Optimizing Conversational Quality in Spoken Dialogue Systems with Reinforcement Learning from AI Feedback

Siddhant Arora[1], **Jinchuan Tian**[1], **Jiatong Shi**[1], **Hayato Futami**[2], **Yosuke Kashiwagi**[2], **Emiru Tsunoo**[2], **Shinji Watanabe**[1]
[1] Carnegie Mellon University, USA
[2] Sony Group Corporation, Japan
{siddhana}@cs.cmu.edu

Abstract

Reinforcement learning from human or AI feedback (RLHF/RLAIF) for speech-in/speech-out dialogue systems (SDS) remains under-explored, with prior work largely limited to single semantic rewards applied at the utterance level. Such setups overlook the multi-dimensional and multi-modal nature of conversational quality, which encompasses semantic coherence, audio naturalness, speaker consistency, emotion alignment, and turn-taking behavior. Moreover, they are fundamentally mismatched with duplex spoken dialogue systems that generate responses incrementally, where agents must make decisions based on partial utterances. We address these limitations with the first multi-reward RLAIF framework for SDS, combining semantic, audio-quality, and emotion-consistency rewards. To align utterance-level preferences with incremental, blockwise decoding in duplex models, we apply turn-level preference sampling and aggregate per-block log-probabilities within a single DPO objective. We present the first systematic study of preference learning for improving SDS quality in both multi-turn Chain-of-Thought and blockwise duplex models, and release a multi-reward DPO dataset to support reproducible research. Experiments show that single-reward RLAIF selectively improves its targeted metric, while joint multi-reward training yields consistent gains across semantic quality and audio naturalness. These results highlight the importance of holistic, multi-reward alignment for practical conversational SDS.

1 Introduction

Spoken dialogue systems (SDS) are rapidly evolving from turn-based voice assistants (Face, 2024; Xie and Wu, 2024) toward real-time, full-duplex conversational agents (Nguyen et al., 2023; Meng et al., 2024; Zhang et al., 2024d; Défossez et al., 2024) capable of listening and speaking simultaneously. Advances in end-to-end architectures, speech foundation models (KimiTeam et al., 2025; Tian et al., 2025a; Arora et al., 2025a), and duplex decoding (Veluri et al., 2024; Zhang et al., 2024c) have enabled more natural interactions, reduced latency, and improved conversational flow. Yet, despite these modeling breakthroughs, achieving human-preferred conversational behavior, semantic coherence, natural prosody, emotionally aligned delivery, and responsive turn-taking, remains an open challenge (Lin et al., 2025a). As systems become more sophisticated, small errors in timing, prosody, or semantic drift can accumulate across turns, directly degrading user experience.

These challenges highlight a growing alignment gap between advances in duplex modeling and the ability to optimize SDS for human-preferred conversational behavior. While there have been efforts to apply reinforcement learning from human or AI feedback (RLHF / RLAIF) as a post-training mechanism in text-based dialogue models and cascaded systems (Yoshida et al., 2025; Su et al., 2016; Serban et al., 2018), systematic investigations of preference learning for end-to-end (E2E) spoken dialogue systems remain scarce. To address this gap, recent work (Chen et al., 2024a; Zhang et al., 2024a; Cao et al., 2025) has begun applying RLHF / RLAIF to speech-based systems, but each explores only a narrow slice of the alignment space. One such work (Wu et al., 2025) shows that large-scale semantic preference learning can improve factuality, safety, and coherence in a deployed duplex SDS by transcribing real user–agent dialogues and using an LLM judge to generate preference data. AlignSLM (Lin et al., 2025b) demonstrates that LLM-based semantic evaluation can effectively guide preference optimization in speech-to-speech language models, producing more coherent and less repetitive continuations. ORISE (Chen et al., 2025) focuses on temporal alignment, using online RL with heuristic, audio-driven rewards to improve turn-taking behavior and responsiveness without

Findings of the Association for Computational Linguistics: ACL 2026, pages 41049–41066

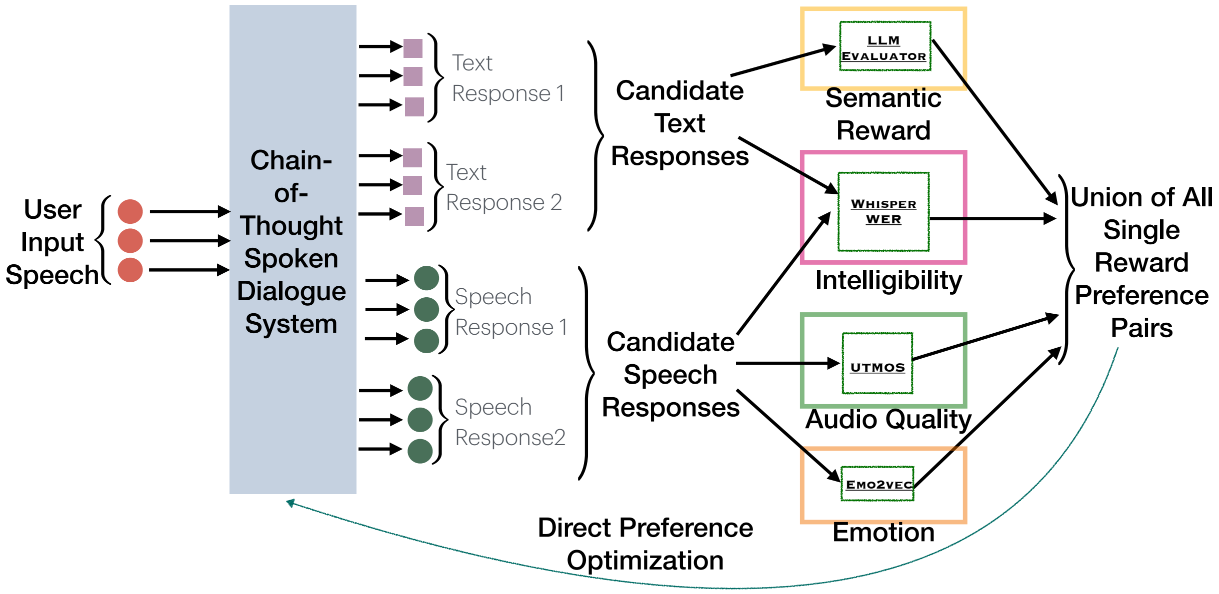

Figure 1: Overview of the proposed dataset-level Multi-Reward RLAIF framework. Unlike prior work that optimizes single rewards (typically semantic quality), we construct four independent preference datasets targeting semantic coherence, audio naturalness, intelligibility, and emotion consistency. These datasets are **jointly sampled** within a single DPO objective, rather than being applied in a sequential manner. The framework supports both turn-by-turn CoT and blockwise duplex SDS by aggregating per-block log-probabilities under utterance-level preferences (S. 4.2)

requiring labeled data. Together, these works illustrate the promise of RLHF for SDS.

However, these approaches still fall short of what real-time SDS demand. First, all existing methods optimize a single reward dimension, typically semantic quality or turn-taking, even though conversational quality is inherently multi-objective, jointly shaped by semantics, audio naturalness, speaker style consistency, emotional tone, and responsiveness. Second, most current RLHF pipelines assume full-utterance feedback, which is incompatible with duplex SDS where the model must commit to partial utterances and make in-flight decisions before sentence boundaries are detected. Third, while audio-based preference learning has recently gained traction in text-to-speech (TTS) systems (Chen et al., 2024a; Tian et al., 2025b), it remains largely unexplored in the context of SDS. Existing SDS-focused RLHF approaches primarily optimize semantic or timing-related objectives and do not incorporate audio-level rewards, such as MOS-based quality estimators or emotion similarity with human reference responses, within a unified dialogue-level learning framework. As a result, the potential of audio-centric rewards to jointly improve conversational quality in SDS has yet to be systematically studied.

To overcome these limitations, ① we propose a multi-reward RLAIF framework for speech-in/speech-out SDS. Rather than introducing a new SDS architecture, our contribution lies in a general alignment strategy that optimizes conversational quality across multiple dimensions, including semantic coherence, audio naturalness, intelligibility, and emotion consistency. Our framework is orthogonal to the underlying model architecture and can be directly applied to stronger or larger SDS backbones without modification. ② We address the mismatch between incremental decoding in duplex SDS and sentence-level rewards by introducing a streaming preference learning formulation that applies utterance-level feedback to partial, blockwise generation, enabling RLAIF to shape in-flight decisions in duplex streaming models (Arora et al., 2025c). Applied to multi-turn Chain-of-Thought SDS models and duplex architectures, our framework delivers consistent improvements across semantic and acoustic metrics. This work establishes a unified foundation for holistic alignment of SDS, moving beyond semantic-only optimization toward agents that sound better and interact more naturally. ③ Finally, we will publicly release the first large-scale multi-reward preference learning dataset for SDS, together with all training and inference code, enabling reproducible RLAIF research for speech-in/speech-out conversational agents.

