SageLM: A Multi-aspect and Explainable Large Language Model for Speech Judgement

Yuan Ge[1,*], Junxiang Zhang[1,*], Xiaoqian Liu[1], Bei Li[2], Xiangnan Ma[1], Chenglong Wang[1], Kaiyang Ye[1], Yangfan Du[1], Linfeng Zhang[3,†], Yuxin Huang[4], Tong Xiao[1,5], Zhengtao Yu[4], Jingbo Zhu[1,5,†]

[1]Northeastern University, Shenyang, China
[2]Meituan, Beijing, China
[3]Shanghai Jiao Tong University, Shanghai, China
[4]Kunming University of Science and Technology, Kunming, China
[5]NiuTrans Research, Shenyang, China
geyuanqaq@gmail.com, zhanglinfeng@sjtu.edu.cn, zhujingbo@mail.neu.edu.cn

Abstract

Speech-to-Speech (S2S) Large Language Models (LLMs) are foundational to natural human-computer interaction, enabling end-to-end spoken dialogue systems. However, evaluating these models remains a fundamental challenge. We propose SageLM, an end-to-end, multi-aspect, and explainable speech LLM for comprehensive S2S LLMs evaluation. First, unlike cascaded approaches that disregard acoustic features, SageLM jointly assesses both semantic and acoustic dimensions. Second, it leverages rationale-based supervision to enhance explainability and guide model learning, achieving superior alignment with evaluation outcomes compared to rule-based reinforcement learning methods. Third, we introduce *SpeechFeedback*, a synthetic preference dataset, and employ a two-stage training paradigm to mitigate the scarcity of speech preference data. Trained on both semantic and acoustic dimensions, SageLM achieves an 82.79% agreement rate with human evaluators, outperforming cascaded and SLM-based baselines by at least 7.42% and 26.20%, respectively.

Project page: https://github.com/IronBeliever/SageLM
More details: https://arxiv.org/pdf/2508.20916

Introduction

The advent of large language models (LLMs) has revolutionized human-computer interaction, yet the ultimate goal remains a seamless, natural dialogue that mirrors human conversation (Clark 1996; Luger and Sellen 2016). The next frontier in this pursuit is speech-to-speech (S2S) interaction, where the nuances of communication extend far beyond mere semantic content (Levitan et al. 2015; Porcheron et al. 2018). In human dialogue, how something is said is often as important as what is said (Pierrehumbert and Hirschberg 1990; Hirschberg 2002). For instance, the query, "You know what? I won a million dollar lottery today!" demands more than a semantically correct response; it calls for an expression of incredulous surprise or shared joy. This inextricable coupling of semantics and acoustics makes evaluating S2S dialogue systems a profound and unsolved challenge.

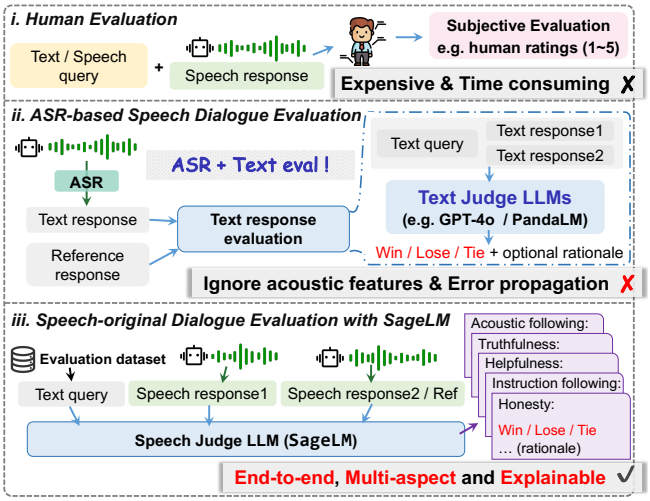


Figure 1: Recent speech-to-speech LLMs evaluation methods rely on human annotations or cascaded pipelines. We propose SageLM, an end-to-end speech dialogue evaluator that provides explainable judgment results across five aspects, including both semantic and acoustic dimensions.

Current evaluation paradigms are fundamentally inadequate for this task. The predominant approach, *a cascaded pipeline* that first transcribes speech with an Automatic Speech Recognition (ASR) model and then uses a text-based LLM for judgment, is critically flawed (Nachmani et al. 2024; Défossez et al. 2024; Zhang et al. 2024; Chen et al. 2024a). Not only is it susceptible to cascading ASR errors, but it is entirely blind to the acoustic dimension, failing to assess the appropriateness of tone, emotion, or prosody. The alternative, *human evaluation*, serves as a gold standard but is prohibitively expensive and slow, rendering it impractical for the rapid, large-scale iteration required to develop advanced S2S models (Veluri et al. 2024; Ding et al. 2025). This leaves a critical gap in the field: the absence of a scal-

*Equal contribution.
†Corresponding author.

able, automated evaluator that can holistically judge both the content and character of spoken dialogue.

To fill this void, we introduce `SageLM`, a multi-**a**spect and expl**a**inable jud**ge** **l**anguage **m**odel designed specifically for the nuanced demands of S2S dialogue evaluation. Building on the "LLM-as-a-judge" paradigm (Zheng et al. 2024; Li et al. 2024; Gu et al. 2024), SageLM operates end-to-end, directly processing speech to render a holistic judgment, thereby bypassing the fragile ASR pipeline. Building such a sophisticated judge, however, presents two foundational obstacles: a data bottleneck and a methodological dilemma.

