MTalk-Bench: Multi-Turn Dialogue Benchmark for Speech-to-Speech Large Language Models

Anonymous ACL submission

Abstract

The rapid advancement of speech-to-speech (S2S) large language models (LLMs) has brought impressive progress in real-time spoken interaction. However, current evaluation methods fall short in assessing their multi-turn dialogue capabilities, especially under realistic and complex communication settings. To fill this gap, we introduce MTalk-Bench, **the first multi-turn S2S benchmark**, specifically designed to evaluate S2S LLMs across 9 high-frequency multi-turn dialogue scenarios. MTalk-Bench adopts a three-tier evaluation framework covering Semantic Information, Paralinguistic Information, and Ambient Sound, reflecting the rich dynamics of human conversation. We conduct both human and LLM-based evaluations, and further analyze the reliability of LLMs as judges. Experimental results demonstrate that GPT-4o-realtime consistently achieves state-of-the-art performance across all tiers, and also exhibits strong reliability when serving as an evaluator. While several S2S LLMs show promising results in semantic comprehension, they still struggle with conversations involving paralinguistic and ambient sound cues. MTalk-Bench offers a standardized and multidimensional evaluation tool to drive future research toward more context-aware, robust S2S dialogue systems.

1 Introduction

The advent of sophisticated S2S-LLMs represents a significant leap forward in human-computer interaction, promising more natural, intuitive, and engaging multimodal and multilingual dialogue experiences (Author and Others, 2023; Borsos et al., 2023). These models, capable of directly processing spoken input and generating spoken output, are poised to revolutionize applications ranging from virtual assistants and customer service agents to educational tools and interactive entertainment. However, the rapid proliferation of these advanced capabilities has outpaced the development of comprehensive and systematic evaluation tools, particularly for assessing their proficiency in complex, multi-turn spoken dialogues.

Despite the rapid progress of S2S-LLMs, there remains a critical gap in their evaluation: the lack of standardized, multi-dimensional benchmarks for multi-turn spoken dialogue. Existing evaluations (e.g., SuperGLUE (Wang et al., 2019) and HELM (Liang et al., 2022) for text, or various single-turn speech tasks (Author and Others, 2022)) fail to capture the challenges unique to extended spoken interactions. These include maintaining contextual coherence across turns, handling paralinguistic cues like prosody and emotion (Author and Others, 2021), and managing real-world audio conditions such as background noise and speaker overlap. Furthermore, little work has explored using LLMs themselves to evaluate such tasks, especially in spoken, multi-turn settings, despite their growing use as general-purpose evaluators.

To address these limitations, we propose MTalk-Bench, the first multi-turn S2S benchmark designed to evaluate S2S LLMs in realistic spoken dialogue. It comprises 270 distinct dialogue samples, systematically distributed across 9 ecologically valid communication scenarios. For each scenario, model performance is assessed along three critical dimensions: *Semantic Information*, *Paralinguistic Information*, and *Ambient Sound*. In addition to human evaluations, we employ multiple LLMs as judges and conduct an in-depth analysis of their reliability. Our framework enables comparative, turn-level analysis, offering a scalable and fine-grained evaluation protocol. MTalk-Bench thus aims to accelerate the development of more coherent, expressive, and context-aware S2S dialogue systems.

The key contributions of this papers are **(I) First Multi-Turn S2S Dialogue Benchmark**: We propose MTalk-Bench, the first benchmark for holis-

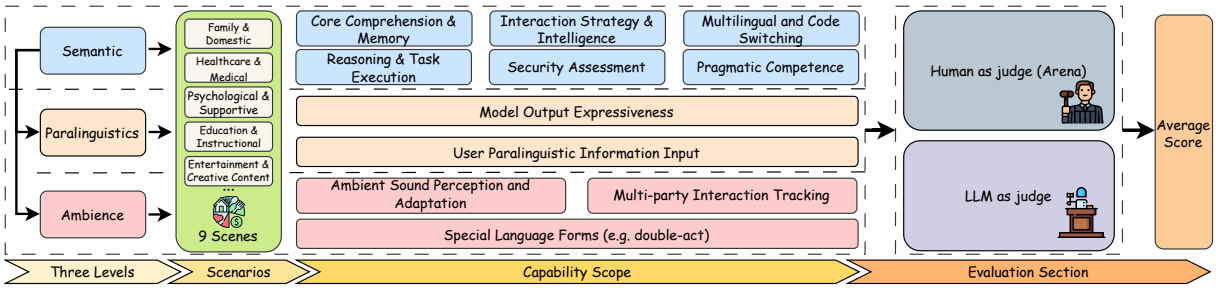

Figure 1: The Overview of MTalk-Bench.

tic evaluation of speech-to-speech (S2S) LLMs in multi-turn dialogues. Its three-level framework (semantic, paralinguistic, and environmental understanding) captures the full complexity of spoken interactions. **(II) User-Centered Scenario/Capability Coverage**: Featuring 27 diverse scenarios grounded in sociolinguistic theory and validated through large-scale user studies, the benchmark aligns with real-world applications and user expectations for intelligent speech agents. **(III) LLM-Driven S2S Evaluation**: We pioneer a dual human-LLM evaluation approach, systematically analyzing LLMs' reliability as automated judges for multi-turn S2S tasks and demonstrating their scalability potential.

2 Related Works

Dialogue and Speech-to-Speech Models Modern dialogue systems have achieved notable fluency through LLMs (Radford et al., 2019; OpenAI, 2023) and conversation-specialized models such as DialoGPT (Zhang et al., 2020). For spoken interaction, while traditional systems use cascaded ASR-LLM-TTS pipelines, emerging end-to-end (E2E) speech-to-speech (S2S) models like Translatotron (Jia et al., 2019), GLM-4-Voice (Zeng et al., 2024), and Qwen2.5 Omni (Li et al., 2025) aim to directly convert speech while preserving paralinguistic features and reducing latency.

Dialogue Benchmarks: from Text to Speech Current benchmarks evaluate diverse facets of dialogue systems and LLMs: MultiWOZ (Budzianowski et al., 2020) remains central for text-based task-oriented dialogue (dialogue state tracking, response generation), while BigBench (Srivastava et al., 2023) and HELM (Liang et al., 2022) assess general LLM capabilities via text inputs but overlook multi-turn interactive dialogue, particularly in speech. SpokenWOZ (Si et al., 2023) and VoiceBench (Chen et al., 2024) advance spoken interaction evaluation, focusing on speech-text task-oriented dialogues and LLM-based voice assistants (e.g., instruction following, perturbation robustness), while multimodal benchmarks (e.g., (Razzhigaev et al., 2024; Sviridov et al., 2025)) integrate text, vision, and audio but lack dedicated evaluation of E2E S2S LLMs in extended conversations. S2S evaluations often rely on component-level metrics (ASR Word Error Rate (OpenAI, 2025), TTS Mean Opinion Score) or task-specific scores (BLEU for speech translation (Popel et al., 2020)), neglecting holistic multi-turn dynamics. Key challenges include maintaining context, consistency, and memory in prolonged interactions (Serban et al., 2016; Sirdeshmukh et al., 2025), alongside effective perception/generation of non-semantic cues (prosody, emotion (Schuller and Batliner, 2013)) and acoustic robustness (OpenAI, 2025) which are rarely systematically assessed in S2S frameworks. MTalk-Bench addresses this gap by holistically evaluating E2E S2S LLMs across core dialogue competencies and interaction dynamics in multi-turn spoken conversations.

3 Framework of MTalk-Bench

MTalk-Bench is designed to address the critical absence of comprehensive frameworks for evaluating multi-turn dialogue capabilities in S2S-LLMs, particularly across semantic, paralinguistic, and ambient sounds dimensions, see Figure 1 for an overview.

3.1 Multi-level Evaluation Framework

The evaluation framework of MTalk-Bench is structured around three interconnected tiers of information processing, each critical for holistic spoken-to-written (S2S) interaction. Drawing on foundational research in linguistics, speech science, and auditory scene analysis (Levelt, 1989; Bregman, 1990; Scherer, 1986), this tiered design offers a coherent framework for assessing the multifaceted capabilities of S2S-LLMs. These tiers reflect progressively

broader aspects of communication:

Tier 1. ***Semantic Information Processing****: The foundational tier focuses on analyzing and generating the core linguistic meaning and dialogue content.*

Building on this foundation, the evaluation ascends to *paralinguistic* information processing, which examines how vocal nuances (e.g., tone, rhythm, and emotional prosody) enhance or alter semantic meaning.

Tier 2. ***Paralinguistic Information Processing****: Building on the semantic tier, this tier addresses expressive vocal nuances (e.g., tone, rhythm, emotion) that modulate meaning and intent.*

Progressing to the broader communicative context, *ambient sounds* information processing encapsulates the model's ability to interpret environmental acoustics, ensuring robustness in real-world settings.

Tier 3. ***Ambient Sounds Information Processing****: The outermost tier contextualizes dialogue by evaluating a model's ability to interpret and adapt to environmental auditory cues (e.g., background noise, acoustic scenes).*

Philosophy: This hierarchical structure mirrors the tiered nature of human communication, enabling systematic evaluation from basic message comprehension to nuanced delivery and environmental robustness. The interdependence of tiers highlights that true communicative competence in S2S-LLMs emerges from integrating capabilities across all levels, rather than excelling at isolated tasks. Each tier is further subdivided into detailed sub-dimensions (see Appendix A.1), ensuring a granular yet holistic performance assessment that transcends traditional metrics.

3.2 User-centric Taxonomy

User Survey Protocol In this work, our evaluation centers on two dimensions: **Scenarios** (specific contexts where interactions occur, e.g., "Family and Domestic Communication" or "Doctor–Patient Communication") and **Capabilities** (functional skills to be evaluated, e.g., "real-time translation" or "contextual memory"). To prioritize these, we employed a pairwise arena-style survey. Participants compared two randomly selected scenarios or capabilities (within the same dimension) and selected the one they deemed more likely to involve interaction with a speech-based AI agent in the near future. This pairwise method, grounded in robust preference elicitation principles (Bradley and Terry, 1952; David, 1963), reduces rating-scale biases and captures nuanced perceptions of relative likelihood.