2 Problem Formulation

As modelled in prior works (Arora et al., 2025c), SDS take a d-dimensional continuous feature stream $X = (\mathbf{x}_t \in \mathbb{R}^d \mid t = 1, \ldots, T)$ as input audio from the user and generate a synchronized spoken response $Y^{\text{sds}} = (\mathbf{y}_t^{\text{sds}} \in \mathbb{R}^d \mid t = 1, \ldots, T)$, where T is the total duration of the conversation, including multiple turns. The objective of an SDS is to generate speech outputs that are coherent, natural, and contextually appropriate, typically modeled by estimating the conditional distribution $P(Y^{\text{sds}}|X, X^{\text{spk}})$, where X^{spk} represents a speaker or style prompt controlling vocal characteristics (e.g., timbre, prosody, or emotion)

3 Spoken dialogue systems

3.1 Turn by turn SDS

Traditional SDS (Glass, 1999) are commonly designed around a turn-based interaction paradigm, in which the system alternates between listening and speaking. Under this formulation, the continuous input–output streams are partitioned into a sequence of turns, i.e. $Y^{\text{sds}} = \{Y_k^{\text{sds}} \mid k = 1, \ldots, K\}$ and $X = \{X_k \mid k = 1, \ldots, K\}$ where K denotes the total number of turns. Given this turn-based interaction, the system models the overall speech output distribution using the causality-based conditional independence (C.I.) assumption:

$$P(Y^{\text{sds}}|X, X^{\text{spk}}) = \prod_k P(Y_k^{\text{sds}}|Y_{1:k-1}^{\text{sds}}, X_{1:k}, X^{\text{spk}}) \tag{1}$$

Thus, the objective of a turn-by-turn SDS is to predict, at each turn, the response that maximizes this turn-level conditional probability.

Prior work (Arora et al., 2025d) has explored incorporating structured intermediate representations, such as ASR transcripts S_k^{asr} and textual responses S_k^{res}, within an E2E SDS framework, yielding Chain-of-Thought (CoT) E2E SDS models. Further details on the mathematical formulation of the CoT SDS proposed in (Arora et al., 2025d) are in S. A.1. We adopt this CoT formulation as the turn-by-turn SDS backbone in our experiments.

3.2 Duplex SDS

A growing body of work has explored duplex, or simultaneous speaking-while-listening, spoken dialogue systems, motivated by the need to support more natural, low-latency interactions that better resemble human conversational behavior (Masumura et al., 2018; Roddy et al., 2018; Skantze, 2017; Meena et al., 2014; Ekstedt and Skantze, 2020).

One common strategy for building duplex SDS is the time-multiplexing approach, which alternates between processing *fixed*-duration segments of user input, and generating spoken output for the subsequent segment (Veluri et al., 2024). Building on earlier turn-by-turn CoT formulations for E2E SDS, recent work, namely SCoT, has investigated how structured intermediate reasoning can be integrated into blockwise duplex systems (Arora et al., 2025c). Further details on the mathematical formulation of SCoT (Arora et al., 2025c) can be found in S. A.2. In this work, we build on SCoT-style duplex models and focus on aligning blockwise generation with utterance-level preference learning.

4 RLAIF Post-training Framework

Having introduced both turn-based and duplex SDS architectures, we now describe how these models are aligned with human-preferred conversational behavior through preference optimization. The specific reward functions employed in this work are detailed later in Section 5.

4.1 Preference Optimization for Turn-by-Turn SDS

While the CoT-based SDS formulation (Arora et al., 2025d) provides a structured autoregressive model for predicting turn-level responses, it does not guarantee that the generated speech aligns with human conversational preferences. To address this limitation, we adopt reinforcement learning from AI feedback (RLAIF) paradigm (Bai et al., 2022; Lee et al., 2023) to simulate preference signals, and apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) as a post-training objective. This enables alignment of the turn-level SDS policy with multi-reward preference signals spanning semantic, acoustic, and stylistic dimensions (see Section 5).

Turn-Level Preference Pairs. For each user turn k, we generate a set of n candidate spoken responses

$$\mathcal{Y}_k = \{Y_{k,1}^{\text{sds}}, \ldots, Y_{k,n}^{\text{sds}}\}.$$

Using multi-reward scoring, we select one or more preferred–dispreferred response pairs from this candidate set, forming preference pairs

$$(Y_k^+, Y_k^-) \in \mathcal{Y}_k \times \mathcal{Y}_k, \qquad Y_k^+ \succ Y_k^-,$$

where Y_k^+ and Y_k^- denote the positive and negative responses according to the reward criteria.

DPO Loss for Turn-Level SDS. Given a preference pair (Y_k^+, Y_k^-) at turn k with turn history

$\mathcal{H}^{(k)} = \left(X_{1:k},\, Y^{\text{sds}}_{1:k-1}, X^{\text{spk}}\right)$, the Direct Preference Optimization (DPO) loss is defined as

$$\mathcal{L}^{(k)}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\Delta_\theta(k) - \Delta_{\text{ref}}(k)\big]\Big), \quad (2)$$

where

$$\Delta_\theta(k) = \log \pi_\theta\Big(Y^+_k \mid \mathcal{H}^{(k)}\Big) - \log \pi_\theta\Big(Y^-_k \mid \mathcal{H}^{(k)}\Big), \quad (3)$$

and $\Delta_{\text{ref}}(k)$ is defined analogously using the frozen reference policy π_{ref}. Here, π_θ denotes the learnable SDS policy being optimized, while π_{ref} is a fixed copy of the same model *before* preference learning (i.e., the supervised fine-tuned SDS checkpoint). The reference policy serves as an anchor that stabilizes optimization by preventing large deviations from the original behavior, following standard DPO practice. The function $\sigma(\cdot)$ denotes the logistic sigmoid, and β controls the sharpness of the preference margin.

4.2 Preference Optimization for Duplex SDS

Although preference labels are assigned only at the *utterance level*, duplex SDS models like SCoT generate partial speech segments in a *blockwise* manner. To ensure that utterance-level DPO training remains compatible with blockwise generation, we decompose each candidate spoken response into fixed-size blocks and compute its log-probability by aggregating blockwise contributions.

Blockwise Representation. Given a turn-level response Y^{sds}_k, we partition it into blocks of size N_{block}, i.e. $Y^{\text{sds}}_k = \{Y^{\text{sds}}_{k,b} \mid b = 1, \dots\}$. This mirrors the block sequence used in duplex SDS for the user speech X as discussed in S. A.2.

Blockwise Factorization of the Utterance Probability. To simplify notation, we define the *turn history* up to block b as

$$\mathcal{H}^{(k)}_b = \mathcal{H}^{(k)} \cup \left(X_{k,1:b-1}, Y^{\text{sds}}_{k,1:b-1}\right)$$

which augments the turn history $\mathcal{H}^{(k)}$ with all user input and system output blocks generated prior to block b within the current turn. Under the same causality-based conditional independence assumption used in time-multiplexed SDS such as SCoT (Arora et al., 2025c), the posterior over the full response factorizes block-by-block as:

$$\log \pi_\theta(Y^{\text{sds}}_k \mid X_{1:k}, Y^{\text{sds}}_{1:k-1}) = \sum_b \log \pi_\theta\Big(Y^{\text{sds}}_{k,b} \mid \mathcal{H}^{(k)}_b\Big), \quad (4)$$

This formulation enables us to use preference pairs (Y^+_k, Y^-_k) are constructed at the utterance level, as in turn-by-turn SDS (S. 4.1), and no reward labels are assigned to individual blocks. During training, the utterance-level preference is applied by substituting the aggregated log-probabilities in Eq. 4 into the DPO objective (Eq. 3). This strategy avoids the need to define reward signals at the partial-utterance level, an aspect that is non-trivial for most conversational and perceptual metrics, while still ensuring that every block contributes to optimizing the preferred response.

5 Reward Functions

In this work, we adopt a dataset-level formulation of multi-reward DPO. Rather than defining a single preference pair using multiple reward constraints simultaneously, we construct independent preference datasets, each targeting a specific dimension of conversational quality (semantic coherence, audio quality, intelligibility, emotion consistency). These datasets are then jointly used during DPO post-training (see S. A.8 for more details), enabling the SDS policy to be optimized across complementary objectives within a unified framework.

