REALTALK-CN: A REALISTIC CHINESE SPEECH TASK-ORIENTED DIALOGUE BENCHMARK WITH CROSS-MODAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in speech large language models (e.g., GPT-4o) have enabled end-to-end spoken interactions, yet their robustness in real-world applications remains unclear, where systems must assist users in completing specific tasks under complex conditions such as multi-turn, ambiguous, and often spontaneous speech, as well as natural alternation between speech and text. Task-oriented dialogue (TOD) offers a realistic scenario to evaluate whether models can effectively help users accomplish such task-oriented goals, but existing benchmarks are mainly text-based, and the few speech datasets are limited to English and often neglect spontaneous disfluencies and speaker diversity. To address this gap, we introduce **RealTalk-CN**, the first Chinese multi-turn, multi-domain speech–text TOD dataset, containing 5.4k dialogues (60K turns, 150 hours) of real human-to-human recordings with detailed annotations for dialogue states, disfluency types, and speaker characteristics. Based on this dataset, we propose a cross-modal interaction task supporting dynamic speech-text switching and a comprehensive evaluation protocol assessing robustness to disfluencies, sensitivity to speaker variation, and cross-domain generalization. Experiments on state-of-the-art models demonstrate the challenges posed by RealTalk-CN and establish its value as a benchmark for developing reliable and fair Speech LLMs in real-world deployments. The dataset and evaluation framework [1] will be open-sourced to encourage further research.

1 INTRODUCTION

Recent years have witnessed significant advances in large language models (LLMs) for multimodal processing. In particular, the speech domain has seen the emergence of end-to-end speech LLMs, such as GPT-4o (Achiam et al., 2023), Qwen2-Audio (Chu et al., 2024), and GLM-4-Voice (Zeng et al., 2024), which can directly process speech input and generate natural language output, enabling fluent human–machine speech interactions.

However, it remains unclear whether these speech LLMs can be truly applied in complex real-world scenarios and effectively assist humans in carrying out task-oriented goals. For example, in in-car voice assistants, users may issue multiple, vague utterances to query destinations; in e-commerce voice assistants, users may first ask product details via speech and then follow up with text input to inquire about discounts. Such goal-driven scenarios involve spontaneous speech (Shriberg, 1994), fuzzy query (Chun et al., 2025), multi-turn interactions and modality switching, posing higher demands on model robustness and practical utility.

In real-world applications, users are primarily concerned with completing specific tasks, such as booking tickets, ordering food, or customer service inquiries. These goal-driven, non-chit-chat interactions constitute task-oriented dialogue (TOD) (Cai et al., 2024), which has long been a core focus of dialogue system research from traditional approaches (Zang et al., 2020; Zhu et al., 2020) to current LLM-based methods (Jang et al., 2022; Dong et al., 2025; Sekulic et al., 2024). However, existing large-scale TOD datasets, such as MultiWOZ (Zang et al., 2020), CrossWOZ (Zhu et al., 2020), and RiSAWOZ (Quan et al., 2020), are almost entirely text-based. This limits the evaluation of speech LLMs on realistic spoken inputs, particularly regarding robustness to spontaneous speech phenomena (e.g., pauses, fillers, repetitions, and self-corrections) (Shriberg, 1994). In addition, real-world usage involves speaker diversity (e.g., gender, age, regional accents) (Krause & Braida, 2004); for instance, (Kulkarni et al., 2024; Chen et al., 2025) show that current speech recognition systems exhibit significant performance degradation for elderly users due to age-related vocal deterioration (Fraser et al., 2015). Analyses of such factors are still largely missing in speech LLM TOD evaluations, though they are critical for ensuring fair and unbiased AI assistance across different speakers (Choi & Choi, 2025).

Resources for speech TOD are scarce. Early datasets such as ATIS (Hemphill et al., 1990) and DSTC2/10 (Henderson et al., 2014; Kim et al., 2021) cover only single-turn or limited scenarios. SpokenWOZ (Si et al., 2023) is

[1] https://anonymous.4open.science/r/RealTalk-7969

the first large-scale multi-turn, multi-domain English speech TOD benchmark, but it lacks annotations for spontaneous speech disfluencies and speaker characteristics, limiting robustness evaluation. In Chinese, there is currently no analogous multi-turn speech TOD dataset, despite the unique linguistic and sociocultural traits of Chinese spoken dialogues (Huang et al., 2023). Meanwhile, speech LLMs supporting Chinese such as Baichuan-Audio (Li et al., 2025), GLM-4-Voice (Zeng et al., 2024), and Qwen-2.5-Omni (Xu et al., 2025) are rapidly advancing, yet corresponding evaluation benchmarks are lacking.

In addition, real-world dialogues often involve modality switching between speech and text. Current studies on speech–text multimodal TOD systems (Si et al., 2023; Li et al., 2024a) typically assume that users provide both modalities simultaneously to improve responses. However, in practice, users naturally alternate between speaking and typing across turns, for example, asking a voice question about product details and then typing a follow-up request for discounts in e-commerce applications. However, existing work lacks attention to such tasks.

To address these real-world challenges, we propose:

- **Realistic Multi-Turn Speech–Text TOD Dataset:** We present RealTalk-CN, the first Chinese multi-turn, multi-domain speech–text task-oriented dialogue dataset, containing 5.4k dialogues (60K turns, 150 hours). All recordings are from real human-to-human conversations and include detailed annotations for dialogue intents, slot values, spontaneous speech disfluencies, and diverse speaker characteristics, reflecting the complexity of real-world usage. Approximately $35,000 was invested in data collection and annotation to ensure high quality and representativeness.

- **Controlled Data Quality Procedures:** Multi-layer quality assurance is implemented, including standardized script design, controlled recording environments, diverse speaker selection, and detailed annotation guidelines, ensuring both reliability and ecological validity.

- **Cross-Modal Interaction Task:** We design a dialogue task that allows users to dynamically switch between speech and text inputs, closely mimicking real-world usage patterns of voice assistants and customer service systems.

- **Robustness and Fairness Evaluation:** A systematic evaluation protocol is introduced to assess model performance under spontaneous speech disfluencies, speaker variability (gender, age, accent), and cross-domain scenarios, providing a comprehensive measure of reliability and fairness in realistic deployments.

2 Related Work

Table 1 summarizes the various aspects of our dataset compared with other related datasets. Related work can be roughly divided into three categories:

Text-based TOD datasets: English resources in this domain include MultiWOZ (Zang et al., 2020), a widely used dataset spanning eight domains with over ten thousand dialogues. For Chinese, notable datasets are CrossWOZ (Zhu et al., 2020), which contains six thousand dialogues and 102 thousand utterances, and RiSAWOZ (Quan et al., 2020), a more extensive collection featuring 11.2 thousand dialogues, 150 thousand utterances, and coverage across twelve domains. These provide rich annotations for dialogue state tracking but lack speech signals.

Spoken language understanding (SLU) datasets: Most English SLU datasets such as SNIPS (Kawar et al., 2021) rely on transcribed text without accounting for speech recognition errors. The largest existing English SLU resource is SLURP (Bastianelli et al., 2020), which covers eighteen domains. In contrast, Chinese research has seen initial progress with CATSLU (Zhu et al., 2019), a multi-domain audio-text dataset introduced during the ICMI 2019 challenge. However, these datasets are only single-turn content understanding tasks.

Speech-based TOD datasets: Existing speech-based task-oriented datasets remain scarce. Early efforts such as DSTC2 (Henderson et al., 2014) and DSTC10 (Kim et al., 2021) provide only small-scale automatic speech recognition outputs. SpokenWOZ (Si et al., 2023) represents the first large-scale English speech-text benchmark but lacks speech disfluency annotation and speaker feature annotation. Moreover, no similarly comprehensive Chinese dataset currently exists, creating a significant gap that impedes research progress in this area.

Other non-task-oriented speech dialogue datasets: Unlike chit-chat datasets such as StyleTalk (Lin et al., 2024), SD-Eval (Ao et al., 2024), VoxDialog (Cheng et al., 2025), and Full-Duplex-Bench (Lin et al., 2025), which focus on paralinguistic cues or full-duplex spoken dialogue, RealTalk-CN is distinguished by its task-oriented dialogue (TOD) nature. Furthermore, our dialogues are created from human-scripted prompts combined with spontaneous impromptu contributions and recorded as real human-to-human speech (non-TTS), in Chinese, enabling comprehensive evaluation of speech-based models' robustness in realistic goal-driven applications that emphasize assisting users in accomplishing tasks.

