SPEECHMEDASSIST: EFFICIENTLY AND EFFECTIVELY ADAPTING SPEECH LANGUAGE MODEL FOR MEDICAL CONSULTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical consultations are inherently speech-based, yet current works focus on fine-tuning large language models (LLMs) to perform patient-unfriendly long-text interaction. While existing speech language models (SpeechLMs) enable efficient speech-based interaction, the scarcity of speech data in medical domain prevents them from being directly fine-tuned for practical applications. In this paper, we propose SpeechMedAssist, a SpeechLM natively capable of conducting multi-turn speech-based interactions with patients. To mitigate data scarcity, we mathematically analyze the architecture of SpeechLMs and decouple one-stage training that requires a large corpus of medical speech data into a two-stage training paradigm. (1) Knowledge&Capability injection: train the LLM core with rewritten and filtered medical text data to inject medical knowledge and endow it with diagnostic and treatment capabilities. (2) Modality alignment: train the SpeechLM using a small amount of synthetic medical speech data that matches patient characteristics to realign the speech and text modalities. After two-stage training, SpeechMedAssist performs excellent on our designed speech-based medical benchmark. Experiments further show that the second stage only requires 10k speech dialogue samples to achieve modality alignment, allowing the knowledge and capabilities acquired in the text modality during the first stage to generalize to the speech modality, which demonstrates the effectiveness and efficiency of our approach.

1 INTRODUCTION

Large Language Models (LLMs) have been widely applied in various vertical domains with their strong language understanding and generation capabilities (Li et al., 2024a). In the medical domain, benefit of 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 clinical report generation (Zhou et al., 2024) after supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF).

However, in interactive scenarios such as medical consultation, which requires a high level of convenience and immediacy, these purely text-based models lacks interactivity and struggles to meet the needs of users across different age groups (Shi et al., 2024). Some works (Huang et al., 2024) transfer text-based interaction to speech using a cascaded system consisting of automatic speech recognition (ASR), an LLM, and text-to-speech (TTS), but incur latency, ASR errors (Binici et al., 2025), and loss of paralinguistic cues, degrading user experience (Ji et al., 2024).

Compared to cascaded systems, end-to-end speech language models (SpeechLMs) offer great potential in medical consultation, as they natively support speech-based interaction, enabling more natural interaction (Adams et al., 2025; Cui et al., 2025). Nevertheless, adapting SpeechLMs to real-world medical consultation remains challenging: (1) existing SpeechLMs are trained on general-purpose data, lacking domain-specific medical knowledge and clinician-like inquiry skills (Clusmann et al., 2023; Ng et al., 2024); (2) the scarcity of medical speech data prevents direct fine-tuning to acquire medical knowledge and capabilities (Banerjee et al., 2024).

In this paper, we propose SpeechMedAssist, a SpeechLM designed for speech-based multi-turn medical consultation, which can natively analyze symptoms, conduct proactive inquiries, and pro-

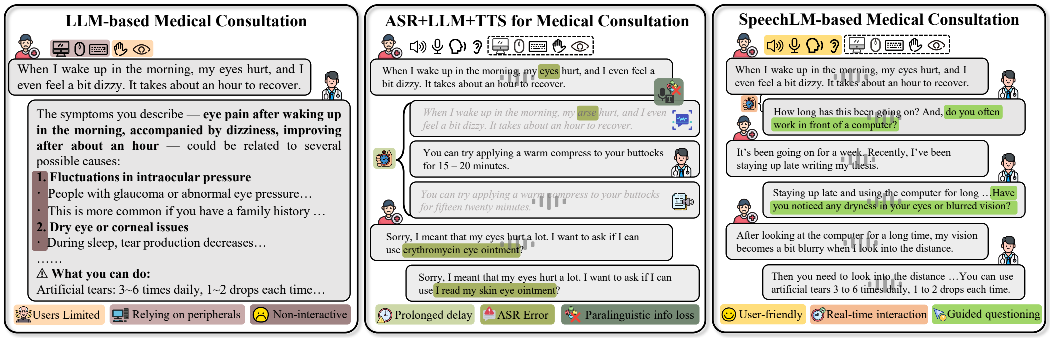

Figure 1: An intuitive illustration highlighting the limitations of text-based interactions and cascaded speech systems, alongside the advantages of end-to-end SpeechLMs in medical consultation.

vide diagnostic and treatment suggestions. To address the scarcity of medical speech data, we analyze the common architecture of SpeechLMs, consisting of a speech encoder, a speech adaptor, an LLM core, and a speech decoder. Inspired by human brain research (Buchweitz et al., 2009), the LLM core in SpeechLMs acting as brain can be treated as a modality-independent network, which means that the knowledge and capabilities acquired from text data can be further used in the speech modality. Based on this insight, we decouple training on speech data into a two-stage paradigm: capability injection from abundant text data and modality alignment with little speech data.