3.2.1 On Scenario Taxonomy

Scenarios	ELO Score
Family and Domestic Communication	**1091**
Doctor–Patient Communication	**1070**
Legal Inquiry and Institutional Interrogation	**1048**
Financial Investment and Advisory Communication	**1045**
Classroom Instruction and Interaction	**1029**
Academic Communication and Research Collaboration	**1027**
Job Interview Communication	**1027**
Workplace Collaboration and Communication	**1017**
Marketing and Customer Relationship Management Communication	**1008**
Informal Social Interaction (Casual Conversation)	1007
Psychological and Supportive Communication	1003
Service-Oriented Interaction	1001
Public Discourse and Interaction	995
Entertainment and Content Creation Communication	980
Negotiation and Conflict Resolution	979
Intercultural and Linguistic Communication	974
Public Affairs and Emergency Response Communication	960
Religious and Spiritual Communication	950
Environmental Advocacy and Policy Promotion	906
Sports and Competitive Communication	873

Table 1: Elo ranking for important **Scenarios** by Pairwise User Survey through 46 questionnaires. Top Nine Scenarios are selected (as in bold).

MTalk-Bench's scenario corpus, designed for authentic communicative challenges, originated from twenty candidates identified in literature reviews across communication studies, linguistics, and HCI (Kuniavsky, 2002; Gumperz, 1982; Clark, 1996; Schegloff, 2007). These candidates were refined via a pairwise survey where users selected scenarios based on perceived likelihood of future S2S-LLM interaction (Bradley and Terry, 1952; Thurstone, 1927; David, 1963). Preference rankings informed the consolidation of related, high-ranking scenarios into nine core contexts for MTalk-Bench, balancing diversity and communicative function, see Table 1.

3.3 On Capability Taxonomy

MTalk-Bench integrates empirical user priorities with theoretical constructs through a three-tiered evaluation framework. Initial S2S capabilities, derived from literature reviews (Jurafsky and Martin, 2000; Wang et al., 2023; Gao et al., 2023a), were prioritized via a pairwise survey to establish a user-informed capability importance (Appendix A.1.2). Grounded in linguistics, speech science, pragmatics, and cognitive science principles (Levelt, 1989; Scherer, 1986; Clark, 1996; Grice, 1975), these capabilities were decomposed into measurable sub-dimensions, complemented by essential communication aspects, and systematically organized into

Capability	ELO Score
Long-Term Conversational Memory	1028
Long-Term Term Instruction Following	1027
Semantic Disambiguation	1020
Logical and Commonsense Reasoning	1019
Task Understanding and Planning	1018
Error Correction and Adaptability	1018
Dialogue Management	1001
Ambient Sound Perception and Adaptation	998
Environmental Audio Information Acquisition	998
Speaker Information Processing	993
Multi-Party Interaction Tracking	989
Emotion Recognition and Perception	989
Pragmatic and Cultural Intelligence	987
Stylistic and Personality Expression Control	986
Cross-Topic Switching Capability	986
Self-Monitoring and Reflective Capability	984
Contextual Adaptation Capability	984
Continual Learning and Self-Optimization	966

Table 2: Elo ranking for important **Capability** by Pairwise User Survey (46 survey responses).

three tiers: Semantic, Paralinguistic, and Ambient Sound Processing, see Sec. 3.1. This structure ensures holistic coverage of S2S interactions, enabling nuanced assessment across sub-dimensions (detailed in Appendix A.1.2 and visualized in Figure 1).

4 Construction of MTalk-Bench

4.1 Dataset Construction

Dataset construction mainly involves dialogue instance generation which were generated via LLM automation and rigorous human refinement. This involves three tiers including as follows.

Tier 1: Dialogue with Semantic Information Semantic dataset creation involved LLMs for: 1) generating contextual multi-turn dialogues (Claude 3.5 Sonnet) (Roller et al., 2021); 2) multi-labeling their semantic capabilities (Claude 3.5 Sonnet); and 3) Primary evaluation dimension inference from these labels (Gemini 2.0 Flash). Critical human refinement subsequently ensured consistency between intended and inferred dimensions, dialogue naturalness, clear testability of the primary capability, and balanced coverage (minimum 10 valid samples per major).

Tier 1 → Tier 2: Dialogue with Paralinguistic Information Derived from the semantic base, the paralinguistic dataset (Schuller and Batliner, 2013) tests: 1) understanding of user inputs augmented with paralinguistic metadata (e.g., tone, emotion); and 2) generation of model outputs with specified expressive features (e.g., varied style).

Tier 1 → Tier 3: Dialogue with Ambient Sound The ambient sound dataset, also from the semantic base, assesses S2S LLM performance in acoustically diverse settings by testing: 1) comprehension of incorporated background sounds (e.g., alarms) (Purohit et al., 2021); 2) tracking of multi-speaker interactions (Chen et al., 2021); and 3) understanding of stylized speech (e.g., crosstalk) (Salesky et al., 2021).

This structured generation yields diverse, targeted challenges for S2S LLM evaluation, as shown in Table 3, which presents examples of dialogue instances across diverse evaluation dimensions. For the detailed construction of the dialogue instances used for MTalk-Bench evaluation, please refer to Appendix A.2, where the generation process and modify progress are further elaborated.

4.2 Evaluation Protocol

MTalk-Bench employs a multi-faceted evaluation protocol to comprehensively assess S2S LLM performance across the defined capability dimensions and information levels. For each capability, specific dialogue scenarios are utilized. These instances may involve pre-defined audio prompts, simulated interaction histories, and explicit instructions or constraints designed to target the capability under evaluation. The protocol leverages a hybrid approach, combining automated LLM-based assessment with human and model-based evaluation on an "Arena" platform, to capture both semantic and acoustic aspects of interaction quality.

4.2.1 LLM-as-Judge

For capabilities primarily assessable through semantic content and logical dialogue structure (predominantly Level 1 capabilities and textual aspects of Level 2), we employ a powerful Large Language Model, specifically GPT-4o (OpenAI, 2023), as an automated judge. GPT-4o evaluates transcribed S2S model interactions against predefined, task-specific rubrics. These rubrics guide the assessment of aspects such as coherence, adherence to instructions, quality of reasoning, and semantic relevance of responses. This methodology offers a scalable and consistent means for evaluating core language understanding and reasoning abilities. We acknowledge, however, potential limitations inherent in LLM-based evaluation, such as susceptibility to inherent model biases, which we aim to mitigate through rubric design and diverse prompt engineering.

4.2.2 Human-based Arena

The Arena platform integrates both human judges and specialized acoustic analysis models. Human judges provide subjective ratings based on

Scenario	Tier1: Semantic Information	Tier2: Paralinguistic Information	Tier3: Ambient Sound
	Speech Questions	Speech Questions	Speech Questions
Medical & Health	Turn 1: You explain the process of getting the flu vaccine to my grandma in simple French. Her hearing is a little poor. Speak slowly. Turn 2: Now, can you tell her the possible side effects? (Capability: Multilingual and Code-Switching)	Turn 1: You explain the process of getting the flu vaccine to my grandma in simple French. <gentle tone, slow pace> Her hearing is a little poor. <slows down speech> Speak slowly. Turn 2: <encouraging> Okay, great <repeats>. Now, <quiet reminder> can you tell her the possible side effects? Keep your voice calm and reassuring. (Capability: Output expressiveness)	Turn 1: <Grandma humming> You explain the process of getting the flu vaccine to my grandma in simple French. Her hearing is a little poor. Speak slowly. Turn 2: <Grandma stops humming, makes a slightly confused sound, and asks a mumbled "What?"> Can you repeat and tell her the possible side effects? (Capability: Ambient Understanding)
Family & Life	Turn 1: I want to invite my family to a weekend barbecue. How can I phrase it in a casual yet persuasive way? Turn 2: Some family members are concerned about the weather. How can I address this and reassure them? (Capability: core comprehension and memory)	Turn 1: <sad and depressed> I want to invite my family to a weekend barbecue. How can I phrase it in a casual yet persuasive way? Turn 2: <torn and distressed> Some family members are concerned about the weather. How can I address this and reassure them? (Capability: Input Paralinguistic understanding)	Turn 1: I want to invite my family to a weekend barbecue. How can I phrase it in a casual yet persuasive way? Turn 2: <sound of wind and distant thunder> Some family members are concerned about the weather. How can I address family members' concerns about the... <sudden loud clap of thunder that briefly drowns out a key word 'weather'>? (Capability: Multi-Party Interaction Tracking)

Table 3: Example of two conversation sets across diverse assessment dimensions.

detailed criteria related to interactional quality, naturalness, and audio fidelity. Concurrently, automated acoustic models can furnish objective metrics, such as emotion classification accuracy from speech, signal-to-noise ratio robustness measures, or speaker diarization accuracy in multi-party contexts.

This dual evaluation approach, combining LLM-as-Judge for content and the Arena platform for acoustic and interactional qualities, provides a holistic assessment of S2S LLM performance. Detailed experimental results, including the performance of various baseline S2S LLMs on MTalk-Bench and in-depth analyses across the different capability dimensions and scenarios, will be presented in Section 5.

5 Experiments

5.1 Evaluation Setup

We adopt a comprehensive evaluation framework to assess the performance of S2S LLMs, centered on human evaluation using an Arena-style pairwise comparison and supplemented by absolute scoring and LLM-as-judge assessments. Unlike existing benchmarks that often convert speech outputs to text for evaluation (Zheng et al., 2023), our setup crucially maintains the speech modality throughout the evaluation process. This preserves vital paralinguistic and acoustic information essential for a holistic assessment of S2S capabilities (Chen et al., 2025; Jiang et al., 2025; Zhang et al., 2023).

To assess capabilities beyond semantic content, such as prosody or noise robustness, we further introduce an absolute scoring protocol. For this, annotators rate individual model responses on a 0–10 scale. A key element of this protocol is the use of a semantically equivalent reference response, which is scored as a baseline of 5.

To complement human evaluation and explore scalable assessment methods, we replicate these pairwise and absolute scoring protocols using state-of-the-art LLMs as judges. The specific methodologies for human and LLM-based evaluations, including data collection and LLM prompting strategies are shown in Appendix B.3. Overall, our framework quantifies S2S-LLM performance using a combination of relative Elo ratings and absolute scores, derived from both human and LLM-based judgments across various dimensions of spoken dialogue.