First, the lack of a large-scale, annotated speech preference dataset has been a primary barrier to progress. To overcome this, we construct *SpeechFeedback*, a comprehensive and diverse preference dataset of 324,774 instances. Each instance contains a speech query, a pair of contrasting responses, and detailed preference annotations covering both semantic relevance and acoustic quality, providing the necessary foundation for training a nuanced speech judge.

Second, even with the right data, the training methodology is paramount. Recent trends have explored rule-based Reinforcement Learning (RL) for tasks where solutions are easily verified (Guo et al. 2025; Swamy et al. 2025; Li et al. 2025a). However, we find this approach fundamentally misaligned with our goal. A simple, rule-based reward signal is insufficient for the complex reasoning required in dialogue evaluation, making the model prone to reward hacking and, crucially, failing to enforce *consistency between its judgments and the explanatory rationales*. We demonstrate that a supervised fine-tuning approach on LLM-annotated rationales is superior enough. This method compels the model not only to predict a judgment but to reason its way to that conclusion, fostering a deeper and more robust understanding that aligns the "what" with the "why." Experimental results demonstrate the success of this approach. SageLM achieves an 82.79% agreement rate with human judgments, significantly outperforming strong cascaded baselines like Whisper + GPT-4o by 7.42% and other end-to-end speech models by a remarkable 26.20%. Our contributions are:

- We construct *SpeechFeedback*, the first large-scale, multi-aspect speech preference dataset to facilitate research in S2S evaluation.
- We demonstrate that for complex judgment tasks, fine-tuning on explicit rationales is a more effective method than rule-based RL, leading to better performance and reasoning consistency.
- We introduce `SageLM`, an end-to-end, explainable judge for S2S dialogue that significantly surpasses existing cascaded and integrated evaluation models, establishing a new state-of-the-art.
- We validate SageLM's effectiveness in rigorous experiments, showing high agreement with human evaluators and superior performance against formidable baselines.

Related Work

Speech Large Language Models

Speech Large Language Models (SLMs) are typically categorized into speech-to-speech (S2S) and speech-to-text (S2T) LLMs. Since GPT-4o, end-to-end S2S LLMs have gained increased attention (Défossez et al. 2024; Zhang et al. 2024; Chen et al. 2024a; Ding et al. 2025; Goel et al. 2025). Despite rapid advancements in S2S LLMs, accurately evaluating their dialogue ability remains a significant challenge. S2T LLMs combine the language modeling capabilities of text-based LLMs with speech understanding capabilities of audio encoders (Chu et al. 2024; Zhang et al. 2025) or codecs (Zhang et al. 2023, 2024; Zhan et al. 2024; Li et al. 2025b). This integration enables S2T LLMs to perform speech dialogue understaning tasks.

Text Large Language Models Evaluation

The evaluation of text-based LLMs in conversational settings primarily assesses their instruction following capabilities, employing either human or model-based evaluation paradigms. The most straightforward approach involves human annotators assigning point-wise or pair-wise subjective labels. However, human evaluation is costly and time-consuming. To address this, the *LLMs-as-a-judge* paradigm leverages powerful models such as GPT-4 to evaluate candidate responses (Zhou et al. 2023; Rafailov et al. 2023; Dubois et al. 2023; Lee et al. 2023). Despite its efficiency, this approach introduces concerns such as model bias (Zheng et al. 2024; Wang et al. 2024a), privacy risks, and computational cost. Consequently, recent open-source efforts have focused on instruction-tuning pretrained LLMs to enhance their evaluation capabilities (Wang et al. 2024b; Li et al. 2023), aiming to mitigate bias, reduce costs, and preserve user privacy. Other works explore lightweight evaluation methods, utilizing smaller models to assess response quality more efficiently (Ge et al. 2024; Sinha et al. 2020; Phy, Zhao, and Aizawa 2020).

S2S Large Language Models Evaluation

Evaluating S2S LLMs remains an open challenge. The modalities supported by S2S LLMs are determined by their architectural design. Typically, S2S models, such as Kimi-Audio (Ding et al. 2025), support both text and speech as inputs and outputs. Evaluation in this comprehensive setting encompasses fundamental speech capabilities, audio understanding, and speech conversation. The evaluation of fundamental speech capabilities includes ASR and TTS tasks, commonly measured using established metrics like Word Error Rate (WER) and Mean Opinion Score (MOS). Audio understanding tasks focus on the model's ability to comprehend and reason about both semantic and acoustic information within audio, utilizing benchmarks such as MMAU (Sakshi et al. 2024), ClothoAQA (Lipping et al. 2022), and AIR-Bench (Yang et al. 2024). However, speech conversation evaluation emphasizes the model's dialog capabilities, including audio-to-text chat and speech-to-speech chat. This demanding task requires not only an understanding of audio semantics and paralinguistic cues, but also competencies in language modeling, reasoning, and speaking style control. Notably, evaluating speech-to-speech conversation remains an open challenge. As shown in Fig. 1, human evaluation, while considered the gold standard, is costly and

