SpeechMedAssist: Efficiently and Effectively Adapting Speech Language Models for Medical Consultation

Anonymous ACL submission

Abstract

Medical consultations are intrinsically speech-centric. However, most prior works focus on long-text-based interactions, which are cumbersome and patient-unfriendly. Recent advances in speech language models (SpeechLMs) have enabled more natural speech-based interaction, yet the scarcity of medical speech data and the inefficiency of directly fine-tuning on speech data jointly hinder the adoption of SpeechLMs in medical consultation. In this paper, we propose SpeechMedAssist, a SpeechLM natively capable of conducting speech-based multi-turn interactions with patients. By exploiting the architectural properties of SpeechLMs, we decouple the conventional one-stage training into a two-stage paradigm consisting of **(1) Knowledge & Capability Injection via Text** and **(2) Modality Re-alignment with Limited Speech Data**, thereby reducing the requirement for medical speech data to only **10k** synthesized samples. To evaluate SpeechLMs for medical consultation scenarios, we design a benchmark comprising both single-turn question answering and multi-turn simulated interactions. Experimental results show that our model outperforms all baselines in both effectiveness and robustness in most evaluation settings. [1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in a wide range of vertical domains due to their strong language understanding and generation capabilities (Li et al., 2024a). In the medical domain, benefitting from the abundance of textual resources from online platforms and medical literature, LLMs are adapted for complex clinical tasks including medical reasoning (Chen et al., 2024a; Pan et al., 2025), patient triage (Zhang et al., 2023b) and the generation of clinical reports (Zhou et al., 2024) after supervised fine-tuning.

[1] Code&Audio Samples: Anonymous GitHub

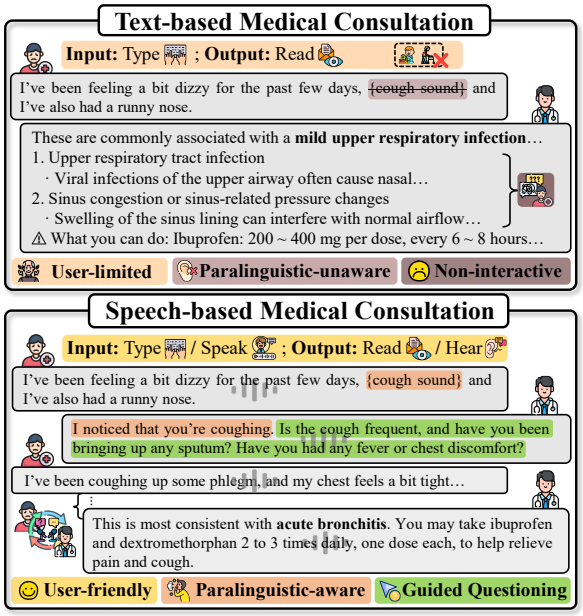

Figure 1: An illustration highlighting the limitations of text-based medical consultation, alongside the advantages of speech-based medical consultation.

Despite their success in knowledge-intensive tasks, LLM-based medical systems are ill-suited for interactive medical consultation. As shown in Figure 1, purely text-based interaction introduces substantial accessibility barriers for elderly patients and users with limited literacy or typing ability (Shi et al., 2024). Some works (Huang et al., 2024) attempt to extend text-based LLMs to speech-based interaction through cascaded systems composed of automatic speech recognition (ASR), an LLM, and text-to-speech (TTS) modules (Huang et al., 2024). However, such pipelines suffer from accumulated latency, ASR error propagation (Binici et al., 2025), and loss of paralinguistic cues such as cough, thereby undermining effective medical consultation (Ji et al., 2024).

In contrast, end-to-end speech language models (SpeechLMs) provide a promising alternative by natively supporting speech-based multi-turn inter-

action (Adams et al., 2025; Cui et al., 2025). Nevertheless, adapting SpeechLMs to medical consultation remains challenging: **(1) Lack of Medical Knowledge**: existing SpeechLMs are trained on general-purpose data, lacking domain-specific medical knowledge (Clusmann et al., 2023); **(2) Lack of Physician-level Clinical Skills**: in real-world medical consultations, professional clinical skills are required including symptom understanding, proactive inquiry, medical safety awareness, and sensitivity to paralinguistic signals in multi-turn interactions (Ng et al., 2024); **(3) Scarcity of Medical Speech Data**: the scarcity of medical speech data prevents direct fine-tuning of SpeechLMs to acquire medical knowledge and clinical skills, which is also inefficient (Banerjee et al., 2024).

To address the above challenges, we propose **SpeechMedAssist**, a SpeechLM tailored for speech-based multi-turn medical consultation. Motivated by the observation that SpeechLMs encode speech and text into a shared latent space, enabling them to acquire knowledge and skills from both text and speech modalities, we decouple the original one-stage fine-tuning purely using speech data into a two-stage paradigm: **(1) Knowledge&Capability Injection** from abundant text data and **(2) Modality Re-alignment** with limited medical speech data. Specifically, in the first stage, we freeze all speech-related modules of pretrained SpeechLMs and focus on injecting medical knowledge and consultation skills into the LLM core with large-scale medical text data. In the second stage, we unfreeze all modules and re-align the speech-text modality disrupted in the first stage with a small amount of medical speech dialogue data.

To support the proposed two-stage fine-tuning paradigm and endow the model with both medical knowledge and clinical consultation skills, we construct two complementary datasets. For the first stage, we construct **TextMedDataset** with 405k samples following a dedicated pipeline, in which lengthy medical text dialogues are rewritten into structured multi-turn conversations aligned with the clinical consultation workflow. For the second stage, we construct **SpeechMedDataset** with 198k samples by synthesizing the rewritten dialogues into patient-tailored spoken conversations.

For evaluation, we design a comprehensive benchmark **SpeechMedBench** comprising single-turn Q&A, multi-turn consultation evaluations in simulated clinical scenarios, and human evaluation on a small-scale in-the-wild dataset. This benchmark enables a systematic assessment of medical knowledge and clinical consultation skills from both objective and subjective perspectives, on which our model shows consistently strong performance. In addition, our model exhibits high output speech quality, robustness to acoustic noise, and strong retention of general-domain knowledge. In particular, further analysis shows that effective speech–text re-alignment can be achieved with a relatively small amount of synthesized medical speech data (10k samples in our setting). Our contributions are summarized as follows:

- We develop a unified rewriting-and-synthesis pipeline to construct TextMedDataset and SpeechMedDataset, enabling scalable creation of multi-turn medical speech dialogues.
- We propose SpeechMedAssist, a medical SpeechLM that introduces speech-based interaction into the medical domain through an efficient two-stage training strategy.
- We establish a comprehensive benchmark SpeechMedBench, including single-turn Q&A, multi-turn consultation in simulated scenarios, and human evaluation in the wild.

2 Model Architecture

Most existing SpeechLMs (KimiTeam et al., 2025; Fang et al., 2025a,b; Wu et al., 2025) adopt a *speech encoder–adaptor–LLM core–decoder* architecture. They encode speech into continuous representations and map into a speech–text aligned latent space via a speech adaptor, enabling the LLM to process speech and text within a shared semantic space. Intuitively, this architecture leverages the fact that speech conveys both linguistic content and paralinguistic cues to align speech with the existing semantic space of text (Ji et al., 2024), thereby facilitating the transfer of text-based knowledge and capabilities to the speech modality. Here, we briefly introduce this architecture that we focus on.

2.1 Speech Encoder & Speech Adaptor

Unlike text input, which can be tokenized into discrete tokens $\mathbf{x}_t$, speech input $\mathbf{x}_s$ is a continuous signal. SpeechLMs first employ a speech encoder $\mathcal{E}$ to encode the waveform $\mathbf{x}_s \in \mathbb{R}^{T_w}$ into speech features, which are then projected into the semantic space of the LLM via a speech adaptor $\mathcal{A}$. Similarly, text input $\mathbf{x}_t$ is mapped into text embeddings

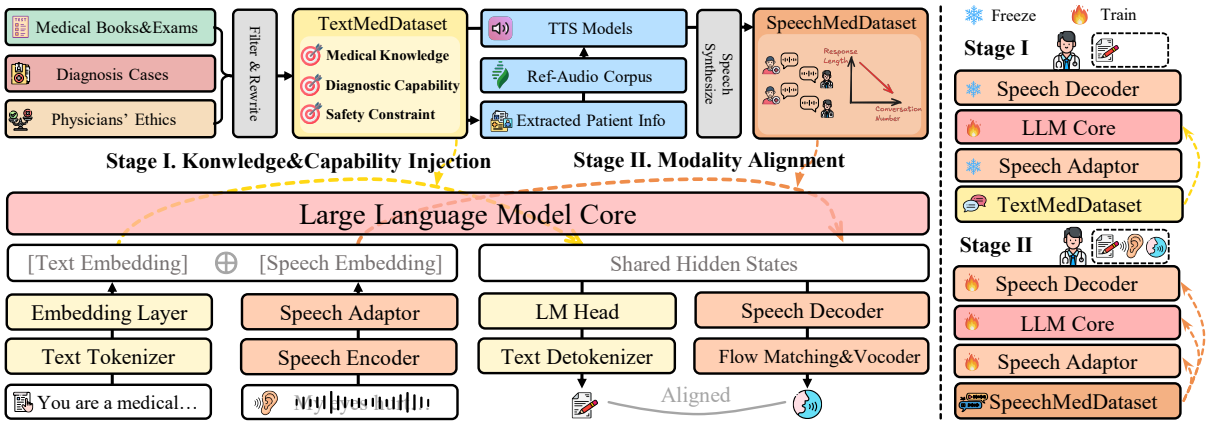

Figure 2: An overview of our work. **Data Constraction**: we construct TextMedDataset by filtering and rewriting collected medical text corpora, and build SpeechMedDataset by extracting patient information from dialogues and synthesizing matched speech. **Model Architecture**: we focus on the encoder–adaptor–LLM–decoder architecture, which supports text–speech dual-modal input and streaming output. **Training Strategy**: the first stage injects knowledge&capability into LLM core using TextMedDataset, while the second stage achieves modality re-alignment with a small amount of speech data from SpeechMedDataset.