5.1 Semantic Quality Reward

To assess the semantic quality of generated spoken responses, we adopt the LLM-based semantic evaluation framework introduced in Align-SLM (Lin et al., 2025b) and adapt it to the SDS setting. For each user turn, the CoT SDS samples multiple candidate intermediate text responses $\hat{S}^{\text{res}}_k$ (Eq. (15) in Appendix A.1 for more details), which are scored by Qwen2.5-72B-Instruct (Qwen et al., 2025) as the LLM judge to construct DPO data (See A.9 for LLM prompt). The judge rates each candidate on coherence, relevance and grounding with dialogue context, producing a scalar semantic score between 0 and 10. In parallel, we follow prior work (Lin et al., 2025b) and also compute an *AutoBLEU* (Lakhotia et al., 2021) score for each candidate, which serves as an automated filter: high AutoBLEU values indicate repetitive phrasing or low-diversity text that lacks meaningful semantic content. Candidates with an LLM coherence score above a positive threshold τ_{pos} *and* AutoBLEU below a repetition threshold δ_{low} are labeled as *positive* samples. Conversely, candidates with LLM scores below a negative threshold τ_{neg} or AutoBLEU exceeding δ_{high} are labeled as *negative* samples. This procedure yields high-quality semantic preference pairs for DPO training.

The semantic judge operates on *text* responses and therefore produces a preference pair $\hat{S}_k^{\text{res},+} \succ \hat{S}_k^{\text{res},-}$ at each turn. To apply DPO (Eq. (2)) using this signal, we relate the turn-level speech policy $\pi_\theta(Y_k^{\text{sds}} \mid X_{1:k}, Y_{1:k-1}^{\text{sds}})$ to the CoT factorization (S. A.1) in which the text response $\hat{S}_k^{\text{res}}$ is an explicit intermediate variable.

Using the chain rule and Viterbi-style approximation, we can write the likelihood of a turn-level spoken response as

$$\log \pi_\theta(Y_k \mid \mathcal{H}^{(k)}) \approx \log P_\theta(\hat{S}_k^{\text{res}} \mid \mathcal{H}^{(k)}) + \log P_\theta(Y_k \mid \mathcal{H}^{(k)}, \hat{S}_k^{\text{res}}). \quad (5)$$

Since the semantic preference signal depends only on the judged text response, we modify Eq. 3 and apply DPO using only the *text-policy term*:

$$\Delta_\theta^{\text{sem}}(k) = \log P_\theta(\hat{S}_k^{\text{res},+} \mid \mathcal{H}^{(k)}) - \log P_\theta(\hat{S}_k^{\text{res},-} \mid \mathcal{H}^{(k)}), \quad (6)$$

(and analogously for π_{ref}), which is then substituted into Eq. (2). Intuitively, this treats semantic DPO as directly increasing the likelihood of the preferred *text response*[1].

5.2 Audio Quality and Intelligibility Rewards

Beyond semantic alignment, high-quality spoken dialogue systems must also produce responses that are acoustically natural and intelligible. To this end, we construct preference learning signals targeting *audio quality* and *intelligibility*, while carefully controlling for semantic content.

Audio Quality. To optimize acoustic naturalness, we use an automatic speech quality estimator $q(\cdot)$ (UTMOS (Saeki et al., 2022)) as a proxy for perceived audio quality. For each dialogue turn k, we generate a set of candidate spoken responses $\mathcal{Y}_k$ as in S. 4.1. Let $q(Y_{k,i}^{\text{sds}})$ denote the predicted quality score for a synthesized utterance $Y_{k,i}^{\text{sds}}$. We construct preference pairs by selecting

$$Y_k^+ = \arg\max_{Y_{k,i}^{\text{sds}} \in \mathcal{Y}_k} q(Y_{k,i}^{\text{sds}}), Y_k^- = \arg\min_{Y_{k,i}^{\text{sds}} \in \mathcal{Y}_k} q(Y_{k,i}^{\text{sds}}), \quad (7)$$

Intelligibility. Let $w(Y_{k,i}^{\text{sds}})$ denote the WER between the hypothesis of synthesized speech $Y_{k,i}^{\text{sds}}$ generated by pre-trained ASR system and the model-predicted text response S_k^{res}. Among candidates satisfying $w(Y_{k,i}^{\text{sds}}) \le \tau_{\text{wer}}$, we select

$$Y_k^+ = \arg\min_{Y_{k,i}^{\text{sds}} \in \mathcal{Y}_k} w(Y_{k,i}^{\text{sds}}). \quad (8)$$

and draw negative samples from candidates with

$$Y_k^- \sim \{Y_{k,i}^{\text{sds}} \in \mathcal{Y}_k \mid w(Y_{k,i}^{\text{sds}}) \ge w(Y_k^+) + \delta_{\text{wer}}\}. \quad (9)$$

where δ_{wer} is a hyperparameter. This margin-based construction avoids ambiguous comparisons and focuses on clearly degraded intelligibility.

For both audio quality and intelligibility rewards, we enforce that the text response S_k^{res} is identical for the positive and negative samples within each DPO pair. Accordingly, the DPO objective for audio-based rewards can be written as

$$\Delta_\theta^{\text{audio}}(k) = \log \pi_\theta\big(Y_k^+ \mid X_{1:k}, Y_{1:k-1}^{\text{sds}}, S_k^{\text{res}}\big) - \log \pi_\theta\big(Y_k^- \mid X_{1:k}, Y_{1:k-1}^{\text{sds}}, S_k^{\text{res}}\big), \quad (10)$$

As a result, the preference signal isolates the quality of the *generated speech output* alone, allowing DPO to directly improve acoustic naturalness and intelligibility without altering the underlying semantic content.

5.3 Emotion-Consistency Reward

In addition to semantic quality and acoustic naturalness, effective spoken dialogue systems should convey emotions consistent with human intent. To encourage emotionally aligned responses, we construct preference pairs based on *emotion similarity* between synthesized speech and human reference speech. Emotion representations are extracted using a pretrained encoder, and a scalar similarity score $e(Y_{k,i}^{\text{sds}})$ is computed for each candidate. We then select preference pairs as

$$Y_k^+ = \arg\max_{Y_{k,i}^{\text{sds}} \in \mathcal{Y}_k} e(Y_{k,i}^{\text{sds}}), Y_k^- = \arg\min_{Y_{k,i}^{\text{sds}} \in \mathcal{Y}_k} e(Y_{k,i}^{\text{sds}}), \quad (11)$$

subject to the constraint that the difference between the maximum and minimum emotion similarity exceeds a threshold δ_{emo}, ensuring that the preference signal reflects a meaningful emotional contrast. As with audio quality and intelligibility rewards, we enforce that the underlying text response S_k^{res} is identical for both Y_k^+ and Y_k^-.

[1]We tried other ablations, including (i) applying DPO jointly to text and speech responses and (ii) augmenting DPO with an additional supervised fine-tuning (SFT) loss on speech outputs. Across these settings, we found that a simple text-only DPO formulation consistently achieved the strongest and most stable improvements, while more complex variants did not provide additional gains.

Table 1: Scale of the constructed preference learning dataset used for multi-reward DPO training.

Reward Type	# DPO Pairs
Semantic Quality	51.1K
Audio Quality	32.0K
Intelligibility	61.0K
Emotion	21.6K
Total	**165.7K**

6 Experiments

6.1 DPO Data Preparation.

To construct high-quality preference data for DPO training, we begin by applying multi-turn CoT SDS model (Arora et al., 2025d) to generate diverse response candidates for real conversational contexts. We use widely adopted human–human dialogue corpora, focusing on the Switchboard dataset (Godfrey et al., 1992), which contains approximately 300 hours of spontaneous telephone conversations. The constructed multi-reward DPO dataset comprises 165.7K high-quality preference pairs, with detailed statistics summarized in Table 1. For semantic preference pairs construction (see S. 5.1), the CoT model generates $n = 10$ candidate text responses per dialogue turn using top-k sampling, ensuring sufficient lexical and semantic diversity. For acoustic and emotion rewards (see S. 5.2 and S. 5.3), we generate $n = 10$ spoken realizations for each candidate text response using top-k sampling. Emotion-consistency preference pairs are constructed using emotion representations extracted by Emo2Vec (Ma et al., 2024) from both synthesized speech and the corresponding ground-truth human speech. Additional details are provided in S. A.3.

