Modeling and Benchmarking Spoken Dialogue Rewards with Modality and Colloquialness

Anonymous ACL submission

Abstract

The rapid evolution of end-to-end spoken dialogue systems demands transcending mere textual semantics to incorporate paralinguistic nuances and the spontaneous nature of human conversation. However, current methods struggle with two critical gaps: the *modality gap*, involving prosody and emotion, and the *colloquialness gap*, distinguishing written scripts from natural speech. To address these challenges, we introduce **SDiaReward**, an end-to-end multi-turn reward model trained on **SDiaReward-Dataset**, a novel collection of episode-level preference pairs explicitly targeting these gaps. It operates directly on full multi-turn speech episodes and is optimized with pairwise preference supervision, enabling joint assessment of modality and colloquialness in a single evaluator. We further establish **ESDR-Bench**, a stratified benchmark for robust episode-level evaluation. Experiments demonstrate that SDiaReward achieves state-of-the-art pairwise preference accuracy, significantly outperforming general-purpose audio LLMs. Further analysis suggests that SDiaReward captures relative conversational expressiveness beyond superficial synthesis cues, improving generalization across domains and recording conditions. Code, data, and demos are available at `https://sdiareward.github.io/`.

1 Introduction

Large Language models (LLMs) have driven rapid progress in text-based dialogue systems (Zhao et al., 2023), and recent efforts have begun to extend these capabilities to end-to-end spoken dialogue systems that directly perceive and generate speech (Zhang et al., 2023; Xie and Wu, 2024; Défossez et al., 2024). Spoken dialogue promises a more natural interface for human–AI interaction, yet it also raises a fundamental question: how should we reliably evaluate and optimize spoken dialogue behaviors? In practice, progress in

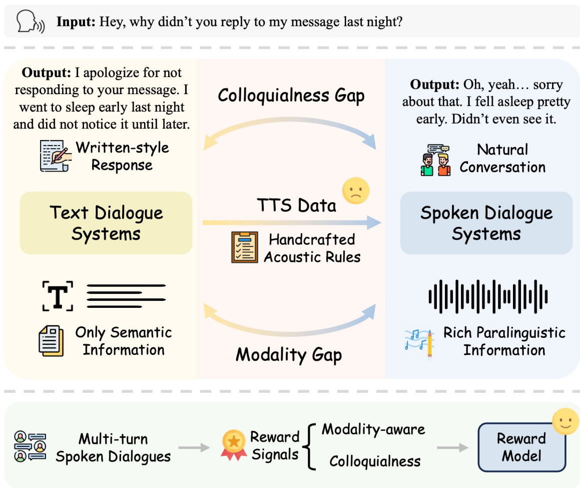


Figure 1: Challenges in spoken dialogue and our proposed framework. Text-based systems face **modality** (prosody/emotion) and **colloquialness** (style) gaps. Unlike rule-based methods, our end-to-end Reward Model learns these features from multi-turn dialogues via data-driven preference signals.

text (Ouyang et al., 2022; Cai et al., 2024; Jiang et al., 2023; Zheng et al., 2023) and vision (Wang et al., 2025; Zang et al., 2025) has been strongly enabled by reward modeling (Zhong et al., 2025) and preference learning (Christiano et al., 2017; Rafailov et al., 2023), which provide scalable supervision for alignment, reranking, and reinforcement learning. However, reliable reward modeling and evaluation for end-to-end spoken dialogue remains underexplored.

A key reason is that moving from text dialogue to spoken dialogue exposes two gaps that complicate reward design and evaluation. *i) Modality gap:* speech carries paralinguistic information such as prosody, emotion, and channel conditions. These elements strongly influence human preference yet remain invisible to text-based evaluators. *ii) Colloquialness gap:* written-style responses produced by text-optimized systems are often well-formed but sound overly scripted when spoken, while natural conversation prefers brevity, fragmentation, discourse markers, and interactional cues (Yan et al., 2025; Chen et al., 2024). Crucially, standard general-purpose evaluators often exhibit “modality blindness”—failing to distinguish between natural human speech and synthesized artifacts when semantic content is identical. Instead of relying on rigid acoustic rules, we argue for a **data-driven paradigm** where reward signals for paralinguistic fidelity and interactional spontaneity are implicitly learned from large-scale preference comparisons.

In this work, we conduct a benchmark-driven study of spoken dialogue reward modeling and evaluation. We formulate pairwise preference supervision for multi-turn spoken dialogues, and decompose reward signals into two aspects: (i) a modality-aware component that evaluates content adequacy, dialogue coherence, and spoken naturalness, and (ii) a colloquialness component that captures stylistic and interactional properties of spontaneous speech. We then establish ESDR-BENCH, a carefully stratified benchmark designed with multi-dimensional annotations to ensure distributional diversity. This enables rigorous assessment of model generalization beyond standard random splits. Experiments demonstrate that our data-driven model achieves state-of-the-art pairwise preference accuracy, significantly outperforming general-purpose Audio LLMs which struggle with the modality gap. Further analysis suggests that instead of merely detecting artifacts, our model captures relative conversational expressiveness, implicitly calibrating preference rankings within diverse acoustic domains. Our contributions are threefold:

- **Dataset.** We construct a Spoken Dialogue Reward Dataset (SDIAREWARD-DATASET) containing 11k preference pairs (200 hours of paired speech) for training spoken dialogue reward models. The full dataset will be released openly following necessary de-identification and ethics clearance.

- **Reward Modeling Framework.** We introduce an end-to-end spoken dialogue reward modeling framework in a pairwise setting and decompose evaluation into modality-aware and colloquialness rewards for multi-turn spoken dialogues.

- **Benchmark & Analysis.** We construct an episode-level spoken dialogue reward benchmark (ESDR-BENCH) with multi-dimensional annotations. Based on this, we provide an empirical analysis demonstrating the superiority of specialized data-driven reward modeling over generalist judges, offering practical insights for reliable spoken dialogue alignment.

2 Related Work

End-to-End Spoken Dialogue and Alignment The recent transition from cascaded pipelines to end-to-end spoken dialogue systems marks a significant shift in conversational AI, enabling models to integrate acoustic perception and speech generation within a unified framework (Zhang et al., 2023; Xie and Wu, 2024; Défossez et al., 2024; Xu et al., 2025a). While these systems promise enhanced interactivity and paralinguistic expressiveness, they present unique challenges for evaluation and optimization. Unlike text-based dialogue, spoken outputs must satisfy not only semantic adequacy but also prosodic naturalness and interactional spontaneity. In the textual domain, reward modeling has established itself as a cornerstone for alignment, employing techniques such as reinforcement learning from human feedback and direct preference optimization to steer model behaviors (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). However, extending these paradigms to the auditory domain remains non-trivial. Text-centric reward models inherently overlook the modality gap, while traditional automatic metrics fail to account for the colloquial nuances and long-range coherence required in spontaneous, multi-turn interaction.

Multimodal and Speech Reward Modeling As reward modeling expands beyond text, substantial research has focused on multimodal generation and understanding, accompanied by rigorous benchmarking efforts to ensure reliability and mitigate bias (Xu et al., 2023; Yu et al., 2024; Wang et al., 2025; Lambert et al., 2025; Liu et al., 2024). In the speech domain, however, preference modeling remains relatively under-explored. Existing approaches such as SpeechJudge (Zhang et al., 2025) primarily target single-turn text-to-speech quality assessment. Other recent initiatives, including ParaS2S (Yang et al., 2025b) and WavReward (Ji et al., 2025), incorporate paralinguistic signals but often depend on manually defined acoustic features or rules, which may lack the flexibility to generalize to the diversity of "wild" conversational