Table 1: Comparison of our dataset with other related datasets. TOD stands for Task-Oriented Dialogue Dataset, SLU is a single-round Spoken Language Understanding dataset. H2H, H2M, M2M stand for human-to-human, human-to-machine, machine-to-machine.

Type	Dataset	Language	Speakers	Dialogues	Avg. turns	Domains	Slots	Audio	Disfluency Annotation	Cross modal task
Text-based TOD	**M2M** (Shah et al., 2018)	EN	M2M	1,500	9.9	2	14	✘	✘	✘
	KVRET (Eric & Manning, 2017)	EN	H2H	2,425	5.3	3	13	✘	✘	✘
	MultiWOZ (Budzianowski et al., 2018)	EN	H2H	8,438	13.7	7	25	✘	✘	✘
	DSTC10 (Kim et al., 2021)	EN	H2H	107	21.4	3	-	✘	✘	✘
	CrossWOZ (Zhu et al., 2020)	ZH	H2H	5,012	16.9	5	72	✘	✘	✘
	RISAWOZ (Quan et al., 2020)	ZH	H2H	10,000	13.5	12	159	✘	✘	✘
Speech-based SLU	**FSC** (Qian et al., 2021)	EN	H	30,043	1	1	-	✓	✘	✘
	SNIPS (Kawar et al., 2021)	EN	H	13,084	1	7	72	✓	✘	✘
	SLURP (Bastianelli et al., 2020)	EN	H	72,277	1	18	55	✓	✘	✘
	CATSLU (Zhu et al., 2019)	ZH	H	16,258	1	4	94	✓	✘	✘
Speech-based TOD	**DSTC2** (Henderson et al., 2014)	EN	H2M	1,612	14.5	1	8	✓	✘	✘
	SpokenWOZ (Si et al., 2023)	EN	H2H	5,700	35.5	26	36	✓	✘	✘
	RealTalk-CN(ours)	**ZH**	H2H	5,400	12.1	58	115	✓	✓	✓

Table 2: Comparison with other non-task-oriented spoken dialogue datasets. *Trainable* indicates whether the dataset has training set; *Disfluency* indicates whether spontaneous speech disfluencies are annotated; *Human+Imp.* indicates a combination of scripted and impromptu human recordings. *GPT* indicates AI-generated text; *TTS* indicates that speech is synthesized using text-to-speech tools.

Dataset	TOD	Lang.	Text	Speech	Disfluency	Cross-Modal	Trainable	Dialogue Style	Multi-Turn
SD-Eval (Ao et al., 2024)	✘	EN	Mixed	Mixed	✘	✘	✓	✘	✘
StyleTalk (Lin et al., 2024)	✘	EN	GPT	TTS	✘	✘	✓	✓	✓
VoxDialog (Cheng et al., 2025)	Mixed	EN/ZH	GPT	TTS	✘	✘	✘	✓	✓
RealTalk-CN (Ours)	✓	ZH	Human+Imp.	Real	✓	✓	✓	✓	✓

3 REALTALK-CN DATA COLLECTION AND QUALITY CONTROL

During the data collection phase, we prioritized speech quality and annotation consistency. The dataset was constructed using pre-written scripts designed to reflect natural spoken language characteristics, including casual grammar, colloquial vocabulary, short sentence structures, and loose syntactic organization (Carter, 1995). The dialogues covered multiple domains while allowing participants to improvise on the recording to maintain conversational authenticity. Crucially, 10% of the collected data intentionally preserved spontaneous speech disfluencies such as repetitions, hesitations, self-corrections, and modal particle drag to simulate real-world conditions.

For speech-text alignment, we implemented a rigorous timestamping mechanism to mark utterance boundaries and dialogue turns. Recording sessions were conducted in quiet indoor environments using both professional microphones and consumer-grade smartphone microphones to ensure device diversity representative of real usage scenarios. Dual recording methods (dedicated recorders and smartphones) were employed to capture authentic acoustic conditions.

Speaker diversity was ensured through 300 volunteers (gender ratio 1:1±10%, ages 18-50 following normal distribution covering young and middle-aged demographics) with predominantly Mandarin proficiency while permitting mild regional accents. The gender, age, and regional distribution are shown in Figure 1. Each participant contributed to 50 dialogue sessions.

Annotation consistency was maintained through multi-round verification with detailed guidelines addressing various Chinese speech phenomena. Transcripts were required to strictly match actual pronunciations while accommodating dialectal variations, such as converting ”Liu nai” to standard ”Niu nai”. Mandarin phonological features including erhua were preserved in transcriptions. Standard references were used to verify proper nouns, while numerical expressions were consistently rendered in Chinese characters. Filler sounds and discourse markers were retained to maintain prosodic authenticity, with special notation applied to the intentionally preserved 10% of data containing disfluencies. Comprehensive quality control measures were implemented throughout the process. Audio clips maintained 0.2-0.3 seconds of silence padding with duration optimized at 5-6 seconds (maximum 12 seconds). A 5% random sampling protocol ensured slot-value annotation accuracy exceeded 95%. The annotation Pipeline incorporated iterative optimization, beginning with pilot annotation of three sample batches to refine guidelines before full-scale implementation. The Ethics Statement of the dataset is described in section 6.

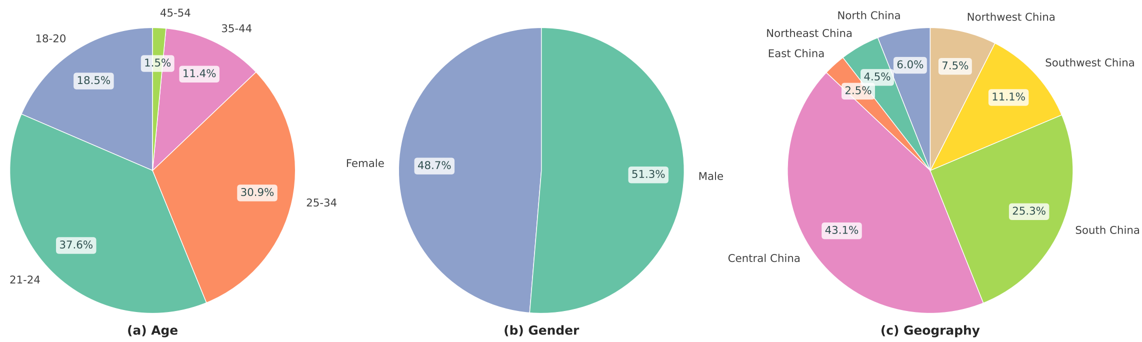

Figure 1: The distribution of Speakers. The dataset covers most age groups. It also has a near-equal gender split. It covers all major regions in China to explore the impact of different regional accents on the speech model. The specific provinces included in each region are in the Appendix A.1.

Table 3: Data statistics of the four subsets. Colloquial means that the text content contains the above-mentioned unfluent spoken language markers, while System means the opposite. and Avg Intent Choices means the average number of candidate intents as answers for each utterance. MD means Multi-Domain, and SD means Single-Domain. M, R, S, and H respectively represent Modal Particle Drag, Repetition, Self-Correction, and Hesitation

Subsets	Samples	Avg Utterance Length	Avg Dialog Rounds	Avg Intent Choices	Avg Disfluency Markers
MD-Col	3,837	27.42	8.54	34.51	M:0.12 R:0.04 S:0.11 H:1.14
MD-Sys	3,837	19.27	7.73	34.77	-
SD-Col	892	25.61	8.14	25.90	M:0.63 R:0.07 S:0.18 H:0.52
SD-Sys	892	20.76	7.58	27.03	-

4 REALTALK-CN DATASET OVERVIEW

RealTalk-CN represents the first Chinese multi-turn, multi-domain speech-text dual-modal TOD dataset, which comprises 5.4k dialogue sessions, including 1.2k single-domain and 4.2k cross-domain conversations, totaling over 60k utterances contributed by 113 speakers. With an average of 12.1 turns per dialogue and 150 hours of validated audio, the dataset covers dozens of task-oriented domains (e.g., dining, transportation, shopping) through authentic human-to-human interactions. Each dialogue is accompanied by comprehensive annotations including dialogue states (slots), intents, transcriptions, utterance-level timestamps, speaker metadata, and labels for spontaneous speech disfluencies phenomena (e.g., filled pauses, repetitions, self-corrections).