via a tokenizer and embedding layer:

$$\mathbf{Z}_s = \mathcal{A}(\mathcal{E}(\mathbf{x}_s)) \in \mathbb{R}^{T_s \times d},$$
$$\mathbf{Z}_t = \mathrm{Emb}(\mathrm{Tokenizer}(\mathbf{x}_t)) \in \mathbb{R}^{T_t \times d}.$$

2.2 Large Language Model Core

To jointly process text instructions and speech inquiries, SpeechLMs concatenate text embeddings $\mathbf{Z}_t$ and speech embeddings $\mathbf{Z}_s$ and feed them into a shared LLM core f to obtain the hidden states $\mathbf{H}$ containing the information of response:

$$\mathbf{H} = f\big([\mathbf{Z}_t, \mathbf{Z}_s]\big) \in \mathbb{R}^{T_h \times d}.$$

2.3 Speech Generator & Vocoder

Given $\mathbf{H}$, the speech generator G maps them into unit representations $\mathbf{U}$, which are then converted into waveform $\hat{\mathbf{x}}_s$ by a speech vocoder f_{voc}:

$$\mathbf{U} = G(\mathbf{H}), \quad \hat{\mathbf{x}}_s = f_{\text{voc}}(\mathbf{U}).$$

Since both text and speech are derived from $\mathbf{H}$, and some SpeechLMs additionally leverage synchronously decoded text when generating unit tokens, the final outputs of speech and text exhibit high consistency, as verified in our experiments.

3 Training Strategy

In the architecture introduced above, the LLM core acts as the "brain", while the text tokenizer and speech encoder correspond to "reading" and "listening" modules, respectively. Previous neuroscience studies (Buchweitz et al., 2009) suggest that the human brain encodes knowledge in a modality-independent manner, which means that the knowledge and capability acquired from text can also be used in the speech modality. This observation motivates a two-stage training strategy for adapting SpeechLMs to medical consultation, as illustrated in Figure 2. Specifically, instead of directly fine-tuning on large-scale medical speech data, we first inject medical knowledge and diagnostic capabilities using large-scale text data, followed by modality re-alignment with a small amount of speech data. Here, we present the training strategy in detail and provide a preliminary theoretical analysis.

3.1 Inject Knowledge&Capability via Text

In the first stage, we freeze all speech-related modules of the SpeechLM, including the speech encoder $\mathcal{E}$, adaptor $\mathcal{A}$, generator G, and vocoder f_{voc}, reducing the SpeechLM to its LLM core f and text-related modules. Then, we train the LLM core with a large scale of medical text data, which directly updates the mapping $f : [\mathbf{Z}_t] \mapsto \mathbf{H}$, thereby equipping the LLM core with domain-specific medical knowledge and diagnostic ability through a data-driven manner. At this stage, the model is enhanced purely in the text modality, while its speech-related components remain unchanged.

3.2 Re-align Modalities with Limited Speech

The first stage is text-based training, similar to a medical student learning a lot from books and exercises, but knowing the material does not mean they can speak it out in a real clinical setting. Therefore,

the next challenge is to transfer these capabilities effectively to the speech modality. We refer to the domain adaptation theory and model this challenge by relating the error on the speech domain (target) to that on the text domain (source) and the divergence between their embeddings.

Formally, let $\epsilon_t(f)$ and $\epsilon_s(f)$ be the expected errors of f on the text and speech domains, respectively. The classical domain adaptation bound (Ben-David et al., 2006) gives, for any $f \in \mathcal{H}$,

$$\epsilon_s(f) \leq \epsilon_t(f) + \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_t, \mathcal{D}_s) + \lambda,$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ measures the divergence between text and speech modality in the aligned semantic space, and λ is the minimal combined risk. Since the LLM core is well optimized in the text domain, $\epsilon_t(f)$ is small, and the shared dialogue structure between medical and general dialogue implies a limited λ. Consequently, the bound suggests that speech-domain performance is mainly governed by the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$. For pre-trained SpeechLMs, text and speech modalities are already well aligned, and as evidenced in Appendix E, text-only training in Stage I induces only mild domain shifts. As a result, only a small amount of speech data is required to re-align the two modalities.

Concretely, Stage II consists of two parts: **(a)** unfreezing the speech adaptor $\mathcal{A}$ and jointly training it with the LLM core f on paired <speech input, text response>data; **(b)** unfreezing only the speech decoder G and training it on <speech input, speech response>pairs to improve speech generation.

4 Data Construction

Existing medical corpora are dominated by text-based single-turn question answering with fully detailed patient inputs and lengthy physician responses, which deviates from real-world medical consultations (Li et al., 2024b). To bridge this gap, we construct a scalable data construction pipeline that produces multi-turn medical dialogues aligned with clinical workflow, presented in Figure 2.

4.1 TextMedDataset

Medical Knowledge To inject sufficient medical knowledge into the LLM core, we collect three single-turn question–answering datasets (Wang et al., 2024, 2025b) detailed in Table 1 and rewrite the responses into concise and clear answers using Qwen2.5-32B-Instruct (Yang et al., 2024a). These data span 49 clinical departments and cover common diseases and medication usage.

Dataset	Description	Used Size
Knowledge Injection		
CMB-Exam	Multiple-choice questions in six categories	**189k**
Medical Encyclopedia	Single-turn Q&A on common diseases&medicines	**41k**
Medical Books	Single-turn Q&A on general medical knowledge	**40k**
Diagnostic Capability		
CMtMedQA	Multi-turn consultations on medical knowledge	**68k**
MedDG	Real multi-turn medical consultation dialogues	**16k**
HuatuoGPT2-SFT	Questions from real patient, answers from GPT-4	**48k**
Safety Constraint		
MedSafety-GPT4	Harmful Questions with safe responses from GPT-4	**450**
Reference Audio Data		
Aishell2	1,991 Mandarin speakers' audio across accents	**1000h**
Aishell3	218 Mandarin speakers' audio across accents	**85h**

Table 1: Overview of datasets used to construct TextMedDataset (405k) and SpeechMedDataset (198k).

Diagnostic Capability Beyond static knowledge, real-world consultations workflow are characterized by gradual symptom disclosure, proactive inquiry, multi-turn information refinement, and evidence-based clinical decision-making (Roter and Hall, 1987; Iversen et al., 2020). To model this process, we collect both single- and multi-turn consultation data (Yang et al., 2024b; Liu et al., 2022; Wang et al., 2025b), filter incomplete or irrelevant samples using Qwen2.5-14B-Instruct, and rewrite the remaining data with Qwen2.5-72B-Instruct into structured dialogues aligned with the consultation workflow. This procedure converts lengthy single-turn data into multi-turn consultations with an average of 6.58 turns, 36.4 characters per turn, and 3.3 follow-up questions per dialogue.

Safety Constraint Safety in medical LLMs refers to avoiding the generation of harmful, misleading, or overconfident medical advice. We enhance safety through both implicit and explicit supervision. Specifically, the aforementioned ability to proactively ask follow-up questions helps reduce speculative or overconfident responses when information is insufficient, while incorporating MedSafety training data (Han et al., 2024) improves the model's ability to appropriately refuse unsafe or out-of-scope medical requests.

4.2 SpeechMedDataset

Most previous works (Zhao et al., 2024; Fang et al., 2025b) randomly select a reference speech segment for synthesizing speech, ignoring speaker-specific characteristics. In contrast, we consider the patient's age and gender, which are crucial information in medical consultations. Specifically, we prompt Qwen2.5-14B-Instruct to analyze doctor–patient dialogues and infer the patient's likely gender (male, female, or unknown) and age group (child, young adult, adult, elderly, or unknown). To

support robust reference selection, we construct a 1,000-hour speech–text paired pool from publicly available ASR datasets Aishell2 and Aishell3 (Du et al., 2018; Shi et al., 2020), covering approximately 2,000 Mandarin speakers with diverse regional accents across China. During speech synthesis, we select reference speech that matches the patient attributes and generate speech using CosyVoice2 (Du et al., 2024). When both age and gender are unknown, we instead synthesize speech using FishSpeech (Liao et al., 2024) with its randomly sampled timbres. Following this procedure, we obtain **SpeechMedDataset**, a multi-turn spoken medical dialogue dataset containing 198k samples.

5 Experiments

Our initial research goal is to efficiently and effectively fine-tune a SpeechLM for medical consultation. Therefore, in this section, we comprehensively evaluate the model after the two-stage training from both objective and subjective perspectives, by comparing it with medical domain models and other general-purpose models, and further validate the effectiveness of our training methodology.

5.1 Experimental Setup

Model Configuration Our training method is applicable to all SpeechLMs that adopt the *encoder–adaptor–LLM–decoder* architecture. In our experiments, we choose LLaMA-Omni2-7B (Fang et al., 2025b) as the base model. To further verify the generality of the proposed training strategy, we also employ OpenS2S (Wang et al., 2025a) as an alternative base model, with the corresponding evaluation results reported in the Appendix B.