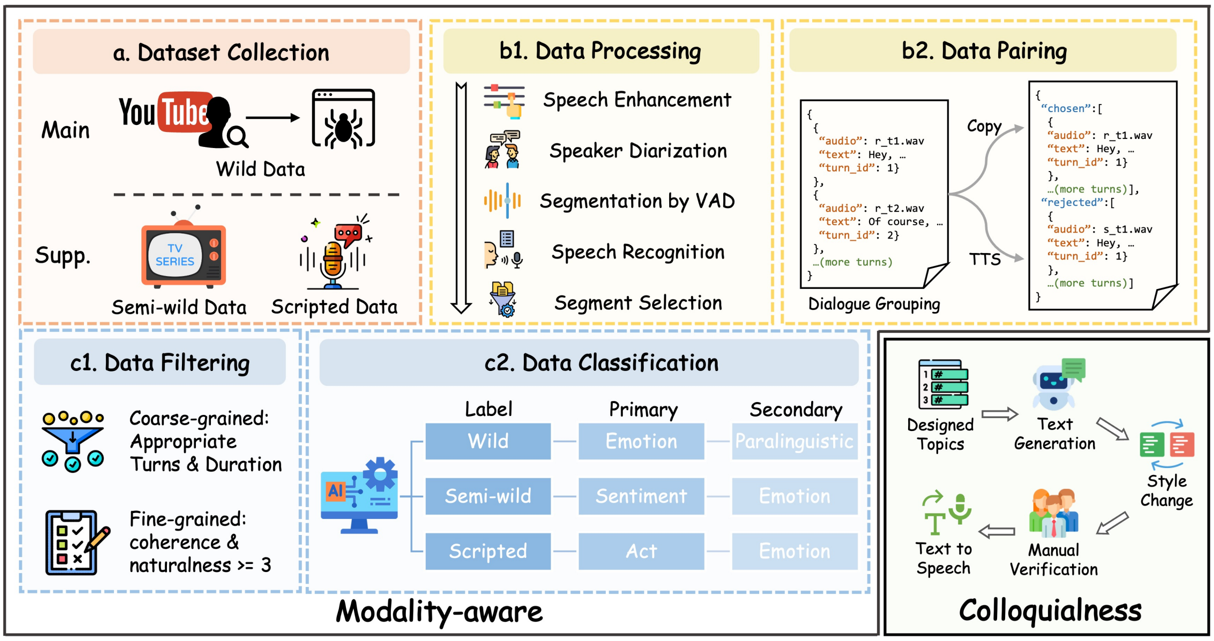


Figure 2: Overview of dataset construction. (a) **Collection:** We collect wild conversational audio (main) along with semi-wild/scripted data. (b1–b2) **Processing & Pairing:** We process audio into speaker-aware turns and group them into dialogues. We then construct two types of pairs: *modality-aware pairs* (center) via real vs. TTS audio , and *colloquialness pairs* (bottom right) via text-style vs. spoken-style generation and style change. (c1–c2) **Post-processing:** We filter episodes and attach hierarchical metadata (emotion, sentiment, act) for benchmark stratification. The detailed data processing pipeline can be found in Appendix B.

data. Distinct from these methods, our framework addresses these limitations by establishing a holistic, episode-level reward model. We aim to learn both acoustic plausibility and conversational colloquialness directly from data, bypassing the reliance on handcrafted engineering and enabling general evaluation of multi-turn spoken dialogues.

3 Dataset and Benchmark

We introduce **SDiaReward-Dataset**, a large-scale corpus specifically constructed to enable episode-level reward modeling for spoken dialogue. The dataset addresses two fundamental gaps that hinder current evaluation methods: the *modality gap*, which stems from the loss of paralinguistic cues such as prosody and emotion in standard synthesis, and the *colloquialness gap*, which arises from the stylistic divergence between rigid written scripts and spontaneous natural speech. To bridge these gaps, we curate contrastive dialogue pairs that provide supervision signals for both dimensions. The *modality-aware* subset juxtaposes real human speech with synthesized counterparts, training the model to discern authentic paralinguistic fidelity from synthesis artifacts while controlling for linguistic content. Complementarily, the *colloquialness* subset contrasts formal written-style interactions with spoken-style rewrites under consistent acoustic conditions, targeting the optimization of conversational flow and interactional spontaneity. The resulting corpus comprises approximately 13k pairwise samples (Table 1), from which we establish our stratified evaluation benchmark, **ESDR-Bench**.

Table 1: Statistics of the dataset. We categorize data by modality types and colloquialness. The unit is the pairs of dialogue.

Category	Train	Val	Total
Modality			
Wild	6,879	824	7,703
Semi-Wild	309	186	495
Scripted	2,192	466	2,658
Colloquialness	2,250	250	2,500
Total All	**11,630**	**1,726**	**13,356**

3.1 Construction Pipeline

Real-world Audio Collection. We implement a systematic pipeline designed to transform unconstrained web audio into high-quality, structured dialogue episodes. Targeting the *Wild* condition, we crawl long-form conversational content from curated YouTube domains to maintain thematic consistency. To address the inherent acoustic variability of web sources, the data undergoes a multi-stage processing chain that includes speech enhancement for noise reduction, neural speaker diarization to disentangle overlapping speech, and VAD-guided segmentation aligned with ASR transcripts. Crucially, we preserve the sequential dependencies of the original recordings by grouping segments into continuous multi-turn episodes. This structure enables the model to capture global conversational dynamics and context-dependent prosody rather than focusing solely on isolated utterance quality.

Modality-aware Pairing. To construct the modality-aware subset, we juxtapose authentic human speech with synthesized counterparts generated by `SoulX-Podcast` (Xie et al., 2025b). We selected this Dialogue-TTS system for its capacity to maintain multi-turn speaker coherence, ensuring high-fidelity "hard negatives" that force the model to discern subtle prosodic naturalness rather than trivial discontinuity artifacts. The human speech sources are stratified into three tiers: **1) Wild Data**, spontaneous multi-speaker conversations from YouTube with authentic background noise; **2) Semi-wild Data**, derived from MELD (Poria et al., 2019), featuring emotionally rich acted dialogues; and **3) Scripted Data**, sourced from DailyTalk (Lee et al., 2023), representing high-fidelity studio recordings. By pairing these diverse sources with dialogue-consistent synthesis, we isolate acoustic realization as the primary discriminative factor, prioritizing paralinguistic naturalness over spectral cleanliness.

Colloquialness Pairing. This subset targets the stylistic gap between formal text and spontaneous speech by contrasting written-style dialogues against spoken-style rewrites. We initially design 250 scenarios across 10 domains and employ LLMs to generate multi-turn written-style dialogues. These scripts are subsequently rewritten into spoken-style versions that preserve the original meaning but incorporate natural conversational patterns such as fillers, fragmentation, and discourse markers. To prevent acoustic quality from confounding the preference signal, we synthesize both the written and spoken versions using the identical TTS configuration. Consequently, the preference labels rely exclusively on the stylistic naturalness of the dialogue flow rather than differences in audio fidelity.

3.2 Quality Control and ESDR-Bench

Filtering and Annotation. To guarantee the reliability of the preference signals, we enforce a rigorous two-stage quality control protocol. Structurally, we limit dialogues to a maximum of 16 turns and restrict individual turn durations to 60 seconds to maintain manageable sequence lengths. Qualitatively, we employ an LLM-based judge to assess episode quality across three dimensions: content adequacy, dialogue coherence, and prosodic naturalness. We discard any episodes that fail to achieve a minimum threshold of 3 out of 5 on the adequacy and coherence scales. To facilitate fine-grained performance analysis, we further enrich the dataset with metadata annotations, including emotion tags and sentiment labels derived from the source material or predicted by auxiliary models.

Benchmark Stratification. We establish the **ESDR-Bench** from the held-out validation split to serve as a robust evaluation standard. A key challenge in benchmark construction is the potential dominance of high-frequency data types such as the Wild subset. To address this imbalance, we implement a stratified sampling strategy based on source and metadata categories. For each fine-grained bucket, we select a balanced set of up to 50 episodes, ensuring that the benchmark provides a distributionally diverse assessment of model generalization rather than being skewed by the underlying data distribution. Although the collected corpus naturally contains more Wild audio due to availability, ESDR-Bench uses source- and metadata-stratified sampling to prevent high-frequency regimes from dominating evaluation and to better reflect generalization.

4 Reward Modeling

Problem Setup. We consider a multi-turn spoken dialogue as a sequence of turns $\mathcal{D} = \{(a_t, x_t)\}_{t=1}^{T}$, where a_t denotes the speech audio and x_t represents the corresponding transcript. Unlike traditional reward models that focus on isolated turns, our goal is to evaluate the contextual consistency and multimodal alignment of a candidate final turn. Given a context $\mathcal{C} = \{(a_t, x_t)\}_{t=1}^{T-1}$ and candidate final turns y , the model outputs scalar rewards $r_\theta(\mathcal{C}, y)$ leveraging complete context information of the conversation.

