

C³: A Bilingual Benchmark for Spoken Dialogue Models Exploring Challenges in Complex Conversations

Chengqian Ma^{1*†}, Wei Tao^{2*}, Yiwen Guo^{3‡}

¹Peking University, ²LIGHTSPEED, ³Independent Researcher
chengqianma@yeah.net, wtao@ieee.org, guoyiwen89@gmail.com

Dataset

Website

GitHub Repo

Abstract

Spoken Dialogue Models (SDMs) have recently attracted significant attention for their ability to generate voice responses directly to users' spoken queries. Despite their increasing popularity, there exists a gap in research focused on comprehensively understanding their practical effectiveness in comprehending and emulating human conversations. This is especially true compared to text-based Large Language Models (LLMs), which benefit from extensive benchmarking. Human voice interactions are inherently more complex than text due to characteristics unique to spoken dialogue. Ambiguity poses one challenge, stemming from semantic factors like polysemy, as well as phonological aspects such as heterograph, heteronyms, and stress patterns. Additionally, context-dependency, like omission, coreference, and multi-turn interaction, adds further complexity to human conversational dynamics. To illuminate the current state of SDM development and to address these challenges, we present a benchmark dataset in this paper, which comprises 1,079 instances in English and Chinese. Accompanied by an LLM-based evaluation method that closely aligns with human judgment, this dataset facilitates a comprehensive exploration of the performance of SDMs in tackling these practical challenges.

1 Introduction

Human conversations, particularly spoken dialogues, are inherently complex owing to ambiguous contexts (Solé and Seoane, 2014) that introduce uncertainties in communication. Ambiguity arises from phonological elements like pauses and intonation, as well as semantic factors such as lexical and syntactic ambiguity, as demonstrated in Figure 1(a)

and Figure 1(b). These ambiguities can lead to misinterpretations, necessitating careful understanding and response from participants. Recently, Spoken Dialogue Models (SDMs), such as GPT-4o-Audio-Preview (OpenAI, 2024b) and MooER-Omni (Xu et al., 2024), have become increasingly involved in human interactions. An SDM processes voice input and delivers voice response (Ji et al., 2024), and an effective SDM should be capable of recognizing and addressing challenging ambiguities to produce coherent replies.

Even in contexts without ambiguity, challenges can arise for SDMs. Speakers may omit previously mentioned entities or those understood as common knowledge, as illustrated in Figure 1(c). Additionally, speakers often use pronouns to refer to specific entities, as shown in Figure 1(d). Such context-dependency is significant in multi-turn interaction (Figure 1(e)). This requires SDMs to accurately identify and resolve omissions and coreferences to understand the intent of a speaker.

Despite the importance of handling ambiguity and context-dependency, it is yet unclear whether current SDMs are capable of addressing these challenges. To bridge the gap, we conduct an in-depth empirical study on the complexity of spoken dialogues and propose a novel dataset meticulously designed to study SDMs in handling complex dialogue situations with phonological ambiguity, semantic ambiguity, omission, coreference, and multi-turn interaction. Together with the dataset, we also propose an automatic LLM (Large Language Model)-based evaluation method to test the capability of SDMs, which aligns well with human evaluation results. After studying six popular SDMs, we deliver three findings to the community, including pointing out the different difficulties of five phenomena, two languages in spoken dialogues, and demonstrating the different advantages of the SDMs.

*Equal contribution.

†Work is done during internship at LIGHTSPEED.

‡Corresponding author.

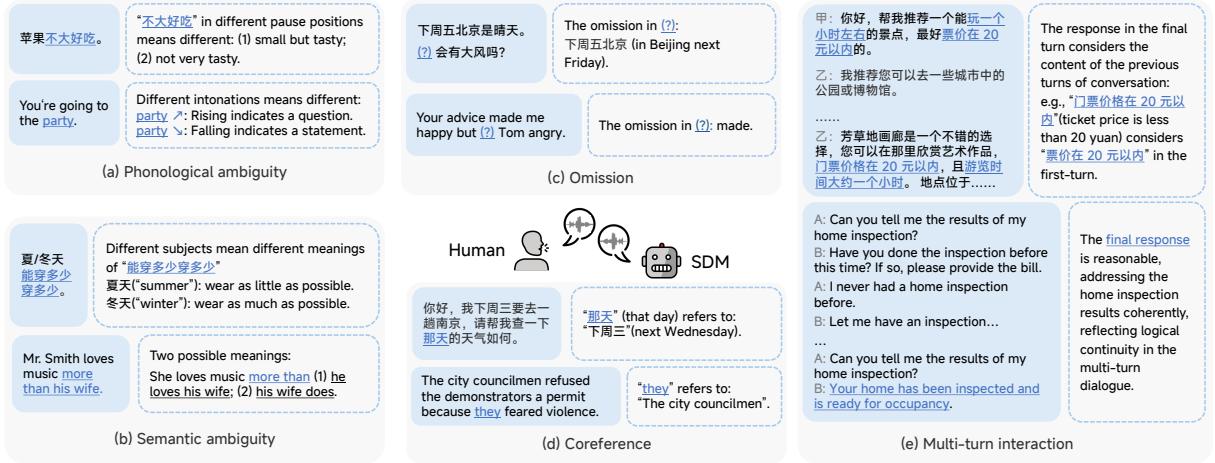


Figure 1: The structure and exemplars within the dataset. The subplots correspond to the sub-datasets of five phenomena. The blue boxes enclose the input for SDM, with some parts of the prompts omitted, while the corresponding outputs are within dashed boxes. Blue underlined text indicates the focal elements of interest, and gray text represents a segment of the prompt. The arrow indicates a rising or falling intonation. The (?) denotes an omitted sentence component. The » points to the referent of the pronoun. The ... represents the omitted dialogue.

2 Related Work

2.1 Spoken Dialogue Models

SDMs can be divided into earlier cascaded models and recent end-to-end models (Ji et al., 2024; Cui et al., 2024). The end-to-end model can directly understand and generate speech representations, while the cascaded model consists of Automatic Speech Recognition (ASR) (Malik et al., 2021; Yu et al., 2021; Hsu et al., 2021), Language Models (LMs), and Text-to-Speech (TTS) modules (Mehta et al., 2024; Popov et al., 2021). Cascaded models lose crucial audio features (e.g., intonation) during ASR processing, forcing LMs to work only on text. This prevents them from interpreting phonetic phenomena in raw audio. Consequently, it is natural that they underperform when there exists ambiguity in human speech. Our evaluation in this paper thus focuses on end-to-end models.

GPT-4o-Audio-Preview (OpenAI, 2024b) is the first end-to-end SDM that can generate fluent voice responses and analyze the emotions and intonations of the audio input. Since the implementation is not public, some open-source works, including LLaMA-Omni (Fang et al., 2024) and Freeze-Omni (Wang et al., 2024b), are explored and proposed. These works achieve low-latency spoken responses based on LLM in English conversation. To achieve real-time full-duplex dialogue capabilities for spoken large language models, Moshi (Défossez et al., 2024) is proposed, and it supports interruptions. To support more lan-

guages’ conversation, MooER-Omni (Xu et al., 2024), GLM-4-Voice (Zeng et al., 2024), VITA-Audio (Long et al., 2025), Step-Audio (Huang et al., 2025), Kimi-Audio (KimiTeam et al., 2025), and Qwen2.5-Omni (Xu et al., 2025) are proposed, and they show great ability in both English and Chinese spoken dialogues. We will study all these mentioned end-to-end SDMs in this paper.

2.2 Benchmarks and Datasets

To evaluate the capacities of SDMs, several benchmarks have been developed, each focusing on different aspects of audio (Hu et al., 2025; Qu et al., 2025). ADU-Bench (Gao et al., 2024) examines the cross-lingual and cross-skill spoken dialogue understanding capabilities of SDMs. Other benchmarks extend beyond language to include additional features. For instance, AIR-Bench (Yang et al., 2024) first evaluates the ability to understand various types of audio signals. SUPERB (Yang et al., 2021) focuses on speaker and emotion recognition. AudioBench (Wang et al., 2024a) assesses the ability to understand speech, audio scenes, and paralinguistic features. SD-Eval (Ao et al., 2024) evaluates SDMs’ responses to utterances with varying emotions, accents, ages, and background sounds. MMAU (Sakshi et al., 2024) includes perception and reasoning tasks across speech, sound, and music. VoiceBench (Chen et al., 2024) focuses on real-world scenarios involving speaker characteristics, environmental conditions, and content factors.

However, these benchmarks have some limita-

tions in four aspects:

(1) Most of the above benchmarks ignore the ambiguity. The only exception, ADU-Bench, considers it but does not cover phonological ambiguities such as press, heterograph, heteronym, and some semantic ambiguities, such as syntactic ambiguities.

(2) None of the aforementioned benchmarks consider comprehension difficulties caused by coreference and omission phenomena.

(3) All of the benchmarks listed include real-world spoken dialogue data from only one language (i.e., English). While ADU-Bench incorporates other languages, these datasets are translated from English, which means they may lack language-specific features, such as tone in Chinese.

(4) These benchmarks focus solely on single-turn dialogues, whereas multi-turn interactions are more common in spoken communication. They do not assess the ability of SDMs to handle multi-turn dialogues.

3 A New Benchmark for SDMs

The field of SDMs is rapidly evolving. Few studies could reveal the limitations and real performance of these models in handling complex ambiguity and context-dependency, which widely exist in human conversations.