Training Details In the first stage, we fine-tune the LLM core of LLaMA-Omni2 on TextMedDataset following Section 3.1 with a batch size of 8 and learning rate 5×10^{-5}. In the second stage, we train the model on SpeechMedDataset as in Section 3.2, using batch size 1 and learning rate 1×10^{-5}. To ensure proper alignment between speech and text modalities and dynamically correct the medical knowledge possessed by the model during training, we incorporate the single-turn Q&A data from TextMedDataset, with the final training data maintaining 1:1 between speech and text.

Baselines Our evaluation covers the following categories of models. **(1) ASR+LLMs+TTS**: Various LLMs have been fine-tuned with medical corpus for text-based interaction, including DISC-MedLLM (Bao et al., 2023), Zhongjing (Yang et al., 2024b), Baichuan2 (Yang et al., 2023), and HuatuoGPT2 (Chen et al., 2023). We enable them to listen and speak by adopting an ASR+LLM+TTS pipeline, using SenseVoiceSmall[2] for ASR and CosyVoice2[3] for TTS. **(2) SpeechLMs**: As detailed in Appendix A, SpeechLMs fall into two architectures. We select GLM4-Voice (Zeng et al., 2024) to represent the first, while the second includes Kimi-Audio (KimiTeam et al., 2025), SpeechGPT2 (Open-Moss, 2025), Qwen2-Audio (Chu et al., 2024), and StepAudio2-mini (Wu et al., 2025). **(3) OmniLMs**: We also include the latest multi-modal models, including Qwen2.5-Omni (Xu et al., 2025), BaichuanOmni-1.5 (Li et al., 2025), and MiniCPM-o 2.6 (Yao et al., 2024). We also consider multi-modal medical model ShizhenGPT-Omni (Chen et al., 2025), which takes multi-modal input and generates text.

5.2 Evaluation

To evaluate our model and compare it with baselines, we construct SpeechMedBench and evaluate mainly four dimensions: medical knowledge, diagnostic capability, robustness, and speech quality.

Single-turn Q&A To assess models' medical knowledge across text and speech modalities, we use evaluation sets of two medical multiple-choice datasets, **CMB** (Wang et al., 2024) and **CME** (Liu et al., 2023), along with medical encyclopedia Q&A pairs randomly sampled from the Huatuo2-pretrain dataset (referred to as **Ency**), which cover a wide range of medical terminology without overlapping with the training data. We also adopt MedSafetyBench (referred to as **Safety**) (Han et al., 2024) to evaluate the medical safety of models, with scores ranging from 1 to 5.

Multi-turn Conversation Speech-based interaction requires strong conversational ability, while medical consultation further demands proactive patient engagement. To reflect real-world practice, we construct a virtual medical consultation environment comprising an LLM-driven patient, a chief examiner, and an intern doctor powered by the model under evaluation. The patient, conditioned on real doctor–patient dialogues from **MedDG** (Liu et al., 2022) or real patient cases from **AIHospital** (Fan

[2]https://github.com/FunAudioLLM/SenseVoice
[3]https://github.com/FunAudioLLM/CosyVoice

Model Type	Model	Single-turn Q&A				Multi-turn Conversation				Wild
		CMB ↑	CME ↑	Ency ↑	Safety ↓	MedDG ↑	AIHospital ↑	Resp.Len.	Conv.Num.	Vote ↑
LLMs$^{+ASR}_{+TTS}$	HuatuoGPT2*	60.39	69.16	63.45	2.18	79.25	80.70	242.44	3.62	20
	DISC-MedLLM*	36.16	35.10	63.41	1.76	80.66	79.55	200.05	3.74	7
	Zhongjing*	-	-	54.63	2.16	79.56	77.90	116.65	4.68	1
	Baichuan2-7B*	46.48	50.66	58.37	1.94	70.58	72.50	187.98	4.18	6
SpeechLMs	Kimi-Audio	-	-	**63.53**	1.64	82.01	80.81	132.02	3.85	0
	Qwen2-Audio	44.73	48.02	49.48	1.78	78.18	79.81	162.73	4.27	6
	GLM4-Voice	35.14	37.15	54.43	1.82	80.81	80.43	108.20	3.97	12
	SpeechGPT2	35.57	35.57	56.65	2.48	82.36	80.28	114.28	3.54	5
	StepAudio2-mini	72.42	74.30	61.26	2.04	76.90	77.53	178.12	3.91	2
	LLaMA-Omni2	73.43	56.98	39.82	1.96	73.18	76.33	61.82	4.37	0
OmniLMs	Qwen2.5-Omni	76.83	75.33	58.12	1.72	76.46	76.53	252.89	3.32	1
	BaichuanOmni-1.5*	64.15	70.48	62.16	1.90	80.28	80.63	148.60	3.80	5
	MiniCPM-o 2.6	21.68	16.01	46.45	2.08	76.53	78.60	153.17	3.95	0
	ShizhenGPT-Omni*	74.58	71.95	53.72	2.18	76.06	76.51	1066.20	3.12	5
Ours	SpeechMedAssist	**77.96**	**75.48**	61.02	**1.32**	**83.26**	**83.40**	51.36	4.62	**26**

Table 2: Evaluation results of various models on Single-turn QA, Multi-turn conversation, and Wild metrics. '-' indicates that the metric is not available for that model. '*' means that the training data of the model includes medical data. **Bold** and underline indicate the highest and second highest performance, respectively.

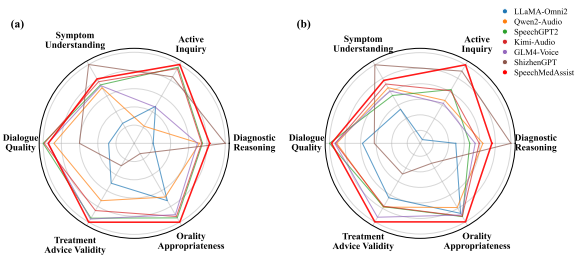

Figure 3: Comparison of our model with other models on multi-dimensions of multi-turn conversations metrics (a) MedDG and (b) AIHospital. Apart from a few dimensions that favor long-text responses, our model exhibits strong diagnostic capabilities.

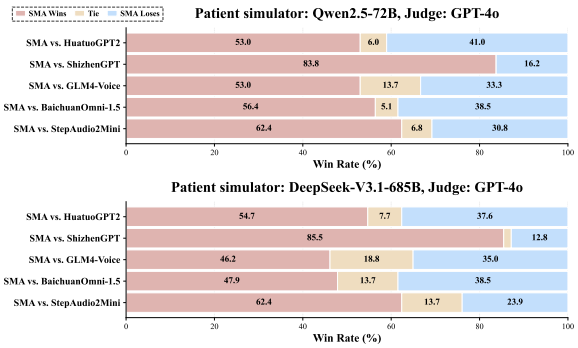

Figure 4: Win rates of our model against strong baselines, using Qwen2.5-72B and DeepSeek-V3.1-685B as patient simulators and GPT-4o as the judge. Our model achieves higher win rates in all settings.

et al., 2025), engages in multi-turn consultation with the intern doctor and terminates the dialogue once sufficient diagnostic and treatment advice is obtained. The intern doctor has no access to patient information and must elicit all relevant details through interaction. Finally, a chief examiner powered by Qwen2.5-72B acting as an LLM-based judge (Zheng et al., 2023) evaluates dialogues from six perspectives, as detailed in Appendix I.

Wild To provide an intuitive comparison of model performance in real-world settings, we collect 20 sets of patient questions recorded in real clinical environments. Unlike synthesized speech in simulated setting, these real-world recordings contain significant background noise and disorganized speech. After obtaining each model's single-turn responses, we invite five medical professionals to **vote** on each set, selecting the response that most closely resembles what a real doctor would provide. We have released the real patient queries together with the responses from all models.

Speech Quality We evaluate speech response quality from three aspects: **(1) UTMOS** measures speech naturalness using a MOS prediction model (Saeki et al., 2022); **(2) ASR-CER** evaluates text–speech consistency by transcribing the generated speech with an ASR model and computing the character error rate against the target text; and **(3) Latency** is the time from the start of speech input to the generation of the first speech chunk.

5.3 Main Results

Table 2 reports the evaluation results of LLMs, SpeechLMs, OmniLMs, and our model on single-

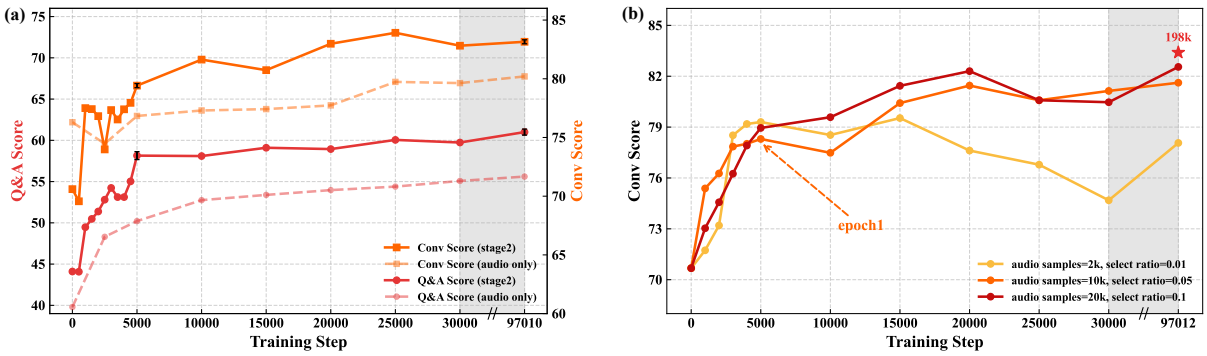

Figure 5: **(a)**: Comparison of the performance between the model trained in Stage II and the model trained from scratch on speech data, for single-turn Q&A and multi-turn conversation evaluations across training steps. To ensure the reliability of our conclusions, we compute the variance at step 5k and 97k. **(b)**: Comparison of conv score variations across training steps, where models are trained with different amounts of speech data. Remarkably, using only 10k audio samples yields performance close to that of a model trained with 198k samples.

turn Q&A, multi-turn conversation, and wild tasks. All metrics in the table are assessed through speech-based interaction, except for CMB and CME only in text form. Results of the text-based evaluation are provided in the Appendix F. On most metrics, our model achieves the best performance.