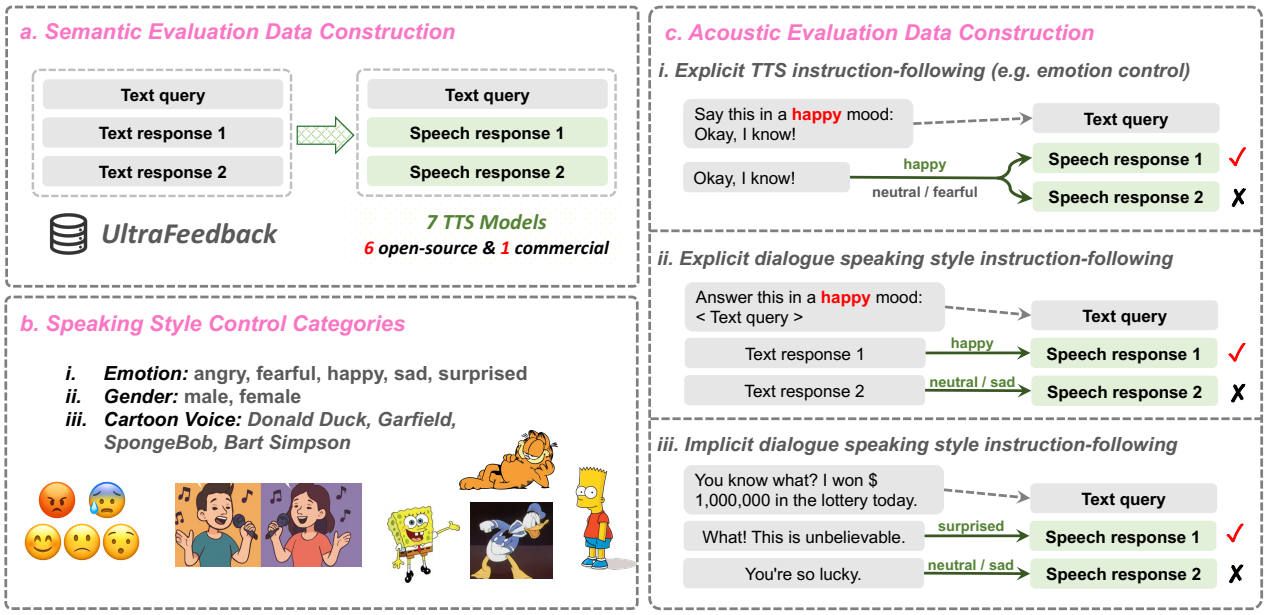


Figure 2: Data construction pipeline of *SpeechFeedback*.

time-consuming, making it impractical to scale comprehensive benchmarking. Furthermore, human evaluation typically overlooks explainability due to the additional cost and time required. ASR-based evaluation, which transcribes the model's speech responses into text followed by conventional text-based LLM conversation evaluation, introduces compounding errors because even state-of-the-art ASR models exhibit non-negligible WERs. Moreover, such cascaded methods disregard paralinguistic features and fail to evaluate the model's ability to adhere to style-related instructions.

Speech Dialogue Judgment

This section defines the speech dialogue judgment task and introduces the *SpeechFeedback* dataset.

Task Definition

The speech dialogue judgment task aims to compare a pair of speech responses based on criteria including truthfulness, honesty, helpfulness, instruction following, and speech instruction following, guided by a textual instruction. Formally, given a judge model J, the pairwise evaluation process can be formulated as:

$$ER = J(Q, R_1, R_2) \quad (1)$$

where Q is the textual query or instruction, R_i represents the i_{th} speech response, and ER is the evaluation result. The evaluation result includes five independent comparison labels C_{a_j} and textual explanations E_{a_j}, where aspect $a_j \in$ {truthfulness, honesty, helpfulness, instruction following, speech instruction following} and $C_{a_j} \in$ {win, lose, tie} represent R_1 is better, R_2 is better, or both responses are of comparable quality respectively.

In our task definition, we utilize textual queries as input instead of spoken audio. This design choice is paramount for ensuring a fair and standardized evaluation across diverse S2S LLMs. Our rationale is twofold. Firstly, existing speech dialogue datasets typically incorporate both textual and spoken queries, thereby facilitating the direct use of text queries without incurring additional overhead. Second, S2S LLMs are trained on synthetic speech generated by heterogeneous TTS models (e.g., SLAM-Omni uses CosyVoice1, while Kimi-Audio uses Kimi-TTS). To mitigate potential distributional discrepancies between training and test audio, we adopt text-based instructions to ensure a fair evaluation across different S2S models.

Data Construction and Pre-processing

SpeechFeedback dataset comprises two components: semantic evaluation and acoustic instruction following evaluation.

Semantic Evaluation Data Construction The semantic dimension speech preference data are synthesized from UltraFeedback (Cui et al. 2023) using 7 TTS models. UltraFeedback is a large-scale, fine-grained, and diverse preference dataset comprising 64k instructions, each with four LLM-generated responses rated by GPT-4 across four dimensions: instruction-following, truthfulness, honesty, and helpfulness, accompanied by scores and rationales. We utilize these textual preference annotations and synthesize corresponding speech responses.

To prepare the corpus for voice-centric interaction and evaluation, we implement a multi-stage filtering pipeline. First, mathematical expressions, code segments, and multilingual prompts are removed, and special characters are sanitized to preserve natural prosody in downstream synthesis. As shown in Fig. 2a, the filtered responses are synthesized by six open-source TTS models and one commercial TTS model—CosyVoice (Du et al. 2024a), CosyVoice2 (Du

et al. 2024b), SparkTTS (Wang et al. 2025b), ChatTTS[1], F5-TTS (Chen et al. 2024b), Index-TTS (Deng et al. 2025), and gpt-4o-mini-tts[2]. Utterances exhibiting a high Whisper ASR word-error rate or a duration shorter than 0.2 s are discarded. Absolute quality scores are subsequently converted to pairwise win–loss labels, and their rationales are rewritten in comparative form via Qwen2.5-32B-Instruct. The combined text–speech sequence is capped at 4,096 tokens. Further preprocessing details appear in the supplementary material.