5.2 Benchmarked Models

We evaluate five state-of-the-art S2S-LLMs with multi-turn dialogue capabilities, selected to represent a range of architectures and development backgrounds. These include both proprietary and open-source systems: **GPT-4o-realtime**[1], **GLM-4-Voice-9B**[2], **Qwen2.5-Omni-7B**[3], **LLaMA-Omni2**[4], and **Westlake-Omni**[5], details about these models can be found in Appendix D.

5.3 Human Evaluation with MTalk-Arena

Human evaluation is conducted on the MTalk-Arena platform (See Appendix B.1), which supports pairwise comparisons and absolute scoring.

[1]Accessed via OpenAI API in May 2025. https://openai.com/index/hello-gpt-4o

[2]Version v1.0, released April 2025. https://github.com/THUDM/GLM-4-Voice

[3]Released May 2025. https://www.alibabacloud.com/blog/alibaba-cloud-releases-qwen2-5-omni-7b-an-end-to-end-multimodal-ai-model_602095

[4]Accessed from ICT NLP Lab GitHub repository, May 2025. https://github.com/ictnlp/LLaMA-Omni

[5]Model details and resources available via Westlake University, accessed May 2025. https://westlake.edu.cn

S2S Model	Overall	Semantic Info.	Paralinguistic Info.	Ambient Sound
GPT-4o-realtime	**1022.67**	**1042.52**	**1011.35**	**1014.13**
GLM4-Voice	1001.63	1001.07	999.67	1004.16
Qwen-Omni	1011.74	1019.71	1009.97	1005.53
LLaMA-Omni	989.21	980.90	992.74	994.00
Westlake-Omni	974.75	955.80	986.27	982.19

Table 4: Aspect-Specific Performance of S2S LLMs based on Human Evaluation

For each task, annotators listen to user prompts and model responses, and follow dimension-specific instructions (semantic, paralinguistic, or environmental) to make a decision.

In the Arena setting, annotators compare two anonymized model responses per turn and select the better one, or indicate a tie. In the absolute scoring setting, each enhanced response is rated on a 0–10 scale, with a stripped-down baseline fixed at 5. This design quantifies the perceived impact of paralinguistic or environmental features in speech.

For Arena-style pairwise comparison, 24 certified annotators produced a total of 258 judgments on the MTalk-Arena platform, comparing model outputs across semantic, paralinguistic, and environmental dimensions. For absolute scoring tasks, two expert annotators independently rated 100 enhanced vs. baseline responses. The inter-annotator agreement, measured by Cohen's Kappa, reached 0.897, indicating almost perfect consistency.

5.4 LLM-as-Judge Evaluation Methodology

To evaluate the feasibility of automatic assessment, we apply the same evaluation tasks to state-of-the-art LLMs acting as judges. Each LLM receives the same prompt-response pairs as human annotators—presented as transcripts and/or linked audio when supported—and performs either pairwise or absolute scoring.

In Arena-style evaluation, the LLM selects the better of two responses or chooses a tie. In absolute scoring, it is asked to rate the enhanced response relative to a neutral baseline. Carefully designed prompts ensure rubric adherence and consistency. Examples of prompt templates are provided in Appendix B.3.

We evaluate multiple LLMs and analyze their consistency with human judgments and internal agreement. This allows us to assess their reliability as scalable evaluators for future speech-based dialogue systems. To this end, we conducted over 10,000 LLM-as-judge evaluation iterations, covering both Arena-style and absolute protocols. This large-scale evaluation enables analysis of ranking stability, correlation with human judgments, and cross-model differentiation.

5.5 Results

We present empirical results based on both human and LLM-based evaluations using the MTalk-Bench protocol.

5.5.1 Overall Performance Ranking

Table 4 reports aspect-specific Elo ratings of all evaluated S2S models based on human judgments from MTalk-Arena. Using 1000 as the baseline, `GPT-4o-realtime` achieves the highest overall score (1022.67), leading across all three dimensions: semantic information (1042.52), paralinguistic cues (1011.35), and ambient sound handling (1014.13). `Qwen-Omni` and `GLM4-Voice` follow, while `Westlake-Omni` consistently ranks lowest. The largest performance gap is observed in semantic understanding, where `GPT-4o` shows a substantial lead.

To evaluate alignment between human and automatic assessments, Table 5 presents Elo scores from human annotators and five LLM-based judges. `GPT-4o-realtime` ranks highest across all LLM judges, with notable margins under `Claude-3-7-Sonnet` (1211.55), `DeepSeek-R1` (1234.34), and `DeepSeek-V3` (1228.52). Correlation between human and LLM scores is strong for top models like `GPT-4o` and `Qwen-Omni`, while lower-ranked models such as `Westlake-Omni` show greater variance, suggesting reduced agreement or higher prompt sensitivity.

5.5.2 Comparative Analysis of Judging Modalities

To assess the effectiveness and consistency of LLMs as evaluators for S2S model outputs, we compared their judgment patterns to those of human annotators, with a particular focus on potential systematic biases. While several LLM judges (e.g., Claude-3-7-Sonnet and GPT-4o-latest) showed moderate-to-strong alignment with human

S2S Model	Human	GPT-4o-latest	Gemini-2.5-Flash	Claude-3-7-Sonnet*	DeepSeek-R1	DeepSeek-V3
GPT-4o-realtime	**1042.52**	**1163.15**	**1110.35**	**1211.55**	**1234.34**	**1228.52**
GLM4-Voice	1001.07	1074.90	1060.75	1102.06	1111.46	1114.24
Qwen-Omni	1019.71	1028.71	1007.01	1020.71	1008.19	991.16
LLaMA-Omni	980.90	929.85	962.47	899.12	884.16	894.31
Westlake-Omni	955.80	803.39	859.42	766.57	761.85	771.78

Table 5: ELO Ratings of S2S LLMs Assessed by Various LLM Judges and Human Evaluators
* This model version is `claude-3-7-sonnet-20250229`.

S2S Models	Paralinguistic Information			Ambient Sound		
	Human	GPT-4o-realtime	Gemini-2.5-pro	Human	GPT-4o-realtime	Gemini-2.5-pro
GPT-4o-realtime	4.64±1.75	6.10±1.44	**6.22**±2.11	4.00±1.32	**5.93**±2.20	**6.21**±2.24
GLM4-Voice	4.86±1.46	**6.23**±1.44	6.22±2.49	4.25±2.05	5.70±2.35	5.24±2.43
Qwen-Omni	**6.00**±1.79	5.71±1.16	6.04±1.95	4.33±1.00	5.51±2.10	5.44±2.26
LLaMA-Omni	4.54±1.05	5.84±1.25	6.06±2.07	**4.64**±1.69	5.10±2.20	4.97±2.37
Westlake-Omni	2.17±2.48	3.11±2.13	3.33±2.26	1.11±0.93	2.30±1.80	2.59±1.77

Table 6: S2S Model Evaluation under Paralinguistic Information and Ambient Sound
Note: The table compares how different S2S models perform under varying paralinguistic cues and ambient sound conditions, as judged by humans and model evaluators.

rankings (detailed in Table 5), we observed measurable biases across all automated evaluators.

As shown in Table 7, certain LLMs exhibit statistically significant *position bias*—a preference for responses placed in a specific order (top or bottom)—as well as *length bias*, favoring longer responses regardless of content quality. For instance, Claude-3-7-Sonnet demonstrates a +7.8% position bias and an 8.6% length bias ($p < 0.001$), while Gemini-2.5-Flash exhibits the strongest position bias (+9.2%). In contrast, GPT-4o-latest and DeepSeek variants show milder or even negative positional biases, though still exhibit notable length preferences (e.g., GPT-4o-latest: 15.8%).

6 Discussion

6.1 Contextual Coherence and Memory in Multi-Turn Dialogues

Table 8 reports the inconsistency rates for each model across one to three turns. The decline in multi-turn coherence reveals a deeper limitation in current S2S LLMs: the absence of persistent dialogue state modeling. Models that degrade sharply across turns likely rely on local context alone, lacking mechanisms for tracking discourse-level goals or constraints.

This is not merely a context length issue, but a structural one. Most architectures treat dialogue as flat token sequences, ignoring the hierarchical and dynamic nature of conversational state. Failures such as semantic drift and contradictory proposals point to a missing inductive bias toward dialogue process modeling.

By contrast, GPT-4o's stability suggests more effective abstraction over dialogue history—possibly through structured memory, turn-level representations, or training regimes that emphasize temporal coherence.

Progress in this area will depend on architectures that reason over dialogues as evolving processes, not static sequences—integrating memory, role alternation, and constraint propagation into the generation loop.

6.2 LLM Judge Biases

While all LLM judges demonstrate statistically significant biases, their bias directions and magnitudes are notably inconsistent, suggesting that such behaviors are not merely side effects, but reflect deeper differences in model alignment and decoding preferences.

Positional Bias Interestingly, GPT-4o and DeepSeek models favor first-presented responses (negative ΔPosition Bias), while Gemini and Claude exhibit the opposite. This polarity implies that bias arises not from prompt format alone, but from learned heuristics—e.g., some models may associate earlier responses with higher salience, while others may defer preference to later content. Such divergence challenges the assumption of "model-agnostic" evaluation setups and raises concerns about reproducibility across judges.