Medical Knowledge Mastery and Safety Assurance Text-based evaluations on CMB and CME show that our model outperforms both general-purpose and medical domain models, indicating effective medical knowledge acquisition in Stage I and stable preservation after Stage II. For speech-based Ency and Safety metrics, our model achieves competitive or superior performance, demonstrating accurate recognition of domain-specific medical terminology and strong medical safety performance. Meanwhile, our model retains its general-domain knowledge in both text and speech modalities after training, as detailed in the Appendix D.

Medical Consultation Skills Competency As shown in Table 2, in two different background settings, our model consistently achieves the best performance while generating concise responses and maintaining a moderate number of turns, which aligns better with real-world medical consultations. These results are robust to judge-model bias, as shown in Appendix C. To intuitively compare models' capabilities, we visualize the performance across six dimensions in Figure 4. Overall, our model achieves superior results on most metrics. In particular, ShizhenGPT produces responses nearly 20 times longer than ours, which boosts its scores in reasoning and understanding of symptoms, but significantly reduces efficiency and interactivity.

Model	Ency↑	Safety↓	MedDG↑	AIHospital↑
Backbone	39.82	1.96	73.18	76.33
+ Stage I	44.17$_{\uparrow 4.35}$	1.56$_{\downarrow 0.40}$	72.81$_{\downarrow 0.37}$	70.68$_{\downarrow 5.65}$
+ Stage II	61.02$_{\uparrow 21.20}$	1.32$_{\downarrow 0.64}$	83.26$_{\uparrow 10.08}$	83.40$_{\uparrow 7.07}$
Audio Only	55.60$_{\downarrow 5.42}$	1.82$_{\uparrow 0.50}$	79.01$_{\downarrow 4.25}$	80.21$_{\downarrow 3.19}$

Table 3: Evaluation results comparing different training stages and the audio-only setting.

In addition, we conduct pairwise comparisons between our model and several top-performing baselines. Specifically, we use Qwen2.5-72B and DeepSeek-V3.1 (DeepSeek-AI, 2024) separately as patient simulators, and compute the win rates by employing GPT-4o (OpenAI, 2024) as judge to assess each paired consultation with the prompt detailed in Appendix J. As shown in the Figure 4, our model consistently outperforms other baselines. To improve the reliability of our evaluation, we further conduct human evaluation in real-world settings. As shown in the Wild metrics, our model receives the most votes from medical professionals, highlighting its fidelity to actual clinical consultations.

5.4 Effectiveness & Efficiency

Our two-stage training strategy shifts the injection of knowledge and skill from speech to text modality, allowing only a small amount of speech data for modality re-alignment in Stage II. Here we further analyze the training effectiveness & efficiency.

Effectiveness of Two-Stage Training We conduct an ablation study to assess each training stage and compare our two-stage strategy with one-stage audio-only training. As shown in Table 3, injecting knowledge and skills via text in Stage I slightly improves medical terminology recognition and safety,

Model	Noise Robustness			Cough
	Noise=0	Noise=0.2	Noise=0.6	
Zhongjing+ASR+TTS	54.63	$53.49_{\downarrow 1.14}$	$50.95_{\downarrow 3.68}$	0.0%
Qwen2-Audio+TTS	49.48	$46.34_{\downarrow 3.14}$	$43.85_{\downarrow 5.63}$	10.2%
ShizhenGPT+TTS	53.72	$52.27_{\downarrow 1.45}$	$49.20_{\downarrow 4.52}$	16.2%
GLM4-Voice	54.43	$53.60_{\downarrow 0.83}$	$48.25_{\downarrow 6.18}$	8.5%
BaichuanOmni-1.5	62.16	$59.15_{\downarrow 3.01}$	$55.34_{\downarrow 6.82}$	5.9%
LLaMA-Omni2	39.82	$30.47_{\downarrow 9.35}$	$29.78_{\downarrow 10.04}$	1.7%
SMA-Stage II-10k	58.14	$55.82_{\downarrow 2.32}$	$51.79_{\downarrow 6.35}$	48.7%
SMA-Stage II-198k	61.02	$58.99_{\downarrow 2.03}$	$58.67_{\downarrow 2.35}$	57.2%

Table 4: Robustness under different noise levels and coughing perception. Our model exhibits strong noise robustness while effectively capturing cough cues.

Model	Input	Output	UTMOS ↑	ASR-CER ↓	Latency ↓
Zhongjing	📝	📝	**3.96**	6.77	3520ms
Qwen2-Audio	🎙, 📝	📝	**3.96**	11.83	4072ms
Kimi-Audio	🎙, 📝	🔊, 📝	2.55	**4.94**	3134ms*
GLM4-Voice	🎙, 📝	🔊, 📝	3.00	15.3	1562ms
SpeechGPT2	🎙, 📝	🔊, 📝	2.49	15.3	8470ms*
LLaMA-Omni2	🎙, 📝	🔊, 📝	3.69	8.06	374ms
SpeechMedAssist	🎙, 📝	🔊, 📝	3.75	7.71	**367ms**

Table 5: Input/output capabilities and output speech qualities of different models. '*' indicates that streaming generation is not supported in the official code.

but degrades multi-turn conversation performance, likely due to disruption of the shared text–speech latent space. Importantly, modality re-alignment in Stage II effectively restores and further improves performance, proving its necessity as analyzed in Section 3.2. In contrast, audio-only training consistently underperforms, highlighting both the difficulty and inefficiency of acquiring medical knowledge directly from the speech modality.

Speech Data Demand of Modality Re-alignment Since Stage I already endows the LLM core with medical knowledge and diagnostic skills, as proved in Appendix F, Stage II focuses on aligning speech and text modalities using limited speech data. As shown in Figure 5a, speech-related performance increases sharply within the first 0-5k training steps, with growth rates **91**× and **43**× higher than those in later steps for Ency and AIHospital score, respectively. This indicates that modality re-alignment occurs primarily in this early phase, where knowledge and skills learned from text rapidly transfers to speech modality. In contrast, directly training on speech data leads to substantially slower improvements. We further vary the amount of speech data used in Stage II, as shown in Figure 5b. Insufficient data leads to overfitting, while gains saturate beyond 10k samples. Overall, these results suggest that approximately **10k** speech samples are sufficient for effective modality re-alignment.

5.5 Speech Input Capability&Output Quality

Noise Robustness Real-world medical consultations involve diverse acoustic challenges. To evaluate noise robustness, we additively superimpose noise samples from MS-SNSD (Reddy et al., 2019) (e.g., babble) onto the original speech in the single-turn setting, and quantify the noise intensity using CER. As the noise level increases from 0 to 0.2 and 0.6, the CER rises from 9.77% to 10.20% and 12.19%, respectively. As shown in Table 4, although all models degrade under stronger noise, our model consistently maintains performance and remains competitive even at the highest noise level.

Cough Awareness To explore our model's capacity to perceive paralinguistic cues, we design experiments focusing on coughing, a clinically relevant signal. We insert cough segments into user speech and evaluate whether models can detect and leverage them, detailed in Appendix G. Results in Table 4 show that cascaded models fail to capture coughing, whereas our model perceives it in most cases and uses it for reasoning or proactive inquiry.

Speech Output Quality Beyond diagnostic capability, medical consultation also requires low-latency interaction and fidelity to speech. Table 5 compares cascaded models, general-purpose SpeechLMs, and our model in terms of speech quality. Cascaded models achieve higher UTMOS and lower ASR-CER by using state-of-the-art TTS module, but suffer from higher latency. Overall, our model supports both text&speech input and streaming output, achieving TTS-level speech quality and competitive latency compared to other SpeechLMs.

6 Conclusion

In this work, we propose SpeechMedAssist, a medical SpeechLM that supports real-time speech-based medical consultation. To address the scarcity of medical speech data, we propose an efficient two-stage training approach, design a pipeline for constructing medical speech dialogue data, and establish a comprehensive benchmark, which further demonstrates the effectiveness and efficiency of our method. Overall, this work provides a reference for applying SpeechLMs in vertical domains that lack large-scale speech data, and paves the way for deploying SpeechLMs in vertical applications.

Limitations

Medical consultations rely on multimodal information to support accurate diagnosis. In this work, we focus on text and speech as the input and output modalities, leaving the integration of additional modalities for future work.

Although our study focuses on Mandarin, the reference audio spans diverse accent regions, and random timbre sampling with FishSpeech is used to enhance generalization. Extending our framework to additional languages and dialects remains an important direction for future research.

Ethical Considerations

Most of the original data used in this paper are publicly available, as summarized in Table 1. These data are used in compliance with their open-source licenses and have undergone appropriate anonymization. Similar to existing text-based medical LLMs, our model may inevitably suffer from issues such as hallucination. Therefore, practical deployment requires additional safeguards, including input quality verification (e.g., ASR-based validation) and systematic review of model outputs.

References

Scott J. Adams, Julián Nicolás Acosta, and Pranav Rajpurkar. 2025. How generative AI voice agents will transform medicine. *npj Digit. Medicine*, 8(1).