Acoustic Evaluation Data Construction Dialogue allows users to explicitly or implicitly control the vocal characteristics of an SLM's response. For instance, a user might explicitly request a faster speaking rate, a response in a different language, or a specific emotional tone. In implicit scenarios, users expect SLMs to exhibit empathy, highlighting the importance of acoustic evaluation in spoken dialogue.

This paper focuses on three primary categories of speaking style control: emotion, gender, and voice, as illustrated in Fig. 2b. Furthermore, as shown in Fig. 2c, we structure acoustic evaluation around three task formats: explicit TTS, explicit dialogue, and implicit dialogue. Each data instance comprises an instruction, two speech responses, and an acoustic evaluation label with its accompanying rationale.

i. Explicit TTS Task

- *Instruction:* The Explicit TTS task utilizes the Ultrafeedback dataset, from which we sample 1,000 instructions for each speaking style category. For each instruction, the text response to be synthesized is randomly selected from four candidate responses. This text is then combined with a prompt template, which is randomly chosen from a set of 20 templates per category generated by GPT-4o.
- *Speech Responses:* We first define *incorrect label set* for each control category. For emotion control, labels are categorized as positive (happy, surprised) or negative (sad, fearful, angry). The incorrect label set for a target emotion comprises emotions from the opposing category and a neutral emotion. For gender control, the incorrect label set consists solely of the other gender label. For cartoon voice control, the incorrect label set includes a neutral voice and all cartoon voices except the target voice. Each sample includes two synthesized speech responses: one with the correct label and one with an incorrect label randomly sampled from the *incorrect label set*, with sampling probabilities of 8:1:1 for correct-correct, correct-incorrect, and incorrect-incorrect pairs, respectively.
- *Acoustic Evaluation Label and Rationale:* The evaluation label is assigned as win, lose, or tie based on the ground truth. The instruction, the two text responses, and the ground truth label are then provided as input to Qwen2.5-32B to generate a corresponding rationale.

ii. Explicit Dialogue Task

We also sample 1,000 instructions for each speaking style category from the Ultrafeedback dataset. The instruction is formed by combining a user query with a randomly sampled speaking style template. Then, two text responses are randomly selected. The remaining procedure for selecting control labels, synthesizing audio, and constructing the evaluation output is identical to *Explicit TTS task*. Additionally, 180 *mixed* samples are created to jointly control emotion and gender, e.g., Respond to `<query>` in a happy female voice.

iii. Implicit Dialogue Task

For the implicit dialogue task, both the instructions and the text responses are generated by `GPT-4o`. We employ self-instruct (Wang et al. 2023) to curate and translate implicit emotion data from Kimi-GenTest (Ding et al. 2025), and use them as seed prompts to generate 500 evaluation samples. The subsequent steps—control label selection, speech synthesis, and label construction—follow the same pipeline as *Explicit TTS task*. However, the rationales are uniquely structured: we prompt `GPT-4o` to explain the implied emotional intent of the query and provide a template-based description of the emotion of each candidate response.

In summary, the overall acoustic evaluation data construction framework incorporates both semantic and acoustic preference annotation.

Preliminary

Training Objective

We explore two prominent fine-tuning approaches: Instruction Tuning (IT) and Reinforcement Learning (RL). IT is a primary method for model fine-tuning, widely employed to learn specific output formats and align with human preferences. On-policy RL methods have recently been shown to be particularly effective for preference learning from a limited number of samples. Moreover, RL tends to perform well where verification is simple, yet generation is complex (Li et al. 2025a; Swamy et al. 2025). Evaluation task, which requires assigning a single discrete label (win, lose, or tie), presents a significant generation-verification gap. This makes it an ideal setting to investigate the effectiveness of RL in the speech modality. The specific training objectives for each method are detailed below.

Instruction Tuning: During the supervised fine-tuning (SFT) stage, the model is trained to minimize the discrepancy between generated responses and reference responses. Specifically, the training objective is to maximize the log-likelihood of the reference output sequence:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\log P_\theta(y_t \mid y_{<t}, x)\right] \quad (2)$$

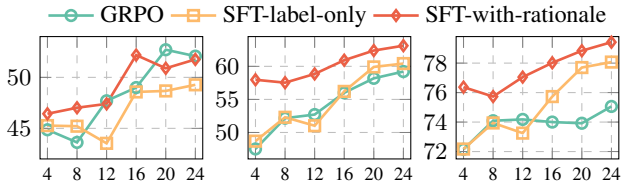
We explore two SFT settings based on the composition of the reference response y. In the SFT (label-only) setting y consists solely of the correct judgment label, while in the SFT (with-rationale) setting y comprises both the label and a corresponding rationale previously generated by GPT-4.

Reinforcement Learning: We leverage the Group Relative Policy Optimization (GRPO) algorithm (Shao et al. 2024), which maximizes a reward-weighted objective using a clipped policy ratio and a KL-divergence penalty to a reference policy. We define a rule-based reward function as a supervision signal:

$$r_i = \alpha \cdot R_a(\hat{s}, s) + \gamma \cdot R_f \quad (3)$$

[1]https://github.com/2noise/ChatTTS

[2]https://platform.openai.com/docs/models/gpt-4o-mini-tts



(a) Score Accuracy (b) Compare Accuracy (c) Agreement

Figure 3: Preliminary: Reinforcement Learning versus Supervised Fine-Tuning on three evaluation metrics while training data scaling up (4k∼24k × 4 aspects).

where R_a is the accuracy reward that measures the discrepancy between the predicted score $\hat{s}$ and the ground truth score s ($1 \leq \hat{s}, s \leq 5$), as defined in Eq. (4). The format reward $R_f \in \{0, 1\}$ evaluates whether the model correctly formats its output using the `<think></think>` and `<answer></answer>` tags. Specifically, the weighted averaging parameters are set to $\alpha = 1.0$ and $\gamma = 0.5$. The advantage estimation and training objective follow the standard GRPO formulation.

