S2S-Arena, Evaluating Speech2Speech Protocols on Instruction Following with Paralinguistic Information

Anonymous ACL submission

Abstract

The rapid development of large language models (LLMs) has brought significant attention to speech models, particularly recent progress in speech2speech protocols supporting speech input and output. However, The existing benchmarks adopt automatic text-based evaluators for evaluating the instruction following ability of these models lack consideration for paralinguistic information in both speech understanding and generation. To address these issues, we introduce S2S-Arena, a novel arena-style S2S benchmark that evaluates instruction-following capabilities with paralinguistic information in both speech-in and speech-out across real-world tasks. We design 154 samples that fused TTS and live recordings in four domains with 21 tasks and manually evaluate existing popular speech models in an arena-style manner. The experimental results show that: (1) in addition to the superior performance of GPT-4o, the speech model of cascaded ASR, LLM, and TTS outperforms the jointly trained model after text-speech alignment in speech2speech protocols; (2) considering paralinguistic information, the knowledgeability of the speech model mainly depends on the LLM backbone, and the multilingual support of that is limited by the speech module; (3) excellent speech models can already understand the paralinguistic information in speech input, but generating appropriate audio with paralinguistic information is still a challenge.

1 Introduction

Voice-based human-computer interaction is one of the most natural forms of communication (Card et al., 1983; Allen et al., 2001). The machine is expected to possess instruction following ability in the speech-to-speech (S2S) protocol—not only understand voice commands issued by users (Chu et al., 2023, 2024; Tang et al., 2024; Ghosh et al., 2024; Hu et al., 2024) but also generate appropriate responses in voice and executing the corresponding tasks (Wang et al., 2024c; Chen et al., 2024c; Liao et al., 2024).

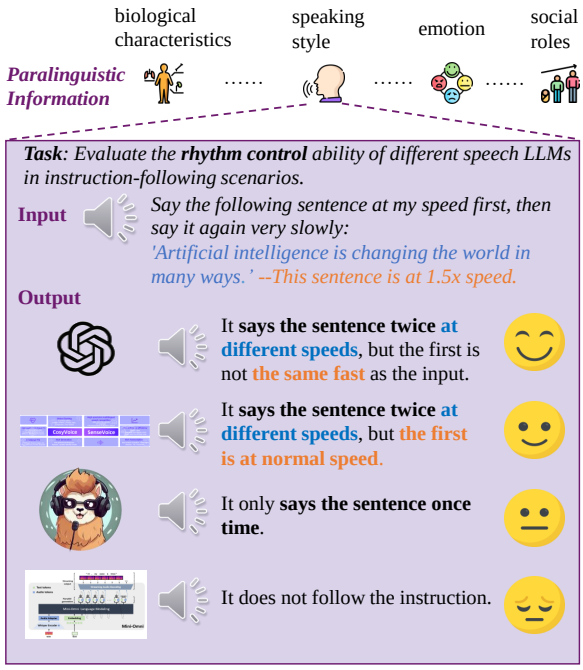

Figure 1: An Example of Evaluating Instruction Following with Rhythm Controlling in Speech-in and Speech-out for Speech Models.

Recent work (Zhang et al., 2023; SpeechTeam, 2024; Fang et al., 2024; Xie and Wu, 2024) has made significant progress by leveraging the powerful semantic understanding capabilities of large-scale language models (LLMs) (Dubey et al., 2024) and shifted to paralinguistic information above the basic semantic information, as illustrated in Figure 1. As a crucial aspect of the more vivid and natural conversation in S2S scenario (Trager, 1958), paralinguistic information encompasses biological characteristics (Schuller et al., 2010), emotion (Batliner et al., 2011), speaking style (Nose et al., 2007), and social roles (Ipgrave, 2009), which can be inferred from pitch, tone, speech rate, and voice quality (Schuller et al., 2013).

Types	Benchmarks for Speech Models	Understanding		Generation		Evaluation	
		Sem.	Par.	Sem.	Par.	Modality	Evaluator
Foundation	Dynamic-SUPERB (Huang et al., 2024)	✓	✓	✓		- *	Auto
	SGAI (Bu et al., 2024)	✓	✓			Text	Auto
	AudioBench (Wang et al., 2024a)	✓	✓			Text	Auto
	MMAU (Sakshi et al., 2024)	✓	✓			Text	Auto
	AV-Odyssey Bench (Gong et al., 2024)	✓	✓			Text	Auto
Chat	SD-Eval (Ao et al., 2024)	✓	✓			Text	Auto
	VoiceBench (Chen et al., 2024b)	✓		✓		Text	Auto
Chat and Foundation	AIR-Bench (Yang et al., 2024)	✓	✓	✓		Text	Auto
	S2S-Arena (Ours)	✓	✓	✓	✓	Speech	Human

Table 1: Comparison of Benchmarks for Speech Models. The star* means that the evaluation modality of the Dynamic-Superb is decided by the tested task. Sem. means the semantics of speech, and Par. means the paralinguistic information of speech.

