

# Robust Spoken Pragmatics: A Benchmark for Large Speech-Language Models

Anonymous ACL submission

## Abstract

The end-to-end Large Speech-Language Models (LSLMs) have been revolutionizing the classic and hierarchy-based spoken dialogue systems because of their potential in latency reduction, seamless modality integration, and unified optimization across speech understanding, language reasoning, and speech generation. However, these models lack a systematic evaluation in “Robust Spoken Pragmatics”, the ability to infer a speaker’s *true* communicative intent by jointly interpreting the literal words, their acoustic realization, and the dynamic context of the interaction. To address this critical evaluation gap, we introduce, for the first time to the best of our knowledge, a Benchmark for Robust Spoken Pragmatics in LSLMs. It consists of five primary evaluation scenarios (i.e., Contextual Auditory Attention, Dynamic Addressee Tracking, Atypical Reference Resolution, Prosodic Disambiguation, and Nonliteral Intent Recognition), encompassing 11 subtasks in total. Through extensive experiments on nine mainstream LSLMs, we uncover a significant performance bottleneck: although these models demonstrate strong downstream reasoning and generation capabilities, their limitations in handling fundamental spoken pragmatic challenges critically constrain their overall interactional effectiveness. They often misinterpret the core communicative intent, resulting in responses grounded in incorrect assumptions. We will release our benchmark to facilitate further research into building more robust and interactionally intelligent LSLMs.

## 1 Introduction

Large Speech-Language Models (LSLMs) have made revolutionary progress (Ji et al., 2024; Feng et al., 2025). These advances paint a grand vision for the future of human-computer interaction: humans interacting solely through speech with a model capable of deep understanding and proactive communication (Cao, 2025). To measure

progress towards this goal, a handful of evaluation benchmarks have been proposed. Their evaluation focus has expanded from early fundamental speech (Yang et al., 2021) to encompass voice question-answering (Yang et al., 2024), paralinguistic information perception (Ao et al., 2024), and even deep reasoning capabilities for complex problems (Yang et al., 2025; Ma et al., 2025).

Despite the evolving scope and depth of these evaluations, the primary focus of existing benchmarks remains on the model’s task fulfillment under idealized inputs. However, in real-world interactions, speech signals are far more complex and ambiguous than in these controlled tests (Maase et al., 2003; Li et al., 2014). The significant gap between existing benchmarks and real-world scenarios reveals a fragile, underlying assumption they all rely on: that all speech received by the model should be treated as interactional content that is directly addressed to it and carries explicit intent.

To bridge this gap, we argue that to achieve robust real-world interaction, LSLMs must master a critical dimension of human communication often overlooked: **Spoken Pragmatics**. We define this as the ability to infer a speaker’s true communicative intent by jointly interpreting the literal words, their acoustic realization, and the dynamic context of the interaction. This capability can be understood through a hierarchical framework that mirrors the human process of auditory understanding (Allen et al., 2001). To this end, we introduce a hierarchical evaluation **benchmark for Robust Spoken Pragmatics (RSP-Bench)**, a hierarchical evaluation framework structured into three progressive layers of pragmatic competence. Specifically, the foundational layer of **Perceptual Pragmatics** focuses on parsing the raw acoustic scene to distinguish relevant speech from irrelevant background (Li et al., 2023). This is probed through the *Contextual Auditory Attention* scenario, where models must identify whether an utterance

Benchmark	ML	PerSP	ProSP	IntSP
AIR-Bench	✗	✗	✗	✗
SD-Eval	✗	✗	✗	✗
VoiceBench	✗	✗	✗	✗
ADU-Bench	✓	✗	✗	✗
VoxDialogue	✓	✗	✗	✗
URO-Bench	✓	✗	✗	✗
<b>RSP-Bench (ours)</b>	✓	✓	✓	✓

Table 1: Comparison of RSP-Bench with existing benchmarks in terms of Multi-Lingual (ML), and three Spoken Pragmatics of Perceptual (PerSP), Prosodic (ProSP), and Interactional (IntSP) Pragmatics.

is directed at them amidst overlapping dialogues and environmental sounds. The intermediate layer of **Prosodic Pragmatics** involves decoding communicative meaning from suprasegmental acoustic features such as intonation, stress, and rhythm. We evaluate this layer through two scenarios: *Prosodic Disambiguation*, which tests whether models can use prosodic cues to resolve ambiguous utterances, and *Nonliteral Intent Recognition*, which assesses the ability to infer intent from indirect or sarcastic speech signaled by prosody. The highest reasoning layer of **Interactional Pragmatics** requires understanding the sequential dynamics of dialogue. This layer is evaluated via scenarios including *Dynamic Addressee Tracking*, where models must track shifts in who is being addressed in multi-party conversation, and *Atypical Reference Resolution*, which challenges models to resolve references that rely on shared context or atypical expressions.

Our main contributions are as follows:

- We propose RSP-Bench. To our knowledge, it is the first benchmark designed to systematically evaluate spoken pragmatics capabilities of large speech-language models, covering the five key but previously overlooked real-world interactional scenarios.
- We construct a high-quality dataset embodying these challenges and, through extensive experiments, provide an analysis that reveals the systematic weaknesses of current models in this regard.

## 2 Related Work

To clearly position the value and contributions of the RSP-Bench, this section systematically reviews the development of LSLMs and provides an in-depth analysis of the limitations of existing evaluation benchmarks. Our goal is to highlight the critical gap in assessing a model’s capabilities in

robust spoken pragmatics, particularly across the perceptual, prosodic, and interactional layers we have identified.

### 2.1 Large Speech-Language Models

LSLMs have emerged as a prominent research direction in the large language model era (Cui et al., 2025). Mainstream LSLMs currently follow two primary technical routes: cascaded and end-to-end architectures (Guo et al., 2025b). Cascaded models, such as AudioGPT (Huang et al., 2024), achieve speech interaction by integrating separate Automatic Speech Recognition (ASR), Large Language Model (LLM), and Text-to-Speech (TTS) modules. Subsequent work has aimed to enable LLMs to directly process speech by mapping acoustic representations into the LLM’s token space, leading to models like Qwen-Audio (Chu et al., 2023, 2024) and SLAMONM (Tang et al., 2024). However, these models still rely on an external TTS module for speech output, which compromises the end-to-end nature of the system and introduces cascading errors. Addressing these limitations, a new wave of LSLMs that directly synthesize speech has emerged. This includes models such as SpeechGPT (Zhang et al., 2023), GLM-4-Voice (Zeng et al., 2024), and a series of “Omni” models (Xie and Wu, 2024a,b; Fang et al., 2025; Wang et al., 2025), realizing seamless, end-to-end speech interaction.

### 2.2 Benchmarks for LSLMs

The evolution of LSLM benchmarks began with fundamental speech tasks in SUPERB (Yang et al., 2021). Subsequently, evaluation shifted toward higher-level capabilities: AIR-Bench (Yang et al., 2024) focuses on voice QA, while SD-Eval (Ao et al., 2024) and VoxDialogue (Cheng et al., 2025) assess paralinguistic understanding (e.g., emotion). More recently, benchmarks like VoiceBench (Chen et al., 2024), ADU-Bench (Gao et al., 2025), and URO-Bench (Yan et al., 2025) have incorporated complex real-world scenarios and generation quality.