In this section, we first empirically study each aspect of conversational complexity. Based on our empirical study, we design the dataset specifically.

3.1 The Complexity of Spoken Dialogues

To investigate the importance of the complex phenomena in spoken dialogue, we conduct a literature review, statistical analysis, and case study. The statistical analysis is performed using datasets in both English and Chinese. For English dialogues, we use CABank (MacWhinney and Wagner, 2010; Yaeger-Dror, 2007; Yaeger-Dror and Beaudrie, 2007). For Chinese dialogues, we use MagicData-RAMC (Yang et al., 2022) as the studied dataset. These datasets are selected because they are constructed based on real-world spoken dialogues rather than text-based dialogues. The reason for not using text-based dialogues is that they differ from spoken dialogues not only in form but also in content (Le Bigot et al., 2004; Placiński and Żywiczyński, 2023). Moreover, these two datasets are used in many top conferences (Guo et al., 2023; Li et al., 2021; Maheshwari et al., 2025) and jour-

nals (Xie et al., 2024; Landini et al., 2024).

3.1.1 Phonological Ambiguity

Phonological ambiguity can be classified into two types: segmental and supra-segmental. The former refers to discrete units that can be identified auditorily in the stream of speech. The latter refers to those features that extend over more than a single unit in an utterance (Ladefoged et al., 2006; Sharma, 2021). To make this section clearer, some terms are clarified as shown in Figure 2.

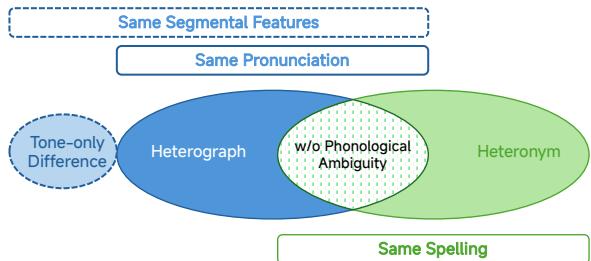


Figure 2: The relation between terms in Section 3.1.1.

Firstly, we investigate the segmental ambiguity.

Tone-only Difference: In spoken dialogue, especially in Chinese, the same segmental features do not convey the same meaning. For example, the Chinese phonetic alphabet *hao* can have four different tones, and each tone refers to a set of Chinese characters. The tone-only difference in pronunciation can lead to ambiguity. We use the tool ([pyp](#)) to count the situation in the dataset. We find that more than 99.25% Chinese characters from real-world dialogues have characters with the same phonetic alphabet but different tones, which can contribute to the ambiguity.

Heterograph: Some words with the same pronunciation may have different spellings. For example, in English, “night” and “knight”, “tail” and “tale” are heterographs¹. We use the tools ([pyp](#); [pro](#)) to count the situation in the dataset and find that there are 7.05% of the English words and 97.94% of the Chinese characters in dialogues are heterographs.

Heteronym: Some words with the same spelling also have different pronunciations. Of the 2,000 most frequently used English words, 9 of them are heteronyms² ([Parent, 2012](#)). A study ([Zhang and Chu, 2002](#)) reveals that there are at least 688 Chinese heteronyms. We use the tool ([pro](#)) to explore

¹A word whose pronunciation is the same, but whose spelling and meaning differ from another’s.

²A word having the same spelling as another but a different meaning, and often a different pronunciation.

English dialogues and find that at least 851 English heteronyms appear more than 42,315 times in real-world spoken dialogues.

The numbers above demonstrate the widespread existence of each phenomenon that can contribute to the segmental phonological ambiguity.

Secondly, we investigate the supra-segmental ambiguity. Pause, intonation, and stress are three supra-segmental features that can lead to ambiguity. Figure 1(a) shows two examples with different pause positions and with different intonations. The placement of stress in English can lead to ambiguity (Haolan, 2025). For example, “a green house” refers to a building with a roof and sides made of glass when the word “green” is stressed, but it denotes a building that is colored green when the word “house” is stressed.

3.1.2 Semantic Ambiguity

As shown in Figure 2, the words that have the same pronunciation and spelling do not have phonological ambiguity, but semantic ambiguity can exist in them. Semantic ambiguity can be classified into two types: lexical and syntactic.

Lexical Ambiguity: It means one word in a sentence can have two or more meanings. For example, in the sentence “They exchanged addresses in darkness”, the term “darkness” can be interpreted as either “in the absence of light” or “secretly”. A study on 11 business articles (Jannah, 2021) identified 27 instances of lexical ambiguity, demonstrating the widespread presence.

Syntactic Ambiguity: This means the situation where a sentence can be interpreted in more than one way due to its grammatical structure. Examples are shown in Figure 1(b). We use the tool (spa) to analyze the dataset and find that there are 15.79% of Chinese and 41.14% of English sentences with syntactic ambiguity in dialogues.

The numbers mentioned above demonstrate that semantic ambiguity often occurs in spoken dialogues.

3.1.3 Omission

Omission (also known as ellipsis) is common in spoken conversations. Two examples are shown in Figure 1. Moreover, subjects, verbs, and pronouns can be omitted in English dialogues (McShane, 2005). Statistically, a study (Glass, 2022) finds that the omission of verb objects is particularly common when describing routines. Another study (Su et al., 2019) shows that 52.4% of Chinese utterances also

have omissions in dialogues.

We use the tools (spa) for analysis and find that the incidence of subject omission (just one type of omission) in the dataset was 2.42% in the English subset and 16.51% in Chinese. It indicates the wide existence of omission in spoken dialogues.

3.1.4 Coreference

Pronouns can be used to refer to what is mentioned before in spoken dialogues, which is called coreference. Two examples are shown in Figure 1. A study (Su et al., 2019) shows that coreference occurs in 33.5% of Chinese daily conversations. Statistically, we use the tools (spa; jie) to count the number of pronouns and find that more than 69.60% English dialogues and 63.67% Chinese ones have coreference. Such high usage of pronouns suggests that coreference is frequent in spoken dialogues, either in English or Chinese.

3.1.5 Multi-turn Interaction

Commonly, one speaker interacts with the other in multiple turns in conversation (Lin et al., 2022). Statistically, in the Chinese dataset collected from human conversations, speakers switch an average of 270 times per dialogue. In the English dataset, the average number of speaker turns per dialogue is 331. Furthermore, the MagicData-RAMC (Yang et al., 2022) dataset, also collected from human conversations, has an average of 135 turns per dialogue. It indicates that multi-turn interactions are important in spoken conversation.

3.2 Benchmark Dataset Design

3.2.1 Pipeline

Firstly, we collect real-world spoken dialogues with each phenomenon mentioned in Section 3.1. To cover as many complex conversations as possible, we determine the standard for collection according to the relevant literature (details can be found in Appendix A.1). With the standard, we collect and extract speech data from web sources and some datasets (Quan et al., 2020; Yu, 2017; Shepherd, 2011; Kocijan et al., 2020; Zhu et al., 2020; Li et al., 2017).

After that, we transfer each real-world spoken dialogue to a unified question instance for the evaluation. We incorporate each dialogue with a prompt for the evaluation. Different instructions are designed for different phenomena. More details can be found in Section 3.2.2.

For example, the incorporated data instance is shown in Figure 3. To avoid the influence of irrelevant factors such as timbre and background music, we re-generate each speech data with the tool (Anastassiou et al., 2024), which makes the dialogue content have a unified timbre and no background noise.

To ensure the quality of the generated speech, we manually check each speech and replace incorrect instances with human voices. The reference answer in each instance is also manually produced.

Table 1: The number for each category of C_{data} . “zh” indicates Chinese, and “en” indicates English.

Category	Subcategory	zh	en
$C_{am-data}$	Phonological	37	29
	Semantic	118	51
$C_{con-data}$	Omission	70	102
	Coreference	60	540
	Multi-turn Interaction	38	34

We divide the C_{data} into $C_{am-data}$ (phonological and semantic ambiguity) to evaluate the ability on ambiguity and $C_{con-data}$ (omission, coreference, and multi-turn interaction) to evaluate the ability on context-dependency (thus the ambiguous dialogues are removed in $C_{con-data}$). The number of each category is presented in Table 1. There are 1,079 instances in the C^3 , comprising 1,586 audio-text paired samples. The number of audio-text pairs exceeds the number of instances because multi-turn dialogues contain multiple samples.

3.2.2 Data Instance Construction

To evaluate SDM’s performance across different complex phenomena, we design specialized instructions for each category. The complete set of instructions and annotation details are provided in Appendix A.2.

Phonological Ambiguity: The phonological ambiguity evaluates both the comprehension and generation capabilities of the SDM. For comprehension assessment, we instruct the SDM that the input contains potentially ambiguous phonological features and request a detailed interpretation. For generation assessment, we explicitly indicate the presence of incorrect phonological features (e.g., pauses, intonation) and prompt the SDM to generate a corrected response with appropriate prosodic markers.

Semantic Ambiguity: We inform the SDM that the meaning of the instance is unclear and instruct the SDM to provide a detailed explanation.

Omission: Our assessment focuses on two capabilities, (1) Detection: Instruct the SDM to identify if there are missing elements in the dialogues. (2) Completion: Inform that some content is omitted and instruct SDM to provide the completed sentence with the omission.

Coreference: We evaluate two related skills, (1) Detection: Instruct the SDM to identify if there is any coreference in the instance. (2) Resolution: Inform that the coreference phenomenon exists in the dialogue and instruct SDM to provide the coreference relationship.