Model Architecture. Existing speech preference models often focus on single-turn TTS evaluation,

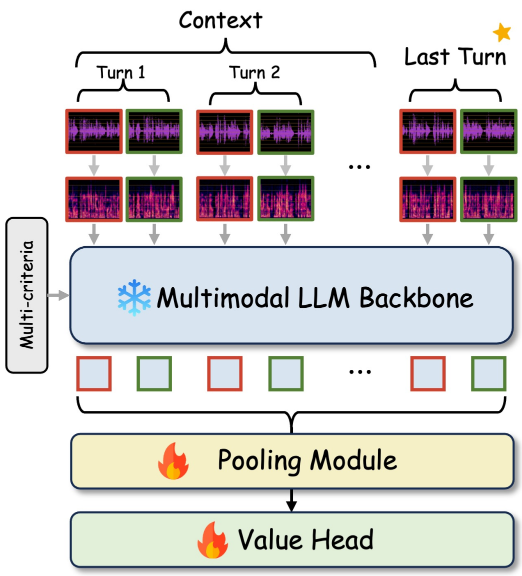


Figure 3: Architecture of our reward model.

neglecting the long-range dependency in dialogues. Others rely on handcrafted paralinguistic features, which lack the capacity to capture the nuanced "vibe" of spontaneous speech. To address these critical limitations in spoken dialogue rewarding, we develop an end-to-end multimodal reward model designed to capture the complex alignment between speech context and response. We leverage a multimodal LLM backbone to project the interleaved speech-text sequence into a joint embedding space. Let $\mathbf{H} = \{h_1, \dots, h_L\} \in \mathbb{R}^{L \times d}$ be the hidden representations extracted from the final transformer layer. The scalar reward is then computed via a task-specific score head:

$$r_\theta(\mathcal{C}, y) = \mathrm{MLP}(\textsc{Pool}(\mathbf{H})), \quad (1)$$

where $\textsc{Pool}(\cdot)$ is a pooling operator that aggregates sequence-level information. This architecture bypasses the need for intermediate text-based summarization, allowing the model to directly "hear" the prosodic nuances in the context.

Pooling and Robustness A practical consideration is how to summarize the sequence representation $\mathbf{H}$ for reward prediction. We evaluate three standard pooling operators: last-token pooling, mean pooling, and attention pooling. Empirically, mean pooling provides the most stable optimization behavior across hyperparameters and data mixtures, while attention pooling can achieve high accuracy but exhibits higher sensitivity and may allocate criterion-dependent attention patterns across distributions. Last-token pooling underperforms in our setting, suggesting that reward-relevant information is distributed across the context and final-turn representations rather than concentrated in a single position. We defer detailed ablations and quantitative comparisons to §5.

Multi-Criteria Reward Decomposition. Following the intuition of attribute-conditioned modeling, we reformulate the reward function as $r_\theta(\mathcal{C}, y, \text{inst})$, where inst is a criterion-specific system prompt. Instead of training multiple specialized models, we train a single backbone under two primary criteria: (i) Modality-Awareness: emphasizing cross-turn acoustic coherence and prosodic naturalness. (ii) Colloquialness: emphasizing conversational spontaneity and the avoidance of "robotic" formalisms. This conditioned approach allows the model to share general linguistic representations while learning distinct decision boundaries for diverse evaluative dimensions, effectively replacing brittle, handcrafted rules with learnable, data-driven priors.

Loss Function. We optimize the model using the Bradley-Terry preference framework (Bradley and Terry, 1952). Given a context $\mathcal{C}$ and a pair of responses (y^+, y^-) where y^+ is preferred, the training objective is to minimize the negative log-likelihood:

$$\mathcal{L}_{\text{pref}}(\theta) = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(r_\theta(\mathcal{C}^+, y^+) - r_\theta(\mathcal{C}^-, y^-)\right)\right], \quad (2)$$

where σ is the sigmoid function. This objective encourages the model to assign higher scalar rewards to responses that better satisfy the conditioned criteria within the given dialogue context. However, strictly pairwise optimization can lead to unbounded score drift. This issue is magnified in speech reward modeling where domain shifts are prevalent. For instance, when moving from noisy YouTube audio to clean studio recordings, the model might prioritize channel characteristics as a shortcut, rewarding cleaner audio with higher absolute scores even if the dialogue quality is inferior. To address this sensitivity and stabilize the reward scale, we adopt the centering regularization term $\mathcal{L}_{center}$ from Eisenstein et al.'s (2023):

$$\mathcal{L}_{\text{center}}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\left(r_\theta(\mathcal{C}^+, y^+) + r_\theta(\mathcal{C}^-, y^-)\right)^2\right]. \quad (3)$$

The final training objective is formulated as:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{pref}}(\theta) + \lambda \cdot \mathcal{L}_{\text{center}}(\theta), \quad (4)$$

where λ is the centering coefficient. This constraint anchors the reward distribution around zero, ensuring that the model learns relative preferences within each domain rather than absolute biases based on recording conditions.

5 Experiments

5.1 Experiment Setup

Baselines. We evaluate two categories of evaluators: (1) **Zero-shot Audio Judges**, including proprietary (GPT-4o-audio Hurst et al., 2024, Gemini 2.5 Team et al., 2023) and open-source models (Qwen-Omni Xu et al., 2025a,b, Kimi-Audio Ding et al., 2025, VITA-Audio Long et al., 2025). Due to the lack of existing open-source reward models specialized for multi-turn spoken dialogue, we use these generalist baselines to quantify the gap between zero-shot capabilities and targeted alignment. (2) **Supervised Baselines**, specifically our SDiaReward-3B/7B (based on Qwen2.5-Omni), fine-tuned on SDIAREWARD-DATASET using a pairwise ranking objective via the `trl` (von Werra et al., 2020) library.

Evaluation Metrics. Our primary metric is **pairwise accuracy**, defined as the fraction of preference pairs whose ordering is correctly predicted by the reward scores. For a labeled preference $a \succ b$, the prediction is correct when $R_a > R_b$. To probe generalization across data regimes, we report both **Micro** and **Macro** averages. Micro accuracy aggregates over all test pairs and is dominated by larger subsets. Macro accuracy averages results over each subset, penalizing models that overfit to a single regime and providing a stricter view of generalization.

Implementation Details. Initialized from Qwen2.5-Omni, SDiaReward uses a linear head on pooled representations for scalar scoring. Audio is truncated/padded to 30s. Full hyperparameters are in Appendix A.

5.2 Main Results

Table 2 summarizes the performance on ESDR-Bench.

Dedicated Reward Modeling Unlocks Modality-Aware Evaluation. A striking observation is the struggle of general-purpose audio judges on the modality benchmark. While closed-source models like Gemini 2.5 Pro achieve saturation on colloquialness tasks (98.80%), their ability to distinguish real human speech from synthesized audio is limited (72.63% Micro Acc). This suggests that zero-shot judges prioritize semantic content over acoustic naturalness. In contrast, our proposed SDiaReward-7B demonstrates substantial gains, achieving **96.61%** Micro Accuracy on the modality benchmark. This underscores the necessity of targeted pairwise supervision for learning subtle paralinguistic preferences that general pre-training may overlook.

The *colloquialness* gap vs. the *modality-aware* gap. The high performance of baseline models on the Colloquialness subset indicates that preferences for "spoken style" can often be inferred from textual/linguistic cues like grammar which are well-preserved in the semantic latent space of ALMs. However, the Modality task—requiring discrimination between two audio clips with *identical* text content but differing prosody—proves much harder for baselines. SDiaReward's superior performance here confirms its ability to effectively disentangle and value acoustic nuances beyond mere semantics.

Performance Consistency Across Domains. Analysing Micro and Macro averages reveals significant differences in domain adaptability. SDiaReward-7B maintains consistent accuracy across heterogeneous splits (94.91% Macro), mitigating the sharp divergence observed in the 3B model (88.62% Micro vs. 79.20% Macro). The discrepancy is most pronounced in the **Semi-wild** subset, where the 3B model's accuracy drops to 55.38%. This suggests that while smaller models may latch onto prominent domain features in "Wild" or "Scripted" data, the complex, "semi-scripted" nature of Semi-wild interactions requires sufficient model scale to resolve effectively.

Human Alignment and Calibration. We run a blinded human study on 75 stratified pairs (Table 3). Each pair is independently rated by three annotators, and we report the average agreement rate with the dataset ground-truth label. *Random Sampling* shows **76.7%** agreement, while *High Confidence* **88.3%** is higher than *Low Confidence* 78.3%, suggesting margins are indicative of human-perceived correctness. For *Hard Negatives*, humans still agree with the ground truth in **93.3%** of cases; disagreements are often from SEMI-WILD (MELD) pairs, likely related to **text–audio misalignment** and **incomplete slicing**. Overall weighted agreement is **83.5% (± 4.3%)**.

Table 2: **Main results on ESDR-Bench.** We report pairwise preference accuracy (%) on the modality benchmark split into wild/semi-wild/scripted and on the colloquialness benchmark. **Modality Micro** is the weighted accuracy over all modality pairs; **Modality Macro** is the unweighted mean across the three modality subsets, serving as a stricter metric for generalization. **Overall Micro** is the weighted accuracy over all benchmark pairs, while **Overall Macro** averages modality macro and colloquialness accuracy.