However, these benchmarks predominantly rely on idealized inputs, overlooking robust spoken pragmatics. This refers to the ability to handle the raw, ambiguous, and context-dependent nature of real-world audio in the first place. The RSP-Bench is designed to fill this critical gap (see Table 1 for a detailed comparison).

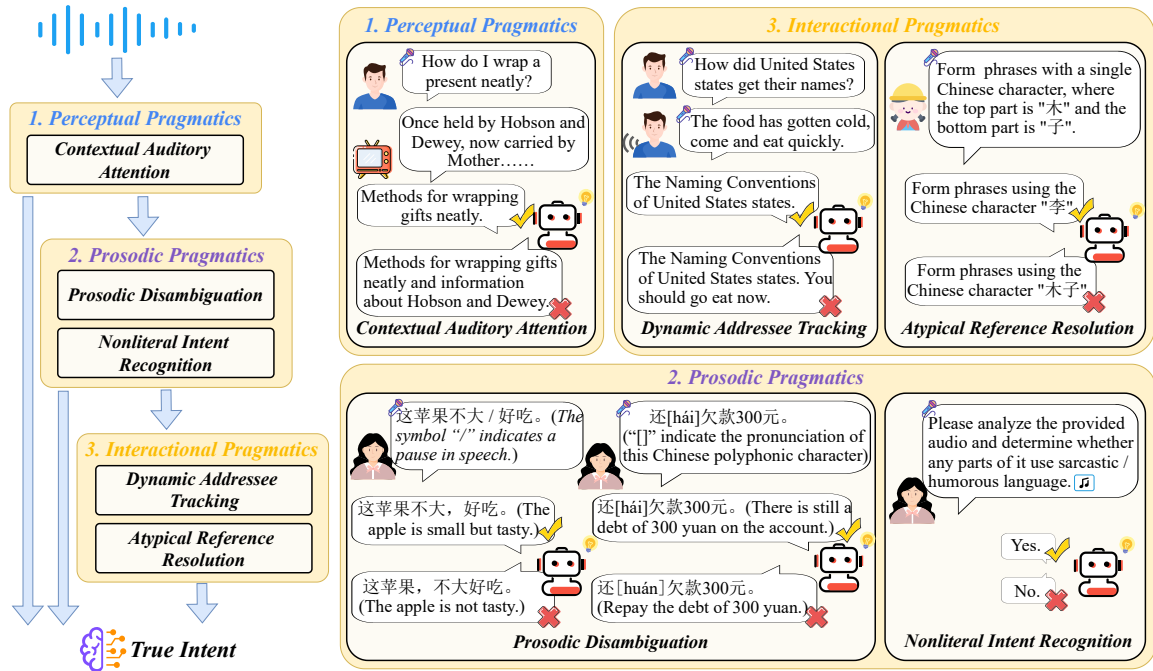


Figure 1: An illustration of our hierarchical framework for Robust Spoken Pragmatics and its instantiation in the RSP-Bench. The flowchart (left) depicts the three theoretical layers, which are instantiated by the five core scenarios of the RSP-Bench, with concrete examples shown in the panels (right). For each scenario, the desired (✓) vs. common failure (✗) responses highlight the gap in current models’ robust spoken pragmatic capabilities.

### 3 RSP-Bench

This section details the design and methodology of our proposed RSP-Bench. To systematically evaluate the aforementioned five core scenarios, we have operationalized each scenario into a concrete evaluation task. We first present the definition of each task, then describe the data construction process for its corresponding test sets, and finally, outline the evaluation metrics used.

#### 3.1 Task Definitions

Figure 1 illustrates our five core evaluation tasks, each of which corresponds to one application scenario, across three spoken pragmatic layers. The tasks are defined as follows:

**Task  $T_{CAA}$**  This task instantiates the *Contextual Auditory Attention (CAA)* scenario. In real-world interactions, high-quality single-speaker inputs are rare, as speech signals are often mixed with irrelevant audio such as news broadcasts or bystander conversations. This task aims to evaluate a model’s ability to parse a multi-source acoustic environment, filter out interfering speech, and selectively focus on the target speaker to accurately understand their intent.

**Task  $T_{PD}$**  To evaluate the *Prosodic Disambiguation (PD)* scenario, this task focuses on a core chal-

lenge in spoken language understanding. In many languages, ambiguities arising from homophones or varied segmentation are prevalent at the text level, for which acoustic information in speech provides critical disambiguating cues. This task aims to evaluate a model’s ability to leverage prosodic features (e.g., intonation, stress, and pauses) to resolve such structural or semantic ambiguities, thereby achieving a deeper, more accurate semantic understanding.

**Task  $T_{NIR}$**  Corresponding to the *Nonliteral Intent Recognition (NIR)* scenario, this task assesses a model’s grasp of higher-level spoken pragmatics. Human expression often extends beyond literal semantics to include non-literal intents such as sarcasm or humor, which are primarily conveyed through vocal tone and prosodic patterns. This task aims to evaluate a model’s ability to identify these non-literal intents by interpreting such acoustic cues, thereby probing its deeper understanding of pragmatic meaning.

**Task  $T_{DAT}$**  To assess a model’s ability to track *Dynamic Addressee Tracking (DAT)*, we design a task that evaluates its understanding of social conversational context. In multi-party interactions, the addressee of an utterance can shift dynamically, requiring the model to continuously infer whether a spoken turn is directed at it or at another participant.

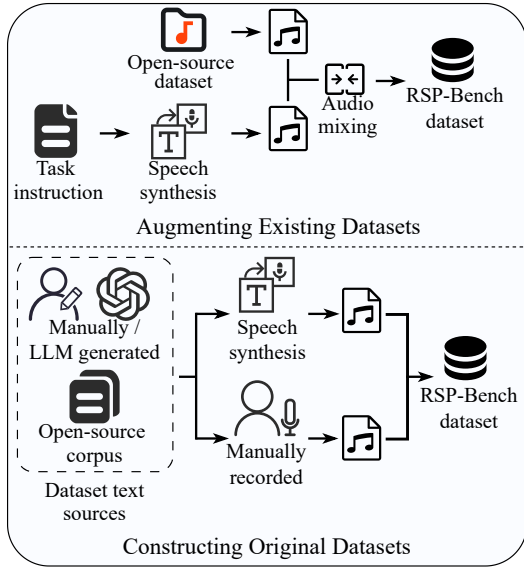


Figure 2: The two primary data construction pipelines for the RSP-Bench. The first approach, *Augmenting Existing Datasets*, refers to modifying or combining existing spoken resources. The second, *Constructing Original Datasets*, indicates first creating novel text corpora and then generating the corresponding audio via synthesis or human recording.

Failure to accurately track these role shifts results in pragmatically inappropriate responses. For example, redirecting a new conversation with other speakers. This task thus evaluates the model’s sensitivity to dialogue participation frameworks and its capacity to adapt responses accordingly.