Multi-turn Interaction:

After the real-world multi-turn dialogues, we repeat the initial question and instruct SDM to provide the identical answer as the previous one.

4 Experiment Settings and Evaluation

4.1 Experimental Settings

We select end-to-end SDMs instead of cascaded ones because the latter are unable to retain the phonological features such as press, pause, and intonation during ASR.

For the SDMs (i.e., Freeze-Omni, LLaMA-Omni, VITA-Audio, and MooER-Omni) that do not natively support multi-turn interaction, we concatenate the dialogue history in sequence before the current input in the evaluation. The real-time full-duplex model (i.e., Moshi) interrupts the input audio when provided with dialogue history, resulting in responses beyond the posed questions. As it cannot be evaluated in the same setting of multi-turn interaction as others, it is not fair to be compared and thus not chosen. Note that some models (i.e., LLaMA-Omni and Moshi) do not support Chinese; therefore, they are evaluated only in English.

4.2 LLM-based Evaluation

Preprocessing Most SDMs output both audio and corresponding text simultaneously. For the model (i.e., Moshi) without generating corresponding text, we convert the audio to text using Whisper (Radford et al., 2023).

Evaluation Method We adopt different methods for different categories in the dataset, C_{data} . For most tasks, except for generating audio with correct phonological features in the phonological ambiguity phenomenon, we evaluate the transcribed text from the audio. This is because phonological features in the response do not affect the comparison results with the reference, so evaluating the text

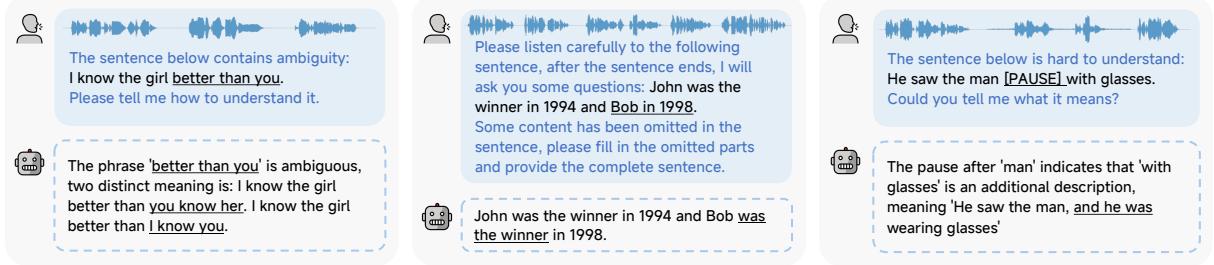


Figure 3: The structure of the data instance. The blue box contains input data in text and audio format, where blue text is the prompt and black text is the dialogue content being questioned. The dashed box contains the reference output, with the underlined portion highlighting the key element. “[PAUSE]” represents the pause in the audio.

Table 2: Accuracy (%) of different SDMs on the Chinese (“zh”) or English (“en”) dialogue data subset of C³.

Category	Freeze-Omni		GLM-4-Voice		GPT-4o- Audio-Prev.		Kimi-Audio		LLaMA- Omni		MooER- Omni		Qwen2.5- Omni		Step-Audio		VITA-Audio		Overall	
	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh	en
Phonological	16.22	8.62	18.92	27.59	29.73	53.45	20.27	46.55	15.52	20.27	18.97	10.34	27.03	48.28	22.97	29.31	8.11	31.03	20.44	28.97
Semantic	1.69	11.76	2.54	15.69	5.93	70.59	4.24	29.41	12.75	2.12	46.08	9.80	6.78	32.35	5.08	21.57	3.39	18.63	3.97	26.86
C _{am-data}	8.96	10.19	10.73	21.64	17.83	62.02	12.25	37.98	14.13	11.19	32.52	10.07	16.90	40.31	14.03	25.44	5.75	24.83	12.21	27.91
Omission	4.29	6.86	5.71	6.37	44.29	16.18	29.29	10.29	5.88	32.14	4.90	2.94	27.86	15.20	17.88	10.78	6.43	7.84	20.98	8.73
Coreference	10.83	47.22	16.67	68.98	54.17	91.11	40.00	87.41	56.94	32.50	36.02	24.63	55.83	68.15	50.83	57.31	33.33	74.81	36.77	61.26
Multi-turn	11.84	44.12	10.53	58.82	13.16	47.06	/	/	55.88	63.16	41.18	/	82.89	95.59	7.89	41.18	63.16	60.29	36.09	55.51
C _{con-data}	8.99	32.73	10.97	44.73	37.20	51.45	34.64	48.85	39.57	42.60	27.37	13.79	55.53	59.64	25.53	36.43	34.31	47.65	31.22	40.22
Overall	8.97	23.72	10.87	35.49	29.45	55.68	23.45	43.42	29.39	30.04	29.43	11.93	40.08	51.91	20.93	32.03	22.88	38.52	23.33	35.15

alone is sufficient. For the task of generating audio with correct phonological features, we evaluate the audio output manually, as it requires examining phonological features that cannot be captured by the transcribed text.

For the evaluation based on transcribed text, we design an automatic LLM-based evaluation method following the paradigm of LLM-as-a-judge (Gu et al., 2024). GPT-4o (OpenAI, 2024a) and DeepSeek-R1 (DeepSeek-AI et al., 2025) are selected as LLM judge due to their great performance in reasoning (DeepSeek-AI et al., 2025). LLM judges are used to compare the SDM output with the reference and determine the correctness. Moreover, we divide the evaluation task into smaller steps, instructing LLM judges with the prompts that are listed in the repository³. For the evaluation based on the audio, three human experts are required to label whether each of the SDM outputs is correct, and we use a voting strategy to make the final decision for each generated response.

The accuracy (i.e., the proportion of instances judged correct out of the total number of instances) is regarded as the metric.

Reliability Analysis To validate the reliability of our designed automatic evaluation method, we

first conduct a human evaluation on the generated responses by GPT-4o-Audio-Preview for C_{data}. Following best practice for the human evaluation (van der Lee et al., 2019), three human experts manually label whether each response is correct. If the labels from all experts are not the same, the majority label is chosen as the reference result.

After the human evaluation, we computed the Pearson (Cohen et al., 2009), Spearman (Xiao et al., 2016), and Kendall (Abdi, 2007) correlation coefficients to quantify the consistency between LLM judges and humans. All the coefficients’ values are more than 0.87 in either the English or Chinese subset, either for DeepSeek-R1 or GPT-4o as LLM judge (detailed numbers can be found in Appendix A.3). It demonstrates that LLM judges have high consistency with humans in each subset for the two LLMs. Moreover, all p-values of the correlation coefficients are less than 0.001, which means the consistency is significant. These statistical results validate the reliability of our automatic evaluation method.

5 Experimental Results and Findings

5.1 Experimental Results

To mitigate bias between DeepSeek-R1 and GPT-4o, we compute the average of their accuracies as

³<https://step-out.github.io/C3-web>

the final result, as shown in Table 2. The SDMs perform differently across different languages and phenomena.

As shown in Table 2, the gap between English and Chinese exceeds 8% across each phenomenon, indicating that SDMs exhibit varying capabilities depending on the language. Meanwhile, in the English subset, GPT-4o-Audio-Preview significantly outperforms other models, achieving an overall accuracy of 55.68%, while the average performance of all SDMs is only 35.15%. In contrast, in the Chinese subset, Qwen2.5-Omni stands out as the top-performing SDM, achieving an overall accuracy of 40.08%, while the average performance of all SDMs is 23.33%. The gap between Chinese and English in top performances and overall scores further highlights the differing strengths of SDMs across languages.

Within the same language, the performance gap between the strongest and weakest phenomena is over 9 times (for Chinese) and 6 times (for English), suggesting that SDMs vary in their strengths across different phenomena.

To illustrate the performance in handling different phenomena, radar charts are presented in Figure 4 and Figure 5. As shown in Figure 4, GPT-4o-Audio-Preview has the largest green area compared to the others, which validates its top performance. In the dimension of multi-turn interaction, GPT-4o-Audio-Preview scores significantly lower than Qwen2.5-Omni, indicating a weakness of the model. Although the overall scores of the top two SDMs, GPT-4o-Audio-Preview at 55.68% and Qwen2.5-Omni at 51.91%, are relatively close, each model exhibits distinct advantages. As shown in Figure 5, Qwen2.5-Omni excels in multi-turn interaction, with a sharp accuracy gap over other SDMs. The performance of the SDMs further highlights their varying strengths across different phenomena. Note that the detailed results from each LLM judge can be found in Appendix A.4.

To further investigate the ability to handle dialogues with omission and coreference, two tasks, including detection and completion (resolution), are provided for the evaluation. The final results of these two tasks are presented in Table 3.

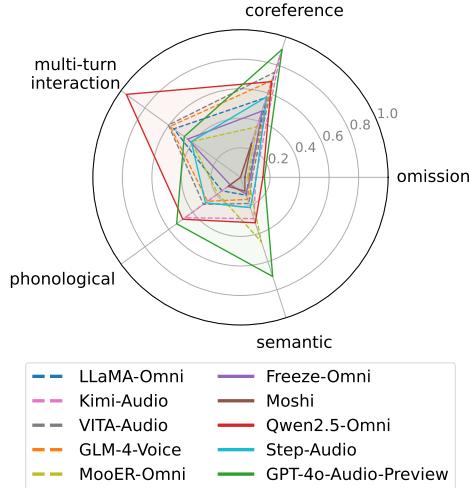


Figure 4: Radar charts depicting the accuracies of each SDM on the English subset of C_{data} .