Specifically, in the first stage, we freeze all speech modules and focus on enhancing the LLM core component with large-scale text data. To equip the LLM core with essential knowledge and abilities for medical consultation, we construct TextMedDataset, which combines single-turn medical QA, multi-turn dialogues, and safety-constrained cases, all rewritten to resemble speech interaction. In the second stage, we unfreeze all modules and perform speech–text modality alignment with a small amount of medical speech dialogue data. In order to obtain medical speech data, we select a subset from TextMedDataset, and synthesize speech dialogues according to patient characteristics, resulting in the first speech-based medical dialogue dataset SpeechMedDataset.

We validate our approach through both objective evaluations (medical examination and speech quality analysis) and subjective assessments in real-world scenarios and simulated virtual consultation environments with three roles: patient, intern doctor, and senior judge, enabling a comprehensive examination of the model's performance. Our contributions can be summarized as follows:

1. We propose SpeechMedAssist, a medical SpeechLM that introduces speech-based interaction into the medical domain through an efficient two-stage training strategy.

2. We develop a pipeline to convert raw medical dialogues into patient-tailored multi-turn speech-based conversations, creating the first speech medical dialog dataset, SpeechMedDataset.

3. We establish a medical consultation benchmark for SpeechLMs by simulating real medical consultation scenarios, which can demonstrate the efficiency and effectiveness of our approach.

2 Model Architecture

Existing SpeechLMs can be divided into two main categories. The first encodes speech into discrete tokens, similar to text tokens, and extends the vocabulary of the LLM to process both speech and text tokens, which requires extensive speech data for training from scratch (Zhang et al., 2023a; 2024; Zeng et al., 2024). The second extracts speech features using a speech encoder and then projects them into a speech-text aligned latent space via a speech adaptor, enabling the LLM to process speech and text simultaneously (KimiTeam et al., 2025; Fang et al., 2025a;b; Wu et al., 2025). Intuitively, the second architecture exploits the fact that speech comprises both textual and paralinguistic information, making it possible to effectively utilize the existing semantic space (Ji et al., 2024). Subsequent analyze and experiments show that this architecture enables the model's capabilities and knowledge learned in the text modality to effectively generalize to the speech modality. Here, we briefly introduce the common architecture of SpeechLMs we adopt.

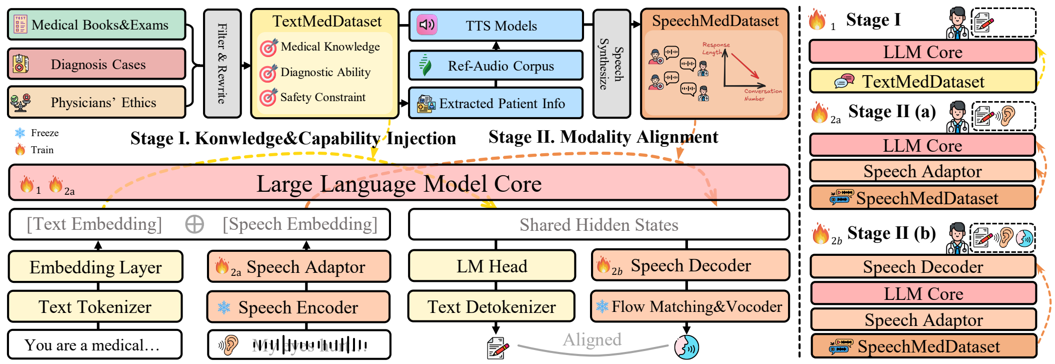

Figure 2: An overview of our work. **Data Constraction**: we construct TextMedDataset by filtering and rewriting existing medical text corpora, and build SpeechMedDataset by extracting patient information from dialogues and synthesizing matched speech. **Model Architecture**: the structure of the SpeechLM we adopt supports dual-modal input and streaming output of text and speech. **Training Strategy**: the first stage injects knowledge&capability using TextMedDataset, while the second stage achieves modality alignment with a small amount of speech data from SpeechMedDataset.