4.1 SPONTANEOUS SPEECH PHENOMENA

As a spoken language-oriented resource, RealTalk-CN captures fundamental distinctions between oral and written communication styles - even within identical semantic contexts, spoken dialogues exhibit casual grammar, colloquial vocabulary, fragmented structures, and loose syntactic organization (Carter, 1995). Our scripting process explicitly mandated conversational language patterns. Crucially, spoken disfluencies (Shriberg, 1994) present additional challenges for language understanding systems. While the English SpokenWOZ dataset (Si et al., 2023) addressed this partially through ASR-derived noise, it lacked explicit annotation of disfluency types. RealTalk-CN advances this through systematic labeling of spontaneous speech disfluencies, enabling the creation of phenomenon-specific subsets for robustness evaluations of speech-based LLMs. This design also facilitates secondary applications like speech disfluency correction. As illustrated in Figure 2, we defined common disfluency categories, instructed speakers to maintain natural conversational flow (including organic production of disfluencies), and implemented rigorous post-hoc annotation protocols.

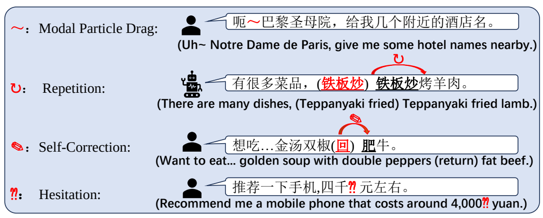

Figure 2: Four speech disfluency types in Chinese dialogues: Modal particle drag, repetition, self-correction, and hesitation. The third example is that in Chinese, "return" and "fat" sound similar.

4.2 Broad Domain Coverage

RealTalk-CN comprehensively encompasses 58 TOD domains, including weather, dining, travel, news, shopping, finance, and healthcare. It also has 55 intents and 115 slot types, which are not available in previous datasets. Detailed intent and slot information can be found in Appendix A.2. The dataset is systematically organized into single-domain and multi-domain dialogues, with the latter involving 2-5 interleaved domains to better simulate real-world scenarios. For example, a travel-related conversation can naturally incorporate weather inquiries and restaurant recommendations. As illustrated in Figure 3, the domain distribution follows a long-tail pattern: high-frequency domains (e.g., travel, weather) cover common daily topics, while mid-to-low frequency domains ensure comprehensive topical diversity. Among multi-domain dialogues, 2,949 sessions involve two domains (representing the majority), followed by 753 sessions with three domains. Additionally, the data set includes complex dialogues that span 4-5 domains.

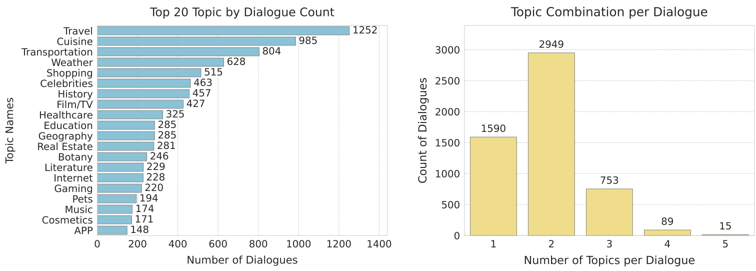

Figure 3: Domain distribution and multi-domain combination distribution of RealTalk-CN.

To enable granular analysis, we partition the dataset into four subsets based on two criteria: (1) single-domain vs. multi-domain composition, and (2) presence of annotated speech disfluencies. Detailed statistics of the test set are provided in Table 3. The dataset splits are detailed in Appendix A.3, and both training data and training framework will also be released. Multi-domain dialogues demonstrate significantly greater complexity than single-domain counterparts, evidenced by higher average intent counts, reflecting more diverse user needs in cross-domain interactions. Linguistically, disfluency-annotated subsets exhibit longer utterance lengths and more dialogue turns due to phenomena such as self-corrections, repetitions, and Modal particle drags. These characteristics mirror authentic speech patterns and present increased robustness challenges for speech-based LLMs in TOD scenarios.

4.3 Innovative Cross-Modal Chat Task

Current research on multimodal dialogue systems focuses primarily on scenarios in which users and systems simultaneously receive and process multiple modalities, such as speech and text. For example, previous work (Si et al., 2023) (Li et al., 2024a) proposed multimodal speech-text dialogue datasets where the evaluation task involves responding to contexts that contain speech and text modalities, aiming to enhance textual representations through aggregated speech embeddings for improved responses. However, this simultaneous multimodal input paradigm rarely occurs in real-world applications. In practical intelligent voice assistant usage, user-system interactions typically span multiple turns with dynamic modality switching. For example, in a restaurant reservation scenario, users might initially inquire via voice and subsequently continue the conversation through text messages or mobile apps, rather than providing identical content through both speech and text simultaneously. A concrete illustration of this pattern is shown in Figure 4.

To address this gap, we propose a novel cross-modal chat task where the conversational context contains mixed speech or text utterances without simultaneous modality presentation. The key challenge lies in the model's

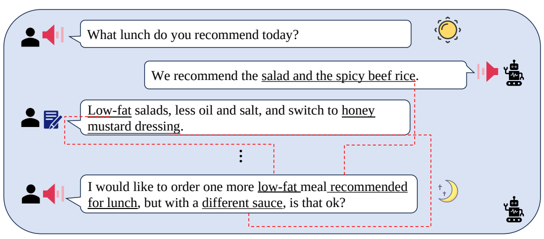

Figure 4: An example of dynamically switching speech-text modality dialogue. In this scenario, the system must integrate the dish recommendations provided via noon voice messages with the user's customized preferences expressed in text, accurately comprehend that "low-fat meal" refers to salad, and correctly identify the user's request to change the dressing.

Table 4: Performance comparison of the models on the intent classification (IC) and slot filling (SF) tasks of the RealTalk-CN dataset. Acc is the accuracy of intent classification, and Pipeline represents Whisper-large-v3 + GPT-4o. GPT-4o-Audio uses the mini version. PAN. represents PANDA score.

Subsets	MD-Col				MD-Sys				SD-Col				SD-Sys				
Tasks	IC		SF		IC		SF		IC		SF		IC		SF		
Metrics	Acc	PAN.	F1	JGA	Acc	PAN.	F1	JGA	Acc	PAN.	F1	JGA	Acc	PAN.	F1	JGA	Average
Pipeline	**53.56**	**53.56**	45.90	26.09	**54.83**	**54.83**	48.81	31.99	**59.75**	**59.75**	38.55	20.68	**62.44**	**62.44**	45.17	28.52	**46.68**
Baichuan-Audio	30.70	30.70	48.60	30.46	28.20	28.20	**54.11**	**40.66**	27.47	27.47	39.96	23.94	30.49	30.49	47.15	**33.80**	34.53
GLM-4-Voice	26.40	26.40	10.48	19.41	19.49	19.49	9.31	39.59	32.51	32.51	9.58	15.31	28.36	28.36	10.64	19.19	21.69
Qwen2-Audio	24.76	24.78	47.67	30.67	18.14	18.26	52.76	25.48	27.47	27.50	38.58	23.78	23.09	23.15	45.69	32.92	30.29
Baichuan-Omni	36.17	36.19	48.06	28.88	34.53	34.54	52.99	39.81	38.68	38.79	39.99	24.42	34.53	34.53	46.34	31.34	37.49
MiniCPM-o	39.74	39.74	46.02	26.56	35.84	35.84	49.91	33.41	41.82	41.82	36.82	20.52	39.01	39.01	44.40	28.52	37.44
Qwen2.5-Omni	24.52	24.54	47.70	30.88	18.17	18.25	52.55	39.75	27.58	27.64	39.57	**24.43**	22.87	22.90	45.67	33.45	31.28
GPT-4o-Audio	46.31	46.31	**51.53**	**31.93**	45.04	45.04	53.65	38.39	48.21	48.21	**43.16**	24.27	49.10	49.10	**48.45**	33.10	43.86

ability to accurately comprehend and track information distributed across different modalities while effectively integrating these heterogeneous inputs to generate consistent and coherent responses. To isolate the impact of modality switching from speech disfluency effects, we specifically employ speech modality for turns containing any of the four disfluency markers, while using text modality otherwise, thereby creating a dynamically switching context.

5 Experimental Design & Evaluation

5.1 Task Design

We designed multiple tasks on the SpokenMMC dataset to fully exploit its potential. We use the same zero-shot evaluation protocol for speech-based LLMs following (Yang et al., 2024; Chen et al., 2024), and we also provide the training set for researchers to use.