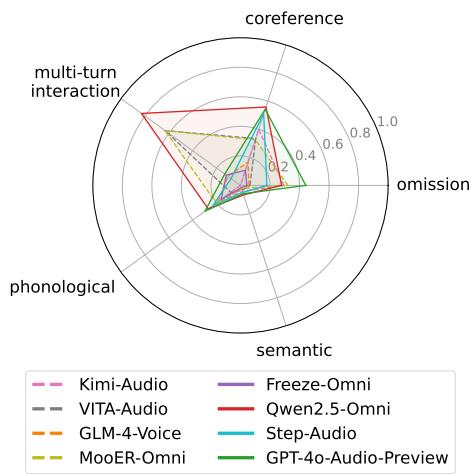


Figure 5: Radar charts depicting the accuracies of each SDM on the Chinese subset of C_{data} .

5.2 Experimental Findings

5.2.1 Ambiguity Is Difficult for SDMs Especially Semantic Ones in Chinese

As shown in Table 2, SDMs achieve overall accuracies of 12.21% (Chinese) and 27.91% (English) on $C_{\text{am-data}}$, significantly lower than the 31.22% (Chinese) and 40.22% (English) observed on $C_{\text{con-data}}$. The performance gap of over 10 percentage points in both languages suggests that ambiguity presents greater challenges for SDMs. Specifically, the overall accuracy in semantic ambiguity is only 3.97% in Chinese, compared to 26.86% in English. This pronounced disparity (exceeding a six-fold difference) underscores the challenges of processing semantic ambiguity in Chinese.

Additionally, the difference in accuracies for phonological ambiguity, 20.44% (Chinese) and

Table 3: Accuracy (%) of omission and coreference phenomena.

Phenomenon	Ability	Lang	Freeze-Omni	GLM-4-Voice	GPT-4o-Audio-Prev.	Kimi-Audio	LLaMA-Omni	MooER-Omni	Moshi	Qwen2.5-Omni	Step-Audio	VITA-Audio	Overall
Omission	Detection	zh	8.57	10.00	82.86	52.86	/	61.43	/	48.57	32.86	8.57	38.65
		en	8.82	4.90	13.73	12.75	6.86	6.86	3.92	11.76	7.84	6.86	8.57
	Completion	zh	0.00	1.43	5.71	5.71	/	2.86	/	7.14	2.86	4.29	3.75
		en	4.90	7.84	18.63	7.84	4.90	2.94	1.96	18.63	13.73	8.82	8.98
Coreference	Detection	zh	20.00	33.33	63.33	60.00	/	58.33	/	86.67	70.00	56.67	56.41
		en	57.59	83.89	95.37	97.04	78.52	25.93	35.37	70.56	63.33	87.59	69.61
	Resolution	zh	1.67	0.00	45.00	20.00	/	6.67	/	25.00	31.67	10.00	17.00
		en	36.85	54.07	86.85	77.78	35.37	46.11	13.89	65.74	51.30	62.04	52.17

28.97% (English), exceeds an 8% gap. The exception is MooER-Omni, which has a gap of less than 1.5 percentage points. This contrast highlights MooER-Omni’s cross-linguistic ability to handle phonological ambiguity.

5.2.2 Processing Omission Is the Most Difficult in Context-Dependency

Table 2 shows that, except for GPT-4o-Audio-Preview and Step-Audio in Chinese, all SDMs have the smallest accuracy when dealing with the omission phenomenon among $C_{\text{con-data}}$. This indicates that omission is the most difficult phenomenon for SDMs to handle in context-dependent dialogues.

Dealing with spoken dialogues with omission or coreference requires both detection and completion (or resolution). To investigate the abilities of SDMs at a granular level, we compare the accuracies of each ability as shown in Table 3. In omission, most SDMs have higher accuracy in detection than in completion. This suggests that although the omission is pointed out, the SDMs could not fully understand and thus complete the missing part. The exception is GLM-4-Voice, GPT-4o-Audio-Preview, Qwen2.5-Omni, Step-Audio, and VITA-Audio in English. With the prompt of the omission phenomenon, these five SDMs can complete more than what they can detect on their own. In coreference, the finding is similar. Most SDMs have higher accuracy in detection than resolution, indicating that although the coreference is pointed out, the SDMs cannot fully understand and resolve it. The exception is MooER-Omni in English, which performs better when pointing out coreference. The above findings teach us that pointing out the phenomenon in dialogue can be helpful for some SDMs, but most of them benefit only slightly.

We also find that most SDMs demonstrated higher accuracy in dealing with coreference resolution than omission completion. The different performances of these two phenomena can be in-

ferred: In the coreference phenomenon, both the pronoun and the antecedent are present in the sentence. The SDM can replace the pronoun with the antecedent by understanding the sentence. However, in the omission phenomenon, the omitted content is not present in the sentence. To complete the omitted parts, the SDM should not only understand each component’s meaning but also generate non-existent components. Therefore, resolving the omission phenomenon is more difficult for SDMs than resolving the coreference phenomenon.

Moreover, we observe that most SDMs exhibit low accuracies (below 65%) in multi-turn interactions, whereas Qwen2.5-Omni achieves significantly higher accuracy, with 82.89% for Chinese and 95.59% for English, outperforming the other models.

5.2.3 Complex Dialogues in Chinese Are More Difficult than Ones in English

As shown in Table 2, the overall accuracies for both $C_{\text{am-data}}$ and $C_{\text{con-data}}$ are higher in English (27.91% and 40.22%) than in Chinese (12.21% and 31.22%). The difference exceeds nine percentage points, indicating that, generally, SDMs perform better in English dialogues.

Specifically, in each phenomenon, the overall accuracy in English is higher, except for omission, suggesting that English phenomena are generally easier for SDMs than their Chinese counterparts.

Specifically, as shown in Table 2, most SDMs demonstrate higher accuracy in English than in Chinese. For instance, Freeze-Omni and GLM-4-Voice achieve accuracies of 23.72% and 35.49% in English, more than double their performance in Chinese (8.97% and 10.87%). This substantial gap highlights the need for enhanced cross-linguistic capabilities in current SDMs.

Summary: These findings suggest that the choice of SDM should depend on the specific situation, such as the phenomenon or language.

6 Conclusion

In this work, we introduce a new benchmark, C^3 , to evaluate SDMs' capabilities in handling various complex conversations. Our empirical study reveals five important phenomena in spoken dialogues that are not fully explored in previous works. With our designed dataset, C_{data} , and LLM-based evaluation method, SDMs can be evaluated more comprehensively. Furthermore, we conduct experiments on six SDMs. The results point out different difficulties in processing these complex phenomena in different languages.

We believe that C^3 , including real and complex challenges in spoken dialogues, is helpful for researchers to achieve natural and intelligent spoken interaction with humans. In the future, we will collect more language dialogues into C_{data} .

Limitations

There are two limitations to this work: First, the five complex phenomena discussed in this paper are not limited to English and Chinese; they have significant potential for other languages. Second, there is potential bias among human experts who evaluate the outputs of SDMs. To mitigate this bias, we employ a voting mechanism.

References

aparrish/pronouncingpy: A simple interface for the cmu pronouncing dictionary. <https://github.com/aparrish/pronouncingpy/>.

fxsjy/jieba: Chinese text segmentation. <https://github.com/fxsjy/jieba>.

mozillazg/python-pinyin: Chinese character pinyin conversion tool (python version). <https://github.com/mozillazg/python-pinyin>.

spacy · industrial-strength natural language processing in python. <https://spacy.io/>.

Hervé Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yiping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. 2024. *Seed-tts: A family of high-quality versatile speech generation models*. *arXiv Preprint*, abs/2406.02430.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. *Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words*. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Weiwei Dai. 2021. On the syntactic structure of chinese ambiguity sentences. *Open Access Library Journal*, 8(10):1–10.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, f Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. *arXiv Preprint*, abs/2501.12948.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard

Grave, and Neil Zeghidour. 2024. *Moshi: a speech-text foundation model for real-time dialogue*. *arXiv Preprint*, abs/2410.00037.

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

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2024. *Benchmarking open-ended audio dialogue understanding for large audio-language models*. *arXiv Preprint*, abs/2412.05167.

Lelia Glass. 2022. English verbs can omit their objects when they describe routines. *English Language & Linguistics*, 26(1):49–73.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. *A survey on llm-as-a-judge*. *arXiv Preprint*, abs/2411.15594.

Zishan Guo, Linhao Yu, Minghui Xu, Renren Jin, and Deyi Xiong. 2023. *CS2W: A chinese spoken-to-written style conversion dataset with multiple conversion types*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3962–3979. Association for Computational Linguistics.

Yang Haolan. 2025. *A brief analysis of phonological ambiguity in language: A comparison between chinese and english*.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. *Hubert: How much can a bad teacher benefit ASR pre-training?* In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6533–6537. IEEE.