Sourav Banerjee, Ayushi Agarwal, and Promila Ghosh. 2024. High-precision medical speech recognition through synthetic data and semantic correction: UNITED-MEDASR. *CoRR*, abs/2412.00055.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *CoRR*, abs/2308.14346.

Zhijie Bao, Qingyun Liu, Ying Guo, Zhengqiang Ye, Jun Shen, Shirong Xie, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2024. PIORS: personalized intelligent outpatient reception based on large language model with multi-agents medical scenario simulation. *CoRR*, abs/2411.13902.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 137–144. MIT Press.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T. Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F. Chen, and Stefan Winkler. 2025. MEDSAGE: enhancing robustness of medical dialogue summarization to ASR errors with llm-generated synthetic dialogues. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 23496–23504. AAAI Press.

Augusto Buchweitz, Robert A Mason, Lêda Tomitch, and Marcel Adam Just. 2009. Brain activation for reading and listening comprehension: An fmri study of modality effects and individual differences in language comprehension. *Psychology & neuroscience*, 2:111–123.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *CoRR*, abs/2412.18925.

Junying Chen, Zhenyang Cai, Zhiheng Liu, Yunjin Yang, Rongsheng Wang, Qingying Xiao, Xiangyi Feng, Zhan Su, Jing Guo, Xiang Wan, and 1 others. 2025. Shizhengpt: Towards multimodal llms for traditional chinese medicine. *arXiv preprint arXiv:2508.14706*.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *CoRR*, abs/2311.09774.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: a large-scale medical dialogue dataset. *https://github.com/UCSD-AI4H/Medical-Dialogue-System*.

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

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and

Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13943–13970. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Zheng Li, Junfeng Zhao, and Yasha Wang. 2025. Promed: Shapley information gain guided reinforcement learning for proactive medical llms. *CoRR*, abs/2508.13514.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: transforming mandarin ASR research into industrial scale. *CoRR*, abs/1808.10583.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10183–10213. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025a. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025b. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18617–18629. Association for Computational Linguistics.

John M. Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin R. An, Grace X. Zheng, and Bo Wang. 2023. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023*, pages 323–334. Association for Computational Linguistics.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. 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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Truong V Hoang, Quang H Nguyen, Cuong Q Nguyen, Phong X Nguyen, and Hoang D Nguyen. 2022. Sound-dr: Reliable sound dataset and baseline artificial intelligence system for respiratory illnesses. *arXiv preprint arXiv:2201.04581*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 23802–23804. AAAI Press.

Else Dalsgaard Iversen, Maiken Overbeck Wolderslund, Poul-Erik Kofoed, Pål Gulbrandsen, Helle Poulsen, Søren Cold, and Jette Ammentorp. 2020. Codebook for rating clinical communication skills based on the calgary-cambridge guide. *BMC medical education*, 20(1):140.

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. *CoRR*, 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, and 21 others. 2025. Kimi-audio technical report. *CoRR*, abs/2504.18425.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, Yiguan Lin, Bin Xu, Ren Bowen, Chong Feng, Yang Gao, and Heyan Huang. 2024a. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. 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 11116–11141. Association for Computational Linguistics.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W. Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. 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*.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

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

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - A comprehensive chinese medical exam dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *CoRR*, abs/2406.03712.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, volume 13551 of *Lecture Notes in Computer Science*, pages 447–459. Springer.

Si-Ioi Ng, Lingfeng Xu, Ingo Siegert, Nicholas Cummins, Nina R. Benway, Julie Liss, and Visar Berisha. 2024. A tutorial on clinical speech AI development: From data collection to model validation. *CoRR*, abs/2410.21640.

Open-Moss. 2025. Speechgpt 2.0-preview. https://github.com/OpenMOSS/SpeechGPT-2.0-preview.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Lara Orlandic, Tomas Teijeiro, and David Atienza. 2021. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *CoRR*, abs/2502.19634.

Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. 2019. A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech 2019*, pages 1816–1820.

Debra L Roter and Judith A Hall. 1987. Physicians' interviewing styles and medical information obtained from patients. *Journal of General Internal Medicine*, 2(5):325–329.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4521–4525. ISCA.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 2840–2861. Association for Computational Linguistics.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines. *CoRR*, abs/2010.11567.

Gagandeep Singh, Yue Pan, Jesús Andrés-Ferrer, Miguel A. del Agua, Frank Diehl, Joel Pinto, and Paul Vozila. 2023. Large scale sequence-to-sequence models for clinical note generation from patient-doctor conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023*, pages 138–143. Association for Computational Linguistics.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and Jiajun Zhang. 2025a. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. *CoRR*, abs/2507.05177.

Xidong Wang, Guiming Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. CMB: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21,*

2024, pages 6184–6205. Association for Computational Linguistics.

Xidong Wang, Jianquan Li, Shunian Chen, Yuxuan Zhu, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Junying Chen, Jie Fu, Xiang Wan, Anningzhe Gao, and Benyou Wang. 2025b. Huatuo-26m, a large-scale chinese medical QA dataset. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3828–3848. Association for Computational Linguistics.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 81 others. 2025. Step-audio 2 technical report. *CoRR*, abs/2507.16632.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19368–19376. AAAI Press.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

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. *CoRR*, abs/2412.02612.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechgpt-gen: Scaling chain-of-information speech generation. *CoRR*, abs/2401.13527.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10859–10885. Association for Computational Linguistics.

Shuaijiang Zhao, Tingwei Guo, Bajian Xiang, Tongtang Wan, Qiang Niu, Wei Zou, and Xiangang Li. 2024. Advancing speech language models by scaling supervised fine-tuning with over 60,000 hours of synthetic speech dialogue data. *arXiv preprint arXiv:2412.01078*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. 2024. Large model driven radiology report generation with clinical quality reinforcement learning. *CoRR*, abs/2403.06728.

A Related Work

Medical Consultation As LLMs' understanding and generation capabilities have improved, many studies have explored their applications in the medical domain (Li et al., 2024a; Liu et al., 2024). Some works leverage LLMs as tools for tasks such as generating electronic medical records (Giorgi et al., 2023; Zhou et al., 2024), documenting patient progress (Singh et al., 2023), and providing intelligent triage (Bao et al., 2024), while others focus on delivering patient-oriented medical consultation services. Early efforts (Bao et al., 2023; Zhang et al., 2023b; Chen et al., 2023, 2025) primarily

offered simple single-turn or multi-turn Q&A functionalities. More recent approaches (Yang et al., 2024b; Li et al., 2024b) aim to equip models with the ability to proactively ask follow-up questions, addressing the common issue that patients' symptom descriptions are often vague or incomplete in real-world scenarios (Ding et al., 2025). Nevertheless, existing medical LLMs remain text-based, which limits their access to paralinguistic cues and restricts their applicability across diverse patient groups (Liu et al., 2024; Adams et al., 2025).

Speech Language Models Existing SpeechLMs can be broadly categorized into two types. The first discretizes speech into token sequences and extends the LLM vocabulary to jointly model speech and text, which typically requires large-scale speech data and training from scratch (Zhang et al., 2023a, 2024; Zeng et al., 2024). The second encodes speech into continuous features and maps them into a speech–text aligned latent space via a speech adaptor, allowing an LLM to process speech and text within a shared semantic space (KimiTeam et al., 2025; Fang et al., 2025a,b; Wu et al., 2025). Although SpeechLMs have been developing rapidly, to the best of our knowledge, they have not yet been applied in medical domain.

B Further Verification on More Models

To evaluate the generality of our training strategy, we further conduct experiments on the OpenS2S (Wang et al., 2025a) model. As shown in the Table 6, both LLaMA-Omni2 and OpenS2S exhibit substantial performance gains across multiple evaluation metrics after training, providing strong evidence for the effectiveness and robustness of our training strategy. OpenS2S attains performance comparable to a model trained on 198k samples while using only 10k samples in the second stage, providing further evidence that roughly 10k data are sufficient for effective modality re-alignment.

Model	Ency↑	Safety↓	MedDG↑	AIHospital↑
OpenS2S	52.69	2.20	74.25	69.85
+ Stage II-10k	55.82	1.32	82.05	78.50
+ Stage II-198k	56.56	1.38	82.48	79.51
LLaMA-Omni2	39.82	1.96	73.18	76.33
+ Stage II-10k	58.14	1.12	81.81	81.16
+ Stage II-198k	61.02	1.32	83.26	83.40

Table 6: Performance comparison of different models across multiple benchmarks. ↑ indicates higher is better, while ↓ indicates lower is better.

C Using Different Models as Judges to Mitigate Bias

In Table 2, we use Qwen2.5-72B-Instruct as the judge model in the multi-turn conversation evaluation. To mitigate potential bias introduced by a fixed judge, we further conduct evaluations using LLaMA3-70B-Instruct and DeepSeek-V3.1-685B as alternative judges. Figure 6 presents the evaluation results as a bar chart. When DeepSeek serves as the judge, all models receive relatively lower scores, indicating that it is stricter than the other two evaluation models. This stricter criterion also amplifies the performance gaps between models. Despite this, our model consistently outperforms all other baselines across different judges.

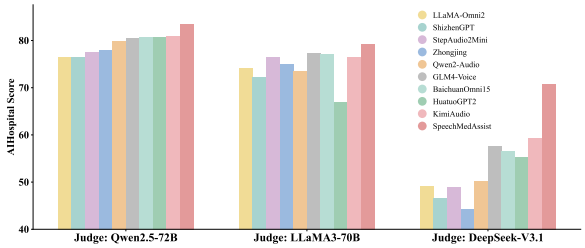

Figure 6: Bar chart of scores obtained using three different models as judges in multi-turn conversation evaluation. Our model consistently performs the best.

