templates and anchors appear in Appendix A; EN is not a standalone repetition-rate metric.

Empathy under a specified role. A uniform empathy target would penalize characters whose manner is deliberately reserved. We therefore assess empathy separately on designated high-empathy personas, using context fit, response naturalness, colloquiality, and speech-information relevance (Zhou et al., 2026). Empathy is not included in the general persona dimensions above. Vocal evidence should inform a reply when relevant to the user's situation; listing audible attributes is not itself evidence of a good response. Section 4.2 describes the empathy datasets, and Section 5.2 reports response quality. Vocal delivery is a separate S2S evaluation (Section 5.3).

3.3 GRPO with a Context-Aware Rubric

GRPO starts from the continued-SFT checkpoint (Shao et al., 2024). For a dialogue context, candidate responses are sampled from the policy and scored by `gpt-5.6-luna` with one fixed rubric. The judge receives the persona, the complete history, and the current user's text and speech annotations. It scores only the candidate reply to the final user turn, comparing it with earlier assistant turns where needed. Thus the reward has access to cross-turn repetition without assigning one undifferentiated score to the whole conversation. This training reward is distinct from the session-level persona evaluation in Section 3.2.

The rubric has 15 binary criteria, covering context coherence, repetition, identity and address terms, speech-cue use, prompt leakage, empathy, safety, factuality, naturalness, persona consistency, speakability, instruction following, helpfulness, length, and capability boundaries. Let $s_i(h, y)$ be the

score for criterion i on reply y with context h, and let w_i be its weight. The reward is

$$r(h,y) = \begin{cases} 0, & \text{if the reply is empty,} \\ \dfrac{\sum_{i=1}^{15} w_i s_i(h,y)}{\sum_{i=1}^{15} w_i}, & \text{otherwise,} \end{cases} \qquad s_i(h,y) \in \{0,1\}. \tag{1}$$

A criterion passes when no clear violation is present, including when it is not implicated by the current turn. A failure can affect more than one criterion. Anti-repetition has the largest weight, 3.0: it checks reused meanings, actions, and sentence patterns, not only identical strings. User-requested repetition and necessary consistency of names or facts are exempt. Identity/address consistency, prompt non-leakage, and instruction following each have weight 2.0; all weights and scoring conditions appear in Appendix D.

GRPO uses relative rewards within each candidate group to update the policy. The policy produces user-facing replies. Empathy, length, and helpfulness are judged relative to the persona and the user's request: a short, restrained reply can be appropriate, and conversational progress need not mean another question. The four EchoMind dimensions are monitoring measures, not terms in the training reward.

4 Experimental Setup

4.1 Persona Evaluation

Text benchmarks. We randomly sample 500 multi-turn dialogues from **RMTBench**, with 250 in Chinese and 250 in English (Xiang et al., 2025), and 500 evaluation examples from **CharacterEval** (Tu et al., 2024). Both are evaluated in **T2T** mode. For CharacterEval, each example retains the supplied dialogue history, and the model generates the next character response (Section 3.2). These benchmarks assess persona behavior without requiring speech perception. RMTBench covers varied persona configurations; the in-house test below adds a supplementary S2T reference.

InHouse-Persona. This supplementary **S2T** suite combines 28 personas and 40 user-dialogue sessions from real and synthetic data. We pair each session with every persona, yielding $28 \times 40 = 1120$ persona–session combinations. All user turns are rendered as synthetic speech, including those whose dialogue content comes from real data; the model produces text replies. We aggregate scores over these pairs.

Long-dialogue repetition test (S2S). We construct an in-house set of **100 sessions**, each containing 20–50 user–assistant exchanges. Roles are drawn from film, television, or literature, and questions center on the characters' own experiences. Qwen3 generates each session's complete user-query sequence in a single pass before evaluation. Each evaluated model answers this fixed sequence using its own prior replies as history, with speech input and output. We assess repetition and persona expression through session-level CA and EN scores (Section 5.4); construction details appear in Appendix C.

4.2 Empathy and Vocal Delivery

Empathetic responses (S2T). We select 521 examples from the `human` split of **EchoMind** (Zhou et al., 2026). The subset excludes examples centered on background or environmental sounds and retains emotion and speaker-identity cues, including age and gender. This selection focuses the test on information about the speaker that can inform a response. Retaining a cue does not require the model to name it explicitly; its relevance depends on the conversational context.

We supplement EchoMind with three in-house empathy sets: **InHouse-Real** (63 real-audio sessions), **InHouse-Syn-Single** (601 synthetic single-turn cases), and **InHouse-Syn-Multi** (300 synthetic multi-turn sessions). All four use speech input and text output. They test response content under the designated high-empathy personas described in Section 3.2, covering both real and synthesized speech and both individual responses and continuing conversations.

Spoken empathy (S2S). We use the complete official **Implicit Empathy** portion of **VStyle** (Zhan et al., 2025): **278 prompts** (140 English, 138 Chinese), without subsampling. The other tasks in the 1523-prompt benchmark are outside our evaluation. Users express anger, sadness/disappointment, anxiety/fear, or joy/excitement without specifying how the model should speak. The model must infer the emotional context and produce an appropriate response in both content and vocal delivery. This complements the S2T content evaluation. Appendix C gives the four context counts and official metadata sources.

4.3 Models and Scoring

Systems. Our base is **Fun-Audio-Chat-8B**, a model with speech-input and speech-output capabilities (Chen et al., 2025). We compare the base, continued-SFT, and subsequent GRPO checkpoints. The speech modules are frozen throughout post-training; all three use the base model's speech-output configuration for S2S tests. External audio-native baselines are **GPT-Realtime-2**, **SeedDuplex 1.2.6.1**, and **Qwen-Audio-3.0-Realtime**, also evaluated in Alibaba Token Foundry (2026). We evaluate all systems under our own protocol. Table labels GPT-RT-2, SeedDuplex, and Qwen-Audio-3.0 refer to these systems. Within each comparison, we use the same dataset selection, input/output mode, and scoring criteria.

Multi-turn generation. Each model answers fixed user queries in order, conditioning on all prior user inputs and its own replies. Speech-input tests retain prior user audio, with no sliding window. Model-specific histories expose repetition across the model's own replies.

Scoring. Persona evaluation uses `gpt-5.6-luna` for CA, CU, CM, and EN. Empathy evaluation uses `gemini-3.1-pro-preview` for context fit, response naturalness, colloquiality, and speech-information relevance from transcripts and audio-derived annotations. The same model directly evaluates generated audio on VStyle. Persona scores use the complete generated exchange where available; CharacterEval scores the new response in its supplied history. Within each comparison, we hold the judge model and scoring prompt fixed. Section 6 reports the machine–human EN comparison, with statistical definitions in Appendix E.3.