Standard task-oriented dialogue tasks, including dialogue intent classification, slot filling, and end-to-end chat. Following (Chen et al., 2024), for the intent classification task, we compute Accuracy and the PANDA discriminant (Li et al., 2024c) estimation method which has a strong correlation with human evaluation. For the slot filling task, we use the classic evaluation metrics F1 and joint goal accuracy (JGA) (Budzianowski et al., 2018). Specific examples of the tasks and our evaluation process can be found in Appendix A.7. For the chat task, since traditional metrics have demonstrated a weak correlation with human judgment (Liu et al., 2023), we implemented GPT-4-based automatic evaluation following (Chen et al., 2024) (Liu et al., 2023) (Yang et al., 2024). All evaluations are conducted using GPT-4o-mini[2], including the sum of the scores of the evaluation without reference and the evaluation with reference. The evaluation prompts can be found in Appendix A.8

Cross-modal chat task, as described earlier, users and assistants dynamically switch between speech and text modalities during conversations.

Robustness evaluation task leverages the annotated speech disfluencies to examine models' tolerance to conversational incoherence, using performance differences on the Colloquial subset as the evaluation metric.

5.2 Baselines

We evaluated several end-to-end speech-based LLMs, including Qwen2-Audio-7B-Instruct (Chu et al., 2024), Baichuan-Audio-Instruct (Li et al., 2025), GLM-4-Voice-9B (Zeng et al., 2024), along with recent Omni-modal foundation models (MiniCPM-o (Yao et al., 2024), Baichuan-Omni-1d5 (Li et al., 2024b), Qwen2.5-Omni-7B (Xu et al., 2025)). For comparison, we also included **Pipeline** approaches combining Whisper-Large-V3 (Radford et al., 2023) with text-only LLMs (GPT-4o) and GPT-4o-Audio-mini[3], aiming to measure performance gaps between current open/closed-source voice LLMs and traditional Pipeline methods. We evaluated the models based on the code in (Chen et al., 2024).

5.3 Results & Discussion

Speech disfluency affects slot filling and chat tasks. In Table 4, on the Colloquial subsets, the performance of most models in the slot filling task dropped significantly, such as the JGA value of the Pipeline method dropped from 31.99 to 26.09, and the Baichuan-Audio dropped from 40.66 to 30.46 on the MD-Col subset, while the intent classification task was not significantly affected. This difference stems from the difference in the performance of

[2]GPT-4o-mini-2024-07-18

[3]GPT-4o-mini-audio-preview

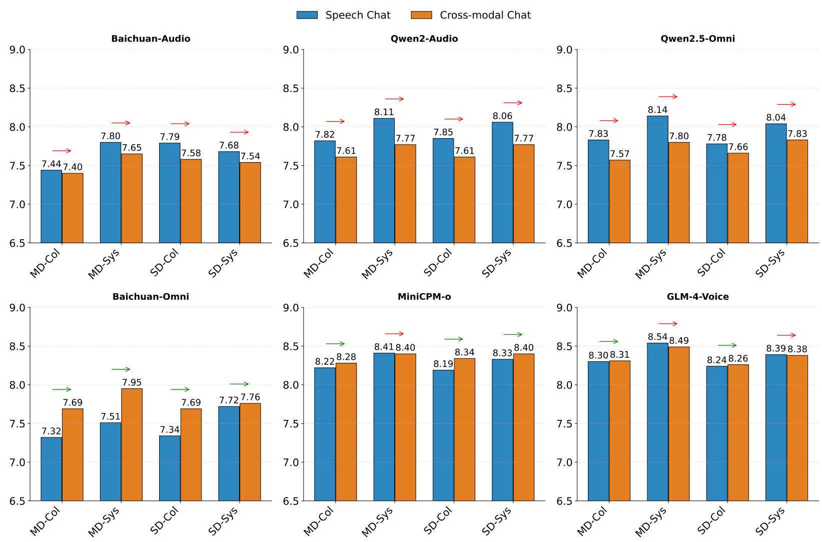

Figure 5: Performance comparison between pure speech chat tasks and Cross-modal chat tasks. The results were analyzed using paired t-tests (Student, 1908) ($p < 0.05$), and the tasks with significant differences were Qwen2-Audio, Baichuan-Audio, Baichuan-Omni, and Qwen2.5-Omni, while those with insignificant differences were MiniCPM-o and GLM-4-Voice. The detailed process can be found in Appendix A.10.

the models in processing speech features and semantic information. The intent classification task mainly relies on capturing the core semantics in the user's sentence, even in the case of unfluent speech expression, the model can maintain high accuracy through contextual semantics. However, in the slot filling task, the model needs to accurately identify and extract specific slot information in the continuous speech stream, which relies on the model's ability to capture key information. When there are self-repetitions, grammatical errors, or non-standard expressions in the speech, the model is easily confused in slot boundaries and content extraction, which puts higher requirements on the model. The chat task (shown in Table 5) requires obtaining the core semantics, extracting some key information, and responding after integration. The performance also declines on the Colloquial subsets.

Table 5: Performance of the model on the chat task of the RealTalk-CN dataset. The Pipeline represents Whisper-large-v3 + GPT-4o, and the score is the score of GPT-4o-mini, with a full score of 10. *Note that GPT-4o-Audio-mini does not support speech mode on the assistant side during input.

Models	MD-Col	MD-Sys	SD-Col	SD-Sys	Avg
Pipline	**8.92**	**9.12**	**8.84**	**9.12**	**9.00**
Baichuan-Audio	7.44	7.80	7.79	7.68	7.67
GLM-4-Voice	8.30	8.54	8.24	8.39	8.37
Qwen2-Audio	7.82	8.11	7.85	8.06	7.96
Baichuan-Omni	7.32	7.51	7.34	7.72	7.47
MiniCPM-o	8.22	8.41	8.19	8.33	8.29
Qwen2.5-Omni	7.83	8.14	7.78	8.04	7.95
Gpt-4o-Audio-mini	8.66	8.79	8.71	8.77	8.73

The Pipeline method and the end-to-end model have their own advantages and disadvantages: in the intent classification task shown in Table 4 and the chat task shown in Table 5, the Pipeline method performs better than the end-to-end model. This result can be attributed to the fact that in the Pipeline method, after Whisper-large-v3 converts speech into text, GPT-4o is responsible for text semantic understanding and intent classification, making full use of GPT-4o's powerful ability in text understanding, and these two tasks rely more on semantics. In contrast, although the end-to-end model can capture speech and text information directly from the speech input, it often declines in semantic understanding ability. However, in the slot-filling task, the end-to-end model outperforms the Pipeline method. In Table 4, Many models have higher F1 and JGA values than the Pipeline method. Tasks such as slot filling rely more on details in speech, while the end-to-end model can be more adaptable to speech quality and expression clarity (such as disfluency) and is more conducive to capturing detailed information of speech. In addition, after calculating the average of their performance for all tasks, the Pipeline method maintained its leading position, while GPT-4o-Audio-mini was second best, and was generally ahead of other end-to-end models in multiple tasks, indicating that it has stronger capabilities in speech understanding and multimodal feature fusion.

Multi-domain complexity mainly affects intent classification capabilities On the Multi-Domain subset, the performance of intent classification tasks is significantly lower than that of the Single-Domain subset. For example, in Table 4, compared with SD-Col, the PANDA of intent classification of MD-Col is generally reduced by 2-5 points, while the performance of slot filling tasks is not significantly affected. This difference reflects the limitations of the model in dealing with semantic diversity and context switching. The intent classification task essentially relies on the model to correctly classify user intent in the semantic space, and multi-domain scenarios involve multiple tasks and contexts, so the model needs to have stronger cross-domain semantic generalization capabilities. However, current end-to-end speech models often lack semantic representation and context adaptability when faced with domain switching, and may mistakenly confuse the semantics of different domains, thus affecting the accuracy of intent classification. In contrast, the slot filling task performs more stably in multi-domain scenarios because it relies on the recognition of specific slots. The model only needs to recognize predefined slot information, and domain changes have little impact on the definition of these slots.

Performance Divergence in Cross-Modal Chat Tasks. Figure 5 illustrates the varied performance of speech foundation models across pure speech-based chat tasks and cross-modal chat tasks, revealing distinct model behaviors. The first category comprises models exhibiting performance degradation, including Baichuan-Audio, Qwen2-Audio, and Qwen2.5-Omni. These models show consistent metric declines in cross-modal scenarios, exemplified by Qwen2-Audio's MD-Col score decreasing from 7.82 to 7.61. Through a detailed case study, we found that the model did have some problems when integrating and responding to heterogeneous modal information, including forgetting the key information and semantics of the previous context of different modalities, reduced quality of response richness, and response text degeneration. The detailed case can be found in Appendix A.13. The second category features models maintaining stable performance, including GLM-4-Voice and MiniCPM-o. In particular, the third category contains models that achieve performance improvements. Baichuan-Omni shows a significant MD-Col score increase from 7.32 to 7.69, which suggests that this model benefits from text-modality substitutions in dialogue history. Further comparisons of the pure-text modality and the performance comparison among the three modalities are presented in Appendix A.4.