Length Bia. All models prefer longer outputs, but to varying extents. DeepSeek-V3 and GPT-4o show the strongest bias (ΔLength Bias >15%),

	GPT-4o-latest	Gemini-2.5-Flash	Claude-3-7-Sonnet	DeepSeek-R1	DeepSeek-V3
TPR (%)	47.5 [46.2–48.9]	54.6 [53.3–55.9]	53.9 [53.0–54.9]	**49.0 [48.0–49.9]**	48.9 [48.0–49.9]
BPR (%)	52.5 [51.1–53.8]	45.4 [44.1–46.7]	46.1 [45.1–47.0]	**51.0 [50.0–52.0]**	51.1 [50.1–52.0]
ΔPosition Bias	-4.9^{***}	9.2^{***}	7.8^{***}	$\mathbf{-2.1}^{**}$	-2.2^{**}
LPR (%)	57.9 [56.5–59.3]	54.5 [53.2–55.8]	**54.3 [53.4–55.3]**	58.5 [57.5–59.4]	58.7 [57.7–59.6]
SPR (%)	42.1 [40.7–43.5]	45.5 [44.2–46.8]	**45.7 [44.7–46.6]**	41.5 [40.6–42.5]	41.3 [40.4–42.3]
ΔLength Bias	15.8^{***}	9.0^{***}	$\mathbf{8.6}^{***}$	16.9^{***}	17.3^{***}

Table 7: Bias Analysis with Statistical Significance Judged by Different LLM Evaluators
Note: Values in brackets are 95% confidence intervals. ΔBias = difference between top/bottom or long/short preference rates. The detail of computational formula is shown in Appendix C.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (from Permutation Test).

Model	1-turn	2-turn	3-turn
GPT-4o-realtime	2.1	4.3	6.8
Qwen2.5-Omni	4.5	6.2	9.7
GLM4-Voice	4.8	7.9	12.4
LLaMA-Omni	5.1	9.8	14.6
Westlake-Omni	8.3	11.7	18.9

Table 8: Percentage (%) of contextually inconsistent responses as a function of dialogue length (lower is better).

despite often penalizing verbosity in instruction-tuning. This suggests an internal conflict between rewarding surface completeness (length = informativeness) and semantic compactness, where models struggle to reconcile fluency with conciseness. Claude is relatively less length-biased, perhaps due to different training signals emphasizing restraint or minimalism.

Implications The presence of biases in LLM judges, largely absent in human judgment, underscores their current inability to fully emulate fine-grained human evaluation criteria. This necessitates bias-aware calibration (e.g., position randomization, length normalization), cross-model ensembling to mitigate individual model distortions, and a critical reconsideration of LLM-only benchmarks for nuanced evaluation tasks. Ultimately, employing LLMs to evaluate other LLMs introduces inherent modeling assumptions that, if left unexamined, risk obscuring or distorting the true assessment of system quality.

Model	1-turn	2-turn	3-turn
GPT-4o-realtime	92.1	89.6	85.2
Qwen2.5-Omni	88.7	85.3	80.9
GLM4-Voice	86.3	82.4	77.0
LLaMA-Omni	83.5	79.0	73.2
Westlake-Omni	80.6	75.1	68.9

Table 9: Effective content ratio (%) across increasing dialogue length (higher is better).

6.3 Information Utility and Conversational Efficiency

Table 9 reports the average effective content ratio across one to three dialogue turns, annotated by two experts. The ratio declines over turns, revealing a core inefficiency in current S2S generation: models prioritize surface fluency over informational utility. This leads to hedging, paraphrasing, and filler—especially in later turns—resulting in verbosity without added meaning.

Such behavior reflects alignment with instruction-tuned objectives that favor politeness and completeness over conciseness. Yet human raters consistently preferred shorter, information-dense responses, particularly in task-oriented settings.

The root issue is both architectural and objective-level: models lack an inductive bias toward *conversational efficiency*. Unlike humans, they do not engage in cost-sensitive utterance planning.

Improving efficiency calls for rethinking training objectives—rewarding utility, penalizing redundancy, and modeling dialogue as a bandwidth-constrained decision problem rather than a maximalist generation task.

7 Conclusion

We present MTalk-Bench, a comprehensive benchmark for evaluating multi-turn speech-to-speech large language models across semantic, paralinguistic, and environmental dimensions. Through large-scale human and LLM-based evaluations, we reveal current model strengths—particularly in short-turn semantic understanding—as well as critical gaps in contextual coherence, prosodic expressiveness, and conversational efficiency. Our results highlight the need for future models to go beyond correctness toward more concise, context-aware, and expressive spoken interaction.

Limitations

While MTalk-Bench covers a diverse range of scenarios and dimensions, it currently focuses on English dialogue and assumes clean user input. Real-world deployment may involve more diverse languages, accents, and overlapping speech, which remain underexplored. In addition, although we evaluate with multiple LLM judges, their alignment with human perception is not perfect, and further calibration may be needed for scalable automated evaluation.

References

A. Author and Others. 2023. The landscape of speech-to-speech large language models: A survey. *Journal of Advanced AI Research*, X(Y):1–25.

B. Author and Others. 2022. Evaluating single-turn spoken language understanding: Challenges and new directions. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1001–1005, Incheon, Korea.

C. Author and Others. 2021. The role of prosody and emotion in spoken dialogue systems. *Computer Speech & Language*, 65:101123.

Zalán Borsos, Neil Zeghidour, Adam Polyak, Chris Piper, David He, Kun Lee, Raphaël Rihani, Yossi Adi, Alexandre Défossez, Emmanuel Dupoux, and Others. 2023. AudioLM: a language modeling approach to audio generation. *IEEE Transactions on Audio, Speech, and Language Processing*, 31:1998–2011. Originally arXiv:2209.03143 [cs.SD].

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

Albert S. Bregman. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2020. Multiwoz – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *Preprint*, arXiv:1810.00278.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. 2025. Audio large language models can be descriptive speech quality evaluators. *Preprint*, arXiv:2501.17202.

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

Yulong Chen, Kang Liu, Diff Shen, Zhaochun Zhang, and Dong Yu. 2021. DialogSum: A Real-Life Scenario Multi-Domain Dialogue Summarization Dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1018–1031, Online. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Herbert Aron David. 1963. *The Method of Paired Comparisons*. Griffin, London.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Linyi Gao, Ameet Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023a. Benchmarking large language models: A survey on tasks, datasets, and evaluation methods. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5057–5096. Association for Computational Linguistics.

Xiang Gao, Yoon Kim Kim, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2023b. Human-in-the-loop large language model personalization for dialogue systems. *arXiv preprint arXiv:2305.16683*.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor Books, Garden City, NY.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

John J. Gumperz. 1982. *Discourse Strategies*. Cambridge University Press, Cambridge.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *Preprint*, arXiv:1904.06037.

Feng Jiang, Zhiyu Lin, Fan Bu, Yuhao Du, Benyou Wang, and Haizhou Li. 2025. S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information. *Preprint*, arXiv:2503.05085.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall, Upper Saddle River, NJ.

Patrik N. Juslin and Klaus R. Scherer. 2003. Vocal expression of affect. In Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors, *Handbook of Affective Sciences*, pages 160–190. Oxford University Press, New York.

Hans Kamp. 1981. A theory of truth and semantic representation. *Formal Methods in the Study of Language*, pages 277–322.

Mike Kuniavsky. 2002. *Observing the User Experience: A Practitioner's Guide to User Research*. Morgan Kaufmann, San Francisco.

Peter Ladefoged and Ian Maddieson. 1996. *The Sounds of the World's Languages*. Blackwell Publishers, Oxford.

Willem J. M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge.

Jinze Li, Zhaowen Lin, Keming Lu, Yangyi Lin, Hongxin Wei, Wei Li, Changyuan Jiang, Yang Zhou, Wei Wang, Ruobin Xie, Min GU, An Zhang, Wenhao Chai, Wenbo Wang, Zhipeng Chen, Haodong Zhao, Jingren Zhou, Sinan Tan, Shijie Geng, and 66 others. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Vinay Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Bokan Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher Fogarty, Hattie Wang, Jenny Pan, Kathy Li, and 19 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Jakob Nielsen. 1993. *Usability Engineering*. Academic Press, Boston.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. Introducing our next-generation audio models. https://openai.com/index/introducing-our-next-generation-audio-models/. Accessed: 2025-05-20.

Juan C. Perez, Jose Villalba, and Najim Dehak. 2022. Paralinguistic-aware models for multi-modal dialogue act classification. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 634–641. IEEE.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):4381.

Koichiro Purohit, Seokhwan Kim, Neel Parde, Abhinav Kalia, Dominique Estival, Ji Hedley Kim, Dinesh Sridhar, Yunsub Choe, Yang Liu, Hong-Gee Kim, and Chul Lee. 2021. DSTC9 Track 1: Beyond Domain APIs: Task-oriented Common Dialogue. In *Proceedings of the 9th Dialog System Technology Challenge*, Online.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Anton Razzhigaev, Maxim Kurkin, Elizaveta Goncharova, Irina Abdullaeva, Anastasia Lysenko, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. 2024. OmniDialog: A multimodal benchmark for generalization across text, visual, and audio modalities. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 183–195, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Dong Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 236–255, Online. Association for Computational Linguistics.

Elizabeth Salesky, Ramon Sanabria, Spandana Singh, Alan W Black, and Graham Neubig. 2021. Robust WPM: A whitespace-punctuated metric for evaluation of ASR robustness. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6738–6742. IEEE.

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis, Volume 1*. Cambridge University Press, Cambridge.

Klaus R. Scherer. 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165.

Hannes Schroeter, Timo Rosenkranz, Arne Zeyer, and Hermann Ney. 2023. Audioaug: A framework for audio data augmentation for robust speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Björn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, Chichester, West Sussex, UK.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3776–3783. AAAI Press.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *Preprint*, arXiv:2501.17399.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 2025. 3mdbench: Medical multimodal multi-agent dialogue benchmark. *Preprint*, arXiv:2504.13861.

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

David Traum. 2008. Representing and evaluating multi-party dialogue. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 131–138. ACM.

Tuomas Virtanen, Annamaria Mesaros, Toni Heittola, Konstantinos Drossos, and Travis E. Oliphant. 2018. *Computational Analysis of Sound Scenes and Events*. Springer, Cham.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 3098–3110. Curran Associates, Inc.

Jianing Wang, Yuyang Sun, Hang He, Philip S. Yu, and Chen Shi. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

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. *Preprint*, arXiv:2412.02612.

Dong Zhang, Shimin Yan, Wei Lin, Siwei Liu, Yuxiang Chen, and Long Li. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. `https://arxiv.org/abs/2305.11000`. ArXiv:2305.11000.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 survey methodology

The design of MTalk-Bench, specifically the selection of its constituent communication scenarios and core evaluation dimensions, was guided by a rigorous, data-driven survey methodology. This approach was adopted to ensure that the benchmark reflects real-world interaction patterns and prioritizes capabilities deemed most crucial by potential users, thereby enhancing its ecological validity and relevance. Our methodology involved two distinct arena-style surveys, one for scenario selection and another for identifying key evaluative dimensions.