He Hu, Yucheng Zhou, Lianzhong You, Hongbo Xu, Qianning Wang, Zheng Lian, Fei Richard Yu, Fei Ma, and Laizhong Cui. 2025. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv Preprint*, abs/2502.04424.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, Peng Liu, Ruihang Miao, Wang You, Xi Chen, Xuerui Yang, Yechang Huang, Yuxiang Zhang, Zheng Gong, Zixin Zhang, Hongyu Zhou, Jianjian Sun, Brian Li, Chengting Feng, Changyi Wan, Hanpeng Hu, Jianchang Wu, Jiangjie Zhen, Ranchen Ming, Song Yuan, Xuelin Zhang, Yu Zhou, Bingxin Li, Buyun Ma, Hongyuan Wang, Kang An, Wei Ji, Wen Li, Xuan Wen, Xiangwen Kong, Yuankai Ma, Yuanwei Liang, Yun Mou, Bahtiyar Ahmidi, Bin Wang, Bo Li, Changxin Miao, Chen Xu, Chenrun Wang, Dapeng Shi, Deshan Sun, Dingyuan Hu, Dula Sai, Enle Liu, Guanzhe Huang, Gulin Yan, Heng Wang, Haonan Jia, Haoyang Zhang, Jiahao Gong, Junjing Guo, Jiashuai Liu, Jiahong Liu, Jie Feng, Jie Wu, Jiaoren Wu, Jie Yang, Jinguo Wang, Jingyang Zhang, Junzhe Lin, Kaixiang Li, Lei Xia, Li Zhou, Liang Zhao, Longlong Gu, Mei Chen, Menglin Wu, Ming Li, Mingxiao Li, Mingliang Li, Mingyao Liang, Na Wang, Nie Hao, Qiling Wu, Qinyuan Tan, Ran Sun, Shuai Shuai, Shaoliang Pang, Shiliang Yang, Shuli Gao, Shanshan Yuan, Siqi Liu, Shihong Deng, Shilei Jiang, Sitong Liu, Tiancheng Cao, Tianyu Wang, Wenjin Deng, Wuxun Xie, Weipeng Ming, and Wenqing He. 2025. *Step-audio: Unified understanding and generation in intelligent speech interaction*. *arXiv Preprint*, abs/2502.11946.

Nur Jannah. 2021. *Lexical and syntactic ambiguity in the business news of BBC News*. Ph.D. thesis, Universitas Islam Negeri Maulana Malik Ibrahim.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. *Wavchat: A survey of spoken dialogue models*. *arXiv Preprint*, abs/2411.13577.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. 2025. *Kimi-audio technical report*. *arXiv Preprint*, abs/2504.18425.

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. *A review of winograd schema challenge datasets and approaches*. *arXiv Preprint*, abs/2004.13831.

Peter Ladefoged, Keith Johnson, and Peter Ladefoged. 2006. *A course in phonetics*, volume 3. Thomson Wadsworth Boston.

Federico Landini, Mireia Díez, Themis Stafylakis, and Lukás Burget. 2024. *Diaper: End-to-end neural diarization with perceiver-based attractors*. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:3450–3465.

Rim Mohammed Abdalla Lasheiky. 2024. Semantic ambiguity in english: A review on lexical, structural, and scope challenges in communication. *AJASHSS*, pages 388–395.

Ludovic Le Bigot, Eric Jamet, and Jean-François Rouet. 2004. Searching information with a natural language dialogue system: a comparison of spoken vs. written modalities. *Applied ergonomics*, 35(6):557–564.

Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Man-Wai Mak, Brian Mak, Xunying Liu, and Helen Meng.

2021. A comparative study of acoustic and linguistic features classification for alzheimer’s disease detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6423–6427. IEEE.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. *Dailydialog: A manually labelled multi-turn dialogue dataset*. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. *Duplex conversation: Towards human-like interaction in spoken dialogue systems*. In *KDD ’22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 3299–3308. ACM.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, Haoyu Cao, Ke Li, Rongrong Ji, and Xing Sun. 2025. *Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model*. *arXiv Preprint*, abs/2505.03739.

Brian MacWhinney and Johannes Wagner. 2010. Transcribing, searching and data sharing: The clan software and the talkbank data repository. *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion*, 11:154.

Gaurav Maheshwari, Dmitry Ivanov, Théo Johannet, and Kevin El Haddad. 2025. Asr benchmarking: Need for a more representative conversational dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. *Automatic speech recognition: a survey*. *Multim. Tools Appl.*, 80(6):9411–9457.

Marjorie J McShane. 2005. *A theory of ellipsis*. Oxford University Press.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. *Matcha-tts: A fast TTS architecture with conditional flow matching*. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11341–11345. IEEE.

OpenAI. 2024a. gpt-4o. <https://openai.com/index/hello-gpt-4o/>.

OpenAI. 2024b. Gpt-4o-audio-preview api. <https://platform.openai.com/docs/guides/audio>.

Kevin Parent. 2012. The most frequent english homonyms. *RELC Journal*, 43(1):69–81.

Marek Placiński and Przemysław Żywiczyński. 2023. Modality effect in interactive alignment: Differences between spoken and text-based conversation. *Lingua*, 293:103592.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail A. Kudinov. 2021. *Grad-tts: A diffusion probabilistic model for text-to-speech*. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv Preprint*, abs/2503.21614.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. *Risawoz: A large-scale multi-domain wizard-of-oz dataset with rich semantic annotations for task-oriented dialogue modeling*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 930–940. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. *Robust speech recognition via large-scale weak supervision*. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Jennifer Rodd. 2018. Lexical ambiguity. *Oxford handbook of psycholinguistics*, pages 120–144.

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

Lok Raj Sharma. 2021. Significance of teaching the pronunciation of segmental and suprasegmental features of english. *Interdisciplinary Research in Education*, 6(2):63–78.

A Shepherd. 2011. Want to talk about it? a minimalist analysis of subject omission in colloquial english. *Unpublished MRes thesis, submitted to the University of Southampton*.

Ricard V. Solé and Luís F. Seoane. 2014. *Ambiguity in language networks*. *arXiv Preprint*, abs/1402.4802.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Peng-wei Hu, Cheng Niu, and Jie Zhou. 2019. [Improving multi-turn dialogue modelling with utterance rewriter](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 22–31. Association for Computational Linguistics.

Abdul Karim Taha. 1983. Types of syntactic ambiguity in english. *International Review of Applied Linguistics in Language Teaching*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zuhuan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2024a. [Audиobench: A universal benchmark for audio large language models](#). *arXiv Preprint*, abs/2406.16020.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024b. [Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM](#). *arXiv Preprint*, abs/2411.00774.

Chengwei Xiao, Jiaqi Ye, Rui Máximo Esteves, and Chunming Rong. 2016. [Using spearman’s correlation coefficients for exploratory data analysis on big dataset](#). *Concurr. Comput. Pract. Exp.*, 28(14):3866–3878.

Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2024. [Domain generalization via aggregation and separation for audio deepfake detection](#). *IEEE Trans. Inf. Forensics Secur.*, 19:344–358.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. [Qwen2.5-omni technical report](#). *arXiv Preprint*, abs/2503.20215.

Junhao Xu, Zhenlin Liang, Yi Liu, Yichao Hu, Jian Li, Yajun Zheng, Meng Cai, and Hua Wang. 2024. [Mooer: Llm-based speech recognition and translation models from moore threads](#). *arXiv Preprint*, abs/2408.05101.

Malcah Yaeger-Dror. 2007. [Cabank english callfriend northern us corpus](#).

Malcah Yaeger-Dror and Alan Beaudrie. 2007. [Cabank english callfriend southern us corpus](#).

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. [Air-bench: Benchmarking large audio-language models via generative comprehension](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1979–1998. Association for Computational Linguistics.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. [SUPERB: speech processing universal performance benchmark](#). In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 1194–1198. ISCA.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. 2022. [Open source magicdata-ramc: A rich annotated mandarin conversational\(ramc\) speech dataset](#). In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 1736–1740. ISCA.

FU Yu. 2017. A formal syntactic study of np-ellipsis in mandarin chinese. *Journal of Foreign Languages*, 40(1):13–23.

Jiahui Yu, Chung-Cheng Chiu, Bo Li, Shuo-Yiin Chang, Tara N. Sainath, Yan Zhang He, Arun Narayanan, Wei Han, Anmol Gulati, Yonghui Wu, and Ruoming Pang. 2021. [Fastemit: Low-latency streaming ASR with sequence-level emission regularization](#). In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6004–6008. IEEE.

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

Zirong Zhang and Min Chu. 2002. A statistical approach for grapheme-to-phoneme conversion in chinese. *JOURNAL OF CHINESE INFORMATION PROCESSING*, 16(3):40–46.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. [Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset](#). *Trans. Assoc. Comput. Linguistics*, 8:281–295.

A Appendix

A.1 Deatils of Dataset Design

Based on our empirical study, we optimize the data construction process and introduce criteria to

ensure the quality of the dataset (Section 3.2). The specific filtering criteria are as follows:

Phonological Ambiguity: The intended meaning of the instances is ambiguous, caused by phonological features including heteronym, heterograph, stress, intonation, pause, and tone-only difference.

Semantic Ambiguity: The sentence contains lexical or syntactic ambiguities. More specifically, lexical ambiguity means it contains polysemous words, while syntactic ambiguity means the phrase or sentence can be parsed in more than one way grammatically.

Omission: The instance omits part of the utterance, and the omission must be inferred from the surrounding context or common knowledge. More specifically, the omission can be the word: subject, verb, or object, commonly understood by speakers.

Coreference: The instance uses pronouns (e.g., he, she, that) or phrases (e.g., the former, the boy) to refer to specific entities mentioned in the dialogue.

Multi-turn Interaction: The instance includes at least five turns with speaker alternation, and semantic dependencies are across dialogue turns.

The dataset design process for each phenomenon is described in detail below.