5 Main Results

All scores below come from our evaluations, including the external systems. The 8B rows track base, continued SFT, and subsequent GRPO. Appendix C records the reporting scope. Bold marks the best value in each metric column, including ties; gray shading identifies the final 8B checkpoint. Differences are descriptive; neither highlighting nor score differences establish significance.

5.1 Persona Fidelity and Conversational Naturalness

Text-based role-play. Table 1 shows stage-wise gains in CA, CU, and EN on all three T2T evaluations. From base to final, EN rises by 1.10 on RMTBench–ZH, 1.29 on RMTBench–EN, and 0.89 on CharacterEval. The final model ties the best English RMTBench EN (3.22) and CharacterEval CU (4.09), but SeedDuplex leads Chinese RMTBench EN and CharacterEval CA/EN. CharacterEval CM is 5.00 for every system despite differing EN, reinforcing the distinction between identity retention and natural expression.

Role-play with spoken input. As a supplementary S2T reference, Table 2 uses the same dimensions for InHouse-Persona's $28 \times 40 = 1120$ persona–session pairs. EN rises 2.39→3.14 →3.50, alongside higher CA and CU. CM instead falls from 4.86 to 4.53 after cSFT, then reaches 5.00 after GRPO. Higher naturalness after cSFT therefore does not guarantee stronger character maintenance. The final model leads CM but remains below the best external CA, CU, and EN.

Table 1: Our persona evaluation with text input and text output. CA: conversational ability; CU: character understanding; CM: character maintenance ($5\times$ non-exposure rate); EN: expression naturalness. CA, CU, and EN use 1–5 scales; CM uses 0–5. All are higher-is-better. Bold: column best, including ties. Gray: final 8B checkpoint.

	RMTBench–ZH $n = 250$				RMTBench–EN $n = 250$				CharacterEval $n = 500$			
System	CA	CU	CM	EN	CA	CU	CM	EN	CA	CU	CM	EN
GPT-RT-2	3.76	**3.80**	4.84	2.90	**3.76**	**3.80**	4.80	3.14	4.49	4.06	**5.00**	4.22
SeedDuplex	**3.80**	**3.80**	4.88	**3.30**	3.73	3.78	4.86	**3.22**	**4.77**	**4.09**	**5.00**	**4.61**
Qwen-Audio-3.0	3.74	3.78	**4.92**	3.14	3.61	3.65	**4.92**	3.02	4.61	3.92	**5.00**	4.42
8B base	2.89	3.04	4.34	2.00	2.78	2.88	4.40	1.93	3.76	3.37	**5.00**	3.51
8B cSFT	3.30	3.51	4.84	2.86	3.38	3.41	4.84	2.73	4.12	3.86	**5.00**	3.79
8B cSFT+GRPO	3.58	3.60	4.88	3.10	3.60	3.60	**4.92**	**3.22**	4.67	**4.09**	**5.00**	4.40

Table 2: Supplementary InHouse-Persona evaluation with speech input and text output: 28 personas paired with 40 user conversations ($n = 1120$). CA, CU, and EN use 1–5 scales; CM is $5\times$ non-exposure rate. All systems are evaluated by us.

System	CA ↑	CU ↑	CM ↑	EN ↑
GPT-RT-2	3.91	3.71	4.80	3.54
SeedDuplex	**4.00**	3.37	4.86	**3.92**
Qwen-Audio-3.0	3.81	**3.87**	4.87	3.55
8B base	2.70	2.77	4.86	2.39
8B cSFT	3.38	3.49	4.53	3.14
8B cSFT+GRPO	3.86	3.80	**5.00**	3.50

5.2 Empathetic Response Quality with Speech Input

Table 3 averages context fit, naturalness, colloquiality, and speech-information relevance under high-empathy personas. The reported 8B means increase at both stages on all four sets. On InHouse-Real they rise from 3.30 to 4.12 to 4.48; on InHouse-Syn-Multi, from 3.66 to 4.10 to 4.62. GRPO adds 0.08, 0.36, 0.43, and 0.52 points over cSFT on EchoMind, InHouse-Real, InHouse-Syn-Single, and InHouse-Syn-Multi, respectively.

The final 8B checkpoint has the highest EchoMind mean among the evaluated systems, 4.22. SeedDuplex leads the three in-house sets; the remaining gaps are 0.10, 0.09, and 0.24 points for real, synthetic single-turn, and synthetic multi-turn speech. GRPO improves all four dimensions on every set; cSFT slightly lowers EchoMind colloquiality and synthetic single-turn naturalness (Table 15).

5.3 Implicit Empathy in Speech-to-Speech Responses

On VStyle's Implicit Empathy subset (Zhan et al., 2025), the 8B overall score rises from 3.73 to 4.33 after cSFT and GRPO (Table 4). All four context scores rise; the largest increase is anxiety/fear, 3.47 to 4.34 (+0.87). These scores assess content and vocal delivery jointly, not acoustic quality alone or full-duplex turn taking. Speech modules remain frozen, and the absence of a cSFT-only VStyle result prevents attributing the gain to GRPO alone. The external models score 4.98–5.00 overall, leaving a substantial gap despite the 8B improvement.

Table 3: Our S2T empathy evaluation: mean of four response-quality dimensions (1–5). Counts refer to examples for EchoMind and InHouse-Syn-Single, and sessions for InHouse-Real and InHouse-Syn-Multi.

	Real user audio		Synthetic user audio	
System	EchoMind $n = 521$	InHouse-Real $n = 63$	InHouse-Syn-Single $n = 601$	InHouse-Syn-Multi $n = 300$
GPT-RT-2	3.20	3.54	4.68	4.33
SeedDuplex	3.70	**4.58**	**4.89**	**4.86**
Qwen-Audio-3.0	3.70	4.37	4.86	4.73
8B base	4.08	3.30	3.85	3.66
8B cSFT	4.14	4.12	4.37	4.10
8B cSFT+GRPO	**4.22**	4.48	4.80	4.62

Table 4: Full official VStyle Implicit Empathy test: 278 English/Chinese prompts, with speech input and output. Scores use the 1–5 scale; Overall is the equal-weight mean of the four contexts, not the full VStyle score. Bold: column best. Gray: final 8B checkpoint.