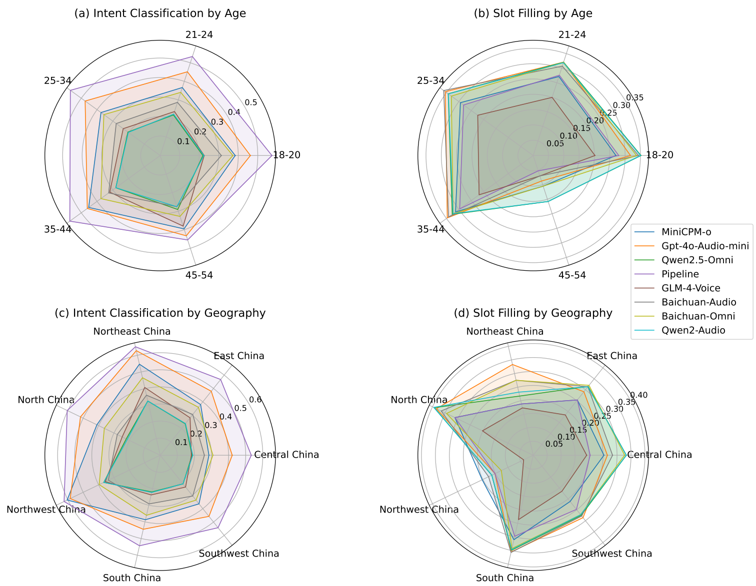

Figure 6: Radar chart showing the impact of speaker's age and geography on dialogue intent classification and slot filling tasks. The result is the average performance of the model on the entire RealTalk-CN dataset, grouped by age and geography (which refers to ancestral origin). We performed an overall Kruskal-Wallis H test (Kruskal & Wallis, 1952) ($p < 0.05$) and a comparative Mann-Whitney U test (Mann & Whitney, 1947) ($p < 0.05$) for significance, and calculated Cohen's d effect size (Cohen, 2013). Details are available in Appendix A.9.

5.4 Speaker variations

Speaker-related attributes such as age (Kulkarni et al., 2024) (Chen et al., 2025) and regional accent (Chen et al., 2024) significantly influence speech model performance. Existing studies (Kulkarni et al., 2024) (Chen et al., 2025) demonstrate that current speech models exhibit notable performance degradation when processing elderly users' speech due to age-related vocal deterioration (Fraser et al., 2015). While (Chen et al., 2024) investigated global English accent variations' impact on speech LLMs, comparable research remains scarce for Chinese speech LLMs, particularly in task-oriented dialogue domains. Given China's vast geographical distribution with diverse Mandarin accents and broad age demographics, we systematically analyze age and regional accent effects on speech-based LLMs in Chinese task-oriented dialogue Scenarios, with results shown in Figure 6.

From a task perspective, models exhibit differential sensitivity to speaker characteristics during intent classification and slot filling. Fewer models showed statistically significant differences in intent classification, particularly across age groups, except for notable variations in Northwest and Northeast China's geographical distribution. In contrast, slot filling demonstrated pronounced susceptibility to both age and regional factors ($p < 0.05$), reinforcing our findings in Section 5.3 that fine-grained semantic parsing tasks are more vulnerable to speech variability.

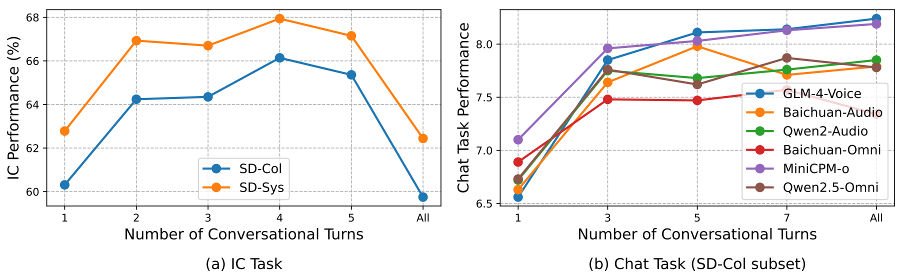

Figure 7: Performance trends as the number of conversational history turns increases. (a) Pipline on IC task. (b) Chat task on the SD-Col subset.

Demographic analysis revealed that SF performance for 45-54-year-old users decreased significantly, confirming the compounded challenges of vocal aging and distributional mismatch with training data. Regional comparisons further indicated superior performance for South and North China users compared to Southwest or Northwest regions, showing that standard Mandarin proximity in training data better serves linguistically central regions, while stronger dialectal features in peripheral areas degrade model generalization.

Model adaptation varied across tasks. In slot filling, Baichuan-Audio and GPT-4-Audio-mini performed well for mainstream speakers but dropped over 20% for elderly users, showing limited robustness to vocal aging. Qwen models remained consistent across age groups. For regional adaptation, MiniCPM-o, GPT-4-Audio-mini, and Qwen showed strong cross-regional generalization. These results highlight key directions for improving model fairness and stability across diverse users.

5.5 Disfluency-Type-Specific Evaluation

To further explore how different types of disfluencies affect model performance, we primarily present results on the slot filling task, as it involves extracting key information. The results for other tasks can be found in Appendix A.5. We obseverd that performance drops are more pronounced on Repetition, followed by Modal Particle Drag, indicating confusion in semantic grounding under repeated phrasing. GPT-4o-Audio achieves the lowest standard deviation, suggesting the best robustness across all disfluency types.

Table 6: SF F1 (%) across disfluency types on SD-Col. Std. Dev. means standard deviation.

Model	Hesitation	Self-Correction	Repetition	Modal Particle Drag	SD-Sys	Std. Dev.
Pipeline	46.56	51.76	39.43	42.52	48.81	5.33
Baichuan-Audio	49.57	54.47	39.06	44.58	54.11	6.62
GLM-4-Voice	38.81	38.79	33.09	30.63	40.00	4.13
Qwen2-Audio	48.18	50.51	41.12	47.58	52.76	4.02
Baichuan-Omni	48.78	53.06	42.68	47.67	52.99	4.27
MiniCPM-o	47.05	48.32	38.29	42.35	49.91	4.59
Qwen2.5-Omni	48.34	49.64	40.76	48.20	52.55	4.04
GPT-4o-Audio	52.11	54.23	45.40	47.97	53.65	3.98

5.6 Impact of Dialogue History Length

Due to space constraints, we report results for two representative tasks. Figure 7 shows intent classification (IC) performance on the SD subsets. We observe that the required contextual information is relatively short: most models reach near-peak accuracy within 3–5 turns, indicating that IC benefits from local context but does not heavily depend on long dialogue histories.

In contrast, Figure 7 presents chat task results on the SD-Col subset, where performance steadily improves as more historical turns are included, with the highest scores achieved when the full context is available.

6 Conclusion

In this paper, we introduce RealTalk-CN, the first large-scale Chinese speech-text dual-modal dialogue benchmark that comprehensively captures speech disfluencies, diverse speaker characteristics, and cross-modal interactions. Our evaluations demonstrate the dataset's effectiveness in benchmarking models on speech robustness, speaker adaptation, and cross-modal consistency. The proposed cross-modal chat task further reveals models' limitations in handling dynamic modality switching. RealTalk-CN sets a new standard for Chinese multimodal dialogue research, providing a critical resource for advancing speech-based language models.

Ethics Statement

The collection and use of data in this study were conducted in accordance with ethical guidelines for research involving human participants. All participants provided informed consent prior to their involvement, with clear explanations of the study's purpose, data usage, and their rights to withdraw at any time. Personal identifiers were anonymized to protect privacy, and participants were compensated fairly for their time, adhering to local minimum wage standards. We paid participants approximately $35,000 for the construction of the dataset. The dataset was designed to promote inclusivity, with balanced representation across gender, age, and regional backgrounds. However, we acknowledge potential biases in speech recognition performance for certain demographic groups (e.g., elderly speakers or regional accents), as highlighted in our analysis. These limitations are documented to encourage future work toward equitable AI systems. While RealTalk-CN aims to advance robust speech-based LLMs, we recognize potential misuse risks and we have adopted a gated access mechanism and conduct strict audits to ensure that the data will not be abused.