A.1.1 Phonological Ambiguities

Understanding In the Chinese subset, ambiguities arise from four types of characteristics: pause, heteronym, heterograph, and syllable with different tones. In the English subset, ambiguities result from four types of characteristics: heterograph, pause, stress, and intonation. During the manual review of the TTS-generated audio, we find that the audio quality in the Chinese dataset is poor for the pause and heteronym characteristics, and in the English dataset for the pause and stress characteristics. Consequently, these four parts of the data are re-recorded manually. The final bilingual dataset contains ambiguous questions with audio and text modalities, and the corresponding textual reference answers. The form of the other dataset is also the same.

Generating In addition, we develop data that tests SDM’s ability to generate dialogues with phonetic characteristics. This data is derived from the understanding phonological ambiguities dataset. An exception is heterographs, which are not included in the generation evaluation, as the reference audio for homophones remains the same. The ambiguous sentences remain unchanged, but the

prompts are changed in different characteristics and different languages. For pauses, stresses, and intonations, the evaluation involves inputting the meaning of the ambiguous sentence and assessing whether the SDM produces these phonological features appropriately. For heteronyms and syllables with different tones, the evaluation involves inputting sentences with incorrect pronunciations and assessing whether the SDM can correct them based on context.

A.1.2 Semantic Ambiguities

This subset of $C_{am\text{-}data}$ examines the ability of SDM to process semantic ambiguities. We first identify the types of semantic ambiguities to collect data from relevant literature (Rodd, 2018; Taha, 1983; Dai, 2021; Lasheiky, 2024). Subsequently, we manually gather data from various websites, including ambiguous sentences and their interpretations. The data are then organized using a standardized prompt, instructing the SDM to provide interpretations of the ambiguous sentences. Finally, the data are converted from text to audio and checked manually for quality.

The Chinese dataset encompasses ambiguities arising from unclear pronominal reference, polysemy, unclear modification scope, unclear part of speech, and unclear subject-object relationship. To ensure optimal audio quality, the first two instances are enhanced with human-voiced recordings, and the remaining are generated by TTS. The English dataset includes lexical ambiguities stemming from unclear parts of speech and polysemy, as well as syntactic ambiguities resulting from unclear pronominal reference and unclear modification scope.

A.1.3 Omission

This section of the dataset examines SDM’s ability to understand comprehension difficulties in dialogues caused by the omission phenomenon. The Chinese dataset is based on the RISAWOZ dataset (Quan et al., 2020), a text dataset specifically designed to study coreference and omission phenomena. The selected portion of the RISAWOZ dataset contains multi-turn dialogues with 1, 3, or 5 sentences, and provides annotations for omission and coreference in each dialogue. We retain the segments of each multi-turn dialogue from the beginning up to the point where omission occurs and add prompts to query SDM to construct the dataset. The English portion is manually extracted

from relevant literature (Yu, 2017; Shepherd, 2011), and data containing the omission phenomenon is constructed with corresponding prompts to query SDM. Unlike the Chinese dataset, which is in the form of multi-turn dialogues, the English dataset is in the form of a single sentence.

For both the Chinese and English datasets, reference answers that supplement the omitted content are provided to enable comparison with SDM’s responses. We construct two questions in the prompt for the data: The first question asks the SDM to determine whether there is an omission phenomenon in the input audio, and the second question informs the SDM of the existence of an omission phenomenon in the input and requests the SDM to complete the omitted content. The two questions are independent of each other. To prevent overlap with the ambiguous contexts dataset, we exclude the omission phenomenon that would cause ambiguity during data selection.

A.1.4 Coreference

This section of the dataset assesses SDM’s ability to comprehend difficulties in dialogues arising from the coreference phenomenon. The Chinese dataset is based on the RISAWOZ dataset (Quan et al., 2020) and employs a similar methodology to the omission section, resulting in multi-turn dialogue data instances, each comprising 1, 3, or 5 sentences. The dataset includes reference answers that resolve coreference by replacing pronouns with their referents, thereby eliminating the coreference phenomenon, to serve as a standard for comparing SDM’s responses.

The English dataset is constructed based on the Winograd Schema Challenge dataset (Kocijan et al., 2020). Each data instance comprises a sentence and a multiple-choice question targeting the referent of a pronoun, with two potential answers provided. The dataset also includes the correct answer to each coreference question. The referents of these pronouns are easily confused, necessitating a deep understanding of the sentence’s meaning as well as robust commonsense knowledge and reasoning capacity to determine the correct answer. To ensure the dataset’s quality and clarity regarding the pronouns in question, we filter out instances where the pronoun appears more than once in the sentence.

Consistent with the omission dataset, to avoid overlap with ambiguous contexts, we exclude coreference phenomena that could introduce ambiguity. We then task the SDM with addressing two dis-

tinct queries: first, to verify the presence of the coreference phenomenon, and second, to deliver the outcomes following coreference resolution.

A.1.5 Multi-turn Interaction

To evaluate the model’s ability to track conversation history, we ensured that the assessment of SDM is conducted in a multi-turn conversational format. The criteria for collecting data are that the dialogues must be multi-turn. The Chinese dataset is based on the CrossWoz dataset (Zhu et al., 2020), which covers multiple domains including tourist attractions, hotels, restaurants, subways, and taxis. The English dataset is derived from the DailyDialog dataset (Li et al., 2017), which is artificially constructed with minimal noise and encompasses a variety of everyday conversational scenarios. Since our method of evaluating SDM involves posing the first question in the dialogue, we ensured that the first sentence of each dialogue is a question when filtering out the dataset. Defining a single input to the SDM and its corresponding response as one turn of dialogue, the Chinese dataset features dialogues with a maximum of 16 turns and an average of 9.68 turns, whereas the English dataset has a maximum of 9 turns and an average of 6.21 turns.

In the dataset, only the content input by the user to SDM is provided, while the responses of SDM are generated by the SDM being evaluated. This subset of $C_{\text{con-data}}$ examines SDM’s ability to remember the content of multi-turn dialogues and to utilize the conversation history to generate current responses when processing dialogues. Therefore, after the dialogue concludes, we revisit the first question in the dialogue and request that SDM respond to that question again. If the final response provided by SDM is consistent with the initial response and the intervening question-and-answer content, it is considered to have good capability to process multi-turn interaction. During the evaluation process, if the SDM being evaluated only provides single-turn dialogue capability, we concatenate the previous question-and-answer pairs to manually construct the conversation history for each input.

A.2 Detailed Structure and Exemplars of Dataset

The annotation details for each phenomenon are as follows:

Phonological Ambiguity: Different meanings are annotated for each sentence, along with the correct

phonological features, including pronunciation, intonation, stress position, and pause position.

Semantic Ambiguity: Semantic ambiguity is divided into lexical and syntactic ambiguity. For lexical ambiguity, different meanings of the same word are annotated. For syntactic ambiguity, different interpretations of the same semantic structure are annotated.

Omission: The omitted parts are annotated based on context and common sense.

Coreference: The word or phrase referred to by the pronoun is annotated based on context and common sense.

Multi-turn Interaction: Multi-turn dialogues do not require annotation, as the reference answer is determined by the SDM’s output.

To provide a more detailed illustration of the contents of each subset within the dataset, Figure 6 - 9 have been presented. The gray text denotes the invariant segments integral to the dataset’s construction, immutable irrespective of variations in the data samples. The underlined blue-highlighted segments indicate the focal areas examined by the SDM, while the non-underlined blue-highlighted portions distinguish the roles of different participants in multi-turn dialogues.

A.3 Correlation Analysis

To further illustrate the correlation between LLMs and human evaluations, Table 4 presents three correlation coefficients, while Table 5 shows their corresponding p - values.

A.4 Detailed Evaluation Results for DeepSeek-R1 and GPT-4o

To illustrate the experimental results of different SDMs on the C³, evaluated separately by DeepSeek-R1 and GPT-4o, Table 6 and Table 7 present the detailed results corresponding to those summarized in Table 2.

To present the radar charts of evaluation results for different SDMs on the Chinese and English sections of C³, evaluated respectively by DeepSeek-R1 and GPT-4o, we include Figure 10 - 13 , which correspond to the summaries shown in Figure 4 and Figure 5.

To illustrate the experimental results of different SDMs on the omission and coreference sections, evaluated respectively by DeepSeek-R1 and GPT-4o, Table 8 and Table 9 are presented, corresponding to the summary in Table 3.

Category		Data instance	Reference answer
Ambiguity	Phonological ambiguity	Generation	Here is a sentence: <u>He is your teacher?</u> The intended meaning is: <u>Is he your teacher?</u> Please read it out with the correct intonation.
		Heterograph	The sentence below is hard to understand: <u>I am a nice man.</u> I always help others and spread kindness. What is meaning of the phrase 'a nice man'?
		Intonation	The sentence below is hard to understand: <u>You're going to the party?</u> Could you tell me what it means?
		Pause	The sentence below with pause is hard to understand: <u>He saw the man / with glasses.</u> Could you tell me what it means?
		Stress	The sentence below with emphasis is hard to understand: <u>I <u>can</u> help you.</u> Could you tell me what it means?
		Lexical	The sentence below contains ambiguity: <u>They exchanged addresses in darkness.</u> Please tell me how to understand it.
	Semantic ambiguity	Syntactic	The sentence below contains ambiguity: <u>Mr. Smith loves music <u>more than his wife.</u></u> Please tell me how to understand it.

Figure 6: The figure delineates the structure and exemplars of the English Ambiguous subset within the dataset.