6.2 Evaluation Data and Metrics

We evaluate all baselines and proposed models on the Eval2000 dataset similar to (Arora et al., 2025c,d). Semantic quality is evaluated by transcribing synthesized speech Y^{sds} using Whisper large (Radford et al., 2023). We report ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) scores against human reference responses, along with perplexity (Jelinek et al., 1977) computed using GPT-2 (Radford et al., 2019). To better capture conversational relevance beyond n-gram overlap, we additionally evaluate responses using Qwen2.5-7B-Instruct as an *LLM judge*, following recent evaluation protocols (Zhang et al., 2024c) (Prompt in S. A.9). We also report *Auto-BLEU*, as a proxy for repetition and degeneration. For RLAIF-trained models only, we further report a *win rate* relative to the no-RLAIF baseline. Given a set of evaluated samples, the win rate is defined

$$\text{WinRate} = \frac{\#(s > s_{\text{base}}) + 0.5 \times \#(s = s_{\text{base}})}{N}, \tag{12}$$

where s denotes the evaluation score for a given sample, s_{base} is the corresponding score from the no-RLAIF baseline, and N is the total number of evaluated samples. This metric reflects the proportion of responses that outperform (or tie with) the baseline under the same evaluation criterion. We utilize the VERSA toolkit (Shi et al., 2025), measuring intelligibility through Whisper hypotheses and evaluating audio quality using UTMOS. We measure intelligibility using word error rate (WER), computed between the ASR transcript of the synthesized speech and the model-predicted text response. Importantly, WER measures *speech realization fidelity*, i.e., whether the generated speech accurately reflects the model's intended text—rather than the semantic correctness of the response itself. We evaluate speaking-style consistency within entire conversations using Emo2Vec (Ma et al., 2024). We rank all SDSs based on their emotional alignment and compute the average rank ("Emotion Rank" (Arora et al., 2025d)) across all utterances.

6.3 Models

Following prior work (Arora et al., 2025d), we first report results on a single-turn E2E ("Direct E2E") spoken dialogue system. We further benchmark our approach against Moshi (Défossez et al., 2024), a strong 7B dual-channel duplex SDS baseline[2]. In addition, we evaluate a multi-turn Chain-of-Thought E2E model (Multi-turn CoT E2E) introduced in prior work (Arora et al., 2025c), which serves as a strong reasoning-aware baseline. We then extend our framework to the duplex setting and compare against the blockwise duplex SCoT-Response model from (Arora et al., 2025c). For both the multi-turn CoT and duplex SCoT-Response models, we apply RLAIF post-training using our proposed preference-learning framework.

We conduct RLAIF post-training under multiple configurations to disentangle the effects of different reward signals. In RLAIF (Single-Reward) settings, models are trained using DPO pairs de-

[2]kyutai/moshiko-pytorch-bf16

Table 2: Semantic Quality Evaluation of RLAIF-Post-Trained Turn-by-Turn SDS. * Statistical significant difference based on Wilcoxon signed rank and Paired Bootstrap Resampling tests with p-value<0.01. Values in bracket indicate the percentage of LLM judge responses receiving low scores (< 5). Win rate is defined only for RLAIF models and measures the proportion of samples that outperform the corresponding no-RLAIF base model.

Model	**ROUGE-L (↑)**	**Perplexity (↓)**	**AutoBLEU (↓)**	**LLM judge (↑)**	**Win Rate (↑)**
Direct E2E (Arora et al., 2025d)	8.4	302.2	51.5	5.50 (24.2%)	✗
Moshi (Défossez et al., 2024)	8.1	136.5	57.8	5.71 (21.0%)	✗
Multi turn CoT E2E (Arora et al., 2025c)	12.1	21.2	68.3	6.18 (10.2%)	✗
+ RLAIF (Single-Reward)	11.9	19.9	**56.5**	**6.33*** (7.1%)	55.4
+ RLAIF (Joint-Reward-v1)	11.8	**19.6**	61.3	6.29* (8.5%)	52.6
+ RLAIF (Joint-Reward-v2)	11.9	19.9	59.9	6.33* (7.5%)	54.4

rived from a single reward signal, namely semantic quality, audio quality, intelligibility or emotion consistency, with the objective of selectively improving the corresponding metric. In contrast, we consider two RLAIF (Joint-Reward) configurations: (i) (Joint-Reward-v1) a joint semantic–audio setting that combines semantic, audio quality, and intelligibility preference data, and (ii) (Joint-Reward-v2) a joint semantic–audio–emotion setting that additionally incorporates speaking-style (emotion-consistency) preferences. This design allows for a systematic evaluation of how individual and combined preference signals contribute to overall system performance. Additional implementation and training details are provided in the Appendix A.4.

7 Results

7.1 RLAIF with Semantic Quality Reward

Table 2 reports semantic-quality evaluation for turn-by-turn SDS, including general E2E baselines and the strong duplex SDS baseline Moshi. Among non-RLHF systems, the Multi-turn CoT E2E model substantially outperforms both the standard E2E and Moshi baselines in terms of semantic metrics, establishing a strong foundation for preference-based post-training.

Building on this baseline, RLAIF (Single-Reward) post-training with semantic preference data yields a statistically significant improvement in LLM-judge scores over the Multi-turn CoT E2E model ($6.18 \rightarrow 6.33$, $p < 0.01$ under both Wilcoxon signed-rank and paired bootstrap resampling tests), with a corresponding win rate of 55.4%. Importantly, beyond average score improvements, RLAIF induces a meaningful distributional shift: the proportion of low-quality responses (LLM-judge < 5) is reduced from 10.2% to 7.1%, corresponding to a 28.5% relative reduction in poor responses. This behavior is consistent with prior RLAIF findings, where improvements are often driven by suppressing degenerate or incoherent generations rather than uniformly increasing mean scores.

RLAIF post-training also improves auxiliary indicators of response quality. In particular, semantic RLAIF substantially reduces repetition, as reflected by a large drop in AutoBLEU ($68.3 \rightarrow 56.5$), and yields more coherent responses as evidenced by lower perplexity ($21.2 \rightarrow 19.9$). When extending to multi-objective optimization, RLAIF (Joint-Reward-v1), which combines semantic, audio quality, and intelligibility preferences, maintains strong LLM-judge performance (6.29). Incorporating additional emotion-consistency preferences in RLAIF (Joint-Reward-v2) further improves LLM-judge scores (6.33) and reduces the fraction of low-scoring responses (7.5%), demonstrating that jointly leveraging various preferences leads to consistent overall gains without degrading semantic quality [3]. To mitigate potential judge bias, we additionally evaluate using independent LLM judges (Gemini 2.5 Flash and Gemini 3 Pro Preview), which are not used during training; results are consistent with statistically significant gains for RLAIF (p<0.01) as discussed in S. A.7.

Together, these results show that RLAIF post-training effectively improves conversational quality in end-to-end SDS, primarily by reducing low-quality and repetitive responses while preserving semantic coherence. Based on qualitative inspection of model outputs, RLAIF post-training consistently improves direct question answering, reduces topic drift, and suppresses repetitive or degenerate response patterns, aligning well with the quantitative gains observed in LLM-judge scores and AutoBLEU. At the same time, we observe occasional failure cases where RLAIF-trained models pro-

[3]For perplexity, the improvement using RLAIF (i.e. + RLAIF (Joint-Reward-v2)) is statistically significant under the Wilcoxon signed-rank test, although it does not reach significance under the paired t-test. For AutoBLEU, the improvement using RLAIF (i.e. + RLAIF (Joint-Reward-v2)) is statistically significant under both the Wilcoxon signed-rank test and the paired t-test.

duce responses that are overly safe or generic, and thus do not sufficiently advance the conversation. We provide representative qualitative examples in T. 7 and 8 in Appendix. We additionally compare against Qwen-Omni (Xu et al., 2025), a significantly larger 7.5B-parameter model in S. A.6. Despite the scale difference, our 1.7B RLAIF-aligned model achieves competitive conversational quality, indicating that the proposed alignment framework is complementary to stronger backbones rather than tied to a specific architecture.

7.2 RLAIF with Acoustic Quality Reward

Table 3 reports audio quality and intelligibility evaluation across general E2E SDS baselines, the strong duplex baseline Moshi, and RLAIF-trained variants. Among non-RLAIF systems, Moshi achieves the highest absolute UTMOS scores, largely attributable to its use of high-quality assistant voice prompts. In contrast, our models are conditioned on lower-fidelity Switchboard speaker prompts, resulting in lower UTMOS scores.