However, existing benchmarks for these models are struggling to keep up with the rapid development of the speech models, as shown in Table 1. Although some benchmarks (Huang et al., 2024; Wang et al., 2024a; Ao et al., 2024; Bu et al., 2024) akin to FLAN (Wei et al., 2022) in text models, designed for speech models, they primarily focus on assessing models with speech understanding capabilities (Lyu et al., 2023; Chu et al., 2023; Shu et al., 2023; Chu et al., 2024; Liu et al., 2024; Tang et al., 2024), overlooking models' speech generation abilities, particularly in chat scenarios. Recent works (Chen et al., 2024b; Yang et al., 2024) take evaluating model's speech generation capabilities into consideration, but the evaluation are still conducted in the text modality, failing to account for whether models are capable of generating speech with paralinguistic information (Ji et al., 2024).

To fill this gap, we propose the S2S-Arena, a novel benchmark assessing the instruction-following capabilities of speech models in speech2speech protocols incorporating paralinguistic information, and an example shown in Figure 1. It requires speech models to not only understand paralinguistic cues (such as rhythm) in speech input but also to follow semantic instructions for generating speech output that preserves paralinguistic features.

Specifically, we adopt a three-stage construction process to build this benchmark: task definition, instruction design, and sample recording (detailed in Section 3). We select the most popular four practical domains with 21 tasks of speech models and carefully design the testing samples at four different levels by considering paralinguistics information in speech understanding and generation. We then collect 94 Text-to-Speech (TTS) synthesis samples and 60 human recordings. Since the speech model as the judge is unreliable (see Section 5.4), we implement a manual arena-style pairwise comparison among the popular four classes of S2S models (see Section 4). We obtain initial comparative results for the current models after 400 evaluations across 22 individuals. Additionally, we provide an in-depth analysis of key aspects, including semantic inconsistency between speech and text, language consistency, reasons for instruction-following failures, and position and length bias (see Section 5). Our contributions are as follows:

We introduce S2S-Arena, a novel arena-style benchmark to evaluate the instruction-following capabilities of speech models within speech-to-speech protocols, incorporating paralinguistic information.

We design 154 TTS and manual recording samples across four popular domains, with 21 tasks, and perform an arena-style manual comprehensive comparison of four different types of speech models based on their speech out [1].

Our findings suggest that the design of future speech models should give more consideration to multimodal and multilingual support, particularly in the context of speech generation involving paralinguistic information. Additionally, we discuss the unique biases in speech model evaluations in contrast to those observed in LLMs.

2 Related Work

2.1 Speech Models in Speech2Speech Protocols

Commercial speech models represented by GPT-4o-real-time [2] can naturally interact with humans in

[1] We release the dataset and comparison result at <Anonymous URL>.

[2] gpt-4o-realtime-preview-2024-10-01.

Model Type	Model Name	Input Form	Backbone	Output Form
Unknown	GPT-4o-realtime	Unknown	Unknown	Unknown
Cascade	FunAudioLLM (SpeechTeam, 2024)	Text /o Special Tokens	Qwen-2 72B	Text /o Special Tokens
Speech-token.	SpeechGPT (Zhang et al., 2023)	Speech Tokens	LLaMA 7B	Speech Tokens
	AnyGPT (Zhan et al., 2024)	Speech Tokens	LLaMA-2 7B	Speech Tokens
	LSLM (Ma et al., 2024)	Speech Tokens	Transformers	Speech Tokens
	Westlake-Omni	Speech Tokens	Qwen-2 0.5B	Speech Tokens
	GLM-4-Voice (Zeng et al., 2024)	Speech Tokens	GLM-4 9B	Speech Tokens
Speech-embed.	Mini-Omni (Xie and Wu, 2024)	Speech Embeddings	Qwen-2 0.5B	Speech Tokens
	LLaMA-Omni (Fang et al., 2024)	Speech Embeddings	LLaMA-3.1 8B	Speech Tokens
	Moshi (Défossez et al., 2024)	Speech Embeddings	Transformers	Speech Embeddings
	Freeze-Omni (Wang et al., 2024b)	Speech Embeddings	Qwen-2 7B	Speech Tokens

Table 2: Comparison of Speech2Speech Models. Speech-token. means Speech-token-based model and Speech-embed. means Speech-embedding-based model. Note that we do not compare LSLM and Moshi for a fair comparison because they do not use LLMs as backbones.

the form of speech-in and speech-out with paralinguistic information such as emotion and speaking-style, but the model architecture and training details have not been publicly disclosed.