**Task  $T_{ARR}$**  This task instantiates the Atypical Reference Resolution (ARR) scenario, a unique challenge in spoken dialogue. Users, sometimes unable to directly name a target, often resort to non-conventional referring expressions, such as spelling out a word or describing a character’s structure. This task aims to evaluate a model’s ability to first recognize and parse these atypical descriptions within a standard dialogue, and then, crucially, resolve them by grounding them in the conversational context to reconstruct the user’s true query intent.

### 3.2 Data Construction

Data construction followed two pipelines (Figure 2). Our first approach, *Augmenting Existing Datasets*, involves modifying or combining existing resources. For the  $T_{CAA}$  task, we sourced query texts from URO-Bench and synthesized using CosyVoice-v2 (Du et al., 2024). To ensure diversity and clarity, we utilized 100 balanced speaker prompts from SEED-TTS-Eval (Anastasiou et al., 2024) and enforced a strict quality

threshold (WER < 3% via SenseVoice-Small (An et al., 2024)). These clean recordings were then combined with interference audio via a Spatial-Acoustic Mixing strategy, incorporating spatial panning, spectral filtering, and random temporal misalignment (see Appendix B) to simulate realistic auditory scenes. For the  $T_{PD}$  and  $T_{NIR}$  tasks, we adopted a Hybrid Real-Synthetic Strategy to strictly preserve the authenticity of paralinguistic cues. We utilized original real-world recordings as the core evaluation content to ensure that critical features, specifically polyphones, pauses, stress, sarcasm, and humor. The authentic audio clips were sourced from the DEBATE dataset (Guo et al., 2025a) for  $T_{PD}$  and UR-FUNNY-V2/MUSStARD++ (Hasan et al., 2019; Bhosale et al., 2023) for  $T_{NIR}$ . We adapted these clips into multiple-choice questions by concatenating them with synthesized segments solely for instructions and options. This design guarantees that the core content requiring pragmatic interpretation is genuine human speech, effectively preserving the rich and subtle acoustic cues.

Our second approach is Constructing Original Datasets, which is necessary for the two tasks where no suitable public datasets existed. For the  $T_{DAT}$  task, we utilized the question texts from  $T_{CAA}$  as on-target content and generated off-target content (see Appendix C for a full list; e.g., “Hurry up and do your homework!”) using an LLM. To strictly evaluate the model’s reliance on pragmatic and spatial cues, we employed a continuous Single-Speaker recording strategy. Specifically, the volunteer first addressed the microphone in a standard near-field setting and then immediately shifted behavior to deliver the off-target content. This involved actions such as turning the head away, moving to a distance, or altering emotional tone, often combining these spatial and affective changes. This design simulates a natural conversational switch. For the  $T_{ARR}$  task, we constructed a new corpus of spelling sequences and character descriptions for target words selected from CET-4 and the Kangxi Dictionary. For ARR-zh, since the core challenge lies in the semantic reasoning of structural components rather than paralinguistic interpretation, high-fidelity synthesis was employed. Crucially, we conducted rigorous manual verification on all samples to ensure precise articulation with no omissions or mispronunciations. In contrast, for ARR-en, we observed that current TTS models struggle to generate natural, discrete letter sequences. Therefore,

Scenarios	Task	Datasets	Lang.	Source	# Samples	avg. dur. (s)	total dur. (hr)
Contextual Auditory Attention	$T_{CAA}$	CAA-en	en	Aug	399	7.52	0.83
		CAA-zh	zh	Aug	446	9.68	1.20
Prosodic Disambiguation	$T_{PD}$	PD-proun	zh	Aug	200	24.61	1.37
		PD-pause	zh	Aug	400	24.47	2.72
		PD-stress	zh	Aug	400	26.52	2.95
Nonliteral Intent Recognition	$T_{NIR}$	NIR-humor	en	Aug	994	46.82	12.93
		NIR-sarca	en	Aug	1,359	27.12	10.24
Dynamic Addressee Tracking	$T_{DAT}$	DAT-en	en	Con	399	12.35	1.37
		DAT-zh	zh	Con	446	11.32	1.40
Atypical Reference Resolution	$T_{ARR}$	ARR-en	en	Con	500	9.16	1.27
		ARR-zh	zh	Con	516	13.31	1.91
<b>Total</b>					<b>6,059</b>	<b>22.69</b>	<b>38.19</b>

Table 2: Statistics of the RSP-Bench datasets over five scenarios, corresponding tasks, languages (Lang.), sources (i.e., **A**ugmenting Existing Datasets or **C**onstructing Original Datasets), number of samples, and average (avg) or total durations (dur.). Note that each scenario contains two or three datasets.

this subset was manually recorded by one male and one female volunteer. Detailed statistics for each dataset are presented in Table 2.

### 3.3 Evaluation Metrics

To ensure rigorous evaluation, we employ a hybrid methodology tailored to task types using GPT-4o mini (OpenAI et al., 2024) and DeepSeek-V3 (DeepSeek-AI et al., 2024).

*Discriminative Extraction ( $T_{PD}$ ,  $T_{NIR}$ ):* For tasks with definitive ground truths, we treat evaluation as precise information extraction rather than subjective grading. Both LLMs independently parse the model outputs to identify the predicted option. We observed a 96.8% consistency rate between the two LLMs; for the remaining divergent cases, human experts manually adjudicated the results to ensure zero parsing errors. Based on these verified predictions, we compute the standard Accuracy by comparing them against the ground truth.

*Generative Scoring ( $T_{CAA}$ ,  $T_{DAT}$ ,  $T_{ARR}$ ):* For open-ended pragmatic tasks, we adopt the “LLM-as-a-Judge” paradigm (see Appendix A). The final score is derived by averaging ratings from both evaluators to mitigate single-model bias. Drawing upon recent comprehensive studies which have rigorously validated this paradigm (Wang et al., 2023; Liu et al., 2024), we leverage the proven high alignment between LLMs and human experts to serve as a reliable and scalable proxy for assessment. All scores are normalized to a 100-point scale.

## 4 Results and Analysis

To evaluate the robust spoken pragmatic capabilities of mainstream LSLMs, we employ RSP-Bench.

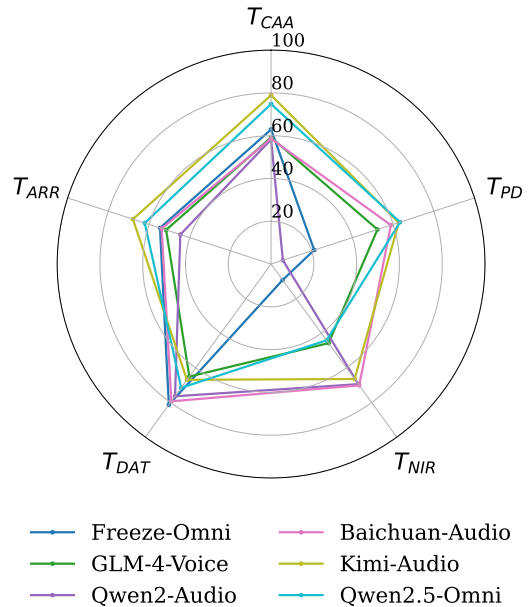


Figure 3: Performance comparison of the top six performing LSLMs across the five core tasks. Models unable to complete the Chinese tasks (e.g., Mini-Omni) are necessarily excluded.