$$R_a(\hat{s}, s) = \exp\left(-\frac{(\hat{s} - s)^2}{2\sigma^2}\right) \tag{4}$$

RL or SFT? Empirical Analysis on Judgment Task

DeepSeek R1 demonstrates that the reasoning capabilities of LLMs can be effectively enhanced via pure GRPO (Guo et al. 2025). Recent studies further suggest that RL performs well in tasks where verification is easy but generation is challenging, such as open-domain QA (Swamy et al. 2025) and audio multi-choice QA (Li et al. 2025a). The task of evaluating speech responses exemplifies this generation-verification gap: verifying a judgment (win/lose/tie) is straightforward, whereas generating a high-quality judgment is non-trivial. This motivates our investigation into the comparative performance of GRPO and full-parameter SFT on speech response judgment task. In our preliminary experiments, both RL and SFT were trained for one epoch, with all other settings consistent with main experiments.

As shown in Fig. 3(a), GRPO outperforms SFT (label-only) in scoring accuracy. WaveReward (Ji et al. 2025) also find this conclusion, showing that RL-based post-training surpasses LoRA-finetuned Qwen2.5-Omni. However, scoring accuracy is too strict to evaluation task, as even humans struggle to assign precise scores. In contrast, pair-wised comparison both better reflects human preferences and is easier to annotate. Fig. 3(b) and (c) indicate that GRPO and SFT perform comparably on compare accuracy, with SFT slightly outperforming GRPO on compare agreement.

We hypothesize that this difference stems from their learning objectives. GRPO's format reward encourages a reasoning process, whereas SFT (label-only) optimizes solely for the final label. To bridge this gap, we introduce an SFT-with-rationale setting, augmenting labels with explanations generated by GPT-4. As shown in Figure 3 (SFT-with-rationale), the inclusion of rationales improves the model's understanding of evaluation results. Additionally, SFT with rationale outperform all other methods.

Furthermore, we observe that GRPO's lacks of supervision over the reasoning process leads to *inconsistency between rationale and final result*. In a manual analysis of 100 sampled cases, the explanation contradicted the final score in 39% of instances. This issue severely limits the reliability and explainability of the model as a speech judge.

SageLM: Two-stage Training

To address the scarcity of acoustic preference data, we introduce a two-stage training strategy for `SageLM`. Drawing inspiration from curriculum learning (Bengio et al. 2009; Wang et al. 2025a), this approach progressively builds the model's evaluative capabilities.

Semantic Preference Learning. The first stage utilizes the abundant semantic preference annotations in the *Speech-Feedback* dataset. This stage trains `SageLM` to evaluate responses along four core dimensions—truthfulness, honesty, helpfulness, and instruction following. Concurrently, the model learns to generate structured comparison outputs that include detailed explanatory rationales.

Acoustic Preference Learning. Building upon the semantic foundation, the second stage incorporates a limited amount of acoustic preference data. This introduces a fifth evaluation dimension: speech instruction following. This dimension assesses whether the synthesized speech adheres to acoustic attributes. These attributes, such as emotion, gender, or a specific voice style, may be explicitly requested or implicitly implied by the user's query.

Based on preliminary experiments, both training stages adopt a rationale-augmented SFT method to supervise both the judgment labels and their corresponding explanations.

Experiments

Experimental Setups

Dataset We employ the SpeechFeedback dataset, a synthetic dialogue evaluation corpus detailed previously. The training set consists of 316,544 semantic and 4,270 acoustic feedback instances. The test set contains 728 semantic and 410 acoustic evaluation instances, all manually verified to ensure label order aligns with human preferences.

Training and Generations Details SageLM is trained using 8 NVIDIA A100-SXM4-80GB GPUs. For instruction tuning, we adopt full-parameter supervised fine-tuning with the following hyperparameters: 3 epochs, cutoff length of 4096, batch size of 16, learning rate of 1e-5, cosine learning rate scheduler with a warmup ratio of 0.1, and bf16 precision. Reinforcement learning is implemented using GRPO, which incurs higher training costs. Accordingly, we set: 1 epoch, group size of 8, cutoff length of 8192, batch size of 32 (4 instructions × 8 samples), learning rate of 1e-5, cosine scheduler with a 0.05 warmup ratio, and bf16 precision. Sampling during RL training uses: temperature = 1.0, top-p = 0.99, top-k = 50, and max completion length = 2048. During inference, all models use identical decoding parameters: temperature = 0.95, top-p = 0.7, top-k = 50, and repetition