Reproducibility Statement

We have made extensive efforts to ensure the reproducibility of our work. All inference code, and automatic evaluation code are released at our anonymous repository: `https://anonymous.4open.science/r/RealTalk-7969`. The prompts used for automatic evaluation are provided in Appendix A.8. The proposed dataset, RealTalk-CN, will be fully released to the community to support reproducibility and further scientific research. Theoretical assumptions, significance analyses, and additional methodological details are presented in Appendix A.9 and Appendix A.10. We also carefully controlled random seeds and standardized experimental settings to minimize variance and ensure consistent reproducibility of results.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Junyi Ao et al. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56898–56918, 2024.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*, 2020.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL `https://aclanthology.org/D18-1547/`.

Yucheng Cai, Wentao Ma, Yuchuan Wu, Shuzheng Si, Yuan Shao, Zhijian Ou, and Yongbin Li. UniPCM: Universal pre-trained conversation model with task-aware automatic prompt. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17042–17061, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.lrec-main.1481/`.

Ronald Carter. Spoken grammar: what is it and how can we teach it? michael mccarthy and ronald carter. *ELT journal*, 49:3, 1995.

Yang Chen, Hui Wang, Shiyao Wang, Junyang Chen, Jiabei He, Jiaming Zhou, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. Seniortalk: A chinese conversation dataset with rich annotations for super-aged seniors. *arXiv preprint arXiv:2503.16578*, 2025.

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

Xize Cheng et al. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*, 2025.

Anna Seo Gyeong Choi and Hoon Choi. Fairness of automatic speech recognition: Looking through a philosophical lens. *arXiv preprint arXiv:2508.07143*, 2025. URL `https://arxiv.org/abs/2508.07143`.

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

Changwoo Chun, Daniel Rim, and Juhee Park. Llm contextbridge: A hybrid approach for intent and dialogue understanding in ivsr. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 794–806, 2025. URL https://aclanthology.org/2025.coling-industry.66/.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Wenjie Dong, Sirong Chen, and Yan Yang. ProTOD: Proactive task-oriented dialogue system based on large language model. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9147–9164, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.614/.

Mihail Eric and Christopher D Manning. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*, 2017.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's disease*, 49(2):407–422, 2015.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. URL https://aclanthology.org/H90-1021/.

Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pp. 263–272, 2014.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010, 2023.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. GPT-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=qaxhBG1UUaS.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. "how robust ru?": Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1147–1154. IEEE, 2021.

Jean C Krause and Louis D Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1):362–378, 2004.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

Ajinkya Kulkarni, Anna Tokareva, Rameez Qureshi, and Miguel Couceiro. The balancing act: Unmasking and alleviating ASR biases in Portuguese. In Bharathi Raja Chakravarthi, Bharathi B, Paul Buitelaar, Thenmozhi Durairaj, György Kovács, and Miguel Ángel García Cumbreras (eds.), *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pp. 31–40, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.ltedi-1.4/.

Linqin Li, Dong Zhang, Suyang Zhu, Shoushan Li, and Guodong Zhou. Response generation in multi-modal dialogues with split pre-generation and cross-modal contrasting. *Information Processing & Management*, 61 (1):103581, 2024a.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, et al. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*, 2025.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 3(7), 2024b.

Zongxia Li, Ishani Mondal, Yijun Liang, Huy Nghiem, and Jordan Lee Boyd-Graber. Panda (pedantic answer-correctness determination and adjudication): Improving automatic evaluation for question answering and text generation. *CoRR*, abs/2402.11161, 2024c. URL https://doi.org/10.48550/arXiv.2402.11161.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. *arXiv preprint arXiv:2402.12786*, 2024.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *CoRR*, abs/2503.04721, March 2025. URL https://doi.org/10.48550/arXiv.2503.04721.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

Yao Qian, Ximo Bianv, Yu Shi, Naoyuki Kanda, Leo Shen, Zhen Xiao, and Michael Zeng. Speech-language pre-training for end-to-end spoken language understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7458–7462. IEEE, 2021.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 930–940, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.67. URL https://aclanthology.org/2020.emnlp-main.67/.

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

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. Reliable LLM-based user simulator for task-oriented dialogue systems. In Yvette Graham, Qun Liu, Gerasimos Lampouras, Ignacio Iacobacci, Sinead Madden, Haider Khalid, and Rameez Qureshi (eds.), *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pp. 19–35, St. Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.scichat-1.3/.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*, 2018.

E Shriberg. Preliminaries to a theory of speech disfluency. *PhD Diss, Univ. of California*, 1994.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36:39088–39118, 2023.

Student. The probable error of a mean. *Biometrika*, pp. 1–25, 1908.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah (eds.), *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 109–117, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.13. URL https://aclanthology.org/2020.nlp4convai-1.13/.

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

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295, 2020. doi: 10.1162/tacl_a_00314. URL https://aclanthology.org/2020.tacl-1.19/.

Su Zhu, Zijian Zhao, Tiejun Zhao, Chengqing Zong, and Kai Yu. Catslu: The 1st chinese audio-textual spoken language understanding challenge. In *2019 International Conference on Multimodal Interaction*, pp. 521–525, 2019.

A APPENDIX

A.1 THE SPECIFIC PROVINCES INCLUDED IN EACH REGION.

Table 7: Region Mapping of Specific Provinces

Province	Region	Province	Region
Liaoning	Northeast China	Hebei	North China
Beijing	North China	Shanxi	North China
Inner Mongolia	North China	Shandong	East China
Shanghai	East China	Jiangsu	East China
Anhui	East China	Jiangxi	East China
Fujian	East China	Taiwan	East China
Henan	Central China	Hubei	Central China
Hunan	Central China	Guangdong	South China
Guangxi	South China	Hainan	South China
Chongqing	Southwest China	Sichuan	Southwest China
Guizhou	Southwest China	Gansu	Northwest China
Xinjiang	Northwest China	Tibet	Southwest China

A.2 INTENTS AND SLOTS INFORMATION

All Intents and Slots information has been translated into English. The Chinese version can be found on our dataset website.

A.2.1 INTENTS

Table 8: All intents labels in RealTalk-CN. The original Chinese data were translated into English.

Intent 1	Intent 2	Intent 3	Intent 4
None	Introduce Person	Introduce Work	Introduce Plot
Introduce History	Introduce Geography	Introduce Astronomy	Introduce Film
Introduce Constellation	Introduce Scenic Spot	Introduce Plant	Introduce Game
Introduce Cuisine	Seek Recommendation	Recommend Book	Recommend Product
Recommend Brand	Recommend Location	Recommend Film	Recommend Scenic Spot
Recommend Game	Recommend Destination	Recommend Cuisine	Recommend Restaurant
Provide Location	Provide Channel	Provide Type	Provide Options
Provide Approach	Raise Question	Express Opinion	Ask Question
State Demand	Explain Reason	Explain Precautions	Inquire about Person
Inquire about Price	Inquire about Preparation	Inquire about Transport	Inquire about History
Inquire about Reason	Inquire about Location	Inquire about Weather	Inquire about Quantity
Inquire about Time	Inquire about Scenic Spot	Inquire about Precautions	Inquire about Features
Inquire about Type	Inquire about Cuisine	Inquire about Route	Inquire about Approach
Explain Precautions			

A.2.2 SLOTS

Table 9: All slots labels in RealTalk-CN. The original Chinese data were translated into English.

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5
O	APP	Book	Internet	Transport
Transport Route	Person Name	Number of People	Character	Group of People
Price	Price Range	Sports Event	Work	Company
Relationship	Departure Place	Mode of Transport	Function	Animal
Cosmetics	Hospital	Historical Event	Famous Person	Brand
Variety	Factor	Country	Region	Address
Location	Venue	City	Celestial Body	Astronomical Phenomenon
Weather	Award	Name	Season	School
Subject	Religion	Pet	Residential Area	Mountain Range
Tool	Platform	Age	Store	Building
Ingredient	Mobile Phone Brand	Mobile Phone Model	Skin Care Product	Measure
Digital Product	Quantity	Stationery	Tourist Attraction	Date
Period	Time	Time Range	Constellation	Scenic Spot
Clothing	Attire	Dynasty	Organization	Material
Plant	Frequency	Song	Fruit	Activity
Channel	Temperature	Temperature Range	Game	Game Type
Item	Feature	Toy	Movie	TV Series
Disease	Symptom	Destination	Province	Ticket
Station	Type	Cuisine	Occupation	Art
Festival	Cuisine Style	Vegetable	Planet	Document
Course	Distance	Identity	Software	Sport
Communication Tool	Hotel	Snack	Field	Color
Style	Ingredient	Food	Restaurant	Beverage

A.3 TRAINING, VALIDATION, AND TEST SET DIVISIONS

Table 10: Dataset Comparison Across Training, Validation, and Testing Sets

Category	Samples	Avg Utterance Length	Avg Dialog Rounds
		Training Set	
MD-Col	6,269	27.60	8.54
MD-Sys	28,363	19.36	7.74
SD-Col	1,458	25.56	8.23
SD-Sys	5,848	28.90	7.58
		Validation Set	
MD-Col	2,687	27.62	8.54
MD-Sys	8,728	19.51	7.72
SD-Col	626	25.00	8.17
SD-Sys	2,504	20.89	7.75
		Testing Set	
MD-Col	3,837	27.42	8.54
MD-Sys	3,837	19.27	7.73
SD-Col	892	25.61	8.14
SD-Sys	892	20.76	7.58

A.4 Tri-modality evaluation on SD subsets

Table 11: Tri-modality evaluation on SD subsets. Columns represent pure-text, speech, and cross-modal settings.