Model	**Modality Acc**			**Modality**	**Modality**	**Colloq.**	**Overall**	**Overall**
	Wild	Semi-wild	Scripted	**Micro**	**Macro**	**Acc**	**Micro**	**Macro**
Closed-source Judges								
Gemini 2.5 Pro	76.12	65.60	69.77	72.63	70.50	98.80	76.42	84.65
Gemini 2.5 Flash	56.73	53.67	53.91	55.41	54.77	99.60	61.81	77.19
GPT-4o Audio	52.61	49.54	49.26	51.12	50.47	98.00	57.91	74.23
Open-source Audio LMs								
Kimi-Audio	71.27	62.84	56.03	65.30	63.38	66.00	65.40	64.69
Qwen 3 Omni 30B	64.48	54.59	48.84	58.18	55.97	97.20	63.83	76.59
Qwen 2.5 Omni 7B	51.64	51.38	52.43	51.85	51.82	49.20	51.47	50.51
Qwen 2.5 Omni 3B	51.88	44.50	47.15	49.34	47.84	52.00	49.73	49.92
VITA-Audio	46.79	44.95	52.64	48.35	48.13	50.80	48.70	49.47
Ours								
SDiaReward 7B	100.00	92.47	92.27	**96.61**	**94.91**	97.20	**96.70**	**96.06**
SDiaReward 3B	99.39	55.38	82.83	88.62	79.20	92.00	89.11	85.60

Table 3: **Human Verification Results.** We evaluate 75 stratified samples with averaged multi-annotator ratings. *Human Agree.* denotes agreement with the dataset ground truth; *Avg. Margin* is the model margin on each subset, and *SE* reports the standard error.

Sample Subset	**Count**	**Avg. Margin**	**Human Agree.**	**SE**
Validation of Model Capability				
High Confidence	20	1.65	88.3%	± 7.2%
Low Confidence	20	0.06	78.3%	± 9.2%
Random Sampling	20	0.77	76.7%	± 9.5%
Error Analysis (Hard Negatives)				
Model Wrong†	15	-0.19	**93.3%**	± 6.5%
Overall (Weighted)	**75**	**0.62**	**83.5%**	**± 4.3%**

†Model predicts the opposite label (mis-ranked pairs).

Table 4: **Ablation Study.** Performance comparison across pooling strategies and model scales. The *Mean* strategy with center loss achieves the best trade-off between stability and accuracy.

Setting	**Modality**	**Colloq.**	**Overall**
Pooling Strategy (3B Backbone)			
Last Hidden	63.75	48.80	61.59
Attention	87.94	93.60	88.76
Mean	**88.62**	**92.00**	**89.10**
Pooling Strategy (7B Backbone)			
Last Hidden	51.83	40.00	50.12
Attention	70.60	55.20	68.37
Mean	**96.61**	**97.20**	**96.70**
Loss Formulation (7B-Mean)			
w/o Center Loss	95.05	97.20	95.37
w/ Center Loss	**96.61**	**97.20**	**96.70**

5.3 Ablation Experiments

We conduct a comprehensive ablation study to validate our architectural choices, focusing on feature aggregation, model scaling, and loss regularization.

Feature Aggregation and Scalability. We compare three pooling strategies: (1) *Last Hidden State*, (2) *Attention Pooling*, and (3) *Mean Pooling*. As shown in Table 4, ***Mean Pooling*** consistently outperforms others. We posit that while *Last* pooling is sensitive to local boundary noise, *Mean* pooling aggregates the holistic episode-level context, yielding a more linearly separable representation. Scaling from 3B to 7B further boosts performance, with 7B-Mean achieving state-of-the-art results.

Impact of Center Loss Regularization. Standard reward modeling often suffers from unbounded score drift. Figure 4(a) illustrates this issue: the baseline model's average chosen reward drifts to $\mu \approx 5.03$. By introducing center loss, we align the global average to $\mu \approx 0.32$ (Orange curve) without compromising the discriminative margin (Fig. 4(b)). This calibration not only stabilizes training but also slightly improves accuracy ($95.37\% \rightarrow 96.70\%$) by preventing logit saturation.

Analysis of Domain-Specific Bias. Despite the high global accuracy, a granular analysis of score distributions reveals intrinsic domain biases. Figure 4(c) and (d) decompose the scores by data source: **i) High Confidence in Wild Data**: The model exhibits high certainty on *Wild* data, with chosen scores tightly clustered around $+0.8$ and a clear separation from rejected samples. **ii) Adaptive Decision Boundaries**: Interestingly, for

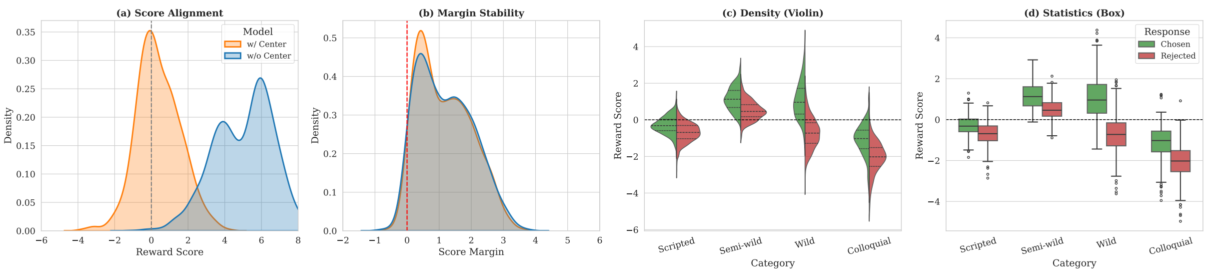


Figure 4: **Ablation Analysis on SDiaReward Model (7B).** (a) **Score Alignment**: The proposed center loss (Orange) effectively anchors the chosen reward distribution to $\mu \approx 0.32$, whereas the baseline (Blue) suffers from significant drift ($\mu > 5.0$). (b) **Margin Stability**: The discriminative margin remains robust. (c) **Density Modes**: Split violin plots visualize reward density, showing high confidence in *Wild* data. (d) **Statistical Ranges**: Box plots reveal domain-dependent decision boundaries; notably, *Scripted* responses receive lower absolute scores despite being correct choices.

Scripted and *Colloquial* data, we observe a negative shift in the score distribution. As shown in the box plots (Figure 4(d)), the median chosen score for *Scripted* data is negative (≈ -0.24), yet the model maintains high classification accuracy. This phenomenon indicates that the Reward Model implicitly learns a *relative* ranking function calibrated to the specific difficulty or style of each domain, rather than a globally absolute metric. While Center Loss normalizes the global mean, these local offsets suggest that future work should explore domain-invariant alignment techniques to further standardize reward scales.

6 Discussion

The Asymmetry of Spoken Dialogue Gaps. Our empirical analysis reveals a fundamental divergence where the *colloquialness gap* is effectively bridged by the linguistic priors of LLMs, whereas the *modality gap* remains the primary technical bottleneck. General-purpose audio models struggle to distinguish prosodic naturalness from synthesis artifacts and often perform near chance levels. SDIAREWARD resolves this by integrating modality-based supervision to ensure high-scoring responses possess both grammatical spontaneity and acoustic authenticity. This unified approach prevents the optimization pipeline from regressing into "scripted synthesis" where responses sound textually informal but prosodically rigid.

Reward as Relative Expressiveness. SDIAREWARD goes beyond simple artifact detection to acquire a metric of *relative expressiveness*. As shown in Figure 4(d), correctly ranked pairs in the *Scripted* domain consistently receive lower absolute scores compared to the *Wild* domain. This pattern indicates that the model implicitly calibrates to the dynamic range inherent to each domain. Such calibration is vital for reinforcement learning as it encourages the generation of emotionally rich interactive behaviors rather than spectrally clean but monotonic audio.

7 Conclusion

In this work, we take a step toward better implicitly reward modeling and evaluation for end-to-end spoken dialogue systems. We introduce SDIAREWARD-DATASET, a comprehensive pairwise preference corpus, and ESDR-BENCH for general episode-level benchmarking. Our end-to-end reward model achieves state-of-the-art accuracy, effectively distinguishing paralinguistic naturalness and conversational spontaneity where general-purpose models fail. Crucially, our analysis suggests that the model learns a general measure of relative expressiveness rather than simple artifact detection. However, we also observe domain-dependent offsets in absolute reward scores. Future work should focus on deriving more general reward signals by refining data diversity and exploring domain-invariant modeling objectives. Such advancements are essential to standardize reward scales across heterogeneous sources, paving the way for stable and scalable reinforcement learning in next-generation spoken dialogue systems.

Limitations

While SDiaReward achieves state-of-the-art performance, our current dataset prioritizes "in-the-wild" recordings to target the complexity of real-world acoustic environments. Future iterations could further enhance robustness by incorporating a broader spectrum of high-quality acted speech and diverse synthesis engines. Additionally, while our human verification confirms high alignment with model predictions, larger-scale studies exploring fine-grained subjective preferences remain a promising direction for future research.