Category		Data instance	Reference answer
Ambiguity	Phonological ambiguity	Heteronym	下面这个句子中有一个字的发音存在错误: <u>母亲背 bēi</u> 着儿子去捡破烂。这个句子想表达的意思是: 母亲把年幼的孩子背在背上 <u>去捡破烂</u> 。请用正确的读音读出这句话。
		Pause	下面有一个断句可能不正确的句子: <u>这个苹果不大好吃</u> 。这个句子想表达的意思是: <u>这个苹果虽然小,但是好吃</u> 。
		Syllable with different tones	下面句子中部分发音存在错误: 伐木工人在森林里辛苦 <u>看</u> 书, 是为了更好地养家糊口。请用正确的语音读出这句话。
		Heteronym	可以请你帮我一个忙吗? 我不太理解下面这段话, 请你帮我给出合理的解释: 小时候 <u>母亲背 bēi</u> 着儿子去捡破烂, 长大了以后 <u>母亲背 bēi</u> 着儿子去捡破烂。
		Heterograph	可以请你帮我一个忙吗? 我不太理解下面这段话, 请你帮我给出合理的解释: 没有凳子时应该 <u>站着</u> 吃饺子; 有醋时应该 <u>蘸着</u> 吃饺子。
		Pause	可以请你帮我一个忙吗? 我不太理解下面这段话, 请你帮我给出合理的解释: 小张必须 <u>5月5日, 前去报到</u> 。小王必须 <u>5月5日前, 去报到</u> 。
		Syllable with different tones	可以请你帮我一个忙吗? 我不太理解下面这段话, 请你帮我给出合理的解释: 兄弟俩人哥哥在 <u>花市</u> 里工作, 弟弟在 <u>画室</u> 里工作。
		Semantic ambiguity	可以请你帮我一个忙吗? 我不太理解下面这段话, 请你帮我给出合理的解释: 重庆的夏天 <u>能穿多少穿多少</u> , 重庆的冬天 <u>能穿多少穿多少</u> 。
			冬天多穿, 夏天少穿。

Figure 7: The figure delineates the structure and exemplars of the Chinese Ambiguous subset within the dataset.

Table 4: Correlation coefficients between LLM evaluation results and human assessment results

Model	Pearson	Spearman	Kendall	Language
DeepSeek-R1	0.8969	0.8969	0.8969	Chinese
GPT-4o	0.8886	0.8886	0.8886	Chinese
DeepSeek-R1	0.8739	0.8739	0.8739	English
GPT-4o	0.8940	0.8940	0.8940	English

Table 5: p-values for correlation coefficients

Model	Pearson p-value	Spearman p-value	Kendall p-value	Language
DeepSeek-R1	$< 10^{-115}$	$< 10^{-115}$	$< 10^{-57}$	Chinese
GPT-4o	$< 10^{-109}$	$< 10^{-109}$	$< 10^{-56}$	Chinese
DeepSeek-R1	$< 10^{-237}$	$< 10^{-237}$	$< 10^{-126}$	English
GPT-4o	$< 10^{-264}$	$< 10^{-264}$	$< 10^{-132}$	English

Category	Data instance	Reference answer
Context-dependency	Omission-detection Please listen carefully to the following sentence, after the sentence ends, I will ask you some questions: <u>Your advice made me happy but Tom angry.</u> Is there any omission of content in the sentence? Please listen carefully to the following sentence, after the sentence ends, I will ask you some questions: <u>Your advice made me happy but Tom angry.</u>	Yes.
	Omission-completion Some content has been omitted in the sentence, please fill in the omitted parts and provide the complete sentence. Please listen carefully to the following sentence, after the sentence ends, I will ask you some questions: <u>The city councilmen refused the demonstrators a permit because they feared violence.</u>	<u>Your advice made me happy but your advice made Tom angry.</u>
	Coreference-detection Are there any instances of pronouns being used to refer to nouns or noun phrases in the sentences above? <u>The city councilmen refused the demonstrators a permit because they feared violence.</u>	Yes, there's a pronoun used in the sentence.
	Coreference-resolution What does the word 'they' refer to: The city councilmen or The demonstrators?	The city councilmen.
	Multi-turn interaction A: Do you have any <u>plan to buy a house</u> in this city? B: I <u>don't have personal plans</u> or interests, so I wouldn't be buying a house. ... A: Do you have <u>any</u> plan to <u>buy a house</u> in this city? If this question has been raised before, please directly provide the previous answer.	I don't have personal plans, but I can provide information or help find resources related to buying a house in the city if you need it.

Figure 8: The figure delineates the structure and exemplars of the English Context-Dependency subset within the dataset.

Category	Data instance	Reference answer
Context-dependency	Omission-detection 我有一道考察语意理解的题目如下, 请听完后, 回答我的问题: 甲说: “下周五北京是晴天吗?” 乙说: “那天是多云转晴的天气。” 甲说: “ <u>会有大风吗?</u> ” 请问对话的最后一句中, 是否存在省略现象?	存在。
	Omission-completion 我有一道考察语意理解的题目如下, 请听完后, 回答我的问题: 甲说: “下周五北京是晴天吗?” 乙说: “那天是多云转晴的天气。” 甲说: “ <u>会有大风吗?</u> ” 上面这段对话的最后一句中存在省略现象, 请补全被省略的内容, 给出省略内容被补全后的完整语句。	下周五北京会有大风吗?
	Coreference-detection 我有一道考察语意理解的题目如下, 请听完后, 回答我的问题: 甲说: “你好, 我下下周三要去一趟南京, 请帮我查一下 <u>那天</u> 的天气如何?” 请问对话的最后一句中, 是否存在使用代词指代具体内容的现象?	存在。
	Coreference-resolution 我有一道考察语意理解的题目如下, 请听完后, 回答我的问题: 甲说: “你好, 我下下周三要去一趟南京, 请帮我查一下 <u>那天</u> 的天气如何?” 上面这段对话的最后一句中, 存在代词指代具体内容的现象, 请使用代词指代的内容替换代词, 给出代词被替换后的完整语句。 甲说: 我想在 <u>北京</u> 玩一玩, 可以给我 <u>推荐</u> 一个 <u>景点</u> 吗? 乙说: 推荐你去 <u>南锣鼓巷</u> , 那是北京的著名 <u>胡同</u> , 可以让你体验北京传统文化和特色小吃。 甲说: 我想在 <u>北京</u> 玩一玩, 可以给我 <u>推荐</u> 一个 <u>景点</u> 吗? 如果之前出现过该问题, 则直接给出之前的答案。	你好, 我下下周三要去一趟南京, 请帮我查一下下周三的天气如何? 推荐你去南锣鼓巷, 这里是北京最著名的胡同之一, 是体验北京传统文化和特色小吃的好地方。
	Multi-turn interaction	

Figure 9: The figure delineates the structure and exemplars of the Chinese Context-Dependency subset within the dataset.

Table 6: Accuracy (%) of different SDMs on the Chinese (“zh”) or English (“en”) dialogue data subset of C³ (DeepSeek-R1).

Category	Freeze-Omni		GLM-4-Voice		GPT-4o- Audio-Prev.		Kimi-Audio		LLaMA- Omni		MooER- Omni		Moshi		Qwen2.5- Omni		Step-Audio		VITA-Audio		Overall	
	zh	en	zh	en	zh	en	zh	en	en	zh	en	en	zh	en	zh	en	zh	en	zh	en	zh	en
Phonological	16.22	6.90	18.92	20.69	29.73	44.83	18.92	44.83	17.24	18.92	20.69	10.34	27.03	37.93	21.62	27.59	8.11	27.59	19.93	25.86		
Semantic	1.69	11.76	1.69	11.76	4.24	68.63	2.54	19.61	9.80	2.54	37.25	7.84	5.93	21.57	5.93	17.65	2.54	17.65	3.39	22.35		
C_{am}-data	8.96	9.33	10.31	16.23	16.98	56.73	10.73	32.22	13.52	10.73	28.97	9.09	16.48	29.75	13.78	22.62	5.33	22.62	11.66	24.11		
Omission	4.29	7.84	4.29	6.86	45.71	16.67	27.14	12.75	6.86	32.86	7.84	4.90	25.71	14.71	17.14	11.76	5.71	7.84	20.36	9.80		
Coreference	13.33	48.15	20.00	67.96	55.00	89.81	40.00	85.56	55.00	28.33	35.74	29.81	50.00	65.74	53.33	55.74	33.33	73.70	36.67	60.72		
Multi-turn	7.89	32.35	10.53	58.82	10.53	47.06	/	/	47.06	60.53	38.24	/	84.21	97.06	5.26	32.35	52.63	52.94	33.08	50.74		
C_{con}-data	8.50	29.45	11.60	44.55	37.08	51.18	33.57	49.15	36.31	40.57	27.27	17.36	53.31	59.17	25.25	33.29	30.56	44.83	30.06	39.26		
Overall	8.68	21.40	11.09	33.22	29.04	53.40	22.15	40.68	27.19	28.64	27.95	13.23	38.58	47.40	20.66	29.02	20.47	35.94	22.41	32.94		

Table 7: Accuracy (%) of different SDMs on the Chinese (“zh”) or English (“en”) dialogue data subset of C³ (GPT-4o).