A.1.1 Scenario Selection

To identify high-frequency human communication scenarios suitable for benchmarking Speech-to-Speech Large Language Models (S2S-LLMs), we initiated our process with an extensive review of interdisciplinary literature, drawing insights from fields such as linguistics, sociology, communication studies, and human-computer interaction (HCI) (Gumperz, 1982; Schegloff, 2007; Clark, 1996). This review yielded a comprehensive list of potential real-life communication contexts.

Subsequently, we employed a pairwise arena-style survey methodology to refine this list and prioritize scenarios based on their perceived likelihood of involving interactions with a speech-based agent. In this survey, participants were presented with two randomly selected candidate scenarios at a time and were asked to choose which one they believed was more likely to involve interaction with an AI speech agent in the near future. This pairwise preference elicitation technique is known for its robustness in capturing relative importance or likelihood, mitigating biases often found in direct rating scales (Bradley and Terry, 1952; David, 1963).

The collected pairwise preference data was then used to construct a directed preference graph, where each node represented a scenario and a directed edge from scenario A to scenario B indicated that A was preferred over B. We then applied transitive logic (e.g., if $A > B$ and $B > C$, then $A > C$) to resolve any inconsistencies and to derive a globally ranked chain of scenario categories. This process effectively transformed the pairwise comparisons into a linear ordering, reflecting the collective judgment of the survey participants. The full preference chain derived from this process is as follows:

Family and Domestic Communication > Doctor–Patient Communication > Legal Inquiry and Institutional Interrogation > Financial Investment and Advisory Communication > Classroom Instruction and Interaction > Academic Communication and Research Collaboration > Job Interview Communication > Workplace Collaboration and Communication > Marketing and Customer Relationship Management Communication > Informal Social Interaction (Casual Conversation) > Psychological and Supportive Communication > Service-Oriented Interaction > Public Discourse and Interaction (Political Debates, Media Interviews, Public Speeches) > Entertainment and Content Creation Communication > Negotiation and Conflict Resolution > Intercultural and Linguistic Communication > Public Affairs and Emergency Response Communication > Religious and Spiritual Communication > Environmental Advocacy and Policy Promotion > Sports and Competitive Communication.

Table 10 presents the raw comparison and selection counts for the evaluated scenarios, illustrating the distribution of preferences.

While the detailed preference chain and survey counts provide granular insights into user perceptions, a direct selection of the top N items might lead to overlapping or overly specific categories for a benchmark. Therefore, to establish a set of distinct yet comprehensive scenarios for MTalk-Bench, we performed a rational consolidation of the high-ranking and thematically related scenarios identified in the survey. This summarization process aimed to abstract broader communicative functions that are highly relevant for S2S-LLM interactions, guided by the overall preference trends and semantic similarities within the survey data. The objective was to ensure that the final scenarios represent diverse, impactful, and frequently encountered communication contexts.

The following nine consolidated high-frequency scenarios were ultimately selected for inclusion in MTalk-Bench:

1. **Family and Domestic Communication** (e.g., coordinating household tasks, family scheduling, managing smart home devices via voice)

2. **Healthcare and Medical Communication** (e.g., initial symptom checking, virtual health assistant consultations, medication reminders, accessing medical information)

Scene	ELO Score
Family and Domestic Communication	1091.9430
Doctor–Patient Communication	1070.0580
Legal Inquiry and Institutional Interrogation	1048.5192
Financial Investment and Advisory Communication	1045.8893
Classroom Instruction and Interaction	1029.5885
Academic Communication and Research Collaboration	1027.4342
Job Interview Communication	1027.3857
Workplace Collaboration and Communication	1017.0773
Marketing and Customer Relationship Management Communication	1008.9531
Informal Social Interaction (Casual Conversation)	1007.5547
Psychological and Supportive Communication	1003.3751
Service-Oriented Interaction	1001.8327
Public Discourse and Interaction (Political Debates, Media Interviews, Public Speeches)	995.7294
Entertainment and Content Creation Communication	980.4584
Negotiation and Conflict Resolution	979.0532
Intercultural and Linguistic Communication	974.5308
Public Affairs and Emergency Response Communication	960.6946
Religious and Spiritual Communication	950.3833
Environmental Advocacy and Policy Promotion	906.4290
Sports and Competitive Communication	873.1106

Table 10: Scenario Comparison and Selection Counts from Pairwise Arena Survey

3. **Institutional Inquiry and Information Exchange** (e.g., querying government services, basic legal information retrieval, financial account inquiries, university helpdesks)

4. **Educational and Instructional Communication** (e.g., AI-powered tutoring, language learning applications, interactive educational Q&A, voice-guided tutorials)

5. **Workplace and Professional Communication** (e.g., meeting dictation and summarization, collaborative task management via voice, professional information lookup, job interview practice)

6. **Entertainment and Creative Content Communication** (e.g., interacting with voice-controlled games, generating stories or scripts via speech, controlling media playback, interactive audio experiences)

7. **Casual and Socio-Emotional Interaction** (e.g., open-domain social chat, companionship with AI, storytelling, expressing feelings and receiving empathetic responses)

8. **Psychological and Supportive Communication** (e.g., AI coaches for well-being, guided mindfulness exercises, initial mental health support and resource navigation)

9. **Service-Oriented Communication** (e.g., customer service inquiries, booking appointments, technical support, retail assistance, travel planning)

This data-driven approach, combining fine-grained preference elicitation with principled consolidation, ensures that MTalk-Bench focuses on scenarios that are both user-validated and pragmatically structured for comprehensive benchmark evaluation. The long tail of more specific or lower-ranked scenarios from the original survey (e.g., "Religious and Spiritual Communication", "Environmental Advocacy and Policy Promotion", "Sports and Competitive Communication" as seen in the full chain and Table 10) provides context but was not prioritized for direct inclusion in the current benchmark version, allowing for a focused yet robust evaluation scope.

A.1.2 Capability Dimension Selection

The identification of core evaluation dimensions for S2S-LLMs was similarly grounded in both existing literature and empirical user feedback. We began by aggregating a comprehensive list of communication capability demands identified in sociology (Goffman, 1967), linguistics (particularly pragmatics and discourse analysis) (Levinson, 1983; Grice, 1975), HCI (Nielsen, 1993), and contemporary research on speech and language models (Wang et al., 2023; Jurafsky and Martin, 2000). These demands were then translated into a set of candidate evaluative dimensions specifically tailored for benchmarking S2S-LLM capabilities.

A second arena-style survey was conducted to prioritize these dimensions. In this survey, participants were presented with two randomly selected capabilities and were asked to indicate which one they considered more important for a speech agent to possess for effective and satisfactory interaction. This approach allowed us to construct a "capability importance chain" through pairwise comparisons, mirroring the methodology used for scenario selec-

tion. The full capability importance chain derived from this survey is as follows:

Long-Term Conversational Memory > Long-Term Instruction Following > Semantic Disambiguation > Logical and Commonsense Reasoning > Task Understanding and Planning > Error Correction and Adaptability > Dialogue Management > Ambient Sound Perception and Adaptation > Environmental Audio Information Acquisition > Speaker Information Processing > Multi-Party Interaction Tracking > Emotion Recognition and Perception > Pragmatic and Cultural Intelligence > Stylistic and Personality Expression Control > Cross-Topic Switching Capability > Self-Monitoring and Reflective Capability > Contextual Adaptation Capability > Continual Learning and Self-Optimization.

Table 11 provides the detailed comparison and selection counts for these capabilities from the survey.

Informed by these user-centered priorities identified through the survey, and further drawing upon established principles from linguistics regarding the tiered nature of language (Chomsky, 1965; Levelt, 1989), semantics and pragmatics which govern meaning and use (Grice, 1975; Kamp, 1981; Jurafsky and Martin, 2000), and speech science which distinguishes linguistic content from vocal expression and acoustic context (Ladefoged and Maddieson, 1996; Scherer, 1986), we have structured the capability dimensions for MTalk-Bench into a comprehensive three-tiered framework. This hierarchical organization, which separates core linguistic message processing (Tier 1) from the interpretation and generation of paralinguistic cues (Tier 2) (Ekman, 1992; Juslin and Scherer, 2003) and adaptation to the broader acoustic environment (Tier 3) (Bregman, 1990; Virtanen et al., 2018), facilitates a systematic and multi-faceted evaluation of S2S-LLMs. The detailed structure of these evaluation dimensions is as follows:

Tier 1: Semantic Information Processing This tier focuses on the model's ability to understand, reason about, and generate meaningful content based on textual and semantic aspects of the dialogue. It aligns with traditional NLP tasks focusing on lexical, syntactic, semantic, and discourse-level understanding (Mann and Thompson, 1988; Grosz and Sidner, 1986).

1. **Core Comprehension & Memory**
   (a) Context Comprehension:
      i. Contextual Memory: Ability to remember information from earlier in the dialogue to support subsequent interaction.
      ii. Long-Term Instruction Following: Ability to adhere to complex or multi-step instructions provided by the user across multiple turns.
   (b) Key Information Extraction & Retention:
      i. User Implicit Information Inference: Ability to understand and retain information not explicitly stated but implied by the user (e.g., preferences, background).
      ii. Self-Consistency: Ensuring the model's own utterances are coherent and non-contradictory over time.
   (c) Accurate Recall: Ability to accurately restate or reference prior statements made by the user or itself, as required.
   (d) Semantic Disambiguation:
      i. Reference Resolution: Correctly resolving pronouns (e.g., "it," "they") and deictic expressions (e.g., "this," "there") to their referents.
      ii. Ellipsis and Completion: Understanding and appropriately handling omitted sentence components, performing contextual completion.
      iii. Vague Expression Handling: Effectively processing ambiguous or vague user inputs, potentially through clarification or by responding based on the most probable interpretation.
   (e) Content Generation & Transformation (Text-focused):
      i. Content Paraphrasing: Rephrasing content in different ways according to user requirements.
      ii. Format Adjustment: Converting content into specified formats as instructed.
      iii. Multi-Turn Content Editing: Reliably performing iterative modifications and refinements to content based on previous versions.