Building on the Multi-turn CoT E2E baseline, RLAIF (Single-Reward) post-training demonstrates targeted and effective improvements. Training with audio-quality preferences substantially improves perceived speech naturalness, increasing UTMOS from 2.16 to 3.06. Similarly, RLAIF (Single-Reward) training with intelligibility-focused preferences yields a marked reduction in word error rate, improving turn-level intelligibility from 6.1 to 3.3. Here, WER is computed between the ASR transcript of the synthesized speech, obtained using Whisper, and the model-predicted text response S_k^{res} used during CoT decoding (S. 5.2). These results confirm that audio-centric DPO signals selectively and reliably improve their intended dimensions.

When multiple preference datasets are jointly leveraged using RLAIF (Joint-Reward) training, the model achieves simultaneous gains in both audio quality and intelligibility. In particular, RLAIF (Joint-Reward-v1), which combines semantic, audio quality, and intelligibility preferences, attains a UTMOS of 2.85 while further reducing WER to 1.0, representing the best intelligibility performance among RLAIF-trained models. Incorporating additional emotion-consistency preferences in RLAIF (Joint-Reward-v2) preserves audio quality but results in a slightly higher WER, indicating a mild trade-off between expressive consistency and intelligibility. Together, these findings validate the effectiveness of audio-quality and intelligibility preference data and highlight the benefit of unified multi-reward RLAIF for improving speech naturalness and intelligibility.

Table 3: Audio quality evaluation. ✗: WER is not reported for Direct E2E models since they do not output explicit intermediate text response.

Model	UTMOS (↑)	WER (↓)
Direct E2E (Arora et al., 2025d)	2.03	✗
Moshi (Défossez et al., 2024)	3.34	✗
Multi turn CoT E2E (Arora et al., 2025c)	2.16	6.1
+ RLAIF (Single-Reward)	**3.06**	3.3
+ RLAIF (Joint-Reward-v1)	2.85	**1.0**
+ RLAIF (Joint-Reward-v2)	2.85	1.7

Table 4: Speaking style consistency evaluation.

Model	Emotion Rank (↓)
Direct E2E (Arora et al., 2025d)	2.81
Moshi (Défossez et al., 2024)	4.92
Multi turn CoT E2E (Arora et al., 2025c)	2.29
+ RLAIF (Single-Reward)	1.98
+ RLAIF (Joint-Reward-v2)	3.00

7.3 RLAIF with Emotion Quality Reward

Table 4 reports speaking-style consistency evaluated via emotion rank, where lower values indicate closer alignment with human reference speech. Among non-RLAIF baselines, the Multi-turn CoT E2E model performs best (2.29), outperforming both E2E and Moshi. RLAIF (Single-Reward) post-training with emotion-consistency preferences further improves alignment, reducing the emotion rank to 1.98, demonstrating the effectiveness of emotion-aware preference learning.

In contrast, RLAIF (Joint-Reward-v2) yields a higher emotion rank (3.00), indicating competition between reward signals. In particular, strong semantic rewards from LLM-based judges can encourage safer or more generic responses (T. 8), which may reduce emotional expressiveness. This highlights an inherent trade-off in multi-objective RLAIF settings and motivates future work on better balancing various rewards in SDS.

7.4 Human Evaluation

To complement automatic metrics and assess perceptual conversational quality, we conducted a controlled human evaluation on 100 randomly selected dialogue samples, each evaluated by three independent annotators. Annotators were presented with anonymized outputs with system identities withheld and asked to indicate their preference among responses generated with RLAIF (*Multi-turn CoT E2E + RLAIF (Joint-Reward-v2)*), without RLAIF (*Multi-turn CoT E2E*), or select no preference. In

Table 5: Semantic quality evaluation for duplex model.

Model	ROUGE-L (↑)	Perplexity (↓)	LLM judge (↑)
SCoT-Response	19.8	42.3	5.95
+ RLAIF	**23.1**	25.0	6.00

Table 6: Human evaluation results comparing RLAIF and non-RLAIF systems.

Preference Outcome	Percentage (%)	
Prefer RLAIF	**56.7**	
Prefer No-RLAIF	24.0	
No Preference	19.3	
Metric	**No-RLAIF**	**RLAIF (Ours)**
Audio Quality (1–5) ↑	2.89	**3.39**
Semantic Quality (1–5) ↑	2.81	**3.22**

addition, annotators rated audio quality and semantic quality on a 1–5 Likert scale.

Tab. 6 shows that annotators preferred the RLAIF outputs in the majority of cases, more than doubling the preference rate compared to the non-RLAIF system. The RLAIF system also achieves consistently higher Likert ratings for both audio quality and semantic quality. Furthermore, annotators exhibit moderate agreement, with 65.7% consensus across evaluators, indicating that the observed improvements are stable and perceptually meaningful. Overall, these results confirm that the gains observed in automatic metrics translate to improved human-perceived conversational quality.

7.5 RLAIF for Duplex Models

Table 5 reports semantic-quality evaluation for the duplex SDS. Compared to the SCoT-Response baseline, RLAIF post-training consistently improves semantic performance, yielding higher ROUGE-L (19.8 $\rightarrow$ 23.1), lower perplexity (42.3 $\rightarrow$ 25.0), and improved LLM-judge scores (5.95 $\rightarrow$ 6.00). These results indicate that the proposed RLAIF framework generalizes beyond turn-by-turn settings and remains effective for blockwise duplex spoken dialogue models. In addition to improvements in semantic and acoustic metrics, our system satisfies key practical requirements for real-time deployment. Specifically, our duplex SDS achieves real-time performance (RTF < 1), sub-second latency, and improved overlap prediction accuracy as discussed in S. A.5, demonstrating that the proposed alignment framework not only improves response quality but also preserves efficient and natural duplex interaction behavior.

8 Conclusion

We presented the first multi-reward RLAIF framework for E2E SDS, addressing key limitations of prior work that focuses on single, utterance-level semantic rewards. Our approach jointly optimizes semantic coherence, audio quality, intelligibility, and emotion consistency on turn-by-turn Chain-of-Thought SDS. By applying utterance-level preferences over blockwise decoding, we enable preference optimization on blockwise duplex models without requiring partial-utterance reward definitions. Experimental results demonstrate that single-reward RLAIF selectively improves its targeted dimension, validating the specificity of our constructed preference data, while joint multi-reward training yields consistent gains across semantic and acoustic metrics. Finally, we release the first large-scale multi-reward DPO dataset for spoken dialogue systems to support reproducible research.

9 Acknowledgement

Experiments of this work used the Bridges2 system at PSC and Delta and DeltaAI system at NCSA through allocations CIS210014 and IRI120008P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by National Science Foundation grants #2138259,#:2138286, #:2138307, #:2137603, and #:2138296.

10 Limitations

While our results demonstrate the effectiveness of multi-reward RLAIF for spoken dialogue systems, several limitations remain. First, preference data construction relies on automatic evaluators, which may introduce bias or noise relative to human judgments. Incorporating human-in-the-loop preferences is an important direction for future work. Second, our multi-reward formulation combines reward signals via dataset-level concatenation rather than explicitly modeling trade-offs or interactions between objectives, which may limit optimal balancing in some conversational contexts. Finally, our experiments focus on English conversational datasets; extending the framework to multilingual settings and more diverse conversational domains remains an open challenge.

11 Ethics Impact

We adhere to the ACL Ethics Policy. Our experiments are based on open-source datasets with no violation of privacy, and we will make all our code and models publicly available. Parts of this manuscript were edited for clarity and language using an AI-based writing assistant. The authors take full responsibility for the content.

References

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025a. On the landscape of spoken language models: A comprehensive survey. *Preprint*, arXiv:2504.08528.

Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, and 1 others. 2022. ESPnet-SLU: Advancing spoken language understanding through espnet. In *ICASSP*, pages 7167–7171. IEEE.

Siddhant Arora, Yifan Peng, Jiatong Shi, Jinchuan Tian, William Chen, Shikhar Bharadwaj, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, Shuichiro Shimizu, and 1 others. 2025b. Espnet-sds: Unified toolkit and demo for spoken dialogue systems. In *NAACL 2025*, pages 248–259.

Siddhant Arora, Jinchuan Tian, Hayato Futami, Jiatong Shi, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. 2025c. Chain-of-thought reasoning in streaming full-duplex end-to-end spoken dialogue systems. *arXiv preprint arXiv:2510.02066*.

Siddhant Arora, Jinchuan Tian, Hayato Futami, Jee weon Jung, Jiatong Shi, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. 2025d. Chain-of-Thought Training for Open E2E Spoken Dialogue Systems. In *Interspeech 2025*, pages 4833–4837.