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 via a tokenizer and embedding layer:

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

2.2 Large Language Model Core

To simultaneously input text instructions and speech inquiries, SpeechLMs usually achieve this by concatenating text embeddings $\mathbf{Z}_t$ and speech embeddings $\mathbf{Z}_s$, and then feeding them into the shared LLM core f to get the hidden states 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 hidden states $\mathbf{H}$, 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

For the SpeechLMs architecture we introduced before, the LLM core acts as the "brain", while the text tokenizer and speech encoder correspond to "reading" and "listening" modules, respectively. In neuroscience, fMRI studies on reading and listening comprehension show that the human brain encodes knowledge in a modality-independent manner (Buchweitz et al., 2009), which means that the knowledge and capability acquired from text data can also be used in speech modality. Inspired by this, we design a two-stage training strategy to effectively adapting SpeechLMs for medical consultation as illustrated in Figure 2. Specifically, rather than directly fine-tune SpeechLMs with a large corpus of medical speech data, we first inject medical knowledge and diagnostic capabilities through large-scale text data, and then use a small amount of speech data for modality alignment. Here, we provide a clearer exposition of the training strategy along with mathematical analysis.

3.1 KNOWLEDGE & CAPABILITY INJECTION 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 MODALITY ALIGNMENT WITH LIMITED SPEECH DATA

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 apply it correctly in a real clinical setting. Therefore, the next challenge is to transfer these capabilities effectively to the speech modality. This can be formally analyzed under the framework of domain adaptation theory, which relates the error on the speech domain to that on the text domain and the divergence between their embeddings.

Formally, let $\epsilon_t(f)$ denote the expected error of hypothesis f on the text domain and $\epsilon_s(f)$ denote the error on the speech domain. The classical adaptation bound (Ben-David et al., 2006) states that, for any hypothesis $f \in \mathcal{H}$,

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

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 has already been optimized in the text domain, both $\epsilon_t(f)$ and λ remain relatively small, which means the error in the speech domain $\epsilon_s(f)$ mainly depends on the divergence term $d_{\mathcal{H}\Delta\mathcal{H}}$. For the original pre-trained SpeechLM, the text and speech modalities are already well aligned and since medical consultation is a subset of dialogue tasks, the text space changes little after first-stage training, ensuring that the divergence introduced by text-only training remains limited. As a result, only a small amount of speech data is required to realign the two modalities.

Concretely, the second stage training is two-fold: **(a)** we unfreeze the speech adaptor $\mathcal{A}$ and train it jointly with the LLM core f using paired <speech input, text response>data; **(b)** in order to optimize speech generation capabilities., we only unfreeze the speech decoder G and train it with <speech input, speech response>pairs.

4 DATA CONSTRUCTION

Although a large amount of existing medical text data is available, most of it consists of text-based single-turn Q&A interactions, where patients provide sufficient information in their inquiries and doctors respond with lengthy answers, which do not correspond to real-world scenarios (Li et al., 2024b). To obtain high-quality multi-turn doctor–patient dialogue data, we design a pipeline that first filters and rewrites the original medical corpus into patient-doctor dialogues that conform to the characteristics of spoken conversations, then extracts patient information, synthesizes speech aligned with patient characteristics, and ultimately constructs a speech-based medical dialogue dataset. The overall data construction process is presented in Figure 2.

4.1 TEXTMEDDATASET

In realistic medical consultations, the key characteristics of the interaction between patients and doctors can be summarized as follows: **(a) Gradual information disclosure**: patient typically reveals only partial symptoms in each turn, rather than providing a complete description of their condition. **(b) Proactive inquiry by doctors**: based on the initial information, doctor needs to conduct a preliminary analysis and pose targeted follow-up questions. **(c) Multi-turn exchanges**: patient responds to doctor's questions, progressively elaborating their condition across multiple dialogue turns. **(d) Final decision-making**: once sufficient information is gathered, doctors integrate the details to provide a final diagnosis or treatment advice (Roter & Hall, 1987; Iversen et al., 2020).