2. **Reasoning & Task Execution**
   (a) Task Comprehension & Planning:

Capability	ELO Score
Long-Term Conversational Memory	1028.9649846053856
Long-Term Term Instruction Following	1027.0182845602856
Semantic Disambiguation	1020.7812135830862
Logical and Commonsense Reasoning	1019.3565012861153
Task Understanding and Planning	1018.5992220320628
Error Correction and Adaptability	1018.4190629708203
Dialogue Management	1001.0609758121847
Ambient Sound Perception and Adaptation	998.5492170251707
Environmental Audio Information Acquisition	998.1834604838184
Speaker Information Processing	993.826840977241
Multi-Party Interaction Tracking	989.5048439436567
Emotion Recognition and Perception	989.0079697688203
Pragmatic and Cultural Intelligence	987.7369928230721
Stylistic and Personality Expression Control	986.8073479398374
Cross-Topic Switching Capability	986.6067503503083
Self-Monitoring and Reflective Capability	984.7468072473208
Contextual Adaptation Capability	984.6825824774363
Continual Learning and Self-Optimization	966.1469421133773

Table 11: Capability Comparison and Selection Counts from Pairwise Arena Survey

i. Task-Instruction Distinction: Accurately differentiating between task descriptions, examples, and the actual input to be processed.
ii. Complex Task Decomposition: Understanding tasks that involve multiple steps or conditions.

(b) Logical & Commonsense Reasoning:
i. General Reasoning: Ability to handle reasoning tasks based on logic (e.g., logical puzzles), commonsense, or general knowledge.

(c) Complex Reasoning: Addressing more intricate inferential challenges requiring multi-step deduction or integration of diverse information.

3. **Interaction Strategy & Intelligence**

(a) Dialogue Management:
i. Handling Overlap and Interruptions: Responding appropriately to speech overlap and user interruptions.
ii. Proactive Interaction: Taking initiative to guide the dialogue, ask questions, or advance the process at appropriate junctures to maintain a natural and fluid exchange.
iii. Topic Control: Naturally managing topic shifts or maintaining conversational focus as needed.
iv. Clarification and Confirmation: Actively seeking clarification when user instructions are unclear or its own understanding is uncertain.

(b) Error Handling & Adaptability:
i. Self-Correction: Understanding and correcting previous responses when errors are pointed out by the user.
ii. Response Stability and Self-Affirmation: Maintaining stable and appropriate responses when faced with user skepticism, contradiction, or uncooperative behavior (including adhering to correct information or conceding errors when appropriate).

4. **Security Assessment**: Evaluating security vulnerabilities in the context of multi-turn dia-

logues.

5. **Multilingual and Code-Switching Capabilities**: Assessing performance in multilingual contexts and handling transitions between languages (code-switching).

6. **Pragmatic and Cultural Competence**
   (a) Non-literal Comprehension: Understanding non-literal expressions such as puns, irony, humor, metaphors, and implicatures.
   (b) Pragmatic Interaction Handling: Understanding and naturally participating in specialized interactional forms (e.g., rap battles, fast-paced dialogues, scripted performances, language games).
   (c) Cultural Adaptation: Recognizing and adapting to communication etiquette, value preferences, and linguistic styles across different cultural backgrounds.

Tier 2: Paralinguistic Information Processing This tier assesses the model's capabilities related to the interpretation and generation of non-lexical vocal cues that convey emotional, attitudinal, and pragmatic meaning. Such cues are known to significantly modulate the interpretation of spoken language (Scherer, 1986; Juslin and Scherer, 2003; Schuller and Batliner, 2013).

1. **Emotion Recognition & Regulation**: Correctly identifying vocal emotions (e.g., happiness, anger, fatigue, anxiety) and determining contextually appropriate responses, if any.

2. **Emotional Speech Synthesis & Control**: Simulating and synthesizing speech with various emotions, including control over the type and intensity of the expressed emotion.

3. **Paralinguistic Feature Recognition**: Interpreting prosodic elements such as speech rate, stress, intonation, and pauses as cues for emotional states or pragmatic intent.

4. **Paralinguistic Feature Generation**: Producing speech with appropriate prosody, stress, and rhythm that aligns with semantic content and conversational context.

5. **Personalized Expressive Modeling**: Imitating the distinctive expressive styles of specific speakers (e.g., humorous, formal, measured, animated).

Tier 3: Environmental Sound Processing This tier evaluates the model's ability to perceive, interpret, and adapt to the broader acoustic environment in which the interaction takes place. Effective human communication, and by extension human-machine communication, often depends on contextualizing speech within its acoustic scene (Bregman, 1990; Virtanen et al., 2018).

1. **Environmental Sound Perception & Adaptation**: Identifying and adapting to variations in background noise, reverberation, and far-field acoustic conditions.

2. **Multi-Party Interaction Tracking**: Following speaker turns, topic flow, and interaction dynamics in conversations involving multiple participants, including appropriate timing for speech initiation/switching.

3. **Handling Special Speech Styles**: Adapting to and processing unconventional or performative speech styles (e.g., highly stylized deliveries, participation in scripted interactions).

The pairwise survey (see Table 11 and the importance chain) highlighted the significance of dimensions such as Long-Term Conversational Memory, Long-Term Instruction Following, Semantic Disambiguation, Logical Reasoning, Error Correction, and Dialogue Management. These empirically prioritized areas are comprehensively embedded within the detailed sub-dimensions of the Semantic Information Processing tier. Similarly, the importance of acoustic awareness (e.g., Ambient Sound Perception) and emotional understanding (Emotion Recognition) from the survey are directly addressed in Tiers 2 and 3. This structured framework, rooted in both empirical user preferences and established scientific distinctions in the study of language and speech, thus ensures a thorough evaluation of S2S-LLMs.

A.1.3 Methodological Soundness and Rationale

The design of MTalk-Bench, encompassing both the selection of communication scenarios and the definition of evaluative capability dimensions, was underpinned by a commitment to methodological soundness, user-centered principles, and data-driven decision-making. This approach ensures the benchmark's relevance, comprehensiveness, and robustness for evaluating Speech-to-Speech Large Language Models (S2S-LLMs).

For the **selection of communication scenarios**, the process commenced with an extensive review of interdisciplinary literature to identify a broad spectrum of real-world communication contexts. This foundational work was followed by a pairwise arena-style survey, a technique recognized for its efficacy in eliciting robust user preferences and mitigating biases inherent in direct rating scales (Bradley and Terry, 1952; David, 1963; Thurstone, 1927). The resultant pairwise preference data was systematically transformed into a directed preference graph, from which a globally ranked chain of scenarios was derived using transitive logic. This empirical ranking directly informed the subsequent rational consolidation process, where high-ranking and thematically related scenarios were grouped to form a diverse yet manageable set of nine core scenarios for MTalk-Bench. This two-stage process—fine-grained, data-driven preference elicitation followed by principled, literature-informed consolidation—ensures that the selected scenarios are not only user-validated in terms of perceived relevance for S2S-LLM interaction but also pragmatically structured for comprehensive benchmark coverage.

Similarly, the **definition of capability dimensions** adopted a multi-faceted approach. It began with an aggregation of communication capability demands identified from established research in sociology, linguistics, HCI, and current S2S-LLM studies. A second arena-style survey was then deployed to empirically determine the relative importance of these candidate capabilities from a user perspective, yielding a data-backed capability importance chain. Crucially, these empirically identified user priorities were then integrated with established theoretical frameworks from linguistics (Chomsky, 1965; Levelt, 1989), semantics and pragmatics (Grice, 1975; Kamp, 1981; Jurafsky and Martin, 2000), and speech science (Ladefoged and Maddieson, 1996; Scherer, 1986) to construct a comprehensive three-tiered evaluation framework (Semantic Information Processing, Paralinguistic Information Processing, and Environmental Sound Processing). This hierarchical structure, with detailed sub-dimensions grounded in recognized scientific distinctions (e.g., (Mann and Thompson, 1988; Grosz and Sidner, 1986; Ekman, 1992; Juslin and Scherer, 2003; Bregman, 1990; Virtanen et al., 2018)), provides a systematic and nuanced means of assessing S2S-LLM performance.

Overall, the methodological framework employed for MTalk-Bench emphasizes transparency and reproducibility. By grounding the design choices in both quantitative user preference data and established academic theory, we have strived to move beyond arbitrary heuristics. This ensures that MTalk-Bench is not only robust and comprehensive but also possesses strong ecological validity, reflecting both real-world interaction patterns and scientifically pertinent evaluative criteria. The resulting benchmark is therefore well-positioned to drive meaningful advancements in the development of S2S-LLMs that are effective, versatile, and aligned with human communicative needs (Kuniavsky, 2002).

A.2 Dialogue Instance Generation Details

This appendix provides a detailed technical explanation of the methodologies employed for generating the dialogue instances that constitute the MTalk-Bench datasets. These datasets are designed to evaluate multimodal dialogue systems across semantic-level, paralinguistic-level, and environment-aware dimensions.

A.2.1 Focus-Semantics Dialogue Dataset Construction

The construction of the Focus-Semantics Dialogue Dataset involved a synergistic approach combining scripted generation with LLMs and subsequent human refinement to ensure data quality, relevance, and balanced coverage of evaluation dimensions.

The generation process utilized three distinct scripts:

1. **Script 1 (Dialogue Generation):** This script employed Claude 3.5 Sonnet to generate the initial dialogue instances. The inputs to this script specified the desired dialogue scenario, the primary semantic evaluation dimension to be tested (e.g., contextual understanding, knowledge integration, complex instruction following), and the target number of turns for the conversation. A crucial constraint embedded in the prompt for this script was to ensure that the model's response in the final turn was contingent upon information or context established in earlier turns. This design choice explicitly aims to differentiate the benchmark from single-turn evaluation tasks, emphasizing multi-turn reasoning capabilities.

2. **Script 2 (Dimension Labeling):** Following the generation of dialogue instances, a sec-

ond script, also utilizing Claude 3.5 Sonnet, was applied. This script's function was to label each generated multi-turn dialogue for the presence of each of the six major semantic-level evaluation dimensions. This step provided a comprehensive annotation of the capabilities potentially testable within each dialogue.