Model	Accuracy				Agreement			
	Hel.	Hon.	IF.	Tru.	Hel.	Hon.	IF.	Tru.
I. Baseline of cascaded ASR and text language models								
Whisper + GPT-4o	62.09	55.86	65.75	61.54	75.27	74.27	77.02	74.91
Whisper + Qwen2.5-32B	53.48	40.66	59.89	40.29	67.40	65.20	70.70	58.79
Whisper + PandaLM-7B	54.03	34.43	53.85	41.03	67.58	58.61	64.01	59.16
Whisper + Qwen2.5-omni-7B	35.90	25.82	38.28	31.14	48.35	48.72	47.80	49.36
Whisper + Qwen2.5-omni-3B	25.09	14.65	29.86	18.87	34.25	31.87	39.56	31.69
II. Baseline of Speech-to-text large language models direct inference with prompt								
Qwen2-Audio-Base	7.15	4.95	7.51	7.32	9.52	9.43	8.89	10.53
Qwen2-Audio-Instruct	23.81	15.38	24.73	18.13	34.98	33.43	33.15	32.88
Qwen2.5-omni-3B	37.73	30.04	41.02	37.18	53.48	54.67	54.12	55.86
Qwen2.5-omni-7B	41.76	29.49	44.69	38.83	56.50	55.31	56.50	58.06
III. Different finetuning versions of `SageLM`								
Qwen2-Audio-Instruct-7B-SFT	70.33	68.50	65.57	65.02	82.97	81.32	77.56	78.66
Qwen2.5-omni-3B-SFT	70.33	**73.08**	67.40	68.68	82.60	**85.26**	78.57	81.05
Qwen2.5-omni-7B-SFT (SageLM)	**72.35**	67.40	**72.35**	**73.26**	**83.61**	81.78	**81.68**	**84.07**

Table 1: Semantic evaluation of `SageLM` and competing baselines on a human-annotated test set. We report accuracy and agreement (%) across four dimensions: Helpfulness (Hel.), Honesty (Hon.), Instruction Following (IF.), and Truthfulness (Tru.).

penalty = 1.0. We report results averaged across three runs with seeds {42, 123, 1234} for reproducibility.

Evaluation Metrics We report Accuracy and Agreement as primary metrics for both semantic and acoustic dimension evaluation. Accuracy follows Eq. (1), measuring whether the model's predicted comparison label matches the ground truth (win/lose/tie). Following PandaLM (Wang et al. 2024b), Agreement assigns 1 for complete alignment with human judgment, 0 for complete disagreement, and 0.5 otherwise. Results are averaged over three inference runs.

Experimental Results

Cascaded Baselines Following prior work, we first transcribe speech using `whisper-large-v3-turbo` (Radford et al. 2022), then evaluate the transcribed text using several advanced text-based models: gpt-4o, the evaluation-specialized PandaLM-7B, and the Qwen family including Qwen2.5-32B and Qwen-omni-3B/7B. As shown in Table 1 and 2, experimental results demonstrate that *cascade pipelines serve as strong baselines for semantic-level evaluation*. Specifically, Whisper + gpt-4o pipeline achieves an average accuracy of 61.31% and an average agreement rate of 75.37% across four evaluation aspects. However, this approach lacks the capacity to evaluate acoustic dimensions, limiting its applicability. In contrast, Whisper + Qwen2.5-Omni-3B/7B performs poorly due to lower parameter size and limited instruction-following capabilities.

S2T LLM baselines Inspired by evaluation practices in multi-modality LLMs, a common baseline for assessing generative models is to leverage unimodal understanding models, e.g. vision-LLMs for evaluating text-to-image generation. In the speech domain, S2T LLMs possess comprehensive speech understanding capabilities, making them natural candidates for evaluating spoken dialogue. Thus, we examine the evaluation capability of S2T LLMs, including

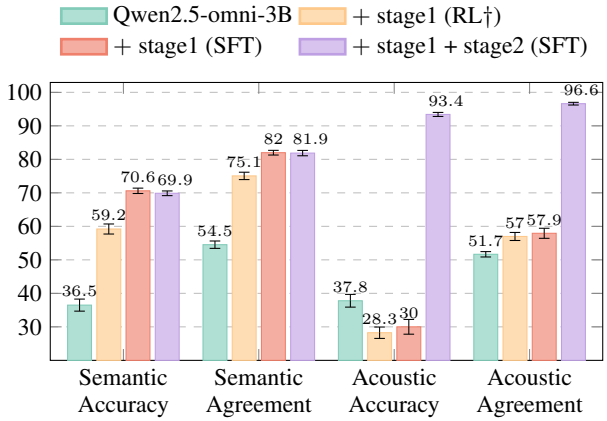


Figure 4: Analysis of the impact of stage1 semantic evaluation training and stage2 acoustic evaluation training.

Qwen-Audio-7B-base&instruct and Qwen2.5-omni-3B/7B, using prompt engineering to elicit dialogue-level judgments. Qwen2-Audio-Base, lacking instruction tuning, fails to follow task instructions in 82% of cases, often repeating prompts or outputting [1|2|Tie]. Although Qwen2.5-Omni outperforms Qwen2-Audio-Instruct-7B in evaluation tasks, it still falls short compared to the Whisper + gpt-4o cascade pipeline. In summary, current S2T models *exhibit limited capability in both semantic and acoustic judgment*, with an average accuracy of only around 40% in acoustic evaluation.

Different Finetuning Versions The performance of the base model significantly influences the performance after post-trained. We compare the outcomes of two-stage training based on three models: Qwen2-Audio-Instruct-7B, Qwen2.5-omni-3B, and Qwen2.5-omni-7B. Experimental results demonstrate that the two-stage post-training en-

Model	Accuracy					Agreement				
	Emo.	Gen.	Voi.	Imp.	Mixed	Emo.	Gen.	Voi.	Imp.	Mixed
I. Baseline of Speech-to-text large language models direct inference with prompt										
Qwen2-Audio-Base	8.05	9.28	7.07	8.11	6.59	9.96	11.86	9.26	12.16	11.32
Qwen2-Audio-Instruct	40.61	41.93	30.97	40.54	29.22	48.28	49.14	43.26	55.41	52.47
Qwen2.5-omni-3B	44.45	41.24	40.74	37.84	24.69	53.64	49.14	53.87	52.70	48.97
Qwen2.5-omni-7B	39.46	40.89	43.43	35.14	28.40	51.53	49.31	56.57	49.55	51.64
***II. Different finetuning versions of* `SageLM`**										
Qwen2-Audio-Instruct-7B-SFT	82.76	94.50	89.23	80.18	72.01	91.38	97.25	94.61	90.09	85.39
Qwen2.5-omni-3B-SFT	91.18	96.22	94.95	94.59	**90.12**	95.59	98.11	97.47	97.47	**94.44**
Qwen2.5-omni-7B-SFT (SageLM)	**95.78**	**98.63**	**99.33**	**97.30**	82.72	**97.89**	**99.31**	**99.66**	**98.65**	91.36