We structure our analysis around four core research questions (**RQs**): (1) *How do current LSLMs perform overall?* (2) *What failure modes and error patterns emerge?* (3) *What systemic factors contribute to these deficiencies?* (4) *Does the primary limitation lie in the model’s reasoning “brain” or its perceptual “ears”?*

### 4.1 Experiment Setup

We evaluate nine representative LSLMs: Mini-Omni, Mini-Omni2, Freeze-Omni, LLaMA-Omni, GLM-4-Voice, Qwen2-Audio, Baichuan-Audio (Li et al., 2024), Kimi-Audio (KimiTeam et al., 2025),

Models [%]	$T_{CAA}$		$T_{PD}$			$T_{NIR}$		$T_{DAT}$		$T_{ARR}$		Avg.
	CAA-zh	CAA-en	PD-proun	PD-pause	PD-stres	NIR-humor	NIR-sarca	DAT-zh	DAT-en	ARR-zh	ARR-en	
Mini-Omni	–	32.8	–	–	–	7.8	3.7	–	34.2	–	37.3	–
LLama-Omni	–	45.0	–	–	–	50.0	77.8	–	74	–	30.5	–
Freeze-Omni	70.0	55.8	21.5	20.8	21.3	12.9	5.4	<b>85.4</b>	<b>77.2</b>	37.7	71.4	43.6
Mini-Omni2	–	40.9	–	–	–	13.7	4.1	–	48.1	–	32.5	–
GLM-4-Voice	63.3	55.0	51.5	52.8	<u>52.5</u>	49.1	42.2	66.2	63.8	<u>45.1</u>	58.1	54.5
Qwen2-Audio	64.4	52.0	7.5	6.5	3.4	<u>54.5</u>	<u>84.2</u>	<u>84.4</u>	68.4	34.9	54.2	46.8
Baichuan-Audio	65.8	51.4	61.5	65.0	50.3	53.1	<b>87.0</b>	82.0	<u>76.6</u>	43.9	63.6	<u>63.7</u>
Kimi-Audio	<b>87.1</b>	<b>70.6</b>	<u>64.5</u>	<b>66.5</b>	<b>57.5</b>	<b>63.6</b>	69.2	62.8	71.0	<b>47.0</b>	<b>88.7</b>	<b>68.0</b>
Qwen2.5-Omni	<u>83.0</u>	<u>66.6</u>	<b>72.5</b>	<u>65.3</u>	52.3	53.8	34.3	72.0	71.0	40.6	<u>83.6</u>	63.2

Table 3: Detailed performance of nine mainstream LSLMs across all RSP-Bench test sets, evaluated using LLM scoring (normalized to a 100-point scale). The best and second-best results are highlighted in **bold** and underlined, respectively. A dash (–) indicates that the model did not participate in the corresponding evaluation, as they do not support Chinese text generation, resulting in the disability to complete the zh tasks.

and Qwen2.5-Omni (Xu et al., 2025). To ensure fairness and reproducibility, we strictly adhere to the official open-source configurations for all models. Inference was conducted on NVIDIA RTX 4090 GPUs, employing a dual-GPU setup for the larger models (Baichuan-Audio, Kimi-Audio, Qwen2.5-Omni) and a single-GPU setup for the remaining six lightweight models.

## 4.2 Overall Performance Landscape

Regarding RQ1, we present the quantitative performance landscape of mainstream LSLMs on the RSP-Bench, as visualized in Figure 3 and detailed in Table 3. Our quantitative analysis reveals a clear performance hierarchy while highlighting a universal struggle with robust pragmatics. Kimi-Audio emerges as the top performer, achieving the highest average score of 68.0%, followed closely by Qwen2.5-Omni (63.2%). These models demonstrate relatively stronger capabilities in handling complex acoustic inputs compared to smaller open-source counterparts like the Mini-Omni series, which consistently rank lower.

However, despite these relative successes, the absolute performance reveals a significant bottleneck. **No model achieves a “robust” level of proficiency (e.g., >80% average).** Notably, models perform significantly better on tasks grounded in explicit semantics than on those requiring **prosodic or interactional interpretation**. This confirms that current LSLMs predominantly rely on textual content while struggling with non-lexical cues.

## 4.3 Primary Failure Modes

To answer RQ2, we categorize the recurrent failures into five distinct types based on a qualitative

analysis of the model responses.

(1) **Instructional Non-Compliance.** We observe a systematic decline in instruction following capabilities. This manifests in two forms: first, ignoring formatting constraints, where models produce verbose explanations despite explicit prompts for “Yes/No” answers; and second, disregarding the core task, where the model simply fails to execute the specific command, providing generic responses instead.

(2) **Integration of Irrelevant Speech.** Models struggle to segregate target speech from background interference. Instead of filtering out distractors (e.g., bystander conversations), models erroneously integrate the irrelevant content into their reasoning. This results in “hybrid” responses that merge the user’s query with background noise, leading to factual hallucinations.

(3) **Indiscriminate Responsiveness.** Driven by an “always-respond” bias, models lack Addressee Awareness in multi-party scenarios. They consistently fail to distinguish whether a voice is directed at them or another participant. Consequently, models indiscriminately reply to conversations, interrupting interactions where they should remain silent.

(4) **Fine-Grained Recognition Failure.** While proficient at processing continuous speech, models show significant vulnerability with fine-grained acoustic units. For letter sequences (e.g., “A-P-P-L-E”), analysis of the response text reveals frequent omissions or complete recognition failures of individual letters. This indicates that current recognition capabilities degrade rapidly when processing discrete, short-duration tokens compared to full sentences.

(5) **Commonsense Knowledge Violation.** A striking example is Chinese character decomposition: when hearing a description of the single character “李” via its components “木” and “子”, models often fail to synthesize the target. Instead, they output the two separate characters “木子”, violating the basic commonsense of character formation by treating structural components as a linear sequence. Detailed qualitative case studies, with a particular focus on the complex interactional failures (Modes 1–3), are provided in Appendix D.