3. **Script 3 (Primary Dimension Inference):** The third script, leveraging Gemini 2.0 Flash, was tasked with inferring the primary evaluation dimension for each dialogue. This inference was based on the comprehensive set of labels produced by Script 2. The output of this script was a single primary dimension deemed most prominent or testable within the dialogue.

A critical step in the quality assurance process involved comparing the primary dimension identified by Script 3 with the originally intended primary dimension specified as input to Script 1. Any inconsistencies flagged during this comparison were subjected to manual review by human annotators. These annotators meticulously examined the dialogues to resolve discrepancies and ensure the assigned primary dimension accurately reflected the dialogue's core challenge.

Further human refinement was conducted to enhance the dataset's overall quality. This involved:

- Ensuring the dialogue flowed naturally and coherently from a human perspective.
- Verifying that the designated primary capability was clearly and unambiguously testable within the dialogue structure.
- Guaranteeing that every major semantic dimension had at least 10 valid, high-quality samples. This was to ensure a balanced representation and robust evaluation across all targeted capabilities. This process sometimes involved modifying existing dialogues or, if necessary, generating new ones to meet the quota for underrepresented dimensions.

This iterative process of LLM-based generation and human-in-the-loop refinement is crucial for creating datasets that are both scalable and reliable for benchmarking advanced dialogue systems (Gao et al., 2023b).

Figure 2 shows the complete pipeline for constructing focus-semantics dialogue dataset.

A.2.2 Focus-Paralinguistic Dataset Construction

The Focus-Paralinguistic Dataset was derived from the refined Focus-Semantics Dialogue Dataset through a series of modifications and augmentations. The goal was to create dialogue instances that specifically test a model's ability to understand and generate paralinguistic cues. This dataset comprises two primary subtypes:

1. **Understanding Input-Side Paralinguistic Signals:** To evaluate the model's perception and adaptation to paralinguistic information from the user, user utterances in the base semantic dialogues were augmented with explicit metadata. This metadata included attributes such as speaker traits (e.g., age, gender, accent), vocal tone (e.g., sarcastic, empathetic, urgent), expressed emotion (e.g., joy, anger, sadness), and other relevant vocal characteristics. The model is then expected to interpret these signals and adjust its responses accordingly. This approach aligns with methodologies for creating richer, more context-aware dialogue interactions (Perez et al., 2022).

2. **Generating Expressive Output Speech:** This subtype focuses on the model's ability to produce expressive and paralinguistically rich spoken responses. The original instructions (or system prompts) associated with the base semantic dialogues were modified. These modifications explicitly required the model to generate responses that varied in terms of emotion, tone, speaking style (e.g., formal, casual), or other specified paralinguistic features. The evaluation would then assess the appropriateness and naturalness of the generated expressive speech.

A.2.3 Focus-Environmental Sound Dataset Construction

Similar to the paralinguistic dataset, the Focus-Environmental Sound Dataset was also derived from the base Focus-Semantics Dialogue Dataset. This dataset aims to evaluate the model's robustness and adaptability in the presence of various environmental auditory cues. It includes three distinct subtypes:

1. **Environmental Sound Understanding:** To test the model's adaptive comprehension in

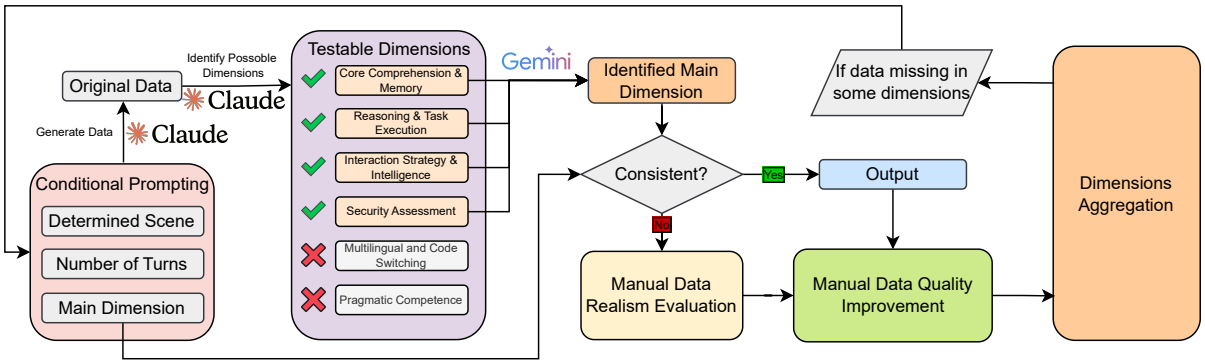

Figure 2: The Benchmark Architecture

noisy or event-rich environments, background sounds were programmatically inserted or described within the dialogue context. Examples include common environmental sounds such as a door knock, a ringing alarm, a public address (PA) announcement, or ambient noise like cafe chatter. The model is expected to recognize or appropriately react to these sounds if they are relevant to the dialogue context. Methodologies for data augmentation with background noise are common in robust speech processing and can be extended to dialogue understanding (Schroeter et al., 2023).

2. **Multi-speaker Interaction Understanding:** This subtype addresses the challenge of tracking and understanding conversations involving multiple participants. The input structure of the base dialogues was modified to clearly delineate utterances from different speakers. This allows for the evaluation of the model's ability to maintain conversational coherence, attribute statements correctly, and manage turn-taking in a multi-party setting. Evaluating dialogue systems in multi-speaker scenarios is crucial for real-world applicability (Traum, 2008).

3. **Stylized/Idiosyncratic Speech Comprehension:** This subtype focuses on the model's capability to understand non-standard speech patterns or stylized language. Dialogues were modified to include special language patterns, such as instances of crosstalk (e.g., simulated comedic duets where speakers might interrupt or speak over each other in a stylized manner) or other forms of overlapping speech and idiosyncratic expressions. This tests the model's robustness to deviations from clear, single-speaker conversational norms.

The systematic construction of these three datasets, with their specific focuses, allows for a comprehensive and granular evaluation of multimodal dialogue systems' capabilities beyond traditional text-based semantic understanding.

B ELO Scores

B.1 MTalk Arena platform

The MTalk Arena planform is shown in Figure 3

B.2 The Detailed Computation Procedure for ELO

To obtain a comparative ranking of S2S models in the S2S-Arena framework, we adopt an Elo rating system, originally developed for chess ranking, to aggregate results from pairwise model comparisons. Below we describe the detailed computation process.

Initialization

Each model is assigned an initial Elo score of 1000. The Elo score will be updated based on the outcomes of pairwise comparisons.

Pairwise Comparison Setup

Let model A and model B be compared on the same evaluation instance. Each pair receives one of the following outcomes based on human or LLM judgment:

- A wins over B: $S_A = 1$, $S_B = 0$
- B wins over A: $S_A = 0$, $S_B = 1$
- Tie: $S_A = 0.5$, $S_B = 0.5$

Score Update Rule

Let R_A and R_B denote the current Elo scores of models A and B, respectively. The expected win probability for A is computed as:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}, \quad E_B = 1 - E_A$$

The Elo scores are then updated using:

$$R'_A = R_A + K(S_A - E_A), \quad R'_B = R_B + K(S_B - E_B)$$

where K is a constant controlling the update rate. We use $K = 32$ in our experiments, following common practice in Elo-based evaluation systems.

Aggregation and Ranking

The final Elo score of each model is computed after all pairwise comparisons are completed across evaluation instances. Models are then ranked in descending order of their final Elo scores.

Stability and Variance

To estimate variance and ensure ranking robustness, we conduct bootstrapping across evaluation instances. Elo scores are recomputed over multiple resampled subsets to derive confidence intervals, as reported in the main results.

B.3 Prompt for LLM-as-Judge

We explore two use cases of large language models (LLMs) as automated judges for evaluating speech-to-speech model outputs. Prompts are designed to elicit fair, fine-grained, and dimension-specific judgments. The prompts are presented in the form of structured instructions and are passed to the LLM alongside transcript content and scoring criteria. Below we provide examples for both scenarios used in our study.

(a) Multi-turn Arena-style Judgment

In this setting, the LLM acts as a judge for pairwise comparisons, following the same turn-by-turn evaluation scheme as human annotators in the MTalk-Arena platform. It is asked to compare two model-generated transcripts based on a specified evaluation dimension (e.g., semantic understanding, paralinguistic cues, or environmental awareness), and to select the better response or indicate a tie. Figure 4 shows the full prompt template used for this task.

(b) Judgment on Enriched Inputs with Paralinguistic or Environmental Features

We further instruct LLMs to evaluate how effectively a model incorporates non-verbal elements—such as emotional tone, prosody, or ambient noise—into its responses. These prompts explicitly highlight the presence of auxiliary cues and ask the LLM to assess expressiveness, realism, and contextual appropriateness. Figure 5 presents the prompt used for this task.

C Statistical Definitions and Inference Methods

We define below the metrics and statistical tests used for analyzing position and length biases in S2S model preferences.

1. Preference Rate (TPR, BPR, LPR, SPR)

For a given preference condition (e.g., top position), we define the preference rate as:

$$\text{Preference Rate} = \frac{n_{\text{preferred}}}{N}$$

where $n_{\text{preferred}}$ is the number of times the preferred category (e.g., top or long) is selected, and N is the total number of evaluation instances.

2. Bias Score (Difference in Preference)

To quantify directional bias, we compute the difference in preference rates between two competing categories:

$$\Delta_{\text{bias}} = p_1 - p_2$$

where p_1 and p_2 are the preference rates for the two categories, such as top vs. bottom (for position bias) or long vs. short (for length bias). A positive Δ_{bias} indicates a bias towards category 1.

3. Confidence Interval (Wilson Score)

The 95% confidence interval for a preference rate $p = \frac{x}{n}$ is calculated using the Wilson Score Interval:

$$\hat{p} = \frac{x + \frac{z^2}{2}}{n + z^2}, \quad z = 1.96$$

$$\text{half-width} = \frac{z \cdot \sqrt{\frac{x(n-x)}{n} + \frac{z^2}{4}}}{n + z^2}$$

$$\text{CI}_{95\%} = \hat{p} \pm \text{half-width}$$

This interval is more accurate than the normal approximation, especially when p is near 0 or 1 or when n is small.