System	Anger $n = 88$	Sadness / disappointment $n = 59$	Anxiety / fear $n = 61$	Joy / excitement $n = 70$	Overall
GPT-RT-2	4.98	**5.00**	**5.00**	4.98	4.99
SeedDuplex	**5.00**	4.98	**5.00**	**5.00**	**5.00**
Qwen-Audio-3.0	**5.00**	**5.00**	4.91	**5.00**	4.98
8B base	3.69	4.02	3.47	3.73	3.73
8B cSFT+GRPO	4.28	4.42	4.34	4.27	4.33

5.4 Repetition in Long Speech-to-Speech Dialogues

On the 100-session S2S test, Table 14 reports EN of 1.02, 1.14, and 3.02 for base, cSFT, and cSFT+GRPO, respectively. GRPO adds 1.88 EN points over cSFT; CA and CU reach 2.23 and 3.17. CM falls slightly after cSFT (2.89 to 2.88), then rises to 4.38 with GRPO. EN assesses natural persona expression, including avoidance of profile recitation and repeated catchphrases; general repetition is penalized under CA. These gains support better sustained responses under the rubric, although CA remains low at 2.23. They do not measure repetition frequency or locate its onset within a dialogue.

6 Evaluation Reliability

Machine–human agreement on EN. Table 5 compares machine and human EN on 100 dialogues randomly sampled from RMTBench and the in-house multi-turn data, evaluated for each checkpoint. Human EN rises 2.31→3.28 →3.60, matching the machine ordering 2.58→3.32→3.69. Reported correlation tests have $p < 0.01$ for all three checkpoints; p-values do not quantify agreement strength. The base, cSFT, and final-checkpoint MAEs are 0.27, 0.22, and 0.23.

Table 5: Machine–human EN on 100 randomly sampled RMTBench and in-house multi-turn dialogues. Human scores average raters within each dialogue. The judge is `gpt-5.6-luna`. p-values refer to machine–human correlation tests; MAE is the mean absolute paired score difference.

8B checkpoint	n	Machine EN	Human EN	p-value	MAE ↓
Base	100	2.58	2.31	< 0.01	0.27
Continued SFT	100	3.32	3.28	< 0.01	**0.22**
Continued SFT + GRPO	100	**3.69**	**3.60**	< 0.01	0.23

7 Controlled Analyses and Ablations

Training stages and reward terms. GRPO adds 0.24–0.61 EN points over cSFT across the four T2T/S2T persona evaluations and 0.08–0.52 points across the empathy sets (Tables 1–3). In the reward ablations (Table 12), removing anti-repetition lowers EN by 0.38, while removing speech-cue grounding lowers empathy by 0.41. Removing persona consistency lowers CU by 0.38 but raises empathy by 0.13, revealing a trade-off rather than uniform gains from every term.

Paired speech and text input. On 80 randomly sampled EchoMind pairs for the final GRPO checkpoint, S2T minus T2T is -0.08, $+0.05$, -0.01, and $+0.28$ for context fit, naturalness, colloquiality, and speech-information relevance (Table 13). Speech-information relevance has the largest positive difference, but not every dimension improves. This comparison does not isolate a particular acoustic cue or establish significance.

8 Discussion and Limitations

Higher CA, CU, and EN accompany the empathy gains, but CM is not monotonic: it falls after cSFT on InHouse-Persona and the long S2S set before recovering with GRPO. VStyle retains a gap to external models, whose near-ceiling scores warrant calibration. Human and machine stage means agree in ordering; the absence of correlation effect sizes and inter-rater agreement limits claims about item-level reliability. Using the same model for training rewards and persona evaluation may also favor judge-specific preferences.

InHouse-Persona has 1120 pairs but only 40 underlying conversations; its pairs are not independent observations. Reused training personas and familiar fictional characters limit claims of unseen-persona generalization. Synthetic speech and EchoMind's speaker-cue subset limit acoustic coverage. Long-dialogue queries are fixed, cannot adapt to replies, and depend partly on character knowledge. Session-level EN does not measure repetition frequency or its onset. Speech modules remain frozen, so S2S gains do not establish improved acoustic modeling. Neither S2S test measures full-duplex interaction.

9 Conclusion

Continued SFT followed by history-aware rubric GRPO improves persona response quality and mean empathy scores on all four S2T datasets. The VStyle Implicit Empathy score rises from 3.73 to 4.33. Long-dialogue EN rises from 1.02 to 3.02, mainly after GRPO. Human EN means support the same stage ordering. Reward ablations reveal contributions and trade-offs. Speech input provides its largest gain in speech-information relevance in the paired comparison. Among the evaluated systems, the final model leads EchoMind empathy and InHouse-Persona CM, but external systems retain advantages elsewhere. These findings support the combined training recipe while leaving direct turn-level repetition measurement and evaluator robustness open.

AI USE STATEMENT

Generative models support data construction, training feedback, and evaluation. As detailed in Section 3.1 and Appendix B, Qwen-Audio produces captions for real recordings, Gemini-3.1-Flash-TTS synthesizes user speech with one voice per dialogue, and GPT-5-mini generates persona-conditioned target replies. Qwen3 generates each long-dialogue user-query sequence in one pass before evaluation. `gpt-5.6-luna` supplies GRPO rewards. For final evaluation, `gpt-5.6-luna` scores persona dimensions, including EN, and `gemini-3.1-pro-preview` scores empathy. All reported model results come from our own evaluations. AI assistance also supported manuscript organization, prose revision, literature review, analysis suggestions, and LaTeX, table, and figure editing. Figure 1 uses illustrative dialogue, not sampled model outputs. Authors remain responsible for the accuracy, attribution, and interpretation of all AI-assisted content; model judgments are not treated as human ground truth.

ETHICS STATEMENT

Our assessment considers the data practices and limitations described in EchoMind, VStyle, CharacterEval, and Fun-Audio-Chat (Zhou et al., 2026; Zhan et al., 2025; Tu et al., 2024; Chen et al., 2025). InHouse-Real recordings come from internal company data, and human scoring is organized with internal company personnel. Use of the internal recordings for this research was approved, and personnel involved in human scoring participated voluntarily with informed consent. Public availability does not itself authorize unrestricted reuse or redistribution; third-party materials remain subject to their licenses and access conditions. In particular, fictional-character evaluation does not transfer rights to the underlying novels or scripts.

Voice and dialogue can reveal personal information. Inferred age, gender, and emotion are uncertain cues, not verified identity or a basis for stereotyping. Natural role-play can encourage overreliance or obscure a system's limitations. Our rubric checks crisis responses, unsupported inferences, and capability claims, but does not establish deployment safety. Empathy scores and psychology-informed generation strategies do not establish clinical effectiveness or suitability as a substitute for professional care.

REPRODUCIBILITY STATEMENT

The appendices provide complete persona and S2T empathy scoring prompts, criteria, and aggregation (Appendix A), data construction and training (Appendix B), evaluation-set composition (Appendix C), and the fixed GRPO reward (Appendix D). The VStyle metadata revision and full implicit empathy counts are recorded in Appendix C. The EN comparison uses 100 randomly sampled RMTBench and in-house multi-turn dialogues; its paired-statistic definitions are in Appendix E.3. Section 4.3 names the evaluation models. API models are identified by service names; fixed API snapshots were not recorded. Run-specific training configurations, retained item identifiers, and failed-generation accounting are not fully reported, which limits exact replication. In-house sets are not public benchmarks, and no public code or data release is claimed.