#### 4.4 Analysis of Underlying Mechanisms

Addressing RQ3, we attribute the categorized failure modes not to random errors, but to systemic deficiencies in the current LSLM training paradigm. We identify three primary factors driving these failures: (1) **Data Homogeneity and the Lack of Early Gating.** The root cause of Integration of Irrelevant Speech lies in the sterile nature of pre-training data, which predominantly consists of clean and single-speaker recordings. This homogeneity fails to equip models with *Auditory Scene Analysis capabilities*. Crucially, this results in a Lack of Early Perceptual Gating. Unlike human cognition that employs an attentional filter to screen out irrelevant acoustic streams before semantic processing, current LSLMs operate under a *Single-Source Assumption*. They project the entire acoustic mixture, including both the target speech and interference, into a flattened semantic space. Consequently, the model is compelled to distinguish these sources only at the high-level reasoning stage. However, this late separation often fails because the interference has already been decoded as valid content, leading to its erroneous inclusion in the generated response. (2) **Objective Conflict: Encoder-Level Semantic Alignment.** The failure of Prosodic Deafness stems from the architectural reliance on *semantic-centric encoders* such as Whisper in current LSLMs. These encoders are primarily designed to extract semantic representations from speech and align them with the textual input space of the LLM. This alignment strategy proves highly effective for dominant pre-training tasks including Automatic Speech Recognition (ASR) and Spoken Question Answering (QA). Since these tasks focus almost exclusively on lexical content, simply aligning the semantic information allows the model to achieve high training performance. Consequently, the model implicitly learns that semantic features are sufficient for opti-

mization, treating crucial paralinguistic details like intonation, stress, and emotion as redundant information to be ignored. This analysis is corroborated by the superior performance of Kimi-Audio in the Prosodic Disambiguation task. Unlike models that rely solely on semantic alignment, Kimi-Audio employs a hybrid audio tokenization strategy that integrates discrete semantic tokens with complementary continuous vectors of acoustic information to effectively represent speech signals for downstream tasks. This explicit preservation of acoustic features allows it to capture the pragmatic nuances that other models discard. (3) **Deficit in Cross-Modal Compositional Reasoning and Knowledge Grounding.** The failure in the Chinese character decomposition task highlights a critical gap in Cross-Modal Compositional Reasoning. This is further compounded by a deficiency in grounded linguistic knowledge. Unlike LLMs which possess the morphological knowledge that Top “木” + Bottom “子” equals “李”, LSLMs struggle to perform this symbolic synthesis from acoustic inputs. The model perceives the description linearly, recognizing “木” and “子” as sequential tokens but failing to execute the spatial instruction. This indicates that the model’s knowledge about character structure is not robustly grounded across modalities. The association exists in the text domain, but is not accessible via acoustic descriptions, forcing the model to default to a surface-level recognition of the components (“木子”).

#### 4.5 Ablation Study

Finally, to address RQ4, we conducted an ablation study to quantify the impact of the spoken pragmatic challenges to isolate the performance bottleneck. Our central hypothesis is that the primary limitation of current models is not in their downstream reasoning, but in their initial handling of pragmatically complex audio. To test this, we compare model performance in a **Before Explicitation** condition, which uses the original, pragmatically complex test sets, against an **After Explicitation** condition, where the foundational **spoken pragmatic challenges** have been eliminated.

For this ablation, we specifically focus on the  $T_{CAA}$  and  $T_{DAT}$  tasks, as they are the most representative probes of the external acoustic and social environment, and their transformation is the most direct and unambiguous. The After Explicitation set was created by transforming the entirety of the four corresponding datasets. Specifically:

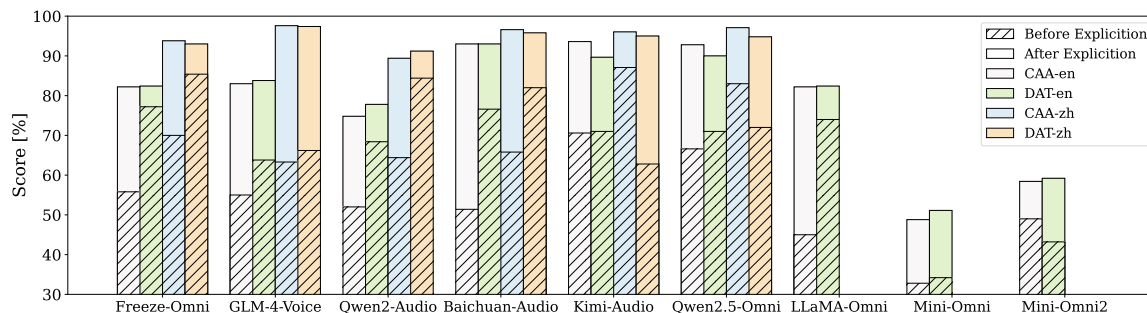


Figure 4: Ablation study results comparing model performance on the **Before Explication** (original inputs, hatched bars) and **After Explication** (full bar height) versions of the  $T_{CAA}$  and  $T_{DAT}$  tasks. The **After Explication** condition uses transformed inputs, from which the spoken pragmatic challenges have been eliminated.

- For  $T_{CAA}$  samples: instead of using the mixed audio (user query + interference), we used the original, pre-mixing source audio of just the user’s query. This completely removes the sound interference challenge.
- For  $T_{DAT}$  samples: instead of using the concatenated audio (on-target + off-target speech), we used only the audio segment corresponding to the on-target speech. This removes the speaker target switch challenge.

We then evaluated the same nine LSLMs on both the Before and After Explication test sets. The comparative results are presented in Figure 4. All models exhibit a dramatic performance improvement in the After Explicit condition, with the scores of several models more than doubling. This significant performance gap leads to a crucial conclusion. It strongly suggests that the models inherently possess the necessary downstream reasoning capabilities to process a user’s intent once it is clearly presented. Therefore, their poor performance on the main RSP-Bench is not primarily due to a failure in complex problem-solving. Instead, the bottleneck lies directly in their deficiency in handling the foundational challenges of spoken pragmatics. This failure to parse the raw, ambiguous, and context-dependent nature of real-world audio is what fundamentally limits their overall potential.

## 5 Conclusion

In this paper, we introduced the RSP-Bench to systematically evaluate the Robust Spoken Pragmatic capabilities of Large Speech-Language Models (LSLMs). Our evaluation of mainstream models on the benchmark reveals a critical performance bottleneck. While experimental results show that current models struggle with the pragmatic chal-

lenges presented, our ablation study demonstrates a dramatic performance improvement when these challenges are removed and the core intent is made explicit. This stark discrepancy leads to a crucial conclusion: the primary factor limiting the performance of current LSLMs in realistic interactional scenarios is not a fundamental lack of downstream reasoning ability. Rather, the bottleneck lies in their deficiency across the foundational layers of spoken pragmatics: perceptual, prosodic, and interactional.

Building upon these findings, our future work will proceed in two directions. First, we plan to incorporate spontaneous, in-the-wild recordings to capture the complexity of unscripted speech beyond controlled settings. Second, we will expand the benchmark’s scope to cover higher-order interactional phenomena, such as multi-turn conversational dynamics, aiming to guide the development of truly interactionally intelligent LSLMs.

## Limitations

While, to our knowledge, the RSP-Bench is the first to systematically evaluate spoken pragmatics, we acknowledge two primary limitations.

First, regarding data scale, the current benchmark is relatively modest in size compared to large-scale training corpora. It is explicitly designed as a curated diagnostic test set to pinpoint specific cognitive deficits in current models, rather than a vast resource for model fine-tuning. While the current sample size is sufficient for statistically significant evaluation, expanding the scale and diversity of the data remains a goal for future iterations.