As one of the representatives of open-source models, FunaudioLLM (SpeechTeam, 2024) adopts the most straightforward and traditional approach, implementing Speech2Speech through a cascade of Automatic Speech Recognition (ASR), LLM, and TTS, while incorporating special tokens to encode and represent the paralinguistic information contained in the speech, as shown in Table 2.

To better capture the paralinguistic information contained in speech, some works integrate it into the LLM process by adopting speech tokenization or embedding via speech-text alignment training.

Speech-token-based models such as SpeechGPT (Zhang et al., 2023), AnyGPT (Zhan et al., 2024), Westlake-Omni[3], and GLM-4-Voice (Zeng et al., 2024) use discrete speech tokens as the input by speech encoders like Whisper (Radford et al., 2023) or Hubert (Hsu et al., 2021). After multi-stage speech-text alignment training based on LLM, these models generate speech tokens for the voice decoder, preserving rich paralinguistic information such as tone and emotion.

Besides, speech-embedding-based models like Mini-Omni (Xie and Wu, 2024), LLaMA-Omni (Fang et al., 2024), and Freeze-Omni (Wang et al., 2024b) convert speech inputs into embeddings instead of discrete tokens, which are then fed into LLM for speech-text alignment training. Once trained, these models enable real-time speech interaction and consider the information contained in the speech.

[3]https://github.com/xinchen-ai/Westlake-Omni

2.2 Benchmarks for Speech Models

The evaluation of speech models is also advancing rapidly, as shown in Table 1. Based on the form of evaluation samples, existing benchmarks can be categorized into three types:

Prior works ((Bu et al., 2024; Wang et al., 2024a; Sakshi et al., 2024; Gong et al., 2024)) focus on evaluating the models on foundation task completion with speech understanding, but with output presented in text. For example, MMAU (Sakshi et al., 2024) emphasizes advanced perception and reasoning with domain-specific knowledge in speech, challenging models to tackle tasks akin to those faced by experts. In Dynamic-SUPERB (Huang et al., 2024), due to the crowdsourcing of tasks, the evaluation modalities are mixed.

Other works focus on evaluating the speech-based chat ability of the speech model. SD-Eval (Ao et al., 2024) evaluates whether the model perceives paralinguistic information such as age, emotion, and surrounding sounds contained in speech, but the output is still presented in text form. VocieBench (Chen et al., 2024b) uses TTS to build the speech-based Alpaca-Eval (Li et al., 2023) for evaluating instruction following ability but lacks consideration for paralinguistic information.

AIR-Bench (Yang et al., 2024) considers the paralinguistic information in the speech input in both chat and foundation task completion but lacks considering paralinguistic information in speech output.

Therefore, existing benchmarks mainly suffer from lacking consideration of paralinguistic information in speech output, inconsistent evaluation modalities (Chen et al., 2024a; Ye et al., 2024; Zhang et al., 2023), and unreliable automatic metrics (Streijl et al., 2016)(see Section 5.4).

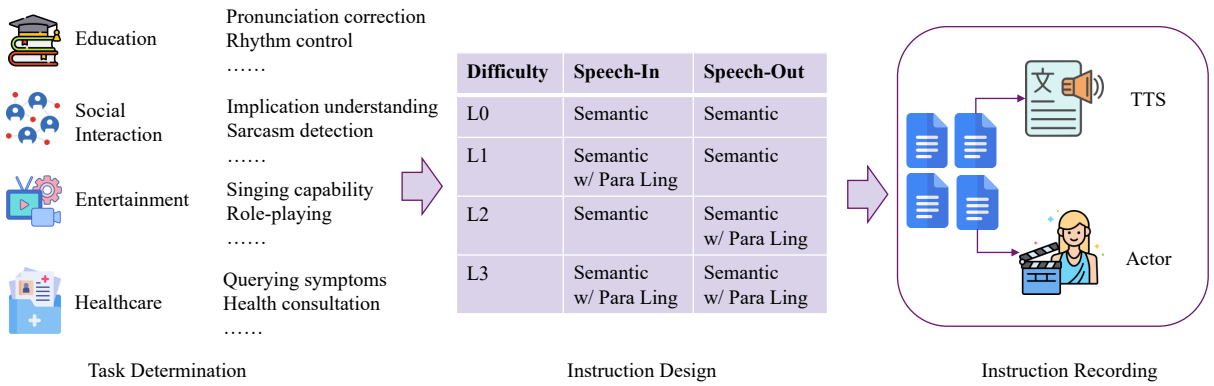

Figure 2: The Three-Stage Process of S2S-Arena Construction: Task Determination, Instruction Design and Instruction Recording.