Category	Freeze-Omni		GLM-4-Voice		GPT-4o- Audio-Prev.		Kimi-Audio		LLaMA- Omni		MooER- Omni		Moshi		Qwen2.5- Omni		Step-Audio		VITA-Audio		Overall	
	zh	en	zh	en	zh	en	zh	en	en	zh	en	en	zh	en	zh	en	zh	en	zh	en	zh	en
Phonological	16.22	10.34	18.92	34.48	29.73	62.07	21.62	48.28	13.79	21.62	17.24	10.34	27.03	58.62	24.32	31.03	8.11	34.48	20.95	32.07		
Semantic	1.69	11.76	3.39	19.61	7.63	72.55	5.93	39.22	15.69	1.69	54.90	11.76	7.63	43.14	4.24	25.49	4.24	19.61	4.56	31.37		
C_{am}-data	8.96	11.05	11.15	27.05	18.68	67.31	13.78	43.75	14.74	11.66	36.07	11.05	17.33	50.88	14.28	28.26	6.17	27.05	12.75	31.72		
Omission	4.29	5.88	7.14	5.88	42.86	15.69	31.43	7.84	4.90	31.43	1.96	0.98	30.00	15.69	18.57	9.80	7.14	7.84	21.61	7.65		
Coreference	8.33	46.30	13.33	70.00	53.33	92.41	40.00	89.26	58.89	36.67	36.30	19.44	61.67	70.56	48.33	58.89	33.33	75.93	36.88	61.80		
Multi-turn	15.79	55.88	10.53	58.82	15.79	47.06	/	/	64.71	65.79	44.12	/	81.58	94.12	10.53	50.00	73.68	67.65	39.10	60.29		
C_{con}-data	9.47	36.02	10.33	44.90	37.33	51.72	35.71	48.55	42.83	44.63	27.46	10.21	57.75	60.12	25.81	39.56	38.05	50.47	32.39	41.19		
Overall	9.26	26.03	10.66	37.76	29.87	57.95	24.75	46.15	31.60	31.44	30.90	10.63	41.58	56.42	21.20	35.04	25.30	41.10	24.26	37.36		

Table 8: Accuracy (%) of omission and coreference phenomena (GPT-4o).

Phenomenon	Ability	Lang	Freeze-Omni	GLM-4-Voice	GPT-4o- Audio-Prev.	Kimi-Audio	LLaMA- Omni	MooER- Omni	Moshi	Qwen2.5- Omni	Step-Audio	VITA-Audio
Omission	Detection	zh	8.57	14.29	82.86	57.14	/	60.00	/	51.43	31.43	8.57
		en	7.84	3.92	11.76	7.84	5.88	1.96	1.96	13.73	5.88	7.84
Coreference	Completion	zh	0.00	0.00	2.86	5.71	/	2.86	/	8.57	5.71	5.71
		en	3.92	7.84	19.61	7.84	3.92	1.96	0.00	17.65	13.73	7.84
Coreference	Detection	zh	16.67	26.67	63.33	56.67	/	60.00	/	93.33	66.67	53.33
		en	52.22	84.44	95.93	98.52	78.89	24.44	25.56	71.11	65.56	87.78
Coreference	Resolution	zh	0.00	0.00	43.33	23.33	/	13.33	/	30.00	30.00	13.33
		en	40.37	55.56	88.89	80.00	38.89	48.15	13.33	70.00	52.22	64.07

Table 9: Accuracy (%) of omission and coreference phenomena (DeepSeek-R1).

Phenomenon	Ability	Lang	Freeze-Omni	GLM-4-Voice	GPT-4o- Audio-Prev.	Kimi-Audio	LLaMA- Omni	MooER- Omni	Moshi	Qwen2.5- Omni	Step-Audio	VITA-Audio
Omission	Detection	zh	8.57	5.71	82.86	48.57	/	62.86	/	45.71	34.29	8.57
		en	9.80	5.88	15.69	17.65	7.84	11.76	5.88	9.80	9.80	5.88
Coreference	Completion	zh	0.00	2.86	8.57	5.71	/	2.86	/	5.71	0.00	2.86
		en	5.88	7.84	17.65	7.84	5.88	3.92	3.92	19.61	13.73	9.80
Coreference	Detection	zh	23.33	40.00	63.33	63.33	/	56.67	/	80.00	73.33	60.00
		en	62.96	83.33	94.81	95.56	78.15	27.41	45.19	70.00	61.11	87.41
Coreference	Resolution	zh	3.33	0.00	46.67	16.67	/	0.00	/	20.00	33.33	6.67
		en	33.33	52.59	84.81	75.56	31.85	44.07	14.44	61.48	50.37	60.00

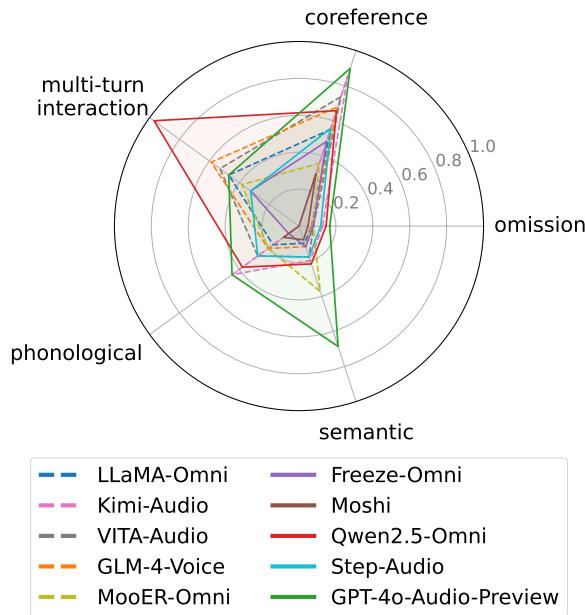


Figure 10: Radar charts depicting the experimental results of each SDM on the English portion of the dataset, assessed using DeepSeek-R1.

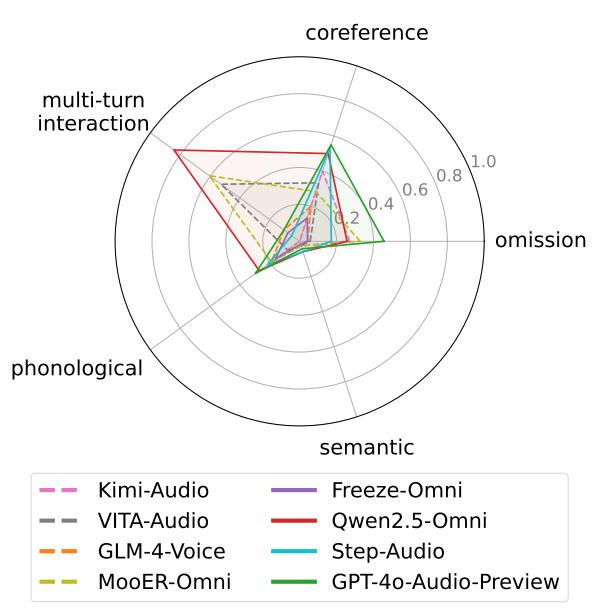


Figure 12: Radar charts depicting the experimental results of each SDM on the Chinese portion of the dataset, assessed using DeepSeek-R1.

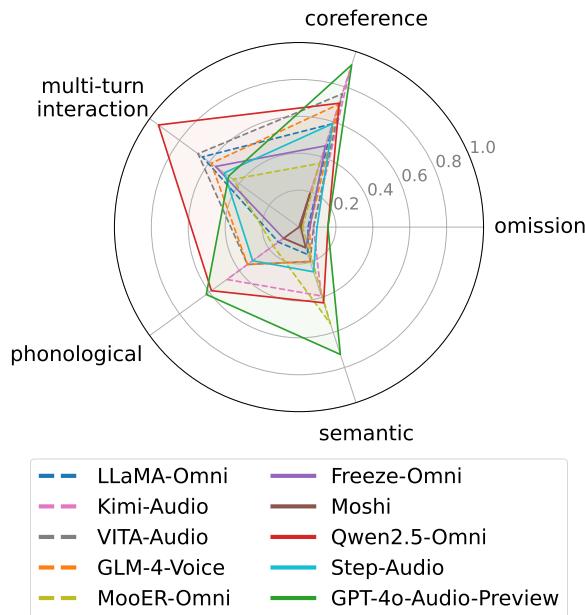


Figure 11: Radar charts depicting the experimental results of each SDM on the English portion of the dataset, assessed using GPT-4o.

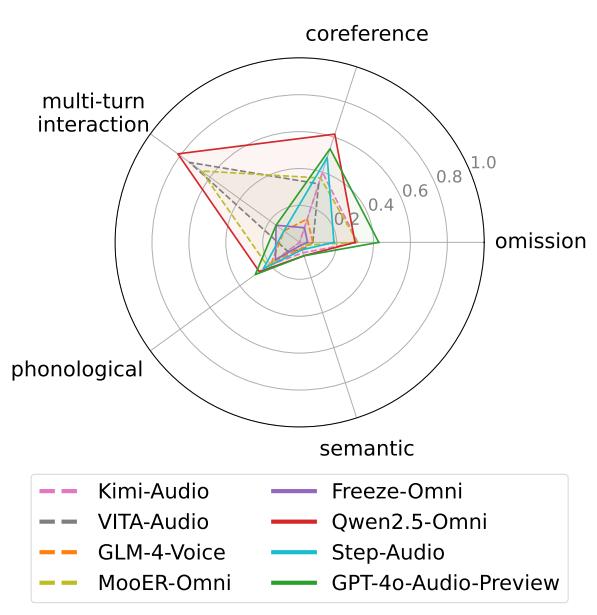


Figure 13: Radar charts depicting the experimental results of each SDM on the Chinese portion of the dataset, assessed using GPT-4o.