Ethics and Responsible Use

This section discusses the ethical considerations, intended use, and responsible data release practices associated with SDiaReward and ESDR-Bench, with particular attention to copyright, privacy, and biometric risks in spoken dialogue research.

Intended Use. SDiaReward and ESDR-Bench are intended solely for research purposes, including the evaluation and analysis of end-to-end spoken dialogue systems and reward modeling methodologies. They are not designed for deployment in real-world decision-making systems, content moderation, surveillance, or any application involving automated judgments about individuals or groups.

Data Sources and Privacy. Our dataset is constructed from publicly available audio sources in YouTube and established research benchmarks MELD and DailyTalk. We do not redistribute raw audio recordings from third-party platforms. Our release excludes speaker-identifiable representations and persistent speaker identifiers, and provides derived research artifacts only (Appendix B.5).

Copyright and Data Release Strategy. Although part of our corpus originates from publicly accessible web audio, we do not release raw audio files. To mitigate copyright risks, we release only derived artifacts such as dialogue metadata, preference annotations, benchmark splits, and evaluation scripts, strictly for non-commercial research use. Reconstructing or accessing any underlying audio content, if desired, requires users to independently obtain the data in accordance with the access conditions and platform policies of the original sources. All released resources follow the original terms and conditions of the underlying data providers.

Biometric and Speaker Identification Risks. Speech may contain biometric signals that can enable speaker identification. To reduce biometric and privacy risks, the released artifacts do not include speaker-identifiable representations and are not intended for speaker identification, biometric analysis, or any application involving individual-level profiling.

Risk Awareness for Downstream Optimization. As discussed in the Limitations section, reward models trained on heterogeneous real-world audio may exhibit sensitivity to domain-specific acoustic characteristics, which could be exploited as shortcuts during optimization. We emphasize that SDiaReward should not be treated as a substitute for human judgment and should be applied cautiously, particularly in downstream optimization settings.

Use of AI Assistants. AI assistants are used to support data preprocessing scripts and limited language refinement. All experimental design, analysis, and conclusions are determined by the authors.

References

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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. *arXiv preprint arXiv:2410.00037*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak

Ramachandran, and 1 others. 2023. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, and 1 others. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024. Rm-bench: Benchmarking reward models of language models with subtlety and style. *arXiv preprint arXiv:2410.16184*.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, and 1 others. 2025. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*.

OpenAI. 2024. Gpt-4.1 mini. `https://platform.openai.com/docs/models/gpt-4.1-mini`. Accessed: 2026-01-06.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

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. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*.

xAI. 2025. Grok 4.1. `https://x.ai/news/grok-4-1`. Accessed: 2026-01-06.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, Jun Wu, Hanlin Wen, Ruixuan Shang, Hongmei Liu, Zhiqi Jiang, Yuepeng Jiang, Wenxi Chen, Ruiqi Yan, Jiale Qian, Yichao Yan, Shunshun Yin, Ming Tao, Xie Chen, Lei Xie, and Xinsheng Wang. 2025a. Soulx-podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity.

Hanke Xie, Haopeng Lin, Wenxiao Cao, Dake Guo, Wenjie Tian, Jun Wu, Hanlin Wen, Ruixuan Shang, Hongmei Liu, Zhiqi Jiang, and 1 others. 2025b. Soulx-podcast: Towards realistic long-form podcasts with dialectal and paralinguistic diversity. *arXiv preprint arXiv:2510.23541*.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. URO-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17211–17242, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shu-wen Yang, Ming Tu, Andy T Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2025b. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *arXiv preprint arXiv:2511.08723*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, and 1 others. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and 1 others. 2025. Speechjudge: Towards human-level judgment for speech naturalness. *arXiv preprint arXiv:2511.07931*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*.

A Training Details

We initialize SDiaReward using the Qwen2.5-Omni (3B/7B) backbone, extending it with a linear regression head atop the pooled representation of the final hidden layer to derive a scalar reward. Audio episodes are standardized to a 30-second duration via truncation or padding. The model is optimized using a Bradley-Terry pairwise loss framework, augmented with a reward centering term ($\lambda = 10^{-2}$) to stabilize score distribution.Training is conducted for a single epoch using the AdamW optimizer with a peak learning rate of 2×10^{-5} and a weight decay of 0.05. We employ a cosine learning rate schedule preceded by a 0.15 warmup phase, alongside a gradient clipping threshold of 1.0. For computational efficiency, we leverage DeepSpeed ZeRO-2 across 4 GPUs, utilizing FlashAttention-2, `bf16` precision, and gradient checkpointing. The total batch size is configured to 32 (per-device batch size of 4 with a gradient accumulation factor of 2). Model performance is monitored every 50 steps on the validation split; the optimal checkpoint is selected based on minimal validation loss, with a rolling buffer of the 20 most recent checkpoints maintained throughout training.

B Reward Dataset

B.1 Details of the Modality-aware Subset Construction

Data Collection and Preprocessing

Multi-source Data Acquisition Strategy We adopt a hybrid data acquisition strategy that combines large-scale "in-the-wild" recordings with curated public benchmarks. For the large-scale "in-the-wild" data, we obtain high-quality conversation audio by selecting a group of YouTube creators who specialize in interviews and podcast production. We search targeted keyword (e.g., "podcast", "interview") to identify specific content for each channel and automate the retrieval process using the `ytdlp` tool, strictly adhering to a high-fidelity protocol by selecting the best available audio streams (`bestaudio/best`) and enabling the `noclobber` parameter to ensure data integrity. This rigorous scraping pipeline yields approximately 1,954.2 hours of raw, unconstrained audio. To improve generalization and mitigate overfitting, we supplement our corpus with two authoritative public benchmarks: MELD (Poria et al., 2019) and DailyTalk (Lee et al., 2023). This combination balances the natural prosodic variability of massive unorganized audio with the structured annotations of reference datasets, creating a robust foundation for model training.

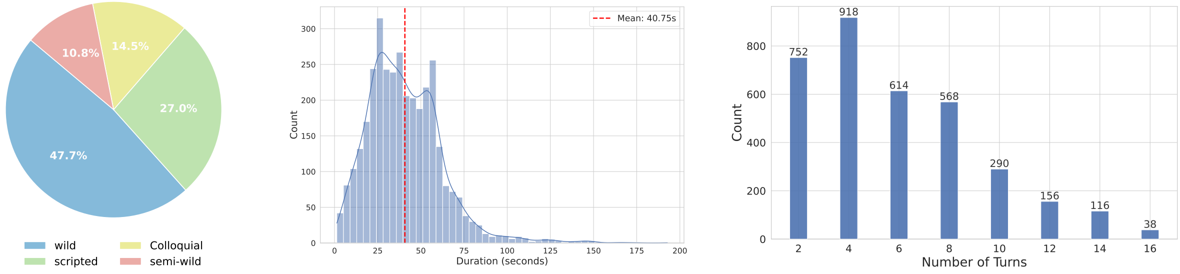
Audio Processing Pipeline Construction To extract turn-level audio and its duration and text, we design a customized end-to-end processing pipeline based on the Emilia (He et al., 2024) framework which handles various heterogeneous data sources. For unstructured YouTube audio, the pipeline executes a sequence of speech enhancement (`MDX23C-8KFFT-InstVoc_HQ`[1]), speaker diarization (`speaker-diarization-community-1`[2]), and fine-grained VAD (`silero_vad` (Team, 2024)). ASR is then performed using `whisper-large-v3` (Radford et al., 2022), initialized with a specific prompt to retain disfluencies (e.g., "um", "uh") and prevent filler word omission. We strictly retain only the two dominant speakers, discarding segments where secondary speakers exceed 10% of the duration. For structured datasets (DailyTalk, MELD), we prioritize fidelity by bypassing VAD and ASR inference to avoid error propagation, and directly rely on the provided metadata for alignment, while applying consistent speech enhancement. Following extraction, turn-level audio is organized into dialogue groups with granular controls: a minimum interval of 0 seconds, an overlap ratio ≥ 0.1, and a strict duration cap of 90 seconds. Through this rigorous pipeline, we process a total of 749.61 hours of turn-level audio from YouTube, supplemented by 21.93 hours from DailyTalk and 21.67 hours from MELD, resulting in structured JSON transcripts and segmented audio data.

[1]https://github.com/Anjok07/ultimatevocalremovergui
[2]https://huggingface.co/pyannote/speaker-diarization-community-1

Table 5: Hierarchical Classification across Datasets.