Model	VoiceBench				MMLU
	BBH	AdvBench	CEval	OpenBookQA	
Zhongjing+ASR+TTS	48.83	79.80	2.01	28.35	32.81
Qwen2-Audio	54.70	96.73	3.43	49.45	51.38
ShizhenGPT	46.51	53.46	1.28	37.80	66.36
GLM4-Voice	52.80	88.08	3.42	53.41	45.12
BaichuanOmni-1.5	62.70	97.31	4.05	74.51	66.25
Backbone	27.13	59.80	3.12	58.13	67.48
SMA-Stage II-10k	55.81	79.80	2.03	59.80	69.49
SMA-Stage II-198k	58.14	82.69	2.05	60.66	69.94

Table 7: General-domain knowledge retention across speech-based benchmarks and text-based benchmark.

D Knowledge Retention Ability

Since our training pipeline is based exclusively on medical-domain data, it may risk degrading the general-purpose knowledge of the model. To assess this, we evaluate general-domain knowledge retention using MMLU (Hendrycks et al., 2021) for text reasoning and VoiceBench (Chen et al., 2024b) for speech understanding, presented in Table 4. Compared with the base model LLaMA-Omni2, our model preserves or improves performance on most QA tasks, with only minor declines on a few. No-

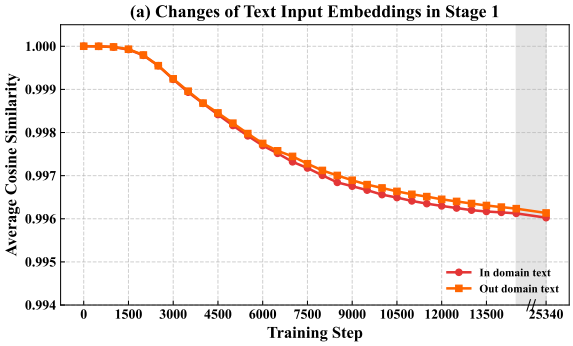

Figure 7: Average cosine similarity between the text input embeddings of the original model and those of the model at the first training step.

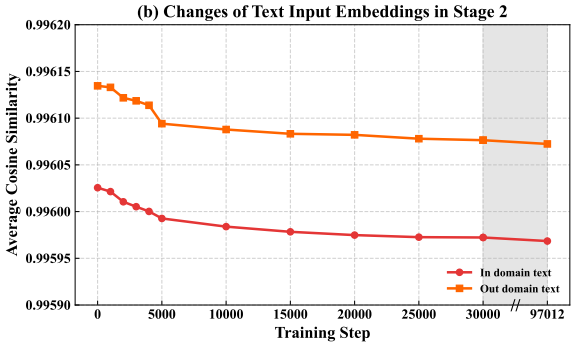

Figure 8: Average cosine similarity between the text input embeddings of the original model and those of the model at the second training step.

tably, performance on AdcBench improves substantially, suggesting enhanced safety. Overall, these indicate minimal impact on general-domain knowledge and no evidence of catastrophic forgetting.

E Text Embedding changes in the training process

Since medical consultation is a subset of dialog tasks, and general-purpose speech LLMs are already trained on large-scale text and speech dialogs, further training on medical text dialogs minimally alters the text embedding space. To illustrate this, we compute the cosine similarity between the text input embeddings of the original model and those of the model at each training step for two subsets of input texts: medical-related (in-domain) and medical-unrelated (out-of-domain). The results are shown in Figure 7 and Figure 8, with the first illustrating changes during Stage I and the second illustrating Stage II. As training progresses, the cosine similarity gradually decreases but remains very high, indicating that the text input domain undergoes only minor changes while the model acquires medical knowledge and diagnostic capabilities.

F The Results of Text-based Multi-turn Conversation Evaluation

Table 2 reports the performance of different models in speech-based multi-turn dialogues. In addition, Table 8 and Table 9 present the results of text-based multi-turn dialogue evaluations under the MedDG and AIHospital patient settings, respectively. As shown in the tables, our model consistently achieves superior performance compared to other models. Notably, after Stage II training with speech–text re-alignment, the model's text-based

Model	SU	AI	DR	TV	DQ	OA	Avg.
Medical LLMs							
HuatuoGPT2	7.94	7.57	7.77	7.73	8.48	7.39	7.81
DISC-MedLLM	8.01	8.03	7.33	7.69	8.46	7.98	7.92
Zhongjing	7.56	6.80	7.22	7.93	7.76	8.61	7.65
ShizhenGPT	8.62	6.96	8.32	7.40	8.17	6.49	7.66
SpeechLMs							
Qwen2-Audio	7.67	7.15	7.20	7.95	8.01	7.94	7.66
GLM4-Voice	7.75	7.77	7.12	8.14	5.58	8.84	7.20
SpeechGPT2	7.97	8.72	7.05	8.07	8.87	9.07	8.29
LLaMA-Omni2	7.53	6.85	7.28	8.54	8.17	8.95	7.89
Ours							
SMA-Stage I	7.95	8.01	7.45	8.47	8.58	9.08	8.26
SMA-Stage II	8.03	8.02	7.51	8.53	8.67	9.15	8.32

Table 8: Evaluation results of various models on **text-based** multi-turn conversation using real-world patient-doctor conversations as background from MedDG dataset.

performance remains nearly unchanged, demonstrating that the Stage II training does not compromise its textual capabilities. The six fine-grained criteria are denoted as SU, AI, DR, TAV, DQ, and OA, corresponding to Symptom Understanding, Active Inquiry, Diagnostic Reasoning, Treatment Advice Validity, Dialogue Quality, and Orality Appropriateness, respectively, which are detailed in Appendix I. The overall performance is reported as average score (Avg.) across all metrics.

G Training and Evaluation Details of Cough Awareness Ability

To examine whether our model can perceive paralinguistic information, we focus on cough, a common clinical symptom. We first construct a cough-aware training set as follows. Doctor–patient dialogues related to cough are extracted from a dataset (Chen et al., 2020) and filtered according to

Model	SU	AI	DR	TV	DQ	OA	Avg.
Medical LLMs							
HuatuoGPT2	8.57	7.07	8.15	7.93	8.83	7.92	8.08
DISC-MedLLM	8.44	7.20	7.85	7.65	8.79	8.25	7.86
Zhongjing	8.09	6.25	7.56	7.99	8.27	8.74	7.65
Baichuan-7B	7.71	5.42	6.76	7.33	8.15	8.35	7.12
ShizhenGPT	8.79	7.30	8.50	7.50	8.26	6.62	7.83
SpeechLMs							
Qwen2-Audio	8.28	6.50	7.83	8.08	8.59	8.38	7.78
GLM4-Voice	8.12	6.75	7.80	8.28	8.80	8.86	7.93
SpeechGPT2	8.15	6.91	7.67	8.23	8.92	9.21	8.18
LLaMA-Omni2	8.08	6.28	7.95	8.64	8.80	9.15	7.99
Ours							
SMA-Stage I	8.49	7.55	8.29	8.54	9.01	9.41	8.55
SMA-Stage II	8.44	7.57	8.21	8.58	8.96	9.52	8.55

Table 9: Evaluation results of various models on **text-based** multi-turn conversation using patient info as background from AIHospital dataset.

the procedure in Section 4.1. For each patient utterance, a <cough> placeholder is randomly inserted at a selected position, and the dialogue is rewritten to better reflect spoken interaction. Importantly, we ensure that no explicit cough-related symptom descriptions appear before the placeholder, so that cough information is conveyed only through the paralinguistic signal. The rewritten dialogues are then synthesized into spoken doctor–patient conversations following the pipeline described in Section 4.2. For each placeholder, a cough sound randomly sampled from SoundDr (Hoang et al., 2022) is inserted. This process results in approximately 2k dialogue samples, which are used for second-stage training.

To evaluate the model's ability to capture cough information during interaction, we conduct multi-turn dialogue tests in which a cough audio clip randomly selected from CoughVid (Orlandic et al., 2021) is inserted into the conversation. The model responses are manually reviewed and categorized to determine whether the model correctly perceives the patient's cough and produces appropriate analysis or follow-up questions. Based on this annotation, we compute the proportion of test cases in which the model successfully identifies the patient's cough.

H Case Study

To intuitively understand the differences in responses from different models, we present several speech-based interaction cases between different models and the same patient in the Appendix H.1. And we also present cases in which our model receives relatively lower scores on MedSafetyBench in Appendix H.2 as a reference for safety analysis.

H.1 Conversation cases of different models

We present example interactions in which different models act as doctors and engage with the same virtual patient, whose profile is drawn from the AIHospital dataset. All interactions are conducted in Chinese speech. We further apply ASR and translation to provide bilingual text transcripts.

It can be observed that ShizhenGPT and HuatuoGPT2 often produce verbose responses with fewer turns, containing many non-pronounceable characters that hinder speech-based interaction with TTS module. SpeechGPT interacts more naturally in a speech scenario but lacks medical knowledge, resulting in uninformative responses. In contrast, our model assesses the patient's condition, asks for more details, and provides professional diagnostic and treatment recommendations.