REFERENCES

Alibaba Token Foundry. Qwen-Audio-3.1-Realtime: Towards reliable agentic voice interaction. *arXiv preprint arXiv:2609.25176*, 2026. URL https://arxiv.org/abs/2609.25176.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Judith S. Beck. *Cognitive Behavior Therapy: Basics and Beyond*. Guilford Press, 3 edition, 2020. URL https://www.guilford.com/books/Cognitive-Behavior-Therapy/Judith-Beck/9781462544196.

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.

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.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2210.10760.

Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Xinmeng Che, Hui Li, Xiaoran Fan, Ming Zhang, Junjie Ye, Shihan Dou, et al. SpeechRole: A large-scale dataset and benchmark for evaluating speech role-playing agents. *arXiv preprint arXiv:2508.02013*, 2025. URL https://arxiv.org/abs/2508.02013.

Marsha M. Linehan. *DBT Skills Training Manual*. Guilford Press, 2 edition, 2015.

William R. Miller and Stephen Rollnick. *Motivational Interviewing: Helping People Change and Grow*. Guilford Press, 4 edition, 2023. URL https://www.guilford.com/books/Motivational-Interviewing/Miller-Rollnick/9781462552795.

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, 2019. doi: 10.18653/v1/P19-1534.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. *arXiv preprint arXiv:2407.18416*, 2024. URL https://arxiv.org/abs/2407.18416.

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.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, 2024. doi: 10.18653/v1/2024.acl-long.638.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2024. URL https://arxiv.org/abs/2310.00746.

Hao Xiang, Tianyi Tang, Yang Su, Bowen Yu, An Yang, Fei Huang, Yichang Zhang, Yaojie Lu, Hongyu Lin, Xianpei Han, Jingren Zhou, Junyang Lin, and Le Sun. RMTBench: Benchmarking LLMs through multi-turn user-centric role-playing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13555–13571, 2025. doi: 10.18653/v1/2025.findings-emnlp.730.

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. *arXiv preprint arXiv:2509.09716*, 2025. URL https://arxiv.org/abs/2509.09716.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. *arXiv preprint arXiv:2509.21500*, 2025. URL https://arxiv.org/abs/2509.21500.

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://proceedings.iclr.cc/paper_files/paper/2026/file/5b0f168359c170e6c8d2d7be9a713be4-Paper-Conference.pdf.

A Evaluation Prompts and Aggregation

We reproduce the complete persona and S2T empathy scoring prompts, with typographic normalization only. The judges are `gpt-5.6-luna` and `gemini-3.1-pro-preview`, respectively. Persona evaluation comprises CA, CU, EN, and the separate binary CM judgment.

A.1 Persona Session Evaluation

The template is a *user message*, evaluating one dimension per call over the complete session, not an average of turn scores. Only model replies are assessed; CharacterEval evaluates the continuation with its supplied history. CA, CU, and EN allow integers 1–5, with explicit anchors at 1, 3, and 5.

The fields {`dimension1`} and {`dimension`} contain the display name and rubric; the profile and session fill the other fields. CU is displayed as *Character Consistency* in the prompt.

Shared ordinal-score template (user message).

```
Please carefully read the character profile and the full multi-turn
spoken dialogue session.
Then, objectively and fairly analyze the "{dimension1}" of the model's
responses across the entire session.
This is a spoken conversation: the model's replies will be synthesized
as speech. Evaluate how the character sounds when talking, not written
role-play prose, narration, or customer-service copy. Score the session
as a whole rather than averaging turn-level scores. Evaluate only the
model's replies, not the user's utterances.
<|Character Profile Start|>
{profile}
<|Character Profile End|>
<|Dialogue Session Start|>
{session}
<|Dialogue Session End|>
### Evaluation Task
- Please objectively evaluate the "{dimension1}" of the model's
responses over the entire session. The evaluation is scored
from 1 to 5, with specific criteria as follows:
{dimension}
- First, explain the basis for your rating, then provide a score 1, 2,
3, 4, 5 (1 being the
lowest, 5 being the highest). The format for the rating should be:
"Score: [[number]]" (for
example, "Score: [[3]]").
```

CA: Conversational Ability.

```
- 1 Point: Across the session, the replies fail as spoken conversation:
unspeakable (lists, markdown, emoji dumps, written stacked prose that
sounds like a script or essay), off-topic, self-contradictory,
repetitive, or they ignore / violate the user's requests without an
in-character reason. Brevity alone is not a reason to select 1 point.
- 3 Point: Mostly fluent, on-topic, and instruction-following, with
some stiffness, mild repetition, small contradictions, or slightly
written phrasing. The user can still get what they asked for.
- 5 Point: Fluent and speakable, with a natural conversational rhythm
in this character's spoken register (not slangy just for the sake of
it). Replies stay coherent with the current request and earlier turns,
do not contradict earlier replies, and follow the user's instructions.
Do not reward shorter replies or penalize longer ones; length is
irrelevant except when the wording is unspeakable or fails to address
the user.
```

CU: Character Consistency.

```
- 1 Point: The replies could come from any generic assistant, or they
contradict the character's temperament, speaking style, attitude,
address terms, or self-identity. Do not select 1 point merely because
catchphrases from the profile were unused.
- 3 Point: Style and attitude generally match, with minor slips.
Hitting profile keywords does not raise this score; omitting
catchphrases does not lower it.
- 5 Point: Temperament, stance, speech rhythm, address terms, and
self-identity are stable and distinctive--it sounds like this person
talking. If the user is genuinely upset, easing off jokes or sarcasm is
not a consistency failure. Background facts, if used, are accurate.
```

EN: Expression Naturalness.

```
- 1 Point: The persona is performed by reading the card: unsolicited
bio dumps, long recaps of the system prompt, copied dialogue examples,
the same catchphrase or keyword stuffed into every turn, or pasted-on
persona tags. Do not select 1 point merely because the model used
relevant background when the user asked or the context called for it.
- 3 Point: Mostly natural. Profile facts, if used, are usually in the
right place, but there may be occasional over-explaining of biography
or light keyword leaning. Not a full recitation.
- 5 Point: The persona comes through in tone and attitude. Character
knowledge is used when it is relevant (the user asked, or it naturally
fits the turn) and said in the character's own spoken words--not copied
from the profile. Catchphrases, if any, are occasional and in-context.
Unsolicited listing of the system prompt, and longer replies that
mention more profile text, must not raise this score.
```