Setting	SD-Col (Text)	SD-Col (Speech)	SD-Col (Cross-Modal)	SD-Sys (Text)	SD-Sys (Speech)	SD-Sys (Cross-Modal)
Baichuan-Audio	6.98	7.79	7.58	7.37	7.68	7.54
GLM-4-Voice	8.44	8.24	8.26	8.38	8.39	8.38
Qwen2-Audio	7.92	7.85	7.61	7.95	8.06	7.77
Baichuan-Omni	7.10	7.34	7.69	7.56	7.72	7.76
MiniCPM-o	7.98	8.19	8.34	8.21	8.33	8.40
Qwen2.5-Omni	7.93	7.78	7.66	7.95	8.04	7.83

A.5 Intent Classification Accuracy Across Disfluency Types

Table 12: Intent Classification Accuracy (%) across disfluency types. Std. Dev.: standard deviation across types.

Model	Hesitation	Self-Correction	Repetition	Modal Particle Drag	SD-Sys	Std. Dev.
Pipeline	52.97	56.10	52.00	59.25	54.83	3.28
Baichuan-Audio	25.33	30.89	25.33	30.82	28.20	2.77
GLM-4-Voice	26.83	27.24	26.67	16.10	19.49	5.41
Qwen2-Audio	25.60	24.39	21.33	14.72	18.14	4.87
Baichuan-Omni	35.80	33.33	37.33	33.33	34.53	3.40
MiniCPM-o	39.83	41.87	37.33	35.96	35.84	2.63
Qwen2.5-Omni	25.43	22.36	21.33	15.07	18.17	4.35
GPT-4o-Audio	46.08	42.68	49.33	42.68	45.04	3.69

A.6 Use of large language models

We have employed a large language model to polish the writing of our paper. The content has been manually verified by us to ensure accuracy, factuality, and the absence of hallucinations.

A.7 TASK EXAMPLES AND EVALUATION METHODS

For each data, we first perform the intent classification task, then input the intent predicted by the model, and require the model to answer a specific slot value for a certain intent, thereby completing the slot filling task. Figure 8 is an example of the intent classification task, and Figure 9 is an example of the slot filling task for the dialogue example in Table 13. The dialogue has been translated into English.

Table 13: Task Example. The original Chinese data were translated into English.

User Query	Model Response
I want to go to the seaside,can you recommend some beaches near Dalian?	
	Bangchu Island Beach, Golden Pebble Beach, and Dachangshan Island.
What other interesting places are nearby?	
	You can also visit Tiger Beach and Dalian Aquarium.
What are the opening hour of Dalian Aquarium?	

Prompt: The above is the current audio of the last paragraph. Just choose the most appropriate answer from the following options as the intention (i.e. purpose or goal) of the last round of speech: {**choices**}. Your answer should strictly follow the following format: Intent is: xx

"choices": "['Introduce works', 'Introduce history', 'Introduce geography', 'Introduce astronomy', 'Introduce attractions', 'Introduce food', 'Seek recommendations', 'Recommend books', 'Recommend products', 'Recommend brands', 'Recommend places', 'Recommend movies and TV shows', 'Recommend attractions', 'Recommend destinations', 'Recommend food', 'Recommend restaurants', 'Provide positioning', 'Provide channels', 'Provide types', 'Provide choices', 'Provide ways', 'Raise questions', 'Raise insights', 'Raise needs', 'Explain reasons', 'Explain precautions', 'Ask about prices', 'Ask about preparations', 'Ask about travel methods', 'Ask about history', 'Ask about reasons', 'Ask about locations', 'Ask about weather', 'Ask about quantity', 'Ask about time', 'Ask about attractions', 'Ask about precautions', 'Ask about channels', 'Ask about features', 'Ask about types', 'Ask about food', 'Ask about routes', 'Ask about ways', 'Explain precautions']"

Answer: Ask about time

Figure 8: Example of the intent classification task. During evaluation, the choices in prompt are replaced with the choices variable, and then the model output is compared with the label to calculate the score. The original Chinese data were translated into English.

Prompt: Please carefully analyze the current audio content and extract key information in combination with the user's intent "{**intent_pre**}".
Among them, "{**k**}" is an important attribute of the intent. Please identify the specific content of "{**k**}" that is only related to the intent "{**intent_pre**}" from the voice.
Please strictly follow the following format when answering:
{**k**} is: [specific value 1, specific value 2,... specific value n]

intent_pre: Ask about time
k: Tourist attractions
Answer: Dalian Oceanarium

Figure 9: An example of the slot filling task, using the model prediction result of the intent classification task as the input "intent_pre", k specifies the slot type, and Answer is the correct label. The original Chinese data were translated into English.

A.8 GPT-4 AUTOMATIC EVALUATION METHOD

Figures 10 and 11 show the prompts we use for reference automatic evaluation and no-reference automatic evaluation, respectively. We refer to the evaluation prompts in [28].

```
meta_prompt_ref = """
    I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output.
    Your task is to rate the model' s responses based on the provided user input transcription [Instruction], the model' s output transcription [Response], and the basic reference answer [Reference].

    Please evaluate the response on a scale of 1 to 5:
    1 point: The response is largely irrelevant, incorrect, or fails to address the user' s query. It may be off-topic or provide incorrect information.
    2 points: The response is somewhat relevant but lacks accuracy, completeness, or alignment with the user' s query. It may only partially answer the question or include extraneous content.
    3 points: The response is generally accurate and relevant, providing information consistent with the reference answer but may lack depth, clarity, or contextual adaptation.
    4 points: The response is accurate, relevant, and contextually appropriate, not only matching the reference but also providing a clear and well-structured answer that aligns well with the user' s query.
    5 points: The response is exceptionally accurate, relevant, and informative. It goes beyond the basic reference answer, offering richer, clearer, or more contextually appropriate information while fully addressing the user' s query.

    Below are the transcription of user' s instruction, reference answer, and model' s response:
    ### [Instruction]: {prompt}
    ### [Reference]: {reference}
    ### [Response]: {response}

    After evaluating, please output the score only without anything else.
    You don' t need to provide any explanations.
    """
```

Figure 10: The prompt that uses gpt4 for automatic evaluation with reference. The prompt provides the conversation context, the reference is the reference answer to the current round of questions, and the response is the actual prediction of the model.

```
meta_prompt_open= """
    I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output.
    Your task is to rate the model's responses based on the provided user input transcription [Instruction] and the model's output transcription [Response].

    Please evaluate the response on a scale of 1 to 5:
    1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information.
    2 points: The response is somewhat relevant but lacks accuracy or completeness. It may only partially answer the user's question or include extraneous information.
    3 points: The response is relevant and mostly accurate, but it may lack conciseness or include unnecessary details that don't contribute to the main point.
    4 points: The response is relevant, accurate, and concise, providing a clear answer to the user's question without unnecessary elaboration.
    5 points: The response is exceptionally relevant, accurate, and to the point. It directly addresses the user's query in a highly effective and efficient manner, providing exactly the information needed.

    Below are the transcription of user's instruction and models' response:
    ### [Instruction]: {prompt}
    ### [Response]: {response}

    After evaluating, please output the score only without anything else.
    You don't need to provide any explanations.
    """
```

Figure 11: The prompt that uses gpt4 for automatic evaluation without reference. The prompt provides the conversation context, and the response is the actual prediction of the model.

A.9 Speaker Significance Test Results

The model's significance test results for the speaker's age and region. We only list the significant groups. The p-value represents the result of the Mann-Whitney U test [35], and the d-value represents Cohen's d effect size [36].