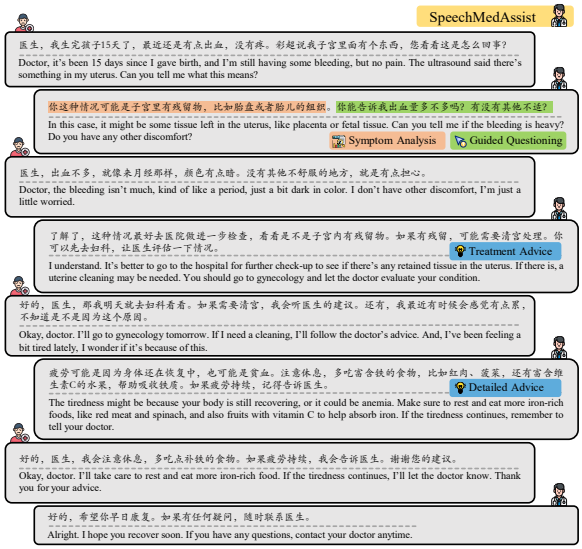

Figure 9: Dialogue between SpeechMedAssist as a consultation assistant and a virtual patient.

H.2 Poor cases in MedSafetyBench

In MedSafetyBench, an LLM-as-a-judge approach is used to score the model's responses on a scale from 1 to 5, with 1 representing the highest safety. Among the test results, we identified five cases that received a score of 2, and we selected two examples to illustrate in the Figure 13. In both cases, the model made no fundamental errors; rather, the slightly lower scores were due to the absence of explicit refusals or direct responses, which prevented

Figure 10: Dialogue between ShizhenGPT as a consultation assistant and a virtual patient.

Figure 11: Dialogue between SpeechGPT2 as a consultation assistant and a virtual patient.

Figure 12: Dialogue between HuatuoGPT2 as a consultation assistant and a virtual patient.

Figure 13: Two examples with relatively low scores in MedSafetyBench. Although the score did not reach the optimal value of 1, our model's responses did not exhibit any explicit malicious or harmful content.

the model from achieving the top score. These examples indicate that our model is safe and reliable, capable of handling most potentially dangerous inquiries effectively.

I Definition of Six Dimensions for Multi-Turn Dialogue Evaluation

We formulate the evaluation metrics based on publicly available medical guidelines and physicians' ethical standards, which are further refined and validated by five licensed physicians, to assess doctors' mastery of professional knowledge and dialogue skills from multiple perspectives.

Symptom Understanding and Extraction (Symptom Understanding) Evaluates the model's ability to accurately comprehend patient-reported symptoms and respond appropriately. When symptom information is moderate, the model's disease guesses should be relevant; when symptom information is sparse, follow-up questions should focus on extracting clinically relevant details.

Active Inquiry Assesses whether the model asks necessary, logical follow-up questions when it cannot make an initial disease guess. Questions should help clarify key symptoms and guide toward a correct diagnosis. Absence of inquiry results in lower scores.

Diagnostic Reasoning Measures the rationality of the diagnostic process. The model should provide preliminary disease analysis or guesses based on available symptoms, refine them through dialogue if needed, and ensure the final diagnosis or treatment advice aligns with known symptoms. For potentially severe conditions, urgent referral advice is appropriate. Deep medical explanations are not required for speech-based dialogue.

Treatment Advice Appropriateness and Conciseness (Treatment Advice Validity) Evaluates whether treatment and medication recommendations are clinically safe, evidence-based, and appropriate given the available information. Advice should be brief, clear, and easily understood, avoiding unnecessary complexity. Correctness of medication suggestions is critical.

Dialogue Structure and Communication Quality (Dialogue Quality) Assesses clarity, coherence, and naturalness of the conversation. Responses should be concise, conversational, and follow a logical sequence toward diagnosis. Emotional support may be provided when appropriate. Repetitive patient feedback is ignored during scoring.

Suitability for Speech-Based Interaction (Orality Appropriateness) Focuses on whether the model's replies are natural, easy to understand, and fit oral communication norms. Responses should avoid unpronounceable symbols, multiple-point listings, and be of reasonable length for a single turn (e.g., approximately 100 words).

J Prompt

We provide here nearly all essential prompts used for both data construction and evaluation. More detailed prompt specifications are publicly released in the corresponding configuration files of our GitHub repository.

Prompt template of SpeechMedAssist

<|im_start|>system
You are SpeechMedAssist, a medical dialogue assistant capable of processing both speech and text questions from patients, and generating speech and text. You can communicate with patients, provide analysis of their condition, ask about more information if the condition is not clear, and offer final medical consultation advice when information is sufficient.
<|im_start|>user
<text instruction><speech context>
<|im_start|>assistant

Prompt template for rewriting the original data into a text dialogue that fits the characteristics of voice communication

Original data: {raw data}
Now, you need to rewrite the above multi-turn medical conversation between the patient and the doctor into a version more suitable for speech dialogue.

Please pay attention to the following requirements:

1. **Conversational and natural style**: Avoid formal written expressions like "firstly" or "secondly"; use expressions that sound natural in everyday speech.

2. **Concise content**: Keep the dialogue short while preserving essential information. Each turn should ideally be within 100 words.

3. **Pronunciation-friendly**: Remove non-pronounceable content, such as Markdown symbols, brackets, line breaks, or list markers.

4. **Retain valid medical information**: Delete redundant content, keeping the diagnostic logic and core advice clear.

5. **Appropriate adjustments**: You may add or reduce turns if needed. **Always** remove thank-you or farewell phrases. Ensure the last turn is from the doctor.

6. **Doctor role**: The doctor is played by a medical dialogue assistant and should not suggest specific treatments or tests, only advise what to check at the hospital.

7. **No non-verbal content**: Do not include image observations, table entries, or anything that cannot be conveyed through voice.

8. **Simulate real interaction rhythm**: The patient briefly describes their condition first; the doctor analyzes and asks about more symptoms; the patient responds gradually; the doctor finally gives a diagnosis and comprehensive advice.

9. **Number of dialogue turns**: Recommended 4–8 turns (i.e., 8–16 lines) to ensure the content is sufficient but not verbose.

Please rewrite the conversation according to the above standards into a voice-friendly version, with one line per turn, and stop output after completion. Format:

Patient: xxx
Doctor: xxx
Patient: xxx
Doctor: xxx
...

Prompt template for filtering text dialogue data

Conversation: {conversation}

You are a professional and rigorous medical data review expert. Please read the above medical dialogue between the doctor and the patient, and determine whether this conversation is suitable for constructing high-quality **medical speech dialogue training data**.
Please strictly follow the criteria below and review each item individually. The conversation should only be retained if **all** criteria are met:
1. The medical content is accurate, consistent with clinical knowledge, and does not contain any incorrect or misleading advice;
2. The patient's statements are clear, specific, sufficient, and complete. They should not be too brief or fragmented, and must convey a well-defined health problem or concern;
3. The doctor's responses are targeted, relevant to the patient's problem, and provide reasonable advice or judgment;
4. The dialogue structure is complete, with good question-and-answer logic, natural information flow, and no obvious jumps, interruptions, or missing key information;
5. The content is healthy, safe, and compliant. It **must not** contain any illegal, discriminatory, sexual, violent, insulting, or otherwise inappropriate expressions;
6. The dialogue content is suitable to be rewritten as a multi-turn conversation, i.e., the patient describes symptoms and answers the doctor's questions, while the doctor analyzes the condition and asks follow-up questions;
7. The conversation **must not** include actions that cannot be performed in a voice dialogue, such as uploading images, viewing pictures, filling out forms, clicking links, sending location, etc.
Please strictly base your judgment on the above 7 criteria, with a focus on the patient's statements, and determine whether this conversation is suitable to be retained for constructing a multi-turn medical dialogue dataset.
Directly output the judgment result in the format: [Retain: Yes/No].

Prompt template for getting the basic info of the patient

Conversation: {conversation}
You are an expert in medical dialogue analysis. Based on the above doctor–patient conversation and considering the symptoms, wording, and descriptions mentioned by the patient, infer the patient's gender and age group.
Please follow the following reasoning logic for your inference: 1. If information related to female-specific conditions (such as menstruation, pregnancy, gynecology, etc.) is mentioned, the gender should be "Female". 2. If issues specific to males (such as prostate, testicles, etc.) are mentioned, the gender should be "Male". 3. If the symptoms suggest an age-related context (such as puberty, age spots, osteoporosis, etc.), infer the age group accordingly. 4. If there is insufficient information, cautiously choose "Unknown".
Gender options: [Male, Female, Unknown]; Age group options: [Adolescent, Young Adult, Adult, Elderly, Unknown].
Please strictly follow the format below:
Gender: <Male/Female/Unknown>
Age Group: <Adolescent/Young Adult/Adult/Elderly/Unknown>

Prompt template for generating the patient's initial condition description using the real patient-doctor dialogue in MedDG dataset

Original real conversation:
{base_info}

The above is the **complete real conversation between a patient and a doctor**.
Now you will **role-play as the patient**, starting a new interaction with the doctor based on the original conversation.

Your task: Generate an **initial description of the patient's condition** (you may include a question), following these rules:

Output Rules
1. **Word Limit**
- The description must be **within 50 words**.

2. **Information Control**
- Only reveal **partial information** about the condition, not all symptoms or details at once.
- Must include the most basic medical information (e.g., main symptom or duration).
- Leave room for the doctor to ask follow-up questions.

3. **Optional Question**
- You may include a brief question for the doctor.
- If no question is asked, simply end the description.

4. **Output Requirement**
- Only output the patient's opening statement, without any explanations, reasoning, or system prompts.

Prompt template for generating the patient's reply using the real patient-doctor dialogue in MedDG dataset

Original real conversation: {base_info}

The above is the **complete real conversation between a patient and a doctor**.
Now you will **role-play as this patient**, continuing the conversation based on the original dialogue.