3 S2S-Arena

To evaluate the ability of speech models to interact with humans in real-world speech-to-speech protocols, we introduce S2S-Arena, a benchmark that includes samples of varying difficulty levels to assess the instruction-following capabilities of speech models, considering paralinguistic information during both speech understanding and generation.

Unlike previous works, our benchmark incorporates two data sources—TTS and human-recorded speech—across two scenarios: foundation task completion and chat conversation. Additionally, we include a manual arena-style comparison of the speech modality instead of auto text-based evaluation to provide a more comprehensive and realistic evaluation. The three-stage process of S2S-Arena construction is shown in Figure 2.

3.1 Task Determination

Considering the widespread usage scenarios of the speech model in speech2speech protocols (Ao et al., 2024; Yang et al., 2024), we choose Education, Social Interaction, Entertainment, and Medical Consultation as evaluation domains. In each domain, we further design multiple fine-grained tasks, such as pronunciation correction and rhythm control in education, implication understanding, and sarcasm detection in social interaction, as shown in Table 3.

3.2 Instruction Design

For the sample design in each task, we consider a combination of TTS (Chen et al., 2024b) and human recording in both task completion and chat scenarios. Moreover, we divided each sample into four difficulty levels based on whether or not to consider the paralinguistic information in speech understanding and generation processes.

L0: Considering Instruction Following. It assesses only the model's ability to follow instructions without considering the paralinguistic information in speech-in and speech-out. For example, in the *Querying symptoms* task, the model receives the instruction, "I have a headache. What could be the cause?" The model is expected to provide possible causes of headaches by simply following instructions without any paralinguistic information.

L1: Considering Speech-in Paraglinguistic information. It further evaluates whether the model produces corresponding speech output by understanding paralinguistic information embedded in speech-in. For example, in the *Identity-based response* task, the model is given a spoken input from a child asking, "If it rains tomorrow, how should I plan my day?" The model is expected to discern the speaker's age using paralinguistic information and respond with suggestions suitable for children rather than adults.

L2: Considering Speech-out Paralinguistic Information. It evaluates whether the model generates the speech with paralinguistic information following speech instruction-following requirements. It is similar to TTS evaluation (such as TTS-Arena [4]) but uses speech-based instruction input without paralinguistic information. For example, one of the instructions in the *Tongue twisters* task is "Recite a tongue twister at three different speeds: fast, medium, and slow." The model's speech response should not only recite a tongue twister but also demonstrate each recitation at three different speeds.

L3: Considering Both Speech-in and Speech-out Paralinguistic Information. It assesses whether the model understands the speech-in par-

[4]https://huggingface.co/spaces/TTS-AGI/TTS-Arena

Domain	Task	Evaluation Target
Education	Pronunciation correction	Can the model correct inaccurate pronunciations?
	Emphasis control	Can the model understand stress emphasis and emphasize specific content with the right stress?
	Rhythm control	Can the model adjust the output pace, speaking faster or slower as required?
	Polyphonic word comprehension	Can the model accurately understand polyphonic word?
	Pause and segmentation	Can the model accurately pause and segment in ambiguous cases?
	Cross-lingual emotional translation	Can the model accurately convey emotions during translation?
	Language consistency	Does the model respond in the same language as the query when asked in different languages?
Social Companionship	Implication understanding	Can the model respond humorously, understanding implied meanings?
	Sarcasm detection	Can the model detect sarcasm in phrases like "You're amazing!"?
	Identity-based response	Can the model adapt responses based on the user's age (child, adult, elderly) and handle identity-based queries?
	Emotion recognition and expression	Can the model recognize emotions and provide appropriate responses based on different emotions?
Entertainment	Singing capability	Can the model sing a song upon request?
	Natural sound simulation	Can the model simulate certain natural sounds?
	Poetry recitation	Can the model recite poems?
	Role-playing	Can the model simulate a character with specific age, gender, accent, and voice tone?
	Storytelling	Can the model narrate a story with emotional depth?
	Tongue twisters	Can the model correctly pronounce a given tongue twister?
	Stand-up comedy/skit performance	Can the model perform a skit, playing both roles in a comedic dialogue?
Medical Consultation	Querying symptoms	Can the model answer questions related to symptoms?
	Health consultation	Can the model provide general health advice?
	Psychological comfort	Can the model provide comforting psychological support?

Table 3: Task Description Across Four Domains.

alinguistic information and generates speech with paralinguistic information properly. This highest level closely approximates real-world speech-to-speech scenarios. For example, in the *Cross-lingual emotional translation* task, one prompt with a happy emotion is "Help me tell him in Chinese that Mike is coming to my house tomorrow for a week." The model should fully recognize the expressed happiness and translate the message into Chinese with an equivalent emotional tone.