4. Permutation Test for Significance of Bias

To assess whether the observed bias Δ_{obs} is statistically significant, we conduct a non-parametric permutation test:

1. Combine all preference labels (e.g., "top" and "bottom") into a single set of size N.
2. Randomly shuffle the labels and reassign them into two groups of sizes n_1 and n_2.
3. For each permutation $i \in \{1, \ldots, M\}$, compute the permuted bias score:

$$\Delta^{(i)} = \hat{p}_1^{(i)} - \hat{p}_2^{(i)}$$

4. Estimate the two-tailed p-value:

$$p = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left(\left|\Delta^{(i)}\right| \geq |\Delta_{\text{obs}}|\right)$$

where $\mathbb{I}(\cdot)$ is the indicator function, and M is the number of permutations (e.g., 10,000).

If $p < 0.05$, we consider the observed bias statistically significant.

D Evaluation Models

- **GPT-4o-realtime**[6]: A multimodal model developed by OpenAI, supporting real-time speech-to-speech interaction with expressive prosody and perception capabilities.
- **GLM-4-Voice-9B**[7]: An open-source end-to-end S2S model developed by Zhipu AI and Tsinghua University, optimized for bilingual multi-turn speech interaction.
- **Qwen2.5-Omni-7B**[8]: A fully multimodal model from Alibaba Cloud capable of processing and generating audio, text, image, and video, supporting real-time dialogue.
- **LLaMA-Omni2**[9]: A speech-augmented variant of Meta's LLaMA model, extended for conversational audio tasks.
- **Westlake-Omni**[10]: A multimodal conversational model designed for prosodic and emotion-aware speech interaction, developed by Westlake University.

E Annotator Characteristics, AI Usage, and Artifact Information

Annotator Characteristics

All annotation tasks in this study were performed by individuals with the following characteristics:

- **Affiliation:** All annotators were undergraduate students at the time of participation.
- **Language Proficiency:** Each annotator had an IELTS score of 6.5 or above, indicating strong English reading and comprehension skills.
- **Training:** Annotators received standardized instructions and examples prior to the annotation process to ensure consistency and accuracy.

AI Usage Statement

Artificial intelligence tools were used during the writing process of this paper to assist with language refinement and structural organization. However, we affirm the following:

- All research data were collected and processed by human researchers.
- All annotations, analyses, and conclusions were independently produced by the authors.
- The use of AI did not influence the substantive content of the study and served solely as a writing aid.

Ethical Considerations and Artifact Information

Potential Risks. This work does not pose significant foreseeable risks. However, as with any benchmark, there is potential for misuse, such as drawing

[6] Accessed via OpenAI API in May 2025. https://openai.com/index/hello-gpt-4o

[7] Version v1.0, released April 2025. https://github.com/THUDM/GLM-4-Voice

[8] Released May 2025. https://www.alibabacloud.com/blog/alibaba-cloud-releases-qwen2-5-omni-7b-an-end-to-end-multimodal-ai-model_602095

[9] Accessed from ICT NLP Lab GitHub repository, May 2025. https://github.com/ictnlp/LLaMA-Omni

[10] Model details and resources available via Westlake University, accessed May 2025. https://westlake.edu.cn

unfair comparisons or relying excessively on automated metrics without human judgment.

Use or Creation of Scientific Artifacts. We introduce and release scientific artifacts, including a benchmark dataset, evaluation scripts, and analysis tools, to support reproducibility and further research.

License for Artifacts. All artifacts are made available under an open-source license (e.g., CC BY 4.0 or MIT), allowing use, modification, and redistribution with appropriate credit.

Consistency with Intended Use. The released artifacts are intended strictly for research and educational purposes. Commercial use or deployment in high-stakes settings without further validation is not encouraged.

Data Safety and Sensitivity. The dataset does not contain personally identifiable information (PII) or deliberately offensive content. Still, as it includes model-generated dialogue, users should exercise caution and perform content screening as necessary.

Documentation. Comprehensive documentation is provided for all artifacts, covering data schema, usage instructions, and evaluation guidelines to ensure transparency and facilitate adoption by the community.

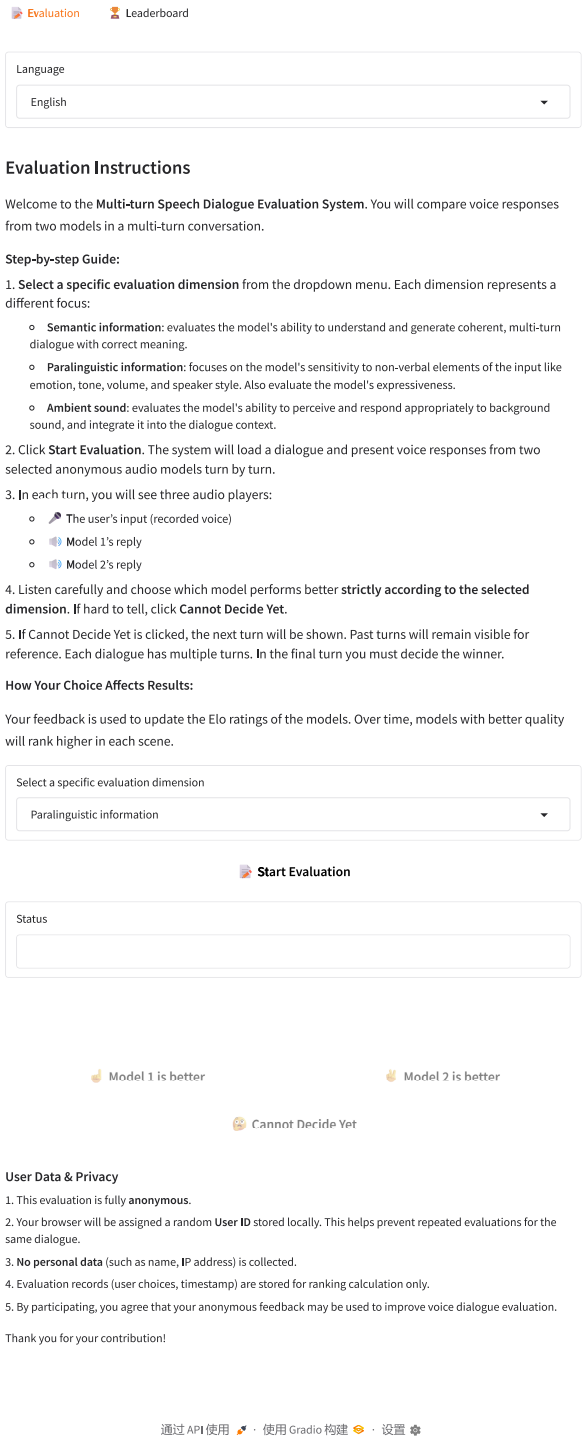

Figure 3: The MTalk-Arena human evaluation interface. Annotators compare model responses across multiple dialogue turns based on a selected evaluation dimension (e.g., semantic, paralinguistic, or environmental). Feedback is collected anonymously and used for Elo ranking updates.

[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question in a multi-turn dialog displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity,
and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. Only answer with "Model 1" or "Model 2".
Respond in JSON format like this:

```json
{
   "answer": (your answer)
}
```

[User Prompt]
Round 1:
User: [User's question]
Model 1: [Model 1's Answer]
Model 2: [Model 2's Answer]
Round 2:
User: [User's question]
Model 1: [Model 1's Answer]
Model 2: [Model 2's Answer]
(Optional) Round 3:
```

Based on the conversation, choose which model performed better (Model 1 or Model 2).

Figure 4: Prompt template used for LLM-based pairwise comparison in Arena-style evaluation. The LLM receives structured dialogue history and evaluates responses based on a selected dimension.
```

You are an expert evaluator assessing a speech-to-speech language model's ability to understand and respond to paralinguistic cues (emotion, tone, speaking rate, volume, speaker style) in multi-turn spoken conversations.

Input Audio Structure:
Each test sample includes two versions of a dialogue between a user and the model, presented as concatenated audio:
1. Base Dialogue (Semantic-only input) – user utterances contain only neutral, text-level semantic information:
- User turn 1 audio
- Model turn 1 audio
- User turn 2 audio
- Model turn 2 audio
- (Optional) User turn 3 audio
- (Optional) Model turn 3 audio

2. Modified Dialogue (Paralinguistic-enhanced input) – user utterances include paralinguistic features like emotion, hesitation, loudness, etc.:
- User turn 1 audio (with paralinguistic cues)
- Model turn 1 audio
- User turn 2 audio (with paralinguistic cues)
- Model turn 2 audio
- (Optional) User turn 3 audio
- (Optional) Model turn 3 audio

Your Task:
1. Listen to both the Base and Modified dialogues (2–3 turns each).
2. Assume the model's response quality for the Base dialogue is fixed at 5 out of 10.
3. Evaluate the Modified dialogue, focusing on whether the model uses paralinguistic input to produce better, worse, or unchanged responses.
4. Assign a score from 0 to 10, and provide a brief justification.

Evaluation Criteria:
Assess the Modified responses using the following dimensions:
1. Semantic Appropriateness
- Is the model still accurately understanding and responding to the user's intent?
2. Paralinguistic Awareness
- Does the model recognize paralinguistic information, including emotional or tonal cues (e.g., anger, sadness, hesitation)?
- Does it adjust its prosody, wording, or strategy appropriately in response?
3. Conversational Strategy
- Does the model improve user experience through empathy, reassurance, or dynamic response strategies?
- Is the response more natural, human-like, or contextually sensitive?
Scoring Guide:
- 8–10: Strong, meaningful improvement using paralinguistic cues
- 6-7: Noticeable enhancement with some adaptive behavior
- 5: No meaningful change from Base response
- 3–4: Weak or inaccurate handling of paralinguistics
- 0–2: Misinterpretation leading to harmful or inappropriate response
Output Format (required):
You must respond in JSON format like this:
```json
{
 "score": X,
  "justification": "Explain your score using all three criteria: semantic understanding, paralinguistic awareness, and conversational strategy."
}
```

Figure 5: Prompt used for evaluating paralinguistic or environmental integration. The LLM is guided to assess expressive features beyond semantic content.