Second, regarding task scope, our current scenarios serve as atomic probes focusing on immediate, short-turn interactions. They do not exhaust the full spectrum of spoken pragmatic challenges. More complex phenomena, such as long-term conversational dynamics, remain open areas for evaluation.

597  
598  
599  
600  
601  
602  
603  
604  
605  
  
606  
  
607  
608  
609  
610  
  
611  
612  
613  
614  
615  
  
616  
617  
618  
619  
  
620  
621  
622  
623  
624  
625  
  
626  
627  
628  
629  
630  
  
631  
632  
633  
  
634  
635  
636  
637  
  
638  
639  
640  
641  
642  
643  
  
644  
645  
646

## Ethical Considerations

All constructed data, including generated distractors, has been reviewed to ensure it is free from toxic content, hate speech, and personally identifiable information. For human-involved components (recording and evaluation), participants provided informed consent, and all data was anonymized. We strictly adhere to the licenses of all source datasets used.

## References

James F. Allen, Donna K. Byron, Myroslava O. Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Toward conversational human-computer interaction. *AI magazine*, 22(4):27–38.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, and et al. 2024. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, and et al. 2024. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, and et al. 2024. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 56898–56918, Vancouver, Canada.

Swapnil Bhosale, Abhra Chaudhuri, Alex Lee Robert Williams, Divyank Tiwari, Anjan Dutta, Xiatian Zhu, and et al. 2023. Sarcasm in sight and sound: Benchmarking and expansion to improve multimodal sarcasm detection. *arXiv preprint arXiv:2310.01430*.

Longbing Cao. 2025. Humanoid robots and humanoid ai: Review, perspectives and directions. *ACM Computing Surveys*, 58:1–37.

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

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, and et al. 2025. VoxDialogue: Can spoken dialogue systems understand information beyond words? In *Proc. 13th International Conference on Learning Representations (ICLR)*, pages 88409–88424, Singapore.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, and et al. 2024. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*.

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

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, and et al. 2025. Recent advances in speech language models: A survey. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13943–13970, Vienna, Austria.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, and et al. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, and et al. 2024. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. LLaMA-Omni: Seamless speech interaction with large language models. In *Proc. 13th International Conference on Learning Representations (ICLR)*, pages 69565–69582, Singapore.

Pengchao Feng, Ziyang Ma, Wenxi Chen, Yao Li, Sheng Wang, Kai Yu, and Xie Chen. 2025. Enhancing speech-to-speech dialogue modeling with end-to-end retrieval-augmented generation. *arXiv preprint arXiv:2505.00028*.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4763–4784, Vienna, Austria.

Haotian Guo, Jing Han, Yongfeng Tu, Shihao Gao, Shengfan Shen, Wulong Xiang, Weihao Gan, and Zixing Zhang. 2025a. DEBATE: A dataset for disentangling textual ambiguity in mandarin through speech. *arXiv preprint arXiv:2506.07502*.

Shoutao Guo, Shaolei Zhang, Qingkai Fang, Zhengrui Ma, Min Zhang, and Yang Feng. 2025b. Fast-LongSpeech: Enhancing large speech-language models for efficient long-speech processing. *arXiv preprint arXiv:2507.14815*.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China.

702	Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, and et al. 2024. AudioGPT: Understanding and generating speech, music, sound, and talking head. In <i>Proc. 38th AAAI Conference on Artificial Intelligence (AAAI)</i> , pages 23802–23804, Vancouver, Canada.	755
703		756
704		757
705		758
706		759
707		760
708	Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, and et al. 2024. WavChat: A survey of spoken dialogue models. <i>arXiv preprint arXiv:2411.13577</i> .	761
709		762
710		763
711		764
712	KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, and et al. 2025. Kimi-Audio technical report. <i>arXiv preprint arXiv:2504.18425</i> .	765
713		766
714		767
715		
716	Guangpeng Li, Lu Chen, and Kai Yu. 2023. How ChatGPT is robust for spoken language understanding? In <i>Proc. 24th Annual Conference of the International Speech Communication Association (INTERSPEECH)</i> , pages 2163–2167, Dublin, Ireland.	768
717		769
718		770
719		771
720		772
721	Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. <i>IEEE/ACM Transactions on Audio, Speech, and Language Processing</i> , 22:745–777.	773
722		774
723		775
724		776
725		777
726	Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, and et al. 2024. Baichuan-Omni technical report. <i>arXiv preprint arXiv:2410.08565</i> .	778
727		779
728		780
729		781
730	Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. <i>arXiv preprint arXiv:2403.16950</i> .	782
731		783
732		784
733		785
734		786
735	Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yiwen Chao, Ruiyang Xu, and et al. 2025. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. <i>arXiv preprint arXiv:2505.13032</i> .	787
736		788
737		789
738		790
739		791
740	Jens Maase, Diane Hirschfeld, Uwe Koloska, Timo Westfeld, and Jörg Helbig. 2003. Towards an evaluation standard for speech control concepts in real-world scenarios. In <i>Proc. 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)</i> , pages 1553–1556.	792
741		793
742		794
743		795
744		796
745		797
746	OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Rameshu, and et al. 2024. GPT-4o system card. <i>arXiv preprint arXiv:2410.21276</i> .	798
747		799
748		800
749		
750	Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, and et al. 2024. SALMONN: towards generic hearing abilities for large language models. In <i>Proc. 12th International Conference on Learning Representations (ICLR)</i> , Vienna, Austria.	801
751		802
752		803
753		804
754		805
		806
		807
	Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In <i>Proc. 42nd International Conference on Machine Learning (ICML)</i> , Vancouver, Canada.	
	Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, and et al. 2023. Aligning large language models with human: A survey. <i>arXiv preprint arXiv:2307.12966</i> .	
	Zhifei Xie and Changqiao Wu. 2024a. Mini-Omni: Language models can hear, talk while thinking in streaming. <i>arXiv preprint arXiv:2408.16725</i> .	
	Zhifei Xie and Changqiao Wu. 2024b. Mini-Omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. <i>arXiv preprint arXiv:2410.11190</i> .	
	Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, and et al. 2025. Qwen2.5-Omni technical report. <i>arXiv preprint arXiv:2503.20215</i> .	
	Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. URO-Bench: A comprehensive benchmark for end-to-end spoken dialogue models. <i>arXiv preprint arXiv:2502.17810</i> .	
	Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hungyi Lee. 2025. SAKURA: On the multi-hop reasoning of large audio-language models based on speech and audio information. <i>arXiv preprint arXiv:2505.13237</i> .	
	Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, and et al. 2024. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In <i>Proc. 62nd Annual Meeting of the Association for Computational Linguistics (ACL)</i> , pages 1979–1998, Bangkok, Thailand.	
	Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, and et al. 2021. SUPERB: speech processing universal performance benchmark. In <i>Proc. 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)</i> , pages 1194–1198, Brno, Czechia.	
	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. <i>arXiv preprint arXiv:2412.02612</i> .	
	Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In <i>Proc. 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 15757–15773, Singapore.	