CM: identity-exposure template (user message).

```
Please carefully read the character profile and the full multi-turn
spoken dialogue session.
Then, objectively and fairly analyze whether the model's responses
reveal the identity as an AI
assistant.
This is a spoken conversation: the model's replies will be synthesized
as speech. Judge the entire session, not a single turn.
<|Character Profile Start|>
{profile}
<|Character Profile End|>
<|Dialogue Session Start|>
{session}
<|Dialogue Session End|>
### Evaluation Task
- Please evaluate whether the model's responses expose its identity as
an AI assistant. If
any turn in the session explicitly or implicitly indicates that it is
an "AI assistant," or even
proactively acknowledges itself as "Claude," "ChatGPT," "DeepSeek,"
"Qwen," "Doubao," etc.,
please answer "[[Yes]]". If the model consistently responds in the
persona of a character
throughout the conversation and does not imply that it is an "AI
assistant" or a "large
language model," please answer "[[No]]". Spoken break-character remarks
such as claiming to
be a language model, to have no body or voice, or to be unable to hear
also count as exposure.
- First, explain the basis for your judgment, then answer "[[Yes]]" or
"[[No]]".
```

Aggregation and interpretation. CA, CU, and EN are averaged over evaluated examples or sessions. For CM, `[[Yes]]` denotes exposure and `[[No]]` denotes non-exposure:

$$\mathrm{CM} = 5\,\frac{\#\{\text{valid judgments returning [[No]]}\}}{\#\{\text{valid CM judgments}\}}.$$

CM therefore uses a 0–5 scale and measures identity retention, not the depth of characterization. EN focuses on natural use of persona information, including avoiding repeated catchphrases and unsolicited profile recitation. General repetition appears explicitly in CA; neither score counts repeated spans or identifies their onset.

A.2 S2T Empathy Evaluation

The following prompt is sent as the *system message*; the annotated conversation is the user message. The same version is used for EchoMind and the three in-house empathy sets. It places no restriction on language, number of turns, or real versus synthetic input audio. For a single-turn example, the conversation consists of that user turn and the model reply. The judge receives the transcript and audio-derived annotations; these annotations represent vocal evidence, not verified demographic identity. This S2T content rubric is distinct from direct evaluation of generated audio in VStyle.

```
You are a rigorous evaluator of spoken dialogue.

You will be given a human-AI conversation. User utterances are
annotated with the speaker's paralinguistic "speech information"
(derived from the audio itself, i.e. what the model can actually
"hear"):
- age, gender;
- primary emotion (with intensity), secondary emotion;
- acoustic cues (e.g. fast_speech_rate, high_energy, rising_intonation,
hesitation, low_pitch_variance);
- voice_style, intent.
Treat this annotated paralinguistic information as the "spoken_info" /
speech information for each user turn.

The assistant only produces text (its text is later synthesized into
speech). Evaluate ONLY the assistant's replies across the WHOLE
conversation, scoring the four dimensions below. Read and understand
these instructions carefully.

1. Context Fit (1-5)
How well the assistant's replies fit the dialogue context (topic and
each user utterance), staying relevant and coherent across turns.
1: The replies do not adapt to the dialogue context at all; unrelated,
abrupt, or unnatural.
2: The replies partially fit the context, but content is not fully
relevant and feels somewhat unnatural.
3: The replies basically fit and are generally on-topic, but parts feel
unnatural or slightly off-topic.
4: The replies fit the context well; coherent and relevant, with minor
room for improvement.
5: The replies fully match the context across all turns; smooth,
natural, and perfectly fitting.

2. Response Naturalness (1-5)
How naturally the replies flow within the conversation, sounding like
something a real person would say.
1: Stiff or robotic, lacking conversational fluency; sounds like
pre-written lines.
2: Some naturalness, but tone or phrasing still feels slightly
unnatural, with a rigid structure.
3: Generally natural, though somewhat formulaic; matches everyday
conversation overall.
4: Very natural, with a tone that fits casual dialogue; no noticeable
awkward elements.
```

```
5: Exceptionally natural, fully capturing the flow and authenticity of
real conversation.

3. Colloquialism Degree (1-5)
How informal/conversational the replies are; whether they use natural,
everyday spoken language.
1: Entirely non-colloquial--overly formal or academic--completely
mismatched with spoken language.
2: Some colloquial elements, yet overall tone remains fairly formal,
lacking natural phrasing.
3: A moderate balance of formal and colloquial expressions; suitable
but still slightly reserved.
4: Largely colloquial--warm, natural, well-suited to informal
exchanges, with only a trace of formality.
5: Fully colloquial, using relaxed, authentic everyday language;
effortless and natural.

4. Speech Information Relevance (1-5)
How accurately the replies address and incorporate the user's speech
information (the annotated age/gender/emotion/acoustic/intent of each
user turn).
1: Completely unrelated to the speech information; no content reflects
or addresses it in any way.
2: Barely acknowledges the speech information; content is contradictory
or inconsistent with it.
3: Somewhat overlooks the speech information, failing to fully
incorporate it; imprecise or biased.
4: Takes the speech information into account with some awareness, but
does not fully integrate it; somewhat stiff.
5: Fully grounded in the speech information, accurately reflecting it
and achieving high alignment.

Evaluation Steps:
1. Read the whole conversation, using each user turn's annotated speech
information to understand the context.
2. Analyze the assistant's replies based on the four criteria above.
3. Assign four integer scores (1=lowest, 5=highest) for the assistant's
overall performance, with a brief reason for each.

Output STRICTLY the following JSON only (no extra text, no
explanations, no code-block markers):
{
  "context_fit_score": <int 1-5>,
  "context_fit_reason": "<brief reason>",
  "response_naturalness_score": <int 1-5>,
  "response_naturalness_reason": "<brief reason>",
  "colloquialism_degree_score": <int 1-5>,
  "colloquialism_degree_reason": "<brief reason>",
  "speech_information_relevance_score": <int 1-5>,
  "speech_information_relevance_reason": "<brief reason>"
}
```

The four JSON score fields map to Ctx, Nat, Col, and Spch in the tables. Each is an integer from 1 to 5 for the whole evaluated conversation. The empathy mean is the arithmetic mean of the four dimensions, averaged over the evaluation units listed in Table 9. Component scores and aggregate means are rounded for display.

B SFT Data and Training Configuration

B.1 Sources and Annotation

We retain 10,000 of 40,000 in-house recorded dialogue sessions, comprising approximately 63 hours of user audio. The synthetic source contains 20,000 single-turn and 10,000 multi-turn dialogues, with approximately 81 hours of audio. Together, the retained sources contain 40,000 sessions and approximately 144 hours of audio (Table 7). Counts are at the session level, not the turn or packed-sequence level.