3.3 Instruction Recording

We design the instruction text and use Seed-TTS (Anastassiou et al., 2024) to synthetic the speech in some easier tasks such as the *Natural sound simulation* task. Additionally, we sample other normal samples, such as the *emotion recognition and expression* task from the existing popular speech datasets (Livingstone and Russo, 2018). Finally, we manually record the samples for other more difficult tasks such as the *sarcasm detection* and *singing capability* tasks that the TTS model cannot handle. To enhance the robustness of the benchmark, we use different vocal tones and add eight background noises, such as airport background sounds, to simulate diverse acoustic environments.

To ensure the quality of the samples, four native Mandarin speakers (two males and two females) with IELTS scores above 6.5 were recruited to assess data quality due to the samples being mainly English and Chinese. If any participants identified an issue with a particular sample, that would be discarded. In the end, we collected 154 independent speech instruction samples in 21 tasks, and more details can be seen in Appendix A .

4 Experiments

4.1 Experimental Settings

Most existing benchmarks automatically evaluate the speech model's output in the text modality by LLM (Zheng et al., 2023), but this method will lose valuable information in the speech modality (Chen et al., 2024a; Ye et al., 2024; Zhang et al., 2023), particularly paralinguistic information. Moreover, different from text-based automatic evaluation, speech-based automatic evaluation (Streijl et al., 2016; Saeki et al., 2022) is usually unreliable with bias, as demonstrated by our experiments (see Section 5.4).

Therefore, we adopt a manual arena-style approach with ELO ranking (Elo and Sloan, 1978) to more directly and comprehensively evaluate the performance of various speech models. More Details of ELO ranking calculation can be seen in the

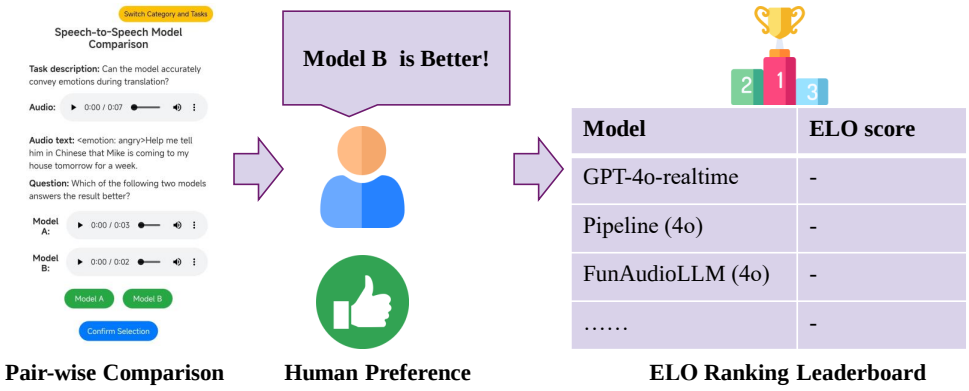

Figure 3: The Evaluation Process of S2S-Arena.

Appendix C.

Followed by Chat-Arena (Chiang et al., 2024) [5], we build a S2S-Arena web-based evaluation tool for evaluators to perform a reference-free comparison. Given a speech as the input, we invite human evaluators to rank two speech outputs generated by different speech models, considering both semantics and speech quality, as shown in Figure 3.

4.2 Benchmarked models

We select the following four categories of representative models for evaluation. **GPT-4o-realtime**[6]: We utilize the speech-enabled API version of GPT-4o instead of the app version. **Cascade Model**: We select the FunAudioLLM (SpeechTeam, 2024) as the strong Cascade model. For the best performance, we utilized the GPT-4o to replace the Qwen2 72B for the LLM module, which is named FunAudioLLM (4o). Besides, we also construct a vanilla Pipeline (4o) with Whisper, GPT-4o, and cosyVoice[7] for comparison. **Speech-Token-Based Model**: We select SpeechGPT [8], an Open-source LLM-based speech model as the representative Speech-Token-Based Model. **Speech-Embedding-Based Model**: We select recent two Omini series models (Mini-Omni (Xie and Wu, 2024) and LLaMA-Omini (Fang et al., 2024)) to represent Speech-Embedding-Based Model.

4.3 Results

We conducted a preliminary experimental investigation in S2S-Arena and received about 400 pair-wise comparison results with over 22 individuals, all of whom are native Mandarin speakers with IELTS scores above 6.5. To verify the evaluation quality, we select 10% of the samples annotated by two different annotators simultaneously, and the agreement between annotators is 83.7%.

4.3.1 Overal ELO Ranking

Table 4 presents the overall ELO rankings of each model based on their performance across all tasks in the four evaluation domains. It can be seen that GPT-4o real-time achieved the best ranking due to its excellent performance in the Education and medical consultation domains that require more knowledge. It also ranks high in social companionship due to its excellent ability to capture paralinguistic information. Surprisingly, it performs poorly in entertainment. After checking the samples, we found it refuses to do the task it does not have the ability to do.