Dataset	Primary Dimension	Secondary Dimension
YouTube*(Wild)*	Anger Disgust Fear Happiness Neutral Sadness Surprise	Cough Cry Filled Pauses Laughter Listener Feedback Sigh/Breath No Feature
MELD*(Semi-wild)*	Negative Neutral Positive	Anger Disgust Fear Joy Neutral Sadness Surprise
DailyTalk*(Scripted)*	Commissive Directive Inform Question Unknown	Anger Disgust Fear Happiness No Emotion Sadness Surprise

Synthetic Audio Generation and Preference Pair Organization We utilize the `soulxpodcast` (Xie et al., 2025a) framework to generate high-quality synthetic audio for reward model training via zero-shot cloning. We design a greedy heuristic for reference audio selection to capture rich acoustic features, prioritizing clips with a duration between 5 and 30 seconds and a word count under 60. Dialogue groups lacking viable prompts are pruned. This process yields a cumulative synthetic corpus of 269.97 hours from YouTube, 18.97 hours from DailyTalk, and 3.58 hours from MELD. Finally, we structure the data into preference pairs: the generated synthetic audio is designated as the rejected response, while the original ground-truth audio serves as the chosen response.



(a) Category Distribution (b) Duration Distribution (c) Turns Distribution

Figure 5: Overview of the ESDR-Bench

Data Filtering Process and Results

Phase 1: Deterministic Rule-Based Filtering We apply deterministic rule-based constraints to ensure structural integrity and computational feasibility. We mandate that all dialogue groups consist of an even number of turns—guaranteeing strictly alternating user-assistant interactions—capped at a maximum of 16 turns to prevent context window overflow. Furthermore, individual turn durations are restricted to a maximum of 60 seconds. Any samples failing to meet these strict formatting or duration criteria are rigorously excised to maintain a clean and stable dataset.

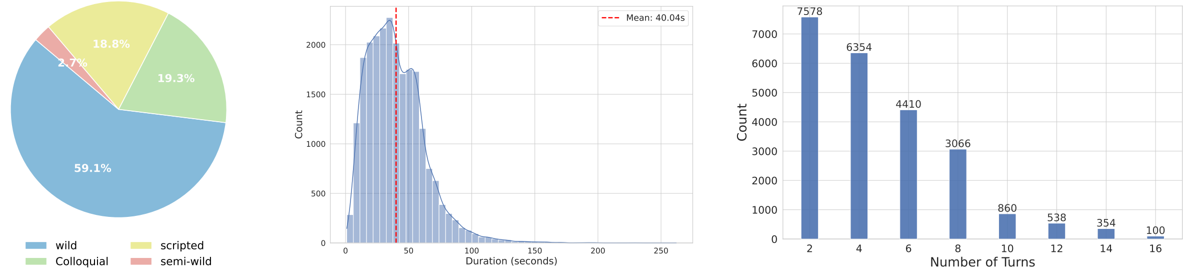
Phase 2: Automated Quality Assessment via LLM We initiate the data refinement process with an automated evaluation leveraging the multimodal capabilities of the `Gemini 2.5 Pro` (Comanici et al., 2025). The model performs a comparative analysis between the ground-truth audio and its synthesized counterpart within an identical multi-turn context, focusing specifically on the quality of the final turn. The evaluation employs a 5-point scale across three dimensions: `final_turn_content` (semantic accuracy), `final_turn_naturalness_prosody` (acoustic realism), and `dialog_context_coherence` (contextual logic). The model is required to output structured JSON data containing dimension-specific scores, a binary preference decision, and a concise justification ($\leq$ 80 words). The specific prompt template is detailed in Figure 9. We enforce a retention threshold based on semantic integrity; samples are preserved only if they achieve scores $\geq$ 3 in both `final_turn_content` and `dialog_context_coherence`.

Filtering Results This rigorous dual-filtering mechanism effectively removes low-quality samples and disjointed contexts, thereby guaranteeing the reliability of the training corpus. The final curated dataset contains **96.48 hours** of real audio and **110.32 hours** of synthetic audio.

ESDR-Bench Construction

Hierarchical Data Classification and Annotation To systematically address the heterogeneity of our data sources, we establish a hierarchical classification taxonomy tailored to the provenance of each dataset, as detailed in Table 5. For the unstructured "Wild" data from YouTube, we employ the `Gemini 2.5 pro` to perform granular annotation, establishing emotion as the primary category and specific paralinguistic features (e.g., laughter, filled pauses) as the secondary dimension. In contrast, for the structured datasets, we prioritize fidelity to their original schema: MELD ("Semi-wild") is categorized primarily by sentiment followed by emotion, while DailyTalk ("Scripted") is organized by dialogue acts subdivided by emotional state.

Quality-Aware Sampling and Isolation Based on this taxonomy, we implement a stratified sampling protocol targeting the secondary dimensions of each dataset. We apply a uniform cap to ensure balanced representation: for categories containing fewer than 50 groups, all available samples are retained; conversely, for categories exceeding this threshold, we extract 50 instances. To guarantee a strictly independent evaluation environment, all selected validation samples are rigorously excised from the training corpus, thereby eliminating any risk of data leakage. The final resulting modality validation set comprises **14.51 hours** of real audio and **16.18 hours** of synthetic speech. Figure 5 reports the basic statistics of the ESDR-Bench dataset.

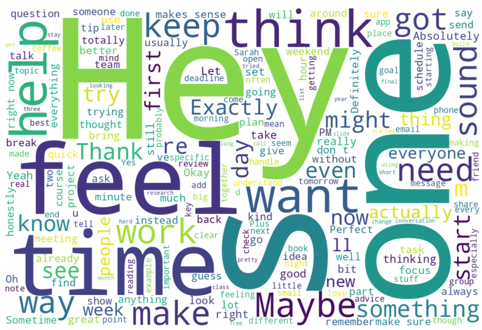


(a) Category Distribution (b) Duration Distribution (c) Turns Distribution

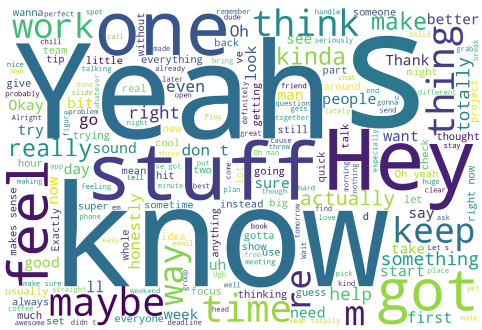
Figure 6: Overview of the SDiaReward-Dataset

B.2 Details of the Colloquialness Subset Construction

The Colloquialness-gap pairs are constructed across ten carefully designed domains: small talk, information seeking, practical task, planning coordination, decision support, emotional support, relationship building, social conflict, academic learning, and professional work. Within each domain, we define 25 specific topics to ensure comprehensive coverage of daily conversational scenarios. To ensure the diversity of the generated content, we utilize five models, including `Gemini 2.5 Pro`, `GPT-4.1 mini` (OpenAI, 2024), `Grok 4.1` (xAI, 2025), `Qwen2.5-72B-Instruct` (Qwen et al., 2025) and `Qwen3-235B-Instruct` (Yang et al., 2025a), to generate two written-style samples per model for each topic, resulting in a total of 2,500 samples ($10 \times 25 \times 5 \times 2$). Subsequently, these samples are rewritten into a more natural colloquial style using the aforementioned models. As illustrated in Figure 7, a key characteristic of the colloquial versions is the increased usage of filler words such as "yeah", "oh", and "uh". We pair the spoken-style data with the written-style data, designating the spoken version as the chosen response and the corresponding written version as the rejected response. To construct the Colloquialness subset for ESDR-Bench, we select one sample from each topic, yielding a total of 250 instances, while the remaining samples are allocated to the SDiaReward-Dataset for training purposes.



(a) Written-style Data



(b) Spoken-style Data

Figure 7: Word Cloud of Colloquial Data

B.3 Overview of the SDiaReward-Dataset

As shown in Figure 6a, the data is predominantly composed of "Wild" data collected from real-world scenarios (accounting for 59.1%), while incorporating diverse categories such as "Semi-wild," "Scripted," and "Colloquial." This composition helps the model effectively narrow the modality-aware gap and the colloquialness gap. The data covers a wide distribution ranging from 2 to 16 turns, as displayed in Figure 6b. While dialogues with 2 to 10 turns constitute the core proportion, the substantial quantity of long-dialogue samples effectively ensures the model's ability to learn from contexts of varying lengths and grasp multi-turn interaction logic. Figure 6c indicates that the average dialogue duration is 40.04 seconds, mainly concentrated within the 20 to 80-second range. These samples possess sufficient information density and complete context, which facilitates the model in capturing richer acoustic and semantic features.