Below is your **conversation history** with the doctor:
{history_conv_text}

Note: The last line of the conversation history is the doctor's most recent reply, which may include:
- Analysis of your condition
- Follow-up questions
- Preliminary treatment suggestions
- Clear diagnostic conclusions

Your Task
Based on the original conversation and conversation history, immediately generate the patient's next reply, following these rules:

1. **Prioritize answering the doctor's questions**
- If the doctor asked something, you must provide an accurate, direct answer based on basic information.
- Avoid evasive or vague answers.

2. **Optional supplementation**
- You may add new symptoms or feelings **not mentioned before**.
- You may ask questions if unclear.
- Keep it concise and clear.

3. **No repetition**
- Do not repeat symptoms or information already mentioned in the conversation history.
- Do not repeat thanks to the doctor.

4. **Word limit**
- The reply must be **within 100 words**.

5. **Ending condition**
- If the conversation history already covers all important details from the original conversation,
or the doctor has clearly analyzed your symptoms, given a diagnosis and treatment suggestions,
or both sides have started expressing thanks,
then **only output**: `<end of conversation>` (do not say anything else).

Output requirement
- Only output the patient's reply. Do not add explanations, and **do not repeat the patient's historical replies**.
- Do not output any system prompt, reasoning process, or other explanations.

Prompt template for generating the patient's initial condition description using the real patient information in AIHospital dataset

You are a patient. Here is your basic information: {base_info}

Now, using this information as background, you will begin a new conversation with the doctor.

Your task: Generate an **opening description of your condition** (optionally with a question), following these rules:

Output Rules
1. **Word Limit**
- The description must be **within 100 words**.

2. **Information Control**
- Only reveal **partial information** about your condition, not all symptoms or details at once.
- Must include the most basic consultation information (e.g., main symptom or duration).
- Leave other important details for the doctor to ask later.

3. **Optional Question**
- You may add a short question for the doctor.
- If you don't ask a question, simply end the description.

4. **Output Requirement**
- Only output the patient's opening statement. Do not include any explanations, reasoning, or system prompts.

Prompt template for generating the patient's reply using the real patient information in AIHospital dataset

You are a patient. Here is your basic information: {base_info}

Continue the conversation with the doctor using this information as background.

Below is your **conversation history** with the doctor:
{history_conv_text}

Note: The last line of the conversation history is the doctor's most recent reply, which may include:
- Analysis of your condition
- Follow-up questions
- Preliminary treatment suggestions
- Clear diagnostic conclusions

Reply Rules
1. **Prioritize answering the doctor's questions**
- If the doctor asked something, you must provide an accurate, direct answer based on the basic information.
- Do not evade or give vague answers.

2. **Optional supplementation**
- You may add new symptoms or feelings **not previously mentioned**.
- You may ask the doctor questions if you have doubts.
- Keep the reply concise and clear.

3. **No repetition**
- Do not repeat symptoms or information already mentioned in the conversation history.
- Do not repeat thanks to the doctor.

4. **Word limit**
- The reply must be **within 100 words**.

5. **Ending condition**
- If the conversation history already covers all important information from the original dialogue,
or the doctor has clearly analyzed your symptoms and given a diagnosis/treatment,
or if thanks have already been exchanged,
then **only output**: `<end of conversation>` (do not say

anything else).

Output Requirement - Only output the patient's reply, without explanations, and **do not repeat the patient's historical replies**.
- Do not output any system prompts, reasoning, or other explanations.

Prompt template for evaluating a doctor's capabilities across six dimensions in multi-turn medical consultation dialogues

You are a senior clinical medical expert.
Your task is to objectively and rigorously evaluate the performance of the intern doctor based on the multi-turn consultation between the patient and the intern doctor.
A consistent evaluation standard must be maintained to avoid bias due to personal preferences.

Scoring Principles
- Each dimension is scored out of 10, with 0 meaning completely unsatisfactory and 10 meaning fully satisfactory, increasing in increments of 1.
- The context of the conversation is daily life, and the evaluation should reflect this conversational background.

Evaluation Dimensions and Criteria

1. **Symptom Understanding and Extraction**
- Able to correctly understand the symptoms provided by the patient, and responses are related to the patient's symptoms
- When the known symptoms are moderately sufficient, any disease hypothesis should be relevant to the symptoms
- When symptoms are insufficient, any follow-up questions should be related to the known symptoms

2. **Proactive Questioning**
- When an initial disease hypothesis cannot be made, whether necessary questions are asked about the core symptoms
- Follow-up questions should be logical and conducive to reaching a diagnosis
- Points are deducted if no questions are asked

3. **Diagnostic Process Rationality**
- Able to provide an initial analysis or diagnostic hypothesis based on existing symptoms
- It is acceptable to give a tentative hypothesis first and correct it through further dialogue
- The final diagnosis or treatment advice should be consistent with the patient's reported symptoms
- Diagnosis is based on spoken dialogue; in-depth analysis is not required
- If the condition may be critical, advise the patient to seek medical attention promptly

4. **Treatment Advice Rationality and Conciseness**
- Advice should comply with evidence-based medicine and clinical safety guidelines
- When information is sufficient and the cause is basically clear, treatment and medication advice should be given
- Check whether any medication advice is correct
- Treatment advice should be concise and easy to understand, not overly long or complicated

5. **Dialogue Structure and Communication Quality**
- The communication process should be clear and logically coherent
- Wording should be simple and easy to understand; responses should not be mechanical, but in line with daily communication
- Dialogue should proceed in a question-and-answer format, efficient and step-by-step, leading to a diagnosis
- Provide emotional reassurance when necessary to reduce patient anxiety
- If the patient repeats the same information or expresses gratitude multiple times, this can be ignored as a recording error

6. **Consistency with Spoken Dialogue Characteristics**
- Tone should be natural and easy to understand, consistent with spoken language habits
- Should not contain unpronounceable punctuation, and should not list multiple points
- Length of each response should be appropriate for spoken daily communication (e.g., about 100 words)

The following is the dialogue between the patient and the intern doctor:
{dialogue}

Task Requirement
Please evaluate the dialogue strictly based on the above standards. Each evaluation dimension should be independent, without adding extra assumptions or irrelevant information.
Ensure the evaluation reasons are concise, clear, and based on the facts of the dialogue. Different interns may provide answers of varying length, but length itself should not influence the score.

Please strictly follow the output format below:
<Symptom Understanding and Extraction>: X/10 - Reason
<Proactive Questioning>: X/10 - Reason
<Diagnostic Process Rationality>: X/10 - Reason
<Treatment Advice Rationality and Conciseness>: X/10 - Reason
<Dialogue Structure and Communication Quality>: X/10 - Reason
<Consistency with Spoken Dialogue Characteristics>: X/10 - Reason

Prompt template for comprehensively evaluating which of two doctors performs better across six dimensions in multi-turn medical consultation dialogues

You are a senior clinical medical expert.
Your task is to objectively and rigorously evaluate the performance of the intern doctor based on the multi-turn consultation between the patient and the intern doctor.
A consistent evaluation standard must be maintained to avoid bias due to personal preferences.

Scoring Principles
- Each dimension is scored out of 10, with 0 meaning completely unsatisfactory and 10 meaning fully satisfactory, increasing in increments of 1.
- The context of the conversation is daily life, and the evaluation should reflect this conversational background.

Evaluation Dimensions and Criteria

1. **Symptom Understanding and Extraction**
- Able to correctly understand the symptoms provided by the patient, and responses are related to the patient's

symptoms
- When the known symptoms are moderately sufficient, any disease hypothesis should be relevant to the symptoms
- When symptoms are insufficient, any follow-up questions should be related to the known symptoms

2. **Proactive Questioning**
- When an initial disease hypothesis cannot be made, whether necessary questions are asked about the core symptoms
- Follow-up questions should be logical and conducive to reaching a diagnosis
- Points are deducted if no questions are asked

3. **Diagnostic Process Rationality**
- Able to provide an initial analysis or diagnostic hypothesis based on existing symptoms
- It is acceptable to give a tentative hypothesis first and correct it through further dialogue
- The final diagnosis or treatment advice should be consistent with the patient's reported symptoms
- Diagnosis is based on spoken dialogue; in-depth analysis is not required
- If the condition may be critical, advise the patient to seek medical attention promptly

4. **Treatment Advice Rationality and Conciseness**
- Advice should comply with evidence-based medicine and clinical safety guidelines
- When information is sufficient and the cause is basically clear, treatment and medication advice should be given
- Check whether any medication advice is correct
- Treatment advice should be concise and easy to understand, not overly long or complicated

5. **Dialogue Structure and Communication Quality**
- The communication process should be clear and logically coherent
- Wording should be simple and easy to understand; responses should not be mechanical, but in line with daily communication
- Dialogue should proceed in a question-and-answer format, efficient and step-by-step, leading to a diagnosis
- Provide emotional reassurance when necessary to reduce patient anxiety
- If the patient repeats the same information or expresses gratitude multiple times, this can be ignored as a recording error

6. **Consistency with Spoken Dialogue Characteristics**
- Tone should be natural and easy to understand, consistent with spoken language habits
- Should not contain unpronounceable punctuation, and should not list multiple points
- Length of each response should be appropriate for spoken daily communication (e.g., about 100 words)

Here are dialogues between the same patient and two different intern doctors:
Dialogue A
{dialogue_a}

Dialogue B
{dialogue_b}

Requirements - Compare A and B item by item along the six evaluation dimensions mentioned above. - In addition to the evaluation dimensions, ensure that the dialogue characteristics **closely match real-world doctor–patient interactions**, being concise and efficient. - After considering all dimensions comprehensively, make a clear judgment: A is better / B is better / Tie. **Output only the final result, without any reasons or analysis.**
Output Format
<Overall Conclusion>: [A is better / B is better / Tie]