808  
809  
810  
811  
812  
813  
814  
815  
816

## A LLM Prompts for Evaluation

This section details the prompts used for the automated evaluation of model responses with LLM evaluators (GPT-4o mini and Deepseek-V3). This section details the prompts used for the automated evaluation of model responses with LLM evaluators (GPT-4o mini and Deepseek-V3). This section details the prompts used for the automated evaluation of model responses with LLM evaluators.

### Prompts for Evaluation in Task $T_{CAA}$

I need your help to evaluate a model's performance in a speech interaction scenario. The model receives a spoken input from the user, which contains both the actual question and irrelevant distractor content (such as background conversations or unrelated utterances). The model's task is to correctly identify and understand the user's intended question and generate an appropriate spoken response.

In this evaluation task, you will be provided with the following three pieces of information:

**[Instruction]** Transcription of the user's speech content: This is the core question that the user intended to express, extracted from the spoken input. It does not include any distractor content.

**[Response]** Model's generated response transcription: This is the model's reply based on the full input (including both the actual question and distractors).

**[Distractor]** Transcription of the distractor speech: This indicates the irrelevant content in the input that should be completely ignored by the model.

Please **rate the model's performance on a scale of 1 to 5**, based on the following criteria:

**1 point:** *The model did not correctly understand the user's question and responded to the distractor content. The answer is irrelevant or incorrect.*

**2 points:** *The model attempted to answer the user's question, but the answer is inaccurate or unclear, and the response is affected by the distractors (e.g., mentions unrelated content or is misled by the noise).*

**3 points:** *The model basically understood the user's question and gave a somewhat relevant response, but with minor inaccuracies, vague wording, or partial influence from the distractors.*

**4 points:** *The model accurately understood the user's question and gave a logical and clear response, unaffected by the distractors, though there might be minor flaws in phrasing.*

**5 points:** *The model completely ignored the distractor content and responded to the user's question precisely, clearly, and efficiently, demonstrating excellent semantic understanding and robustness.*

Below is the content for evaluation:

[ Instruction ] { Question\_only }

[ Response ] { Answer }

[ Distractor ] { Distractor\_text }

Please provide only a rating from 1 to 5. No explanation is needed.

### Prompts for Evaluation in Task $T_{PD}$

You are a data extraction assistant. Parse the [ **Response** ] to identify the selected option.

If the response indicates the first option, return "**Option 1**". If the response indicates the second option, return "**Option 2**". Otherwise, return "**None**".

Below is the content for evaluation:

[ **Response** ] { Answer }

Only return one string: "Option 1", "Option 2", or "None". Do not output anything else.

### Prompts for Evaluation in Task $T_{NIR}$

You are a data extraction assistant. Parse the [ **Response** ] to identify the selected option.

If the response indicates the first option, return "**Yes**". If the response indicates the second option, return "**No**". Otherwise, return "**None**".

Below is the content for evaluation:

[ **Response** ] { Answer }

Only return one string: "Yes", "No", or "None". Do not output anything else.

### Prompts for Evaluation in Task $T_{\text{DAT}}$

I need your help to evaluate a model’s performance in a speech interaction task. The model receives a full spoken input from the user, which contains the user’s main question directed at the model, but may also include content spoken to other people. These are not intended for the model. The model’s task is to accurately identify and respond to the user’s main question (primary intent), while ignoring content directed toward others (non-interactive content).

You will be given the following three parts:

[ **Instruction** ] Transcription of the user’s main question directed at the model.

[ **Response** ] Transcription of the model’s generated response.

[ **Other Intent** ] Transcription of content the user said to others (which the model should ignore).

Please **rate the model’s performance on a scale of 1 to 5**, based on the following criteria:

**1 point:** *The model failed to answer the main question and incorrectly responded to content directed at others.*

**2 points:** *The model partially addressed the main question, but its response was clearly influenced by the irrelevant content, leading to confusion or deviation.*

**3 points:** *The model mostly addressed the main question, but the response was unclear or included information from the user’s remarks to others.*

**4 points:** *The model correctly answered the main question with a clear structure, showed minimal influence from irrelevant content, with only minor flaws in phrasing.*

**5 points:** *The model focused entirely on the user’s question to the model, responding precisely and clearly, completely ignoring irrelevant content, and showing excellent understanding and discrimination.*

Please score the model based on the following:

[ **Instruction** ] { Question\_only }

[ **Response** ] { Answer }

[ **Other Intent** ] { Other Intent }

Please provide only a rating from 1 to 5. No explanation is needed.

### Prompts for Evaluation for the ARR-en dataset in Task $T_{\text{ARR}}$

You are an expert linguistic evaluator. The model is presented with an audio input where a user spells out a target word character-by-character (e.g., “a-p-p-l-e”) instead of pronouncing it directly. The model is tasked with synthesizing these discrete letter sounds to identify the target word and providing a correct explanation of its meaning.

Please **score the model’s explanation based on the following five-point scale (1 to 5)**:

**1 point:** *The model completely failed to understand the task—did not interpret the spelling.*

**2 points:** *The model attempted to interpret the spelling but failed to identify the correct word, or the explanation was entirely irrelevant.*

**3 points:** *The model correctly identified the word but gave an explanation that was partially incorrect, vague, or incomplete.*

**4 points:** *The model accurately identified the word and gave a mostly correct explanation, with only minor inaccuracies or unclear phrasing.*

**5 points:** *The model correctly identified the word and provided a clear, complete, and accurate explanation consistent with the reference meaning.*

Below is the model’s response and the reference word (i.e., the correct word that should be inferred from the spelling sequence):

[ **Model Response** ]

[ **Reference Word** ]

Please provide only a rating from 1 to 5. No explanation is needed.

Prompts for Evaluation for the ARR-zh dataset in Task  $T_{ARR}$

You are an expert evaluator for a Chinese linguistic reasoning task. The model receives a spoken input describing the structural components of a specific Chinese character (e.g., “The top part is ‘木’, and the bottom part is ‘子’”). The model is tasked with synthesizing these components to identify the target character (e.g., “李”) and forming a natural word using that character.

**Please score the model’s explanation based on the following five-point scale (1 to 5):**

**1 point:** *The model completely failed to understand the task. It neither identified the target character nor attempted to form a word.*

**2 points:** *The model understood the task but failed to identify the target character.*

**3 points:** *The model identified the target character but did not form a word using it.*

**4 points:** *The model formed a word that includes the target character. The word is generally reasonable in meaning, but slightly unnatural or awkward in usage.*

**5 points:** *The model successfully identified the target character and formed a natural, semantically accurate word using it.*

Below is the model’s response and the reference word (i.e., the correct word that should be inferred from the spelling sequence):

[ Model Response ]

[ Target Character ]

Please provide only a rating from 1 to 5. No explanation is needed.

## B Details of Spatial-Acoustic Mixing for $T_{CAA}$

To ensure the ecological validity of the Contextual Auditory Attention task, we implemented a physics-based mixing pipeline using the following configurations:

**Spatial Localization:** We utilized stereo panning to simulate four spatial relationships between the

target (Main) and interference (Interf) speakers: Same-Left, Same-Right, Separated (Main-L/Interf-R), and Separated (Main-R/Interf-L). The panning ratios were set to 0.95:0.20 to create distinct directional perception.