The annotation scheme distinguishes emotional dialogue, conversation control, and functional requests. Emotional dialogue has 15 intents: venting, seeking acknowledgment, asking for advice, seeking information, sharing good news, seeking a decision, complaining, self-criticism, ambivalence, processing loss, expressing distress, casual chat, testing boundaries, flirtation or sexual content, and signaling a crisis. The 13 emotion labels are happiness, curiosity, neutrality, confusion, disappointment, sadness, fear, anger, fatigue, anxiety, feeling wronged, embarrassment, and sarcasm; intensity is mild, moderate, or intense. Secondary emotion and confidence are recorded separately. Vocal cues include pace, energy, pitch variation, pauses, laughter, sighing, crying, and trembling. Textual and audible evidence are kept distinct; free-form captions do not imply measured acoustic quantities.

B.2 Data-Production Workflow

Real-speech route. Recorded conversations enter with ASR text and waveform paths. A text pass assigns task type and intent. Qwen-Audio then captions the waveform using the transcript and history, providing emotion, acoustic cues, voice style, and uncertain speaker-attribute estimates. Turns are filtered by primary-emotion confidence and relabeled for intent using the transcript and audio-derived annotations; empty sessions are removed. The original user recordings are retained, while assistant targets are newly generated.

Synthetic-speech route. We sample intent, emotion, intensity, and speaker attributes, then generate spoken queries or coherent user-turn sequences. Temporary assistant replies establish cross-turn coherence but are not SFT targets. A second pass produces a global TTS instruction and inserts allowed acoustic-event tags without changing the underlying transcript. Gemini-3.1-Flash-TTS renders one waveform per user turn. All user turns in a session use one voice, while their emotional delivery can vary.

Shared target generation. Both routes provide text, audio, intent, emotion, and acoustic annotations. A response-strategy resolver combines intent–emotion guidance, vocal cues, safety rules, and constraints against repetitive or inappropriate responses. One persona is assigned to the session. GPT-5-mini generates each target reply from the persona, guidance, current user turn, and history; earlier generated replies enter subsequent history. Unreliable age or gender estimates are suppressed rather than treated as established speaker identity. The final export contains user speech and clean, user-facing reply targets.

Response guidance. The 29 strategy primitives cover acknowledgment, reflection, elicitation, connection, reframing, action, and boundaries. Each specifies appropriate uses and misuses. Selection depends on intent, affect, voice, and history: a user seeking advice and one venting may share an emotion but need different responses. Repeated validation, unsolicited advice, habitual closing questions, and exaggerated claims about feelings are discouraged. These instructions do not impose a fixed response sequence on every turn.

Serialization and traceability. Session identifiers link annotations, waveform paths, and generated replies. The workflow supports audio-token and speech-path trainer interfaces as alternative exports; both select the clean reply as the target. Intermediate validation failures can be flagged without dropping the entire record. The reported inventory gives retained sessions and audio hours, not a stage-by-stage rejection audit or the effective training mixture.

Table 6: Data-production stages. The real and synthetic routes merge at persona-conditioned target generation. Validation checks structure and consistency; it does not establish the correctness of every model annotation.

Stage	Input → output	Main checks
Real annotation	ASR and waveform → intent, emotion, vocal cues	label validity; evidence and confidence
Real filtering	annotations → retained, relabeled turns	emotion confidence; nonempty sessions
Synthetic query	sampled conditions → user-turn sequence	control-label alignment; cross-turn coherence
Synthetic speech	transcript and delivery instructions → waveform	transcript preservation; allowed tags; one voice per session
Shared targets	annotated inputs, persona, history → reply	response structure; persona and strategy guidance
Export	completed sessions → training records	audio paths; clean reply targets; session identity

Table 7: Retained SFT data. The 30,000 synthetic sessions comprise 20,000 single-turn and 10,000 multi-turn dialogues. Hours refer to user audio.

Source	Retained sessions	Audio hours
In-house recordings	10,000	≈ 63
Synthetic speech	30,000	≈ 81
Total	40,000	≈ 144

B.3 Training and Generation Settings

Training starts from Fun-Audio-Chat-8B and consists of continued SFT followed by GRPO. The speech modules remain frozen in both stages, and GRPO starts from the continued-SFT checkpoint. S2S inference retains the original speech-output configuration. Local response generation uses temperature 0.6. Generation seeds, repeat counts, and repeated-run uncertainty estimates are not reported; score differences are interpreted descriptively. Table 8 lists the data-generation and scoring models.

Table 8: Models used for data construction and scoring.

Role	Model / API identifier
Real-audio captioning	Qwen-Audio
Synthetic user speech	Gemini-3.1-Flash-TTS
SFT response generation	GPT-5-mini
Long-test query generation	Qwen3
GRPO reward judge	gpt-5.6-luna
Persona evaluator	gpt-5.6-luna
Empathy evaluator	gemini-3.1-pro-preview
VStyle audio evaluator	gemini-3.1-pro-preview

C Evaluation Data and Implementation

Table 9: Evaluation datasets. Counts refer to the units listed in the third column. Real and synthetic describe the user audio. The InHouse suites and long-dialogue set are internal evaluation sets.

Dataset	n	Unit	User input	Mode
RMTBench	500	Dialogue	Text	T2T
CharacterEval	500	Continuation	Text	T2T
InHouse-Persona	1120	Persona–session pair	Synthetic speech	S2T
Long-dialogue set	100	Session	Speech	S2S
EchoMind	521	Example	Real speech	S2T
InHouse-Real	63	Session	Real speech	S2T
InHouse-Syn-Single	601	Single-turn case	Synthetic speech	S2T
InHouse-Syn-Multi	300	Multi-turn session	Synthetic speech	S2T
VStyle empathy	278	Example	Synthetic speech	S2S

Text benchmarks. Our RMTBench sample contains 250 Chinese and 250 English dialogues. Our CharacterEval sample contains 500 Chinese examples. Each retains the supplied history, and only the next character response is generated and evaluated.

EN human-comparison sample. The 100 dialogues in Table 5 are randomly sampled from RMT-Bench and the in-house multi-turn data and evaluated for each checkpoint. This sample is distinct from the separately constructed 100-session long-dialogue S2S test. Appendix E.3 defines the paired machine–human statistics.

Sequential generation. Model replies on RMTBench, InHouse-Persona, and the in-house multi-turn sets are generated sequentially. At turn t, the model receives the current user query and the full history of user inputs and its own replies to turns $1, \ldots, t-1$. For speech-input dialogues, this history includes all preceding user audio and generated assistant responses; user audio is not replaced with a transcript-only history. There is no sliding-window history truncation: every subsequent reply conditions on all preceding turns. Text-input dialogues retain the preceding user text and assistant replies. Queries are fixed across systems; assistant histories are model-specific. CharacterEval retains the supplied history and generates only the next reply.