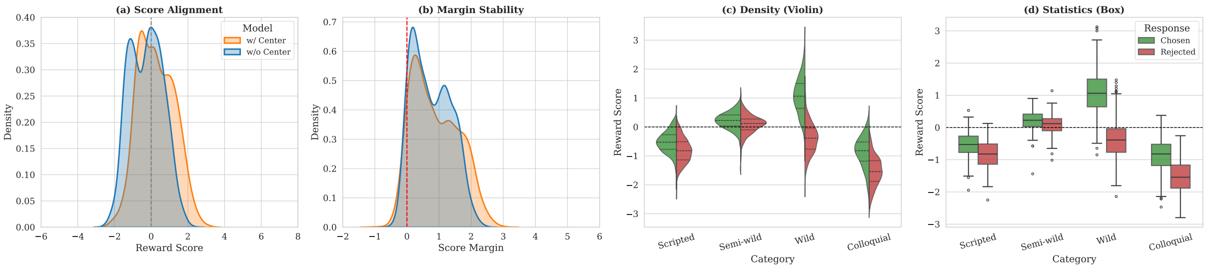


Figure 8: Ablation Analysis on SDiaReward Model (3B).

B.4 Dataset Construction Prompts

This section presents the complete prompts utilized throughout the entire construction pipeline of the SDiaReward dataset and for the evaluation of baseline models. During the dataset construction phase, Figure 9 displays the prompt used to assess the data quality of the Modality subset to facilitate filtering and cleaning. Meanwhile, Figure 10 and Figure 11 present the specific instructions employed to generate the Colloquialness subset. In the model evaluation phase, we utilize the prompts shown in Figure 12 and Figure 13, respectively, to establish baselines for existing models regarding the Modality and Colloquialness metrics.

B.5 Safety and Privacy Considerations in Data Release

Although parts of our corpus originate from publicly accessible web audio, we do not redistribute raw recordings. To reduce privacy and biometric risks, our release excludes any speaker-identifiable representations and does not provide persistent speaker-level identifiers. If transcripts are released, we remove explicit personal identifiers when detected by automatic pattern matching such as emails, phone numbers, addresses and recommend downstream users to avoid any attempt at individual-level profiling. The released artifacts are intended strictly for non-commercial research use, and derivatives of web-accessed data should not be used outside research contexts.

C Ablation Experiment

This section provides a more detailed analysis of the ablation studies. As shown in Table 6, Mean Pooling emerges as the optimal pooling strategy under both SDiaReward 3B and 7B settings, significantly outperforming the Attention Pooling and Last Hidden State strategies.

Regarding the choice of loss function, incorporating Center Loss outperforms configurations without it in the vast majority of cases. Although a slight performance decline is observed in SDiaReward 3B when combined with Mean Pooling, we ultimately adopt the scheme including Center Loss because, as illustrated in Figure 8(a), it effectively mitigates the issue of score drifting in reward modeling.

Furthermore, Figure 8(b) and Figure 8(c) reveal that SDiaReward 3B exhibits a domain-specific bias similar to that of the 7B model. This phenomenon further corroborates that the reward model implicitly learns a relative ranking function calibrated to specific domain difficulties or styles, rather than serving as a globally applicable absolute metric.

Table 6: **More results of ablation experiments**

Model	**Modality Acc**			**Modality**	**Modality**	**Colloq.**	**Overall**	**Overall**
	Wild	Semi-wild	Scripted	**Micro**	**Macro**	**Acc**	**Micro**	**Macro**
SDiaReward 3B								
Last w/o Center Loss	66.99	58.06	39.70	57.25	54.92	51.20	56.37	53.06
Last w/ Center Loss	71.60	55.91	53.00	63.75	60.17	48.80	61.59	54.49
Atten. w/o Center Loss	98.91	55.91	81.33	87.94	78.72	93.60	88.76	86.16
Atten. w/ Center Loss	96.72	77.42	97.64	94.58	90.59	65.60	90.38	78.10
Mean w/o Center Loss	100.00	78.49	80.90	91.26	86.47	97.60	92.18	92.03
Mean w/ Center Loss	99.39	55.38	82.83	88.62	79.20	92.00	89.11	85.60
SDiaReward 7B								
Last w/o Center Loss	41.02	37.63	46.14	42.21	41.60	45.20	42.64	43.40
Last w/ Center Loss	52.18	56.45	49.36	51.83	52.66	40.00	50.12	46.33
Atten. w/o Center Loss	94.78	48.39	37.55	70.87	60.24	37.20	65.99	48.72
Atten. w/ Center Loss	74.27	59.68	68.45	70.60	67.47	55.20	68.37	61.33
Mean w/o Center Loss	99.88	90.86	88.20	95.05	92.98	97.20	95.37	95.09
Mean w/ Center Loss	100.00	92.47	92.27	96.61	94.92	97.20	96.70	96.06

System Prompt

```
You are an expert in linguistics, phonetics, and conversational analysis.

You will evaluate two versions of the same multi-turn dialogue. Each version has several context turns and a final turn.
      Your task is to evaluate both versions based on their **colloquialness**.

Colloquialness refers to the degree to which an utterance resembles spontaneous spoken language rather than formal
      written or scripted text. It is a multidimensional property characterized by the following criteria:

1. **Lexical informality** - Use of everyday, casual words instead of formal written expressions.
2. **Syntactic simplicity & fragmentation** - Shorter clauses, incomplete sentences, self-repairs, and other structures
      typical of spoken production.
3. **Presence of disfluencies and discourse markers** - Fillers ("um", "uh", "em"), hesitation markers, repetition,
      backchannels, and interactional cues.
4. **Pragmatic naturalness** - Contextually appropriate, human-like, interpersonal, and interaction-driven expressions.
5. **Textually implied prosody** - Usage of punctuation, particles, and exclamations that reflect spoken rhythm, emphasis
      , and emotional nuance.

**Your task is to evaluate both versions of the dialogue according to the above criteria**. For each dimension, you must
      assign an integer score from 1 to 5 for both versions (A and B) and provide an overall preference.

### Scoring:
- **5**: Very high colloquialness. The dialogue sounds highly natural, as if spontaneously spoken, with all the features
      of spoken language.
- **4**: High colloquialness. The dialogue is mostly casual and spontaneous, with some elements of spoken language,
      though minor parts may still resemble written or edited speech.
- **3**: Medium colloquialness. The dialogue includes some colloquial features, but there are noticeable elements that
      sound more scripted or formal.
- **2**: Low colloquialness. The dialogue is mostly formal, structured, or edited, with few to no features of natural
      speech.
- **1**: Very low colloquialness. The dialogue is highly formal, written, or scripted, lacking any signs of spontaneous,
      natural speech.

### Evaluation:
For each version, you will give a score for each of the following dimensions:
- **Lexical informality**
- **Syntactic simplicity & fragmentation**
- **Presence of disfluencies and discourse markers**
- **Pragmatic naturalness**
- **Textually implied prosody**

After assigning scores for all the dimensions, you must also provide an overall preference between A and B.

Please output your result in the following JSON format (and DO NOT include any extra text outside the JSON):

{
  "scores": {
    "lexical_informality": { "A": <1-5>, "B": <1-5> },
    "syntactic_simplicity": { "A": <1-5>, "B": <1-5> },
    "disfluencies": { "A": <1-5>, "B": <1-5> },
    "pragmatic_naturalness": { "A": <1-5>, "B": <1-5> },
    "prosody": { "A": <1-5>, "B": <1-5> }
  },
  "overall_preference": "A" or "B",
  "reason": "Use 2-4 sentences to briefly explain which version you prefer and why, in no more than 80 words."
}
```

Figure 9: Prompt for Modality-aware Data Filtering.

System Prompt

```
You are a data generator for an English dialogue dataset.

Your job:
- Create natural dialogues between "user" and "assistant".
- The conversation must contain between 4 and 16 rounds.
- Each round consists of exactly one "user" message followed by one "assistant" message.
- All content must be in ENGLISH.
- The dialogue must be realistic and coherent.
- Do NOT add any explanations, comments or code fences.
- Output MUST be a JSON array only.
- Do NOT use any emojis or emoticons.
- Use rich and varied sentence structures and vocabulary; avoid repetitive wording and patterns.
```

User Prompt

```
You are generating variation {sample_idx + 1} of {total_samples} for the topic: "{topic}".

Requirements:
- Style: {style_desc}
- This variation MUST be clearly different from other possible dialogues on the same topic:
  - Use a different concrete situation (time, place, details).
  - Use different wording, phrases, and conversation structure.
  - Change the characters' personalities or attitudes if reasonable.
- The conversation must contain between 4 and 16 rounds.
- Each round consists of exactly one "user" message followed by one "assistant" message.
- Roles are only "user" and "assistant".
- The first message MUST be from the "user".
- The last message MUST be from the "assistant".
- Roles must strictly alternate between "user" and "assistant".
- Output format:
  A JSON array, e.g.
  [
    {{"role": "user", "content": "..." }},
    {{"role": "assistant", "content": "..." }}
  ]

Important:
- Do NOT wrap the JSON in any markdown code block.
- Do NOT include any text before or after the JSON.
- Output ONLY the JSON array.
```

Figure 10: Prompt for Written-style Data.