Due to the decoupling of ASR, LLM, and TTS in the Cascade model, it performed better than the Speech-Token-Based and Speech-Embedding-Based models with its excellent LLM core (GPT-4o) without considering other factors such as latency and full duplex.

However, although other non-GPT-4o models perform well in the social companionship and entertainment domain, their performance in knowledge-intensive scenarios is significantly reduced due to the limitations of the LLM backbone. It is noted that we did not find a significant difference in performance between the Specch-Token-Based models and the Speech-Embedding models, which can be illustrated in Figure 4.

4.3.2 Pair-wise Comparison

To compare various models directly, we further analyze the win rate between the pairwise speech

[5] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[6] gpt-4o-realtime-preview-2024-10-01.

[7] whisper-large-v3 for ASR, gpt-4o-2024-08-06 for LLM and CosyVoice-300M-Instruct for TTS.

[8] https://github.com/0nutation/SpeechGPT

Model Type	Model	Overall	Edu.	Social Comp.	Enter.	Med.
Unknown	GPT-4o-realtime	**1365**	**1185**	1064	970	**1146**
Cascade	Pipeline (4o)	1207	1065	995	1069	1077
	FunAudioLLM (4o)	1025	1105	**1077**	850	993
Speech-Token-Based	SpeechGPT	849	906	919	**1095**	929
Speech-Embedding-Based	Mini-Omni	841	857	1000	1041	943
	LLaMA-Omni	714	882	945	975	911

Table 4: ELO Rank across Various S2S Models.

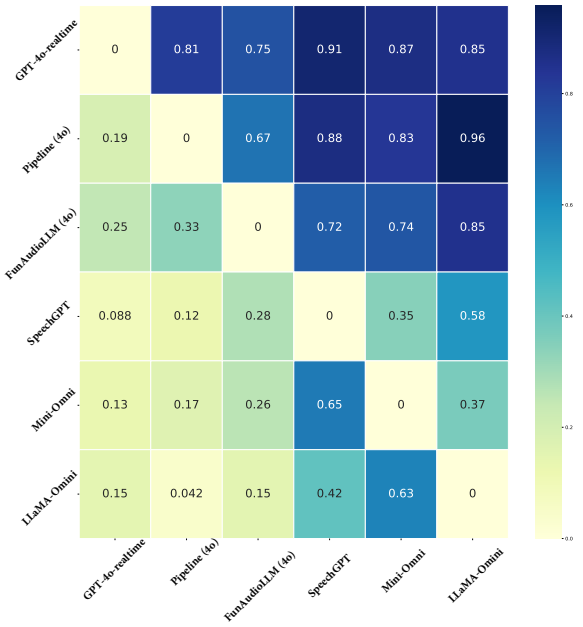

Figure 4: Pair-wise comparison of various models.

models, as shown in Figure 4. It can be seen that the first three GPT-4o-based models (GPT-4o-realtime, Pipeline (4o), and FunAudioLLM (4o)) are significantly better than the last three models (SpeechGPT, Mini-OMini, LLaMA-omni). Moreover, although FunAudioLLM (4o) has the optimal threat power for GPT-4o-realtime, it can be seen that Pipeline (4o) outperforms the other three open-source models with a popular speech encoder (whisper). Besides, although speech-token-based models and speech-embedding-based models take different technology roadmaps, each has their own strengths, making it difficult to determine which one is more outstanding and significant.

4.4 The Causes of Model Failures

We statics the samples where the model failed and summarized the following three reasons: (1) The model follows the instructions but performs worse than other models (37.5%); (2) The model attempts to execute the instructions but fails to complete them (15.4%). (3) The model fails to recognize or understand the given instructions (47.1%).

Notably, higher-performing models (with higher ELO scores) were more prone to Case 1 failures, while lower-performing models (with lower ELO scores) struggled more with Case 3 failures. This suggests that models with stronger speech understanding capabilities still face challenges in speech generation, while weaker models have greater difficulty understanding speech in the first place.

5 Analysis

In this section, we first explore the performance of speech models with support for multimodal and multilingual capabilities. Then, we investigate the position and length biases present in the evaluation and the potential of automated assessment.

5.1 Does Semantic or Paralinguistic Information Dominate?

To analyze which one dominates the model's response, we explore the performance of advanced models (GPT-4o-realtime, Pipeline (4o), and FunAudioLLM (4o)) in Chinese sarcasm detection tasks where the model needs to simultaneously consider paralinguistic and semantic information in speech.