Yuntao Bai, Saurabh Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Zackary Goldstein, Sam McCandlish, Chris Olah, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yifei Cao, Changhao Jiang, Jiabao Zhuang, Jiajun Sun, Ming Zhang, Zhiheng Xi, Hui Li, Shihan Dou, Yuran Wang, Yunke Zhang, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. From scores to preferences: Redefining mos benchmarking for speech quality reward modeling. *arXiv preprint arXiv:2510.00743*.

Chen Chen, Ke Hu, Chao-Han Huck Yang, Ankita Pasad, Edresson Casanova, Weiqing Wang, Szu-Wei Fu, Jason Li, Zhehuai Chen, Jagadeesh Balam, and Boris Ginsburg. 2025. Reinforcement learning enhanced full-duplex spoken dialogue language models for conversational interactions. In *Second Conference on Language Modeling*.

Chen Chen, Yuchen Hu, Wen Wu, Helin Wang, Eng Siong Chng, and Chao Zhang. 2024a. Enhancing zero-shot text-to-speech synthesis with human feedback. *Preprint*, arXiv:2406.00654.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024b. Towards robust speech representation learning for thousands of languages. *arXiv preprint arXiv:2407.00837*.

Jade Copet and 1 others. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. Technical report, Kyutai.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *EMNLP 2020*, Findings of ACL. Association for Computational Linguistics.

Hugging Face. 2024. Speech-to-speech translation toolkit. Accessed: 2024-12-10.

James Glass. 1999. Challenges for spoken dialogue systems. In *Proceedings of the 1999 IEEE ASRU Workshop*, volume 696. MIT Laboratory for Computer Science Cambridge.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024. AudioGPT: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. 2025a. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. *arXiv preprint arXiv:2510.07838*.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. 2025b. Align-SLM: Textless spoken language models with reinforcement learning from AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20395–20411, Vienna, Austria. Association for Computational Linguistics.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. *Proc. ACL 2024 Findings*.

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. 2018. Neural dialogue context online end-of-turn detection. In *SIGdial*, pages 224–228. Association for Computational Linguistics.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a map task dialogue system. *Computer speech & language*, 28(4):903–922.

Ziqiao Meng, Qichao Wang, Wenqian Cui, Yifei Zhang, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. 2024. Parrot: Autoregressive spoken dialogue language modeling with decoder-only transformers. In *NeurIPS Workshop AI-Driven Speech, Music, and Sound Generation*.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguistics*, 11:250–266.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. SpiRit-LM: Interleaved spoken and written language model. *Trans. ACL*, 13:30–52.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Matthew Roddy, Gabriel Skantze, and Naomi Harte1. 2018. Investigating speech features for continuous turn-taking prediction using LSTMs. In *Interspeech*.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab system for voiceMOS challenge 2022. In *Interspeech*, pages 4521–4525.

Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeswar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2018. A deep reinforcement learning chatbot (short version). *arXiv preprint arXiv:1801.06700*.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, Dareen Safar Alharthi, Yichen Huang, Koichi Saito, Jionghao Han, Yiwen Zhao, Chris Donahue, and Shinji Watanabe. 2025. VERSA: A versatile evaluation toolkit for speech, audio, and music. In *"NAACL 2025"*, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, and 1 others. 2024.

ESPnet-Codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. *arXiv preprint arXiv:2409.15897*.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proc. SIGdial Meeting Disc. and Dial.*

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, Berlin, Germany. Association for Computational Linguistics.

Jinchuan Tian, Jiatong Shi, William Chen, Siddhant Arora, Yoshiki Masuyama, Takashi Maekaku, Yihan Wu, Junyi Peng, Shikhar Bharadwaj, Yiwen Zhao, Samuele Cornell, Yifan Peng, Xiang Yue, Chao-Han Huck Yang, Graham Neubig, and Shinji Watanabe. 2025a. Espnet-speechlm: An open speech language model toolkit. *Preprint*, arXiv:2502.15218.

Jinchuan Tian, Chunlei Zhang, Jiatong Shi, Hao Zhang, Jianwei Yu, Shinji Watanabe, and Dong Yu. 2025b. Preference alignment improves language model-based tts. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. *arXiv preprint arXiv:2409.15594*.

Shinji Watanabe and 1 others. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211.

Anne Wu, Laurent Mazaré, Neil Zeghidour, and Alexandre Défossez. 2025. Aligning spoken dialogue models from user interactions. *arXiv preprint arXiv:2506.21463*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Xiong Wang Bin Zhang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Kai Yoshida, Masahiro Mizukami, Seiya Kawano, Canasai Kruengkrai, Hiroaki Sugiyama, and Koichiro Yoshino. 2025. Training dialogue systems by ai feedback for improving overall dialogue impression. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2024a. Speechalign: Aligning speech generation to human preferences. *arXiv preprint arXiv:2404.05600*.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024b. Speechgpt-gen: Scaling chain-of-information speech generation. *arXiv preprint arXiv:2401.13527*.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chaohong Tan, Zhihao Du, and 1 others. 2024c. Omniflatten: An end-to-end GPT model for seamless voice conversation. *arXiv preprint arXiv:2410.17799*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. 2024d. Beyond the turn-based game: Enabling real-time conversations with duplex models. In *Proc. EMNLP*.

A Appendix

A.1 Mathematical Formulation of Chain-of-Thought Turn by turn SDS

More recent E2E approaches replace cascaded architectures (Glass, 1999; Huang et al., 2024) with unified speech–language models (SLMs) that directly generate spoken responses autoregressively (Zhang et al., 2023, 2024b; Nguyen et al., 2025). Although these models mitigate error propagation, they often lack explicit intermediate reasoning structure, which can limit coherence and increase data requirements (Arora et al., 2025d; Défossez et al., 2024). To address this, prior work (Arora et al., 2025d) has explored incorporating structured intermediate representations, such as ASR transcripts S_k^{asr} and textual responses S_k^{res}, within an end-to-end framework, yielding Chain-of-Thought (CoT) SDS models. Using the Viterbi approximation and C.I. assumption, we can modify Eq. (1) to get:

$$\begin{aligned} &P(Y^{\text{sds}}|X, X^{\text{spk}}) \\ &\approx \prod_k P(Y_k^{\text{sds}}|Y_{1:k-1}^{\text{sds}}, X_{1:k}, \hat{S}_{1:k}^{\text{res}}, \hat{S}_{1:k}^{\text{asr}}, X^{\text{spk}}) \end{aligned} \tag{13}$$

where

$$\hat{S}_k^{\text{asr}} = \arg\max_{S_k^{\text{asr}}} P(S_k^{\text{asr}} | Y_{1:k-1}^{\text{sds}}, X_{1:k}, \hat{S}_{1:k-1}^{\text{res}}, \hat{S}_{1:k-1}^{\text{asr}}, \cancel{X^{\text{spk}}}) \quad (14)$$

and

$$\hat{S}_k^{\text{res}} = \arg\max_{S_k^{\text{res}}} P(S_k^{\text{res}} | Y_{1:k-1}^{\text{sds}}, X_{1:k}, \hat{S}_{1:k-1}^{\text{res}}, \hat{S}_{1:k}^{\text{asr}}, \cancel{X^{\text{spk}}}) \quad (15)$$

At each turn, the **CoT model** follows a structured decoding process: where it first infers a transcription S_k^{asr}, then a textual response $\hat{S}_k^{\text{res}}$, and finally synthesizes speech Y_k^{sds} conditioned on these intermediate variables.