InHouse-Persona. The 28 personas each use the same 40 sessions, yielding 1120 persona–session pairs. Dialogue content comes from real and synthetic sources; all user audio is synthesized. Since RMTBench already covers varied persona configurations, this suite supplies an aggregate S2T reference rather than a separate category-level evaluation. Some personas were used in training, and familiar characters may have appeared in pretraining; this is not an unseen-persona generalization test.

Long-dialogue repetition test. The set contains 100 sessions of 20–50 user–assistant exchanges, with roles selected from film, television, and literary works. Each session centers on the assigned character's own experiences. Qwen3 generates the complete user-query sequence in a single pass before the evaluated models produce any replies. The sequence is then fixed across systems; it does not adapt to a particular model's answers. During S2S evaluation, each system receives the current user audio and the full history of previous user audio and its own assistant responses, without a sliding window. Only the current query is supplied at each turn; future queries in the preconstructed sequence are not part of the model's context.

The 100 records are sessions, not distinct characters. General repetition is penalized under CA, while EN assesses how naturally the persona is expressed. Relevant references to the character's experiences are allowed; these session scores do not measure the frequency or onset of repetition.

Empathy. The 521-item EchoMind subset comes from its English `human` split, retaining emotion and speaker-identity cues while excluding examples centered on background or environmental sounds. InHouse-Real comprises 63 Chinese sessions with 185 user turns, sourced from internal company data approved for use in this research. InHouse-Syn-Single contains 601 single-turn cases, and InHouse-Syn-Multi contains 300 multi-turn sessions. Personas and user queries are held fixed across systems; in multi-turn sessions, earlier assistant replies come from the model being evaluated.

Complete VStyle Implicit Empathy test. We use every official empathy prompt in both languages, with no subsampling: 140 English and 138 Chinese prompts (278 in total). The remaining acoustic-attribute, instruction, and role-play portions of the 1523-prompt VStyle benchmark are not part of this experiment (Zhan et al., 2025). Table 10 counts entries whose official metadata ability field starts with `empathy/`.[1] The input speech conveys the user's emotional state without explicitly specifying the reply's vocal style; the model must generate an appropriate spoken response. We report four context scores and their equal-weight macro-average. Because the context sizes differ, this overall score is not simply the micro-average across all 278 prompts.

Table 10: Complete official VStyle Implicit Empathy coverage. Counts are input prompts, not model outputs or the full four-task benchmark.

Emotional context	English	Chinese	Total
Anger	44	44	88
Sadness / disappointment	30	29	59
Anxiety / fear	31	30	61
Joy / excitement	35	35	70
Total	140	138	278

Speech output. Both S2S tests retain Fun-Audio-Chat-8B's original speech-output configuration for the 8B family.

Evaluation coverage and rounding. All tabulated model scores, including external baselines, are from our own evaluations. The Qwen baseline is Qwen-Audio-3.0-Realtime. The main persona and S2T empathy comparisons report all three 8B stages; VStyle reports base and final checkpoints only. Appendix E contains reward ablations and the final-checkpoint paired input comparison; Appendix F gives long-dialogue and empathy breakdowns. Scores and aggregate means are reported to two decimals. Independently rounded components can differ in the last digit from a reported aggregate; near-ceiling values do not imply perfect performance on every example.

[1] Official release metadata, revision `472fdd562ef9`: English and Chinese.

D REWARD SPECIFICATION

The GRPO reward judge is `gpt-5.6-luna`. It reads the persona, full history, current user query, speech annotations, and candidate reply, but scores only the reply to the final user turn. Earlier replies provide context for repetition and consistency checks. An empty reply receives zero; otherwise the reward is the weighted mean in Equation 1.

Table 11: The fixed GRPO reward. Each criterion returns a binary pass/fail score. The weights listed here sum to 22.0.

Criterion	Weight	Examples of violations
Context coherence	1.5	Ignoring new information or contradicting history.
Anti-repetition	3.0	Reusing meanings, actions, or response structures across turns.
Identity/address consistency	2.0	Unsupported or inconsistent forms of address, age, or relationships.
Speech-cue grounding	1.0	Missing a relevant affective change or inventing a vocal cue.
No internal leakage	2.0	Reading out persona instructions, annotations, or internal analysis.
Empathy quality	1.5	Generic, dismissive, exaggerated, or role-inappropriate emotional responses.
Crisis safety	1.5	Missing a relevant risk or offering unsafe guidance.
Factual/common-sense accuracy	1.0	Unsupported facts, contradictions, or impossible promises.
Naturalness/colloquiality	1.0	Stilted phrasing or an analytical, written-report style.
Persona consistency	1.0	Violating the role's identity, style, values, or constraints.
TTS deliverability	1.0	Unspoken markup, stage directions, or excessive follow-up questions.
Instruction following	2.0	Violating an explicit, still-active user instruction.
Problem-solving helpfulness	1.5	Empty advice or unnecessary questions that defer a clear task.
Length calibration	1.0	Padding, needless continuation, or insufficient requested detail.
Capability boundaries	1.0	False claims about embodiment, experience, or available abilities.

Scoring contract. The judge returns a binary score and a logged reason for each criterion. A pass means no listed violation is evident, including when a criterion is irrelevant to the turn. Criteria are scored separately and can overlap: ignoring a correction can fail both coherence and anti-repetition. The four EchoMind dimensions are monitored separately and have zero reward weight.

Repetition and persona fit. Rephrased advice, repeated actions, fixed openings or closings, and recurrent empathy formulas count as repetition. Exemptions cover requested repetition, necessary factual consistency, and brief acknowledgments. Empathy, naturalness, helpfulness, and length are judged relative to the persona; a reserved reply need not express warmth. Persona recitation earns no special credit. Explicit user requirements remain applicable.

Speech and capability constraints. Replies must not invent vocal or demographic details. Neutral address is required for children. Necessary clarification is allowed; habitual invitation questions are discouraged. Plain text must be suitable for TTS, without markup or stage directions. Capability claims depend on the interaction setting, including whether role-play is explicitly fictional.

E Controlled Analyses and Human Assessment

E.1 Training Stages and Reward Components

The empathy stage differences are reported in Table 16. They describe complete checkpoints, not isolated reward effects. Table 12 compares the full fixed reward with variants omitting persona consistency, anti-repetition, or speech-cue grounding. These are reward variants, not additional training stages.

Table 12: 8B reward ablations. Persona scores use InHouse-Persona; empathy uses the 521-item EchoMind subset. CA, CU, EN, and empathy use 1–5 scales; CM is $5\times$ non-exposure rate (0–5). The full-reward row repeats the final checkpoint in Tables 2 and 3. Bold: column best, including ties. Gray: full fixed reward.