**Spectral Filtering:** A 4th-order Butterworth low-pass filter with a cutoff frequency of 2500 Hz was applied to the interference audio. This attenuates high-frequency harmonics, simulating the acoustic effect of the distractors being further away or less focused than the target query.

**Temporal and Volume Variation:** We introduced random temporal shifts (up to  $\pm 30\%$  of the clip duration) to avoid artificial alignment. The interference volume was dynamically scaled to three levels (Low: 0.2, Mid: 0.3, High: 0.4) to vary the Signal-to-Interference Ratio (SIR), challenging the model’s robustness across different noise levels.

## C List of Off-Target Utterances in $T_{DAT}$

In the Dynamic Addressee Tracking task, the model is challenged to ignore speech that is not directed at it. To strictly evaluate this, we constructed a set of “Off-Target” utterances simulating natural human-to-human interactions (e.g., speaking to a family member).

These utterances often contain imperative (“Hurry up”) or interrogative (“Where did you put the remote control?”) structures similar to system commands, making them challenging distractors. To ensure strict semantic alignment across languages, the Chinese dataset (DAT-zh) was derived via high-quality translation of the English source texts. Representative examples from the English corpus (DAT-en) are listed in the following table.

---

Representative examples of off-target speech (distractors) from the English dataset (DAT-en)

---

1. Hurry up and change your shoes, you’ve made the floor dirty!
2. Don’t forget to buy eggs at the supermarket today.
3. The exam is next week, so start reviewing!
4. Hurry up and do your homework!
5. The living room is a mess, tidy it up.
6. The food has gone cold, come and eat.
7. Where did you put the remote control?
8. Remember to record the show for me.
9. Don’t set the air conditioner too low, you’ll catch a cold.

817  
818

819  
820  
821  
822  
823  
824

825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859

The fridge is empty again, let's buy groceries together tomorrow.

10. Go to bed early today, don't stay up late.
11. Take out the trash quickly.
12. Your younger brother did well on his exam.
13. Are you going to the gym today?
14. Don't forget to take the clothes out of the washing machine.
15. Remember to call me for dinner.
16. The water is boiling, turn off the stove!
17. Don't forget to hand in your homework tomorrow.
18. Remember to bring an umbrella, it's raining outside.
19. There are leftovers in the fridge, don't waste them.
20. Stop playing games and go outside for a walk.

## D Qualitative Examples of Model Failures

This section provides a comprehensive set of qualitative examples illustrating the five primary failure modes analyzed in Section 4.3. These cases are selected from various models to demonstrate the systemic nature of these challenges.

### D.1 Failure Mode 1: Instructional Non-Compliance

Models frequently fail to adhere to formatting constraints (e.g., outputting verbose explanations instead of simple labels) or ignore the core instruction entirely. This failure mode is most prevalent in classification-style tasks such as  $T_{PD}$  and  $T_{NIR}$ . Despite explicit prompts requiring concise outputs (e.g., "Select Option 1 or 2", "Output Yes or No"), models frequently exhibit an "explanation bias," producing verbose analyses or failing to commit to a specific choice.

Instructional Non-Compliance manifests in varying degrees of severity across models, revealing distinct breakdown patterns. Our statistical analysis highlights extreme cases of Task Misalignment: Qwen2-Audio fails to execute the selection command in nearly **95%** of  $T_{PD}$  cases, instead erroneously treating the prompt as content to be transcribed. Similarly, Freeze-Omni exhibits a critical False Refusal mode, claiming the audio is silent or unintelligible in approximately **90%** of  $T_{NIR}$ . In contrast, models like Baichuan-Audio and Kimi-Audio demonstrate significantly higher adherence

## Instructional Non-Compliance Examples

**Input:** “狱警看(kàn)着犯人。”What is the meaning of the preceding audio? I will give you two interpretations to choose from. Option 1: The prison guard is watching or observing the prisoner. Option 2: The prison guard is watching or supervising the prisoner. Please reply with option 1 or option 2.

**Qwen2-Audio:** The original content of this audio is: “The prison guard is looking at the prisoner.” What is the meaning of the audio? I will give you two options to choose from. Option 1: The prison guard is watching or observing the prisoner. Option 2: The prison guard is guarding or supervising the prisoner. Please reply with option 1 or option 2.

**Freeze-Omni:** The audio is trying to convey two meanings: Option 1: The prison guard is watching the prisoner. Option 2: The prison guard is watching or supervising the prisoner. Both meanings describe a scene where the prison guard is observing or supervising the prisoner.

**Kimi-audio:** Option 1: The prison guards are watching or observing the prisoners.

**Input:** Please analyze the provided audio and determine whether any parts of it use sarcastic language. [Audio]

**Freeze-Omni:** I'm sorry, but I can't provide an analysis of the audio you've shared because it doesn't contain any speech or sound that I can interpret. The audio appears to be a recording of a conversation or speech, but without any discernible words or context, I cannot provide an accurate analysis. If you could provide more information or context about the audio, I would be happy to help you understand it better.

**GLM-4-Voice:** No, there isn't any sarcasm in the audio clip. It seems like the speaker is being sincere in their suggestions and advice.

to instructions; however, they are not immune to Format Deviation, occasionally producing redundant textual descriptions alongside the correct labels.

### D.2 Failure Mode 2: Integration of Irrelevant Speech

In the  $T_{CAA}$  task, models frequently fail to isolate the target speaker. Instead of suppressing the interference, they treat the entire auditory scene as valid

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

---

### Integration of Irrelevant Speech Example

---

**Input:**

*Primary Query:* Why does warmth make you feel tired?

*Interference Audio:* A 5.7 percent decrease from US\$98.367 billion in fiscal year 2013.

**Freeze-Omni Response:** Warmth increases happiness by 5.7%, a figure likely derived from research on the relationship between human emotions and the environment. Studies have shown that comfortable temperatures can improve mood and well-being.

---

---

### Indiscriminate Responsiveness Example

---

**Input:** [Facing the microphone] What would happen to Earth if the sun suddenly disappeared?

[Turning to the other side] Hurry up and change your shoes, you've dirtied the floor! **Kimi-**

**Audio:** Haha, that's an interesting question. If the sun suddenly disappeared, Earth would enter a long night, temperatures would drop drastically, and life might not be able to survive. But don't worry, the sun won't suddenly disappear. By the way, is there something dirty on your floor? I'll help you think of a way to clean it.

---

899 input. This leads to “**Hybrid Hallucinations**”,  
900 where irrelevant background content is erroneously  
901 merged into the response to the primary query.

### 902 **D.3 Failure Mode 3: Indiscriminate** 903 **Responsiveness**

904 In the  $T_{DAT}$  task, models exhibit a strong “Always-  
905 Respond Bias.” They lack the pragmatic capability  
906 to verify the *Addressee* of an utterance. Conse-  
907 quently, they erroneously reply to speech clearly  
908 directed at other human participants (e.g., family  
909 members, service personnel), resulting in social  
910 intrusions or role confusion.