A.2 Mathematical Formulation of Chain-of-Thought Duplex SDS

In the time-multiplexing setup (Veluri et al., 2024), the input speech X is divided into a sequence of B blocks, $X = \{X_b \mid b = 1, \dots, B\}$. Similarly, the output speech Y^{sds} is represented as a sequence of B blocks, $Y^{\text{sds}} = \{Y_b^{\text{sds}} \mid b = 1, \dots, B\}$. Rather than generating responses only at turn boundaries or frame-by-frame, the model then generates output block-by-block in a streaming fashion, estimating the posterior using the causality-based C.I. as:

$$P(Y^{\text{sds}} | X, X^{\text{spk}}) = \prod_b P(Y_{b+1}^{\text{sds}} | Y_{1:b}^{\text{sds}}, X_{1:b}, \cancel{X_{b+1:B}}, X^{\text{spk}}). \quad (16)$$

Building on earlier turn-by-turn Chain-of-Thought (CoT) formulations for E2E SDS, recent work, namely SCoT, has investigated how structured intermediate reasoning can be integrated into blockwise duplex systems (Arora et al., 2025c). The key idea is to introduce intermediate representations, such as ASR transcripts and textual responses, that are temporally aligned with the speech signal. Let the corresponding aligned transcript and system response for the b^{th} block (See S. 3.2) be A_b^{asr} and A_b^{res} respectively. Incorporating these alignments, the blockwise dialogue policy (Eq. 16) can be augmented to condition on partial reasoning states (similar to (Eqs. (13)-(15)) as shown :

$$P(Y^{\text{sds}} | X, X^{\text{spk}}) \approx \prod_b P(Y_{b+1}^{\text{sds}} | Y_{1:b}^{\text{sds}}, X_{1:b}, \hat{A}_{1:b}^{\text{asr}}, \hat{A}_{1:b+1}^{\text{res}}, X^{\text{spk}}). \quad (17)$$

The intermediate text response is predicted as:

$$\hat{A}_{b+1}^{\text{res}} = \arg\max P(A_{b+1}^{\text{res}} | Y_{1:b}^{\text{sds}}, X_{1:b}, \hat{A}_{1:b}^{\text{asr}}, \hat{A}_{1:b}^{\text{res}}) \quad (18)$$

and the aligned ASR transcript for each block is:

$$\hat{A}_b^{\text{asr}} = \arg\max P(A_b^{\text{asr}} | Y_{1:b}^{\text{sds}}, X_{1:b}, \hat{A}_{1:b-1}^{\text{asr}}, \hat{A}_{1:b}^{\text{res}}). \quad (19)$$

At each block, SCoT performs a structured three-stage decoding process, similar to S. 3.1.

A.3 DPO Data Preparation.

Semantic Reward: For each dialogue turn, the CoT model produces 10 candidate text responses using top-k sampling ($k = 10$), ensuring sufficient lexical and semantic diversity. We analyze the empirical distribution of the LLM judge scores across all candidates using histograms (Figures 2 and 3 in Appendix), which reveal natural separation between coherent and incoherent responses. Based on this distribution, we define positive samples as those with an LLM score greater than $\tau_{\text{pos}} = 6$ and AutoBLEU less than $\delta_{\text{low}} = 30$, indicating semantically meaningful and non-repetitive outputs. Conversely, negative samples are those with an LLM score below $\tau_{\text{neg}} = 5$ or AutoBLEU exceeding $\delta_{\text{high}} = 30$, capturing incoherent, off-topic, or repetitive candidates.

Acoustic and Emotion Reward: For each candidate text response described above, we generate multiple spoken realizations using top-k sampling ($k = 10$), resulting in a set of candidate speech responses per turn. To construct intelligibility-based preference pairs, we retain samples satisfying an upper intelligibility constraint $\tau_{\text{wer}} = 0.25$, and select negative examples whose WER exceeds that of the positive sample by a margin $\delta_{\text{wer}} = 0.05$. In parallel, we construct audio-quality preference pairs using UTMOS scores. For emotion reward pairs, we use Emo2Vec (Ma et al., 2024) to extract emotion representations from both the generated speech outputs and the corresponding ground-truth human speech. For preference construction, the candidate with the maximum emotion similarity is selected as the positive sample, while the candidate with the minimum similarity is selected as the negative sample, subject to a minimum margin of $\delta_{\text{emo}} = 2\%$ between the two scores.

Table 7: Qualitative examples where RLAIF post-training improves response quality. Compared to the baseline, RLAIF-generated responses better address the user's question, reduce topic drift, and avoid repetitive or incoherent continuations.

Dialogue Context	Reference Answer	w/o RLHF	RLHF
Example 1 User: Do you have class on Friday?	Well, I have an exam eight thirty to ten thirty. Mike is taking the whole day off. Mike wanted to leave at like nine o'clock in the morning, but...	No, we're going to have lunch in the afternoon, but we had to do something with our hair because we got a lot of hair.	No, we don't have class on Friday. We have class on Saturday, so I'm going to go to the wedding.
Example 2 User: We are starting the transition. I told the students what the new plan was...	What do you mean, oh, about you are leaving?	Yeah, I know what you mean.	Yeah, I know, I know. I'm just trying to figure out what to do.
Example 3 User: I have been there for two weeks already.	Oh, my gosh! So, who?	Oh, okay. I mean, you start it on the fourth day if you want.	Oh, well, I'm glad you're doing well.

Table 8: Qualitative examples where RLAIF post-training produces overly generic or safe responses that fail to advance the conversation, despite being coherent.

Dialogue Context	Reference Answer	w/o RLHF	RLHF
Example 1 User: The problem I had when I went into teaching...	What grade were you teaching?	Yeah, I think that's good, too. I think that that should be, that's the ideal teacher...	Yeah, I think that's true.
Example 2 User: That is what my sister is like. I want a meatball sandwich from Wawa. Let us walk up. I am like.	Is your sister doing that thing you know, hosing down the backyard?	I think I would have to go with my sister. I don't go to work with my brother...	Yeah, I know, I know.

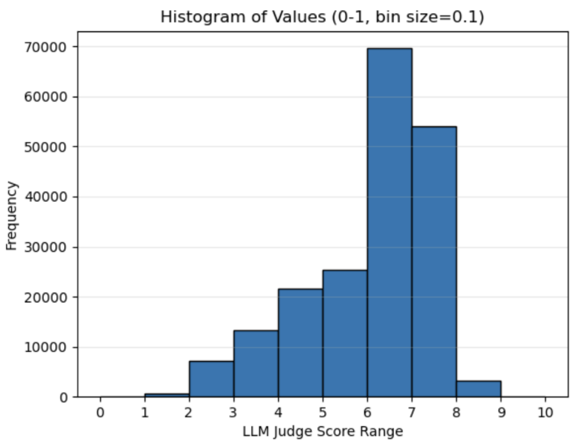

Figure 2: Distribution of LLM judge score over candidate responses

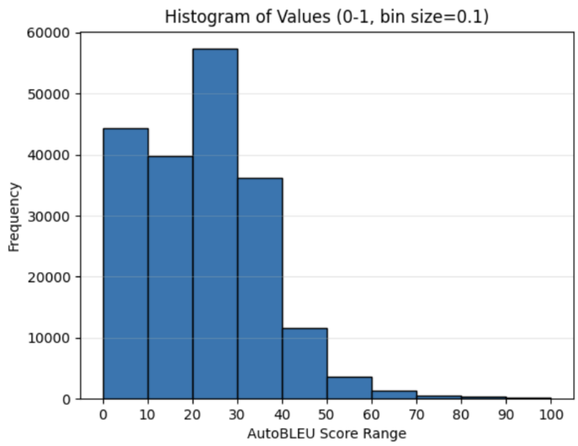

Figure 3: Distribution of AutoBLEU score over candidate responses

A.4 Experiment Setup

Our models are implemented in PyTorch, with all experiments conducted using the ESPnet toolkit (Watanabe et al., 2018; Arora et al., 2022, 2025b). We use the pre-trained SpeechLM from (Tian et al., 2025a), which is initialized from the SmolLM2 1.7B text LLM. We adopt the delay-interleaving architecture (Copet et al., 2024) for multi-stream language modeling.

The SpeechLM is pre-trained for 500k updates using a warm-up–decay learning rate schedule, with each batch containing approximately 3,840 seconds of audio and 1.1M text tokens. Training is performed on a large multimodal corpus comprising 213k hours of speech audio and 292B text tokens. For audio tokenization, we use ESPnet-Codec (Shi et al., 2024)[4] for codec tokenization and XEUS (Chen et al., 2024b)[5] for SSL tokenization. Specifically, codec and SSL tokens are concatenated frame-by-frame.

For decoding, ASR outputs are generated using greedy search followed by hallucination detection and removal, as in (Arora et al., 2025d). Text response generation uses greedy search, with the same post-processing procedure (Arora et al., 2025d). For speech response generation in our CoT models, we again use top-k sampling (consistent with the text setting) and apply an additional intelligibility-based post-processing step following (Arora et al., 2025d). Specifically, we generate ten candidate speech samples and transcribe each using Whisper (Radford et al., 2023). We compute the WER of each transcription against the model's own predicted text response, $\hat{S}_k^{\text{res}}$ (Eq. (15)) for turn-by-turn systems and $\hat{A}_{b+1}^{\text{res}}$ (Eq. 18) for duplex systems, as the reference, and select the candidate with the lowest WER.