A.9.1 Intent Classification by Age

Table 14: Significant Differences in Intent Classification by Age.

Model	Age Group	p-value	d-value	Comparison
minicpm	35-44	0.00	0.16	HIGHER
	21-24	0.04	-0.05	LOWER
qwen2.5_omini	25-34	0.02	-0.06	LOWER
	35-44	0.00	0.16	HIGHER
naive2	21-24	0.01	-0.07	LOWER
	45-54	0.05	-0.20	LOWER
glm	35-44	0.00	0.19	HIGHER
	45-54	0.00	0.32	HIGHER
qwen2	25-34	0.00	-0.08	LOWER
	35-44	0.00	0.15	HIGHER

A.9.2 Intent Classification by Region

Table 15: Significant Differences in Intent Classification by Region.

Model	Region	p-value	d-value	Comparison
minicpm	Northeast China	0.00	0.34	HIGHER
	Northwest China	0.00	0.46	HIGHER
	Central China	0.00	-0.21	LOWER
gpt4o_mini	South China	0.00	-0.07	LOWER
	North China	0.00	0.13	HIGHER
	Northeast China	0.00	0.34	HIGHER
	Northwest China	0.02	0.25	HIGHER
qwen2.5_omini	Northwest China	0.00	0.35	HIGHER
naive2	North China	0.01	0.11	HIGHER
glm	East China	0.00	0.10	HIGHER
	Northeast China	0.00	0.38	HIGHER
	Northwest China	0.02	0.26	HIGHER
	Central China	0.00	-0.16	LOWER
qwen2	Northwest China	0.00	0.35	HIGHER

A.9.3 Slot Filling by Age

Table 16: Significant Differences in Slot Filling by Age.

Model	Age Group	p-value	d-value	Comparison
minicpm	45-54	0.01	-0.40	LOWER
gpt4o_mini	45-54	0.00	-0.53	LOWER
qwen2.5_omini	45-54	0.02	-0.38	LOWER

Table 16: Significant Differences in Slot Filling by Age.

Model	Age Group	p-value	d-value	Comparison
naive2	45-54	0.00	-0.51	LOWER
glm	45-54	0.05	-0.35	LOWER
baichuan_audio	25-34	0.01	0.09	HIGHER
	21-24	0.01	-0.09	LOWER
	45-54	0.00	-0.57	LOWER
baichuan_omni	45-54	0.00	-0.47	LOWER
qwen2	45-54	0.03	-0.37	LOWER

A.9.4 Slot Filling by Region

Table 17: Significant Differences in Slot Filling by Region.

Model	Region	p-value	d-value	Comparison
minicpm	South China	0.00	0.14	HIGHER
	Southwest China	0.00	-0.18	LOWER
gpt4o_mini	South China	0.00	0.11	HIGHER
	North China	0.02	0.12	HIGHER
	Southwest China	0.01	-0.12	LOWER
	East China	0.03	-0.10	LOWER
	Northwest China	0.00	-0.39	LOWER
qwen2.5_omini	South China	0.02	0.08	HIGHER
	North China	0.01	0.14	HIGHER
	Southwest China	0.00	-0.14	LOWER
	Northwest China	0.00	-0.36	LOWER
naive2	South China	0.01	0.10	HIGHER
	Northwest China	0.04	-0.29	LOWER
glm	South China	0.00	0.15	HIGHER
	Southwest China	0.01	-0.13	LOWER
	Northwest China	0.00	-0.43	LOWER
baichuan_audio	South China	0.00	0.11	HIGHER
	Southwest China	0.00	-0.14	LOWER
	Northwest China	0.01	-0.34	LOWER
baichuan_omni	South China	0.01	0.09	HIGHER
	Southwest China	0.01	-0.13	LOWER
	Northwest China	0.00	-0.43	LOWER
qwen2	South China	0.03	0.07	HIGHER
	North China	0.00	0.15	HIGHER
	Southwest China	0.00	-0.15	LOWER
	Northwest China	0.00	-0.36	LOWER

A.10 Cross-Modal Task Significance Test Results

Table 18: Cross-Modal Task Significance Test Results

Model	Test Method	Result	Mean Difference
minicpm	Paired t-test	t=-0.9440, p=0.3452 (not significant)	8.326 vs 8.359
	Wilcoxon test	W=362314.5, p=0.5570 (not significant)	
qwen2.5_omini	Paired t-test	t=6.8330, p=0.0000 (significant)	8.008 vs 7.719
	Wilcoxon test	W=400266.5, p=0.0000 (significant)	
glm	Paired t-test	t=0.6360, p=0.5251 (not significant)	8.429 vs 8.407
	Wilcoxon test	W=412345.0, p=0.3257 (not significant)	
baichuan_audio	Paired t-test	t=2.5170, p=0.0119 (significant)	7.672 vs 7.555
	Wilcoxon test	W=526585.5, p=0.0001 (significant)	
baichuan_omni	Paired t-test	t=-7.5190, p=0.0000 (significant)	7.460 vs 7.832
	Wilcoxon test	W=609023.5, p=0.0000 (significant)	
qwen2	Paired t-test	t=6.9420, p=0.0000 (significant)	7.996 vs 7.707
	Wilcoxon test	W=422586.5, p=0.0000 (significant)	

A.11 Equipment environment

The experiments were conducted using two RTX 4090 GPUs, running on an Ubuntu Linux operating system. To eliminate the effects of randomness and ensure the experiments are reproducible, the temperature was set to 0, and no sampling was performed.

A.12 Limitations

While the RealTalk-CN captures diverse speech disfluencies and speaker variations, it may not fully represent all regional accents and age groups in China's vast linguistic landscape. The cross-modal chat task, though innovative, currently focuses on alternating between speech and text modalities without exploring more complex multimodal interactions. The dataset's size, while substantial, could be expanded to cover even more domains and edge cases. These limitations point to valuable directions for future research to further enhance Chinese speech-based LLM evaluation.

A.13 Case Studies of Cross-Modal Chat Divergence

User: Can you help me find a 4S store nearby?

Assistant: The store has Lavida, Sagitar, Paster, Magotan, Bora and other models.

User: Can you recommend me a car with a price between 100,000 and 170,000 yuan?

Speech Chat response: According to your budget and needs, I recommend you to consider Volkswagen Passat. It is in the price range of 100,000 to 170,000 yuan, has high configuration, good comfort, and Volkswagen brand reputation.

Cross-Modal Chat response: Based on your budget and needs, I recommend you consider the following cars:

1. Mercedes-Benz C-Class: Mercedes-Benz C-Class is a luxury mid-size car with a price range of <u>RMB 400,000 to 500,000</u>. It has rich configurations and a comfortable driving experience, making it suitable for consumers who pursue a quality life...

Figure 12: Forgetting the key information and semantics of the previous context of different modalities in Cross-Modal chat task. The original Chinese data were translated into English.

User: I want to learn about Van Gogh.

Assistant:Van Gogh was a Dutch post-impressionist painter. He showed interest in painting from a very young age, but he started to focus on painting in his twenties.

User: Do you have any recommendations for movies or TV series about Van Gogh?

Speech Chat response: Yes, there are many movies and TV series about Van Gogh. Some movies and TV series about Van Gogh are very popular, such as "L'Éternel Printemps" and "Vincent and the Doctor". These works are all based on Van Gogh and tell about his life and creative career.

Cross-Modal Chat response: Yes, such as "Vincent and Theo", "Starry Night", "Van Gogh's Bridge", "Van Gogh's Ear", "Van Gogh", "Van Gogh's Last Days", "Van Gogh", "Van Gogh"...

Figure 13: Text degradation problem in Cross-Modal chat task. The original Chinese data were translated into English.

User: This professional quality is also very admirable. Can you introduce this anchor?
Assistant: Of course. Li Wenjing, born in Datong, Shanxi in 1974, is a female anchor of CCTV News Channel.
User: What is her hosting experience?
Assistant: Li Wenjing started working at Shanxi TV in 1995.
User: What is Wenjing's hosting style?
Assistant: Unlike ordinary news anchors, Li Wenjing likes to use body language on the screen, and she has a lot of gestures.
Assistant: Her tone is emphatic and very subjective.
User: Has she always been so professional in hosting?
Assistant: Of course not. Many controversial incidents occurred during her broadcasting.
User: It seems that professional professional qualities are gained from experience. “

Speech Chat response: Yes, Li Wenjing has also made steady progress in hosting.
Cross-Modal Chat response: Yes.

Figure 14: Reduced quality of response richness in Cross-Modal chat task. The original Chinese data were translated into English.