Table 2: Acoustic evaluation results against baselines on a human-annotated test set. We report accuracy and agreement (%) across five dimensions: emotion (Emo.), gender (Gen.), voice (Voi.), implicit emotion (Imp.), and Mixed aspects (Mixed.) which jointly evaluate emotion and gender. Results for ASR baselines are not applicable as they ignore acoustic features.

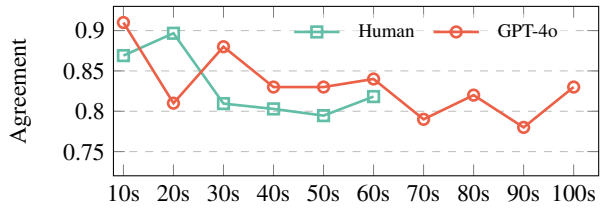


Figure 5: Agreement vs. Combined responses pairs length.

Model	Agreement	Accuracy
Whisper + GPT-4o	69.30%	53.80%
SageLM	87.97%	81.01%

Table 3: Model evaluation agreement with human annotation on unseen AlpacaEval text-dataset.

hances the speech dialogue evaluation ability across all models, with Qwen2.5-omni-7B achieving the best overall performance. Ultimately, *SageLM outperform both cascaded and SLM-based baselines* by at least 7.42% and 26.20% respectively, reaching 82.79% agreement with human judgments. Additional human evaluation confirms that the rationales generated by `SageLM` offer improved explainability and align with final results at 90.89% by human evaluation.

Analysis

Two-stage Training Two-stage post-training enhances the evaluation capability of S2T LLMs on speech responses. Fig. 4 shows that stage1, whether using GRPO or SFT, improves semantic evaluation but degrades acoustic evaluation. In contrast, stage2 incorporates both semantic and acoustic aspect preference data, thereby preserving semantic performance while significantly enhancing acoustic evaluation.

Model Performance of Difference Response Length As shown in Figure 5, our analysis of the test set indicates that SageLM's agreement with human annotations slightly decreases as the combined length of the evaluated response pair increases. To corroborate this finding, we conducted a supplementary analysis on a larger dataset, sampling 100 examples from distinct intervals of combined response length. Experiment results show a similar slight decline in agreement between model's evaluations and GPT-4o annotations, though *the agreement rate remained high at about 80%*.

Out-of-Distribution Generalization on Real S2S LLMs Outputs To assess the generalization of `SageLM` to *unseen dataset* and *real S2S LLMs output distribution*, we evaluate it on AlpacaEval followed by VoiceBench (Chen et al. 2024c), comparing Kimi-Audio with Qwen2.5-omni. As shown in Table 3, `SageLM` surpasses the Whisper + GPT-4o pipeline by 18.67% in agreement and 27.21% in accuracy, validating its capability to evaluate within the true generative distribution of S2S LLMs.

We further analyzed the gap between `SageLM` and cascaded baseline. Whisper introduced substantial transcription errors for Kimi-Audio responses due to speaking rate and prosody, which consequently biased GPT-4o toward favoring Qwen2.5-Omni. In contrast, `SageLM` and human were able to capture the semantic content of the responses and make correct judgments. This finding highlights the issue of error propagation in cascaded systems and underscores the necessity of an end-to-end model for speech judgement.

Conclusions

We investigate evaluation methods for spoken dialogue models and propose `SageLM`, motivated by two main challenges: (1) human-annotated evaluation is costly and difficult to scale, and (2) cascaded ASR-to-text-LLM pipelines suffer from error propagation and loss of acoustic information. To address these issues, we construct SpeechFeedback, a dataset comprising both semantic and acoustic evaluations of speech responses. Preliminary experiments demonstrate that supervised fine-tuning of speech-to-text LLMs using interpretable feedback outperforms both label-only SFT and GRPO, even under a significant generation-verification gap in the evaluation task. We adopt a two-stage SFT strategy to separately enhance semantic and acoustic evaluative capabilities. Results show that `SageLM` achieves 82.79% agreement with human judgments, surpassing cascaded and SLM-based baselines by at least 7.42% and 26.20%, respectively.

Acknowledgments

This work was supported in part by the National Science Foundation of China (Nos. 62276056 and U24A20334), the Yunnan Fundamental Research Projects (No.202401BC070021), the Yunnan Science and Technology Major Project (No. 202502AD080014), the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009), and the Shanghai Science and Technology Program (Grant No. 25ZR1402278).

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Chen, W.; Ma, Z.; Yan, R.; Liang, Y.; Li, X.; Xu, R.; Niu, Z.; Zhu, Y.; Yang, Y.; Liu, Z.; et al. 2024a. SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training. *arXiv preprint arXiv:2412.15649*.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024b. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. *arXiv preprint arXiv:2410.06885*.

Chen, Y.; Yue, X.; Zhang, C.; Gao, X.; Tan, R. T.; and Li, H. 2024c. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Clark, H. H. 1996. *Using language*. Cambridge university press.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377.

Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; and Zeghidour, N. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Deng, W.; Zhou, S.; Shu, J.; Wang, J.; and Wang, L. 2025. IndexTTS: An Industrial-Level Controllable and Efficient Zero-Shot Text-To-Speech System. *arXiv preprint arXiv:2502.05512*.

Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Dubois, Y.; Li, C. X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P. S.; and Hashimoto, T. B. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36: 30039–30069.

Ge, Y.; Liu, Y.; Hu, C.; Meng, W.; Tao, S.; Zhao, X.; Xia, M.; Li, Z.; Chen, B.; Yang, H.; Li, B.; Xiao, T.; and Zhu, J. 2024. Clustering and Ranking: Diversity-preserved Instruction Selection through Expert-aligned Quality Estimation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 464–478. Miami, Florida, USA: Association for Computational Linguistics.

Goel, A.; Ghosh, S.; Kim, J.; Kumar, S.; Kong, Z.; Lee, S.-g.; Yang, C.-H. H.; Duraiswami, R.; Manocha, D.; Valle, R.; and Catanzaro, B. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *arXiv preprint arXiv:2507.08128*.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hirschberg, J. 2002. The pragmatics of intonational meaning. In *Proceedings of Speech Prosody*, volume 2, 11–13.

Ji, S.; Liang, T.; Li, Y.; Zuo, J.; Fang, M.; He, J.; Chen, Y.; Liu, Z.; Jiang, Z.; Cheng, X.; et al. 2025. WavReward: Spoken Dialogue Models With Generalist Reward Evaluators. *arXiv preprint arXiv:2505.09558*.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K. R.; Mesnard, T.; Ferret, J.; Bishop, C.; Hall, E.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Levitan, R.; Benus, S.; Gravano, A.; and Hirschberg, J. 2015. Entrainment and Turn-Taking in Human-Human Dialogue. In *AAAI Spring Symposia*, 44–51.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Li, G.; Liu, J.; Dinkel, H.; Niu, Y.; Zhang, J.; and Luan, J. 2025a. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Li, J.; Sun, S.; Yuan, W.; Fan, R.-Z.; Zhao, H.; and Liu, P. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Li, T.; Liu, J.; Zhang, T.; Fang, Y.; Pan, D.; Wang, M.; Liang, Z.; Li, Z.; Lin, M.; Dong, G.; et al. 2025b. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.

Lipping, S.; Sudarsanam, P.; Drossos, K.; and Virtanen, T. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, 1140–1144. IEEE.

Luger, E.; and Sellen, A. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of

Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286–5297.

Nachmani, E.; Levkovitch, A.; Hirsch, R.; Salazar, J.; Asawaroengchai, C.; Mariooryad, S.; Rivlin, E.; Skerry-Ryan, R.; and Ramanovich, M. T. 2024. SPOKEN QUESTION ANSWERING AND SPEECH CONTINUATION USING SPECTROGRAM-POWERED LLM. In *12th International Conference on Learning Representations, ICLR 2024*.

Phy, V.; Zhao, Y.; and Aizawa, A. 2020. Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 4164–4178. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Pierrehumbert, J.; and Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. *Intentions in communication*, 271: 311.

Porcheron, M.; Fischer, J. E.; Reeves, S.; and Sharples, S. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sinha, K.; Parthasarathi, P.; Wang, J.; Lowe, R.; Hamilton, W. L.; and Pineau, J. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2430–2441. Online: Association for Computational Linguistics.

Swamy, G.; Choudhury, S.; Sun, W.; Wu, Z. S.; and Bagnell, J. A. 2025. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*.

Veluri, B.; Peloquin, B. N.; Yu, B.; Gong, H.; and Gollakota, S. 2024. Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21390–21402. Miami, Florida, USA: Association for Computational Linguistics.

Wang, C.; Gan, Y.; Huo, Y.; Mu, Y.; Yang, M.; He, Q.; Xiao, T.; Zhang, C.; Liu, T.; and Zhu, J. 2025a. Rovrm: A robust visual reward model optimized via auxiliary textual preference data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25336–25344.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; and Sui, Z. 2024a. Large Language Models are not Fair Evaluators. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. Bangkok, Thailand: Association for Computational Linguistics.

Wang, X.; Jiang, M.; Ma, Z.; Zhang, Z.; Liu, S.; Li, L.; Liang, Z.; Zheng, Q.; Wang, R.; Feng, X.; et al. 2025b. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13484–13508. Toronto, Canada: Association for Computational Linguistics.

Wang, Y.; Yu, Z.; Zeng, Z.; Yang, L.; Wang, C.; Chen, H.; Jiang, C.; Xie, R.; Wang, J.; Xie, X.; Ye, W.; Zhang, S.; and Zhang, Y. 2024b. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization.

Yang, Q.; Xu, J.; Liu, W.; Chu, Y.; Jiang, Z.; Zhou, X.; Leng, Y.; Lv, Y.; Zhao, Z.; Zhou, C.; et al. 2024. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. *arXiv preprint arXiv:2402.07729*.

Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Zhang, Q.; Cheng, L.; Deng, C.; Chen, Q.; Wang, W.; Zheng, S.; Liu, J.; Yu, H.; Tan, C.; Du, Z.; et al. 2024. Omniflatten: An end-to-end gpt model for seamless voice conversation. *arXiv preprint arXiv:2410.17799*.

Zhang, Y.; Liu, Z.; Bu, F.; Zhang, R.; Wang, B.; and Li, H. 2025. Soundwave: Less is More for Speech-Text Alignment in LLMs. *arXiv preprint arXiv:2502.12900*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36: 55006–55021.