During inference in blockwise duplex systems, we additionally constrain the ASR transcript and text response to a maximum of 25 words and limit the speech output to the block duration (i.e., 2 seconds), following SCoT (Arora et al., 2025c). To ensure fair comparison between turn-by-turn and duplex systems, we aggregate speech outputs across blocks within each turn and compute quality metrics on the combined output. Finally, for the baseline Moshi system, we provide each turn-level utterance as input and append 20 seconds of silence to ensure response generation is triggered.

All models are trained using 4 NVIDIA H200 GPUs. Further, we split DPO data shown in Table 1, into 99:1 ratio for training / validation while training our preference learning models. Model hyperparameters are shown in Table 10. All results are reported from a single training run. We will publicly release data processing, training and inference details.

[4] https://huggingface.co/ftshijt/espnet_codec_dac_large_v1.4_360epoch

[5] https://huggingface.co/espnet/xeus , a K-means tokenizer trained on the last-layer representations with 5k clusters

A.5 Latency and Interaction Behavior of Duplex Models

Latency and Real-Time Factor We evaluate the real-time performance of our system using the Real-Time Factor (RTF), defined as the ratio between end-to-end processing latency and input speech duration. An RTF < 1 is required for real-time deployment.

As shown in Tab. 11, our RLAIF duplex SDS achieves an RTF of 0.56, which is well below the real-time threshold and matches the efficiency of the SCoT-Response backbone. Compared to turn-based systems, this represents a substantial speedup over Direct E2E (RTF 0.75) and Multi-turn CoT E2E (RTF 2.95), confirming that our system satisfies the computational constraints required for real-time duplex interaction.

We further evaluate an optimized vLLM-based implementation, achieving a generation speed of approximately 110 tokens per second. This corresponds to a P50 latency of 540 ms and a P90 latency of 1320 ms. These results demonstrate sub-second median latency and are comparable to prior duplex systems such as Moshi (reported P50 latency of $\sim$890 ms under our evaluation setup).

Duplex Interaction Behavior Following prior work (Arora et al., 2025c), we evaluate duplex-native behavior by measuring the precision and recall of predicted speech overlaps relative to human reference overlaps. This metric captures the model's ability to engage in simultaneous speaking and listening, which is not applicable to turn-based systems.

The SCoT-Response model already improves overlap prediction compared to prior duplex systems, and RLAIF further enhances this behavior, increasing precision to 60.4% and recall to 77.7%. These results demonstrate that the proposed multi-reward alignment framework preserves and improves duplex interaction capability, enabling the system to more accurately predict when to speak while listening.

A.6 Comparison with Larger-Scale Models

To contextualize the performance of our alignment framework relative to systems with substantially greater model capacity, we compare against Qwen-Omni (Xu et al., 2025), a 7.5B-parameter speech dialogue model. Since Qwen-Omni operates in a turn-based setting, we compare it against our turn-based Multi-turn CoT E2E models to ensure a

Table 9: Comparison with Qwen-Omni and RLAIF-aligned models across semantic and acoustic metrics.

System	Params	Perplexity ↓	Qwen2.5 Judge ↑	Gemini 2.5 Judge ↑	UTMOS ↑	WER ↓
Qwen-Omni	7.5B	23.3	6.26	5.17	**4.15**	4.70
Multi-turn CoT E2E	1.7B	21.2	6.18	4.00	2.16	6.10
+ RLAIF (Joint-Reward-v2)	1.7B	**19.9**	**6.33**	**5.09**	2.85	**1.70**

Table 10: RLHF Post-training Parameters. Hyperparameters found based on performance on validation set.

Parameter	Value
gradient_accumulation_steps	1
epochs	2
gradient_clipping	100.0
bf16 enabled	true
optimizer type	Adam
optimizer lr	0.0000006
optimizer betas	[0.9, 0.95]
optimizer eps	1e-8
optimizer weight_decay	3e-7
optimizer adam_w_mode	true
scheduler type	"WarmupCosineLR"
scheduler warmup_type	linear
scheduler total_num_steps	9000
scheduler warmup_num_steps	100
scheduler warmup_min_lr	0
scheduler warmup_max_lr	0.00001

Table 11: Real-time factor (RTF) across systems. Lower is better; RTF < 1 indicates real-time capability.

Model / System	Interaction Type	RTF ↓
Direct E2E	Turn-based	0.75
Multi-turn CoT E2E	Turn-based	2.95
Moshi	Duplex	0.18
SCoT-Response	Duplex	0.56
RLAIF Duplex SDS (Ours)	Duplex	0.56

Table 12: Duplex interaction behavior measured via overlap prediction.

Model / System	Precision (%) ↑	Recall (%) ↑
Moshi	45.8	40.9
SCoT-Response	59.5	66.8
RLAIF Duplex SDS (Ours)	**60.4**	**77.7**

consistent evaluation protocol. Two differences in experimental conditions should be noted: (i) Qwen-Omni is trained on significantly larger and more diverse datasets, and (ii) it uses high-quality studio-recorded assistant voice prompts, whereas our models are conditioned on lower-fidelity Switchboard speaker prompts, which directly influences absolute audio quality scores.

As shown in Table 9, the 1.7B RLAIF-aligned model matches or surpasses Qwen-Omni on semantic metrics, achieving lower perplexity, a comparable Qwen2.5-Instruct judge score, and a higher Gemini 2.5 Flash judge score. The RLAIF system also achieves substantially lower WER, reflecting higher speech realization fidelity. On the other hand, Qwen-Omni attains a considerably higher UTMOS score (4.15 vs. 2.85), a gap we attribute primarily to its studio-quality voice prompts rather than a fundamental difference in generation capability.

These results suggest that multi-reward RLAIF can substantially narrow the quality gap introduced by differences in backbone scale and training data, and that the framework can be readily applied to stronger or larger SDS backbones as they become available.

A.7 Evaluation with Independent LLM Judges

We further evaluate model outputs using independent LLM judges, namely Gemini 2.5 Flash and Gemini 3 Pro Preview, to assess robustness against potential judge bias. These models are not used

Table 13: Evaluation with independent LLM judges.

Judge	System	Avg. Score ↑	Win Rate ↑
Gemini 2.5 Flash	Multi-turn CoT E2E	4.00	✗
	+ RLAIF (Joint-Reward-v2)	**5.09**	60.5
Gemini 3 Pro Preview	Multi-turn CoT E2E	3.49	✗
	+ RLAIF (Joint-Reward-v2)	**4.43**	63.3

Table 14: Multi-reward DPO training pipeline.

Step	Description
1. Per-reward construction	Construct independent preference datasets for each reward type (semantic, audio quality, intelligibility, emotion).
2. Dataset-level mixing	Combine all preference datasets into a unified pool and shuffle across reward types.
3. Minibatch sampling	Uniformly sample minibatches from the mixed dataset without reward-specific scheduling.
4. Joint DPO optimization	Apply a single DPO objective over sampled pairs to jointly optimize all reward dimensions.

during preference data construction or training. As shown in Table 13, RLAIF yields consistent improvements over the base model across both judges. The gains are statistically significant ($p<0.01$) under both Wilcoxon signed-rank and paired bootstrap resampling tests, indicating that the observed improvements are not specific to the evaluation model and generalize across different LLM-based evaluators.

A.8 Multi-Reward DPO Data Construction and Sampling

We adopt a dataset-level formulation for multi-reward DPO training as shown in T. 14. Preference pairs are first constructed independently for each reward type (Section 5), and then combined into a unified dataset. During training, preference pairs from different rewards are shuffled and sampled uniformly at the minibatch level, and a single DPO objective is applied without stage-wise scheduling. This design ensures that all reward dimensions are optimized jointly while keeping the training procedure simple and stable.

A.9 LLM Prompt

Prompt Used for LLM-Based Evaluation and DPO Data Construction

```
Please rate the response from the voice dialogue system based on the human reference response
and input user utterance and following criteria (1-10 points), and provide a brief evaluation:
1. Relevance: Is the response relevant to the query? Is the content related?
2. Accuracy: Does the response correctly address the user's query and provide accurate information?
3. Completeness: Does the response comprehensively cover all aspects of the query?
4. Conversational Nature: Is the response easy to understand, concise, clear, and fluent?
Output in JSON format:
{
"Strengths": "Positive aspects of the response",
"Weaknesses": "Negative aspects of the response",
"Overall Evaluation": "Overall assessment of the response",
"Total Score (out of 10, directly provide the score)": ""
}
<Dialogue Context>
Reference: <Reference Response>
Agent: <SDS Response>
```