Among the three tested samples, all of the models understood the sarcasm reflected by the inconsistency between the paralinguistic information and semantic information in speech in 67% of the cases, while in the remaining 33% of the scenarios, they responded with the original semantics. Given this interesting discovery, we added eight additional L2 and L3 samples for sarcasm detection to evaluate whether the model has the ability to express sarcasm. The experimental results show that the success rates of the three models drop to 37.5%, 62.5%, and 37.5%. Therefore, balancing paralinguistic information with inconsistent semantics in speech is a challenge for future research.

5.2 Is the Speech Module or LLM more Important for Multilingual Support?

We further analyze the speech models' multilanguage support ability, which is a crucial factor for their practical deployment. Table 5 shows the results of language support tests on the four languages of Chinese, English, Japanese, and Thai selected from the same sample.

It can be seen that GPT-4o-realtime, with its advanced speech encoder/decoder, supports a wide range of languages. However, other GPT-4o-based model cascade models cannot support Thai as their speech codecs cannot process Thai. Interestingly, LLaMA-Omni, which uses Whisper as its encoder, can understand Chinese but only respond in English due to its LLaMA 3.1 backbone. Therefore, integrating multilingual speech encoders/decoders with multilingual LLM backbones to achieve seamless language support is necessary.

Model	Speech-In	Speech-Out
GPT-4o-realtime	EN, CN, JP, TH	EN, CN, JP, TH
Pipeline (4o)	EN, CN, JP	EN, CN, JP
FunAudioLLM (4o)	EN, CN, JP	EN, CN, JP
LLaMA-Omni	EN, CN	EN
Mini-Omni	EN	EN
SpeechGPT	EN	EN

Table 5: Language Support by Models for Input and Output in Four Languages: English (EN), Chinese (CN), Japanese (JP), and Thai (TH).

5.3 Do Positional and Length Biases Exist in Speech Evaluation?

We then analyze the positional and length biases in speech evaluation to determine whether they exist in the same way as they do in text-based evaluation.

For the positional bias, we found 5 samples with different preferences in 22 samples with manually annotated swapping option positions, accounting for 22.7%. Interestingly, unlike text-based evaluations, where the first candidate is typically favored, 80% of the samples in this case had higher win rates when placed later in the sequence. This could be due to humans' tendency to better remember more recent sounds.

Regarding length bias, we found that longer outputs were often preferred, mirroring trends seen in text-based evaluations. In 63.02% of the total 400 comparisons, the longer output was favored. The average length of the winning output was 16.75 seconds, compared to 12.01 seconds for the losing output.

5.4 Is it Possible to use the Speech Model as the Judge Directly?

To verify its feasibility, we select existing speech models (such as GPT-4o-realtime and Qwen2-Audio) as the judge and convert the evaluation prompt into speech using TTS combined with the test sample as the input for judgment (detailed in Appendix D).

The experimental results are shown in Table 6. The most significant finding is that speech models do not achieve the same level of agreement with human evaluations as LLMs, with scores of 30.2% for GPT-4o-realtime and 25.6% for Qwen2-Audio. Additionally, the consistency across multiple evaluations of the same model is also low, with GPT-4o-realtime achieving the highest consistency at only 58.1%. Furthermore, there are significant positional (over 40%) and length biases (55.8% for GPT-4o-realtime) when evaluated by the speech models. Therefore, different from the existing research on text-based evaluation, directly taking the speech model as the judge is not ready for speech-based evaluation.

Model	Agreement		Bias	
	Inter	Human	Positional	Length
GPT-4o-realtime	58.1%	30.2%	40.9%	55.8%
Qwen2-Audio	48.1%	25.6%	86.4%	48.8%

Table 6: Automatic Evaluation Results of Speech Models.

6 Conclusion

In this paper, we introduce S2S-Arena, a novel benchmark designed to evaluate the instruction-following abilities of Speech2Speech models concerning paralinguistic information. We present a comparative performance analysis of existing speech models across 21 tasks in four domains using a carefully crafted benchmark and arena-style evaluation methodology. Additionally, we examine the challenges and limitations of current speech models in the context of multimodal and multilingual capabilities and discuss positional and length biases in both manual and automatic speech-based evaluations. In future work, we aim to broaden the scope of our evaluation and develop an automatic evaluation framework for speech models in Speech2Speech protocols, offering guidance for the development of large-scale speech models that incorporate paralinguistic information for practical applications.

Limitations

We acknowledge that the current sample data size of S2S Arena is relatively small due to the difficulty in obtaining manual speech instructions in real-world scenarios. We also acknowledge that our model selection scope is limited due to some recent speech models' close sources and limited access. We have started building a more widely open arena website and accepting submissions of samples and models from other researchers to obtain more comprehensive and updated evaluations.

References

James F Allen, Donna K Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI magazine*, 22(4):27–27.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *arXiv preprint arXiv:2406.13340*.