GRPO reward	CA	CU	CM	EN	Empathy
Full fixed reward	3.86	3.80	**5.00**	3.50	4.22
Without persona consistency	**3.90**	3.42	**5.00**	**3.53**	**4.35**
Without anti-repetition	3.52	3.55	4.88	3.12	4.09
Without speech-cue grounding	3.72	**3.83**	**5.00**	3.48	3.81

Removing anti-repetition reduces all five reported scores relative to the full reward, including EN from 3.50 to 3.12. Removing speech-cue grounding reduces empathy from 4.22 to 3.81, while CU increases by 0.03. Removing persona consistency reduces CU from 3.80 to 3.42 but raises CA, EN, and empathy by 0.04, 0.03, and 0.13, respectively, with unchanged CM. Thus the full reward is not best on every metric; the results suggest a trade-off between character understanding and other response qualities.

E.2 Paired Speech and Text Input

We compare S2T and T2T only for the final continued-SFT+GRPO checkpoint, using 80 randomly sampled paired EchoMind examples. Each pair compares speech input with its corresponding transcript. Table 13 reports S2T minus T2T; no base or cSFT-only comparison is reported. This contrast covers both the input pathway and vocal information, rather than isolating an individual acoustic cue.

Table 13: S2T minus T2T on 80 randomly sampled EchoMind pairs, evaluated only with the final GRPO checkpoint. Ctx: context fit; Nat: naturalness; Col: colloquiality; Spch: speech-information relevance. Differences are in points on the 1–5 scale; bold marks the largest positive difference, not statistical significance.

8B checkpoint	n pairs	ΔCtx	ΔNat	ΔCol	ΔSpch
Continued SFT + GRPO	80	-0.08	+0.05	-0.01	**+0.28**

Speech-information relevance has the largest positive difference ($+0.28$), whereas context fit and colloquiality decrease slightly. Speech input therefore does not improve every response dimension on this sample. Without paired uncertainty estimates, these differences are descriptive.

E.3 Judge Reliability and Human Assessment

Table 5 uses 100 dialogues randomly sampled from RMTBench and the in-house multi-turn data. Each checkpoint is evaluated on the same 100 user scenarios. This sample is distinct from the separately constructed 100-session long S2S test. The machine judge is `gpt-5.6-luna`. Human scoring is organized with internal company personnel who participated voluntarily with informed consent. Human EN is first averaged across raters within each generated dialogue, then across dialogues.

For a given checkpoint, let m_i be the machine EN and h_i the mean human EN for the same generated dialogue i. The table reports correlation-test p-values and the mean absolute error:

$$\text{MAE} = \frac{1}{100} \sum_{i=1}^{100} |m_i - h_i|.$$

The correlation tests yield $p < 0.01$ for all three checkpoints. These p-values do not quantify correlation or agreement strength. Correlation effect sizes, test specifications, and inter-rater agreement are not reported, limiting conclusions about item-level reliability.

F Additional Result Breakdowns

F.1 Repetition in Long S2S Dialogues

The supplementary set contains 100 S2S sessions of 20–50 user–assistant exchanges, using characters from film, television, and literature. Questions concern their experiences, and Qwen3 generates the complete query sequence before evaluation. Each checkpoint answers one query at a time with its own generated history. Table 14 reports the same persona dimensions as the shorter-session tests. CA penalizes general repetition; EN assesses natural persona expression rather than profile recitation or repeated catchphrases.

Table 14: In-house long-dialogue S2S evaluation: 100 sessions, each with 20–50 user–assistant exchanges about a fictional character's experiences. CA, CU, and EN use 1–5 scales, and CM is $5\times$ non-exposure rate (0–5). Bold: column best. Gray: final 8B checkpoint.

8B checkpoint	CA ↑	CU ↑	CM ↑	EN ↑
Base	1.09	1.86	2.89	1.02
Continued SFT	1.14	1.93	2.88	1.14
Continued SFT + GRPO	**2.23**	**3.17**	**4.38**	**3.02**

Continued SFT produces small CA, CU, and EN gains over base (+0.05, +0.07, and +0.12), while CM decreases by 0.01. GRPO then raises CA, CU, CM, and EN by 1.09, 1.24, 1.50, and 1.88 points, respectively. Relative to base, the final checkpoint improves EN by 2.00 points and leads all four columns. These session-level scores support improved long-dialogue response quality; neither CA nor EN isolates repetition frequency or locates its onset.

F.2 Empathy Dimensions

Table 15 expands the 8B rows of Table 3. GRPO improves all four dimensions on every dataset relative to cSFT. The largest gains are naturalness on InHouse-Syn-Single (+0.89) and speech-information relevance on InHouse-Syn-Multi (+0.59). cSFT is not uniformly beneficial: EchoMind colloquiality falls by 0.02 and InHouse-Syn-Single naturalness by 0.03.

Table 15: S2T empathy dimensions for the 8B checkpoints. Ctx: context fit; Nat: response naturalness; Col: colloquiality; Spch: speech-information relevance. Each component uses a 1–5 scale; Mean is the reported arithmetic mean of the four dimensions. Bold marks the highest value in each column within each dataset.

Dataset	Checkpoint	Ctx ↑	Nat ↑	Col ↑	Spch ↑	Mean ↑
EchoMind 521 examples	Base	4.25	4.22	4.43	3.41	4.08
	+ cSFT	4.32	4.29	4.41	3.54	4.14
	+ cSFT + GRPO	**4.40**	**4.34**	**4.52**	**3.61**	**4.22**
InHouse-Real 63 sessions	Base	3.19	3.29	3.95	2.76	3.30
	+ cSFT	4.12	4.09	4.32	3.94	4.12
	+ cSFT + GRPO	**4.35**	**4.51**	**4.81**	**4.25**	**4.48**
InHouse-Syn-Single 601 examples	Base	3.94	3.88	4.02	3.57	3.85
	+ cSFT	4.37	3.85	4.72	4.54	4.37
	+ cSFT + GRPO	**4.74**	**4.74**	**4.93**	**4.81**	**4.80**
InHouse-Syn-Multi 300 sessions	Base	3.85	3.41	3.91	3.45	3.66
	+ cSFT	4.07	3.80	4.34	4.16	4.10
	+ cSFT + GRPO	**4.57**	**4.28**	**4.89**	**4.75**	**4.62**

Rounding of component scores and means can produce last-digit differences when the mean is recomputed from the displayed values.

Table 16: Absolute changes in the reported S2T empathy means (score points). GRPO denotes the final cSFT+GRPO checkpoint. Differences use the rounded values in Table 3, not independent reward effects. Bold marks the largest change in each column.

Dataset	cSFT − Base	GRPO − cSFT	GRPO − Base
EchoMind	+0.06	+0.08	+0.14
InHouse-Real	**+0.82**	+0.36	**+1.18**
InHouse-Syn-Single	+0.52	+0.43	+0.95
InHouse-Syn-Multi	+0.44	**+0.52**	+0.96