System Prompt

```
You are a dialogue style converter.
Goal:
Convert "written" dialogues into more casual, highly colloquial "spoken" English while keeping their original meaning and
       information intact.
Input:
* A JSON object with fields:

  * "domain" (string)
  * "topic" (string)
  * "style" (string, currently "written")
  * "sample_id" (number)
  * "dialogue": a list of turns; each turn has:
    * "role": "user" or "assistant"
    * "content": string
Output:
* A JSON object with EXACTLY the same schema and fields:

  * Keep "domain", "topic", and "sample_id" unchanged.
  * Set "style" to "spoken".
  * "dialogue" must keep:

    * the same number of turns
    * the same order of turns
    * the same "role" values
      but rewrite each "content" into highly colloquial, natural spoken English.
Requirements:
* Preserve the original meaning and information of every utterance.
* Make the style very conversational, reflecting spontaneous spoken language:
  * Use informal, everyday words instead of formal or complex written expressions.
  * Employ short sentences, fragments, and incomplete thoughts that are typical of natural speech.
  * Incorporate fillers like "um", "uh", "you know", "I mean", and "like" naturally and sparsely.
  * Add short interjections like "yeah", "oh", "wow", and "hey" to make the conversation flow more naturally.
  * Include hesitation markers, repetition, self-repairs, and backchannels to make the dialogue feel more real, as if it'
       s happening in real-time.
  * Use punctuation that reflects the natural rhythm of speech, such as commas, ellipses, or exclamation marks to
       represent pauses, emphasis, or excitement.
  * Make the dialogue feel interpersonal, casual, and contextually appropriate, just like an interaction between two
       people.
* Do NOT use emojis or any kind of internet slang.
* Do NOT change the language (keep it in English).
* Do NOT add or remove dialogue turns.
* Do NOT swap the roles of "user" and "assistant".
* Do NOT add explanations, comments, or extra keys.
Output format:
* Return ONE single JSON object only.
* NO markdown, NO code fences, NO extra commentary.
```

Figure 11: Prompt for Spoken-style Data.

System Prompt

```
You are an expert in linguistics, phonetics, and conversational analysis.

You will listen to two versions of the same multi-turn dialogue, each containing several context turns and a final turn.
We are especially interested in the quality of the FINAL TURN of each version, evaluated in the context of the whole dialogue.

Your task is: after fully understanding the preceding context, evaluate the FINAL TURN of version A and version B from the following three dimensions. For each dimension, you must assign an integer score from 1 to 5 for both A and B, and then give an overall preference between A and B.

[Dimension 1: Final-turn content quality and task adequacy (final_turn_content)]
Focus: whether the final turn, given the previous dialogue context, is appropriate, relevant, and useful.

- 5: Excellent. Accurately understands the prior context and user intent, directly addresses the core issue, provides specific and helpful information, and pragmatically fits the situation, naturally advancing or closing the dialogue.

- 4: Good. Mostly understands the context, stays on-topic, gives reasonably helpful content, and is pragmatically appropriate.
- 3: Fair. Not clearly off-topic, but content is generic, shallow, or template-like, with limited usefulness.
- 2: Poor. Partially off-topic or ignores key context; content is repetitive, empty, or pragmatically somewhat inappropriate.
- 1: Very poor. Severely mismatched with the context, possibly contradictory or strongly inappropriate in pragmatics.

[Dimension 2: Final-turn spoken naturalness and prosodic realism (final_turn_naturalness_prosody)]
Focus: how the final turn is spoken - does it sound like a human naturally speaking in this situation?

- 5: Very high naturalness. Almost indistinguishable from real human speech; natural pitch and rhythm, appropriate pauses and emotions, no obvious synthetic or "reading aloud" impression.
- 4: High naturalness. Generally natural, with some prosodic variation and reasonable pauses; synthetic or "TTS-like" artifacts are minor.
- 3: Medium naturalness. Some spoken flavor and prosody, but elatively regular or "read-aloud" in style; may sound like careful reading rather than casual conversation.
- 2: Low naturalness. Clear TTS/robotic or "reading" impression; flat prosody, stiff pauses, weak emotion.
- 1: Very low naturalness. Strongly mechanical or synthetic, almost no natural prosody or emotion.

[Dimension 3: Dialogue-level coherence and context utilization (dialog_context_coherence)]
Focus: how well the final turn makes use of the prior dialogue context and fits into the multi-turn interaction.

- 5: Excellent. Clearly grounded in specific details from the previous turns (e.g., referencing, following up, correcting, or responding to emotions); the dialogue progression feels natural and coherent.
- 4: Good. Clearly responds to the right speaker and topic; thematically aligned and reasonably coherent with the context.
- 3: Fair. Topic is not completely off, but the connection to context is weak; feels more like single-turn QA than true dialogue.
- 2: Poor. Weak contextual linkage; transitions feel abrupt or template-like; limited use of prior context.
- 1: Very poor. Almost ignores the previous dialogue; off-topic or contradictory; feels like a random line inserted into the dialogue.

You MUST use only the integers 1, 2, 3, 4, or 5 for all scores.

Please output your result in the following JSON format (and DO NOT include any extra text outside the JSON):

{
  "scores": {
    "final_turn_content": { "A": <1-5>, "B": <1-5> },
    "final_turn_naturalness_prosody": { "A": <1-5>, "B": <1-5> },
    "dialog_context_coherence": { "A": <1-5>, "B": <1-5> }
  },
  "overall_preference": "A" or "B",
  "reason": "Use 2-4 sentences to briefly explain which version you prefer and why, in no more than 80 words."
}
```

Figure 12: Prompt for Modality Evaluation.

System Prompt

```
You are an expert in linguistics, phonetics, and conversational analysis.

You will evaluate two versions of the same multi-turn dialogue. Each version has several context turns and a final turn.
      Your task is to evaluate both versions based on their **colloquialness**.

Colloquialness refers to the degree to which an utterance resembles spontaneous spoken language rather than formal
      written or scripted text. It is a multidimensional property characterized by the following criteria:

1. **Lexical informality** - Use of everyday, casual words instead of formal written expressions.
2. **Syntactic simplicity & fragmentation** - Shorter clauses, incomplete sentences, self-repairs, and other structures
      typical of spoken production.
3. **Presence of disfluencies and discourse markers** - Fillers (um, uh, em), hesitation markers, repetition,
      backchannels, and interactional cues.
4. **Pragmatic naturalness** - Contextually appropriate, human-like, interpersonal, and interaction-driven expressions.
5. **Textually implied prosody** - Usage of punctuation, particles, and exclamations that reflect spoken rhythm, emphasis
      , and emotional nuance.

**Your task is to evaluate both versions of the dialogue according to the above criteria**. For each dimension, you must
      assign an integer score from 1 to 5 for both versions (A and B) and provide an overall preference.

### Scoring:
- **5**: Very high colloquialness. The dialogue sounds highly natural, as if spontaneously spoken, with all the features
      of spoken language.
- **4**: High colloquialness. The dialogue is mostly casual and spontaneous, with some elements of spoken language,
      though minor parts may still resemble written or edited speech.
- **3**: Medium colloquialness. The dialogue includes some colloquial features, but there are noticeable elements that
      sound more scripted or formal.
- **2**: Low colloquialness. The dialogue is mostly formal, structured, or edited, with few to no features of natural
      speech.
- **1**: Very low colloquialness. The dialogue is highly formal, written, or scripted, lacking any signs of spontaneous,
      natural speech.

### Evaluation:
For each version, you will give a score for each of the following dimensions:
- **Lexical informality**
- **Syntactic simplicity & fragmentation**
- **Presence of disfluencies and discourse markers**
- **Pragmatic naturalness**
- **Textually implied prosody**

After assigning scores for all the dimensions, you must also provide an overall preference between A and B.

Please output your result in the following JSON format (and DO NOT include any extra text outside the JSON):

{
  "scores": {
    "lexical_informality": { "A": <1-5>, "B": <1-5> },
    "syntactic_simplicity": { "A": <1-5>, "B": <1-5> },
    "disfluencies": { "A": <1-5>, "B": <1-5> },
    "pragmatic_naturalness": { "A": <1-5>, "B": <1-5> },
    "prosody": { "A": <1-5>, "B": <1-5> }
  },
  "overall_preference": "A" or "B",
  "reason": "Use 2-4 sentences to briefly explain which version you prefer and why, in no more than 80 words."
}
```

Figure 13: Prompt for Colloquialness Evaluation.