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, and Noam Amir. 2011. *The automatic recognition of emotions in speech*. Springer.

Fan Bu, Yuhao Zhang, Xidong Wang, Benyou Wang, Qun Liu, and Haizhou Li. 2024. Roadmap towards superhuman speech understanding using large language models. *arXiv preprint arXiv:2410.13268*.

Stuart K Card, Allen Newell, and Thomas P Moran. 1983. The psychology of human-computer interaction.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024c. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, Sakshi Singh, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. 2024. Avodyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan S. Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-Yi Lee. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12136–12140. IEEE.

Julia Ipgrave. 2009. The language of friendship and identity: Children's communication choices in an interfaith exchange. *British Journal of Religious Education*, 31(3):213–225.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2024. Language model can listen while speaking. *arXiv preprint arXiv:2408.02622*.

Takashi Nose, Yoichi Kato, and Takao Kobayashi. 2007. Style estimation of speech based on multiple regression hidden semi-markov model. In *INTERSPEECH*, pages 2285–2288.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Interspeech 2022*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2794–2797.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. 2023. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930*.

Tongyi SpeechTeam. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Robert C. Streijl, Stefan Winkler, and David S. Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multim. Syst.*, 22(2):213–227.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

George L Trager. 1958. Paralanguage: A first approximation. *Stud. Linguist.*, 13:1–12.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Xiong Wang, Yangze Li, Chaoyou Fu, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024b. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024c. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

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 *Proceedings of EMNLP, 2023*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Distribution of Samples

A total of 154 samples were designed across four domains, with tasks categorized into four levels of complexity (L0 to L3), as shown in Table 7. In the Education domain, tasks are predominantly distributed across L1 and L2, with "Cross-lingual Emotional Translation" showing a higher concentration of L3 samples. The Social Companionship domain is primarily focused on L1 and L3 tasks, particularly with a notable number of samples in "Emotion Recognition and Expression" and "Identity-based Response." In the Entertainment domain, tasks are largely concentrated in L2 and L3, while the Medical Consultation domain exhibits a more balanced distribution across L0 to L2, with samples fairly evenly spread across these levels. Overall, the distribution of samples across L1, L2, and L3 is relatively even, with L0 samples being comparatively fewer.

B Experimental Details

We standardize the samples to a 24,000 Hz sample rate to ensure fairness in testing. However, due to some models' limited support for certain input formats, we are required to use alternative formats. Specifically, for SpeechGPT, we convert the input audio to a 22,500 Hz sample rate.

C ELO Rank Details

Specifically, In our evaluation framework, all models start with an initial ELO rating of 1000. Each comparison round is conducted in a no-tie format, with the winning model's ELO score updated based on its relative performance to the competing model. Specifically, we calculate the expected score E_A for model A against model B using the Eq. (1):

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}} \tag{1}$$

where R_A and R_B are the current ratings of models A and B, respectively. The updated rating R'_A for model A is then computed as Eq. (2):

$$R'_A = R_A + K \times (S_A - E_A) \tag{2}$$

where S_A represents the actual outcome for model A (1 for a win, 0 for a loss), and K is the adjustment factor, set to 32.ll

Domain	Task	L0	L1	L2	L3	Total
Education	Pronunciation correction	0	5	0	0	5
Education	Emphasis control	0	4	0	1	5
Education	Rhythm control	0	0	6	1	7
Education	Polyphonic word comprehension	0	6	0	0	6
Education	Pause and segmentation	0	2	0	2	4
Education	Cross-lingual emotional translation	0	0	2	12	14
Education	Language consistency	4	2	0	0	6
Social Companionship	Implication understanding	4	0	0	0	4
Social Companionship	Sarcasm detection	0	3	0	0	3
Social Companionship	Identity-based response	0	12	0	4	16
Social Companionship	Emotion recognition and expression	0	0	0	24	24
Entertainment	Singing capability	0	0	5	2	7
Entertainment	Natural sound simulation	0	0	5	0	5
Entertainment	Poetry recitation	3	0	1	0	4
Entertainment	Role-playing	0	0	0	4	4
Entertainment	Storytelling	0	0	5	0	5
Entertainment	Tongue twisters	0	0	9	0	9
Entertainment	Stand-up comedy/skit performance	0	0	8	0	8
Medical Consultation	Querying symptoms	0	5	0	1	6
Medical Consultation	Health consultation	7	0	0	0	7
Medical Consultation	Psychological comfort	0	0	5	0	5
	Total	18	39	46	51	154

Table 7: Distribution of Samples.

D Automatic Evaluation Prompt

"I will provide an input audio and two corresponding response audios. Please evaluate which response is better. You only need to reply with 'First one wins' or 'Second one wins.' Here is the input audio: [input audio], the first response: [output audio 1], and the second response: [output audio 2]."