To enable the model to handle real medical consultation scenarios after training, we collect three category datasets: single-turn Q&A data (Wang et al., 2024; 2025) for injecting medical knowledge, multi-turn dialogue data for endowing the model with conversational abilities (Yang et al., 2024b; Liu et al., 2022; Wang et al., 2025), and safety-constraint data for enhancing dialogue safety (Han et al., 2024). Since some of the data were directly crawled from the web, we first filter out relatively low-quality entries using Qwen2.5-14B-Instruct (Yang et al., 2024a), removing incomplete or medically irrelevant samples. Then, we prompt Qwen2.5-72B-Instruct to rewrite the filtered data, producing dialogue that aligns with the real-world medical consultation characteristics described earlier. In addition, to facilitate subsequent speech synthesis, non-pronounceable characters are removed during the rewriting process, and the length of each dialogue turn is constrained. We finally obtain TextMedDataset, which contains 405k samples. The prompts can be found in Appendix G

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 use Qwen2.5-14B-Instruct to analyze physician-patient dialogues and infer the patient's likely gender and age group, with the gender classified as male, female or unknown, and the age group classified as child, young adult, adult, elderly or unknown. To ensure the model's generalization capability, we construct a speech–text paired dataset of 1,000 hours based on publicly available ASR datasets Aishell2 (Du et al., 2018) and Aishell3 (Shi et al., 2020), collected from approximately 2,000 Mandarin speakers from different accent regions across China. During synthesis, we select reference speech that matches the inferred patient characteristics from the speech–text paired dataset and generate speech using the TTS model CosyVoice2 (Du et al., 2024). For patient where both gender and age group are unknown, we synthesize speech with random timbres using another TTS model FishSpeech (Liao et al., 2024). Through the steps above, we ultimately obtained a multi-turn spoken medical dialogue dataset SpeechMedDataset, which contains 198k samples.

Table 1: Overview of datasets used to construct TextMedDataset and SpeechMedDataset. Three different types of datasets are used to endow the model with distinct capabilities, while reference audio datasets are used for Speech synthesis to ensure the model's generalization ability.

Dataset	Description	Used Size
	Knowledge Injection	
CMB-Exam	Single-choice and multiple-choice questions covering 6 major categories	**189k**
Medical Encyclopedia	Single-turn medical Q&A about common diseases and medicines	**41k**
Medical Books	Single-turn medical consultation dialogues on general medical knowledge	**40k**
	Dialogue Ability	
CMtMedQA	Multi-turn medical consultation dialogues on medical knowledge	**68k**
MedDG	Privacy-deidentified real multi-turn medical consultation dialogues	**16k**
HuatuoGPT2-SFT	Q&A pairs of real patient questions with answers generated by GPT-4	**48k**
	Safety Constraint	
MedSafety-GPT4	Single-turn Q&A containing harmful requests with safe GPT-4 responses	**450**
	Reference Audio Data	
Aishell2	Audios from 1,991 Mandarin speakers across different accent regions	**1000h**
Aishell3	Audios speech from 218 Mandarin speakers across different accent regions	**85h**

5 Experiments

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

5.1 Experimental Setup

Model Configuration We use LLaMA-Omni2-7B (Fang et al., 2025b) as the base model, which adapts the architecture introduced in Section 2. For speech processing, it adopts the encoder of

Whisper-large-v3 (Cao et al., 2012), followed by a speech adapter that applies 5× downsampling and an FFN with an intermediate dimension of 2048. Qwen2.5-Instruct-7B serves as the LLM core, while the speech decoder is based on Qwen2.5-0.5B with a read–write strategy ($R = 3, W = 10$). The speech vocoder adopts the CosyVoice2 architecture and supports streaming output.

Training Details In the first stage, we fine-tune the LLM core on TextMedDataset following Section 3.1 with a batch size of 8 and learning rate 5×10^{-5}. In the second stage, we train 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 dataset CMB from TextMedDataset, with the final training data maintaining 1:1 between speech and text.

Baselines Our evaluation covers the following categories of models. (1) **Medical LLMs**: Various medical LLMs have been developed based on different backbone models with different training strategies, including DISC-MedLLM (Bao et al., 2023), Zhongjing (Yang et al., 2024b), Baichuan2 (Yang et al., 2023), and HuatuoGPT2 (Chen et al., 2023). These models demonstrate diverse medical capabilities, but their interactions are all text-based. We enable them to listen and speak by adopting an ASR+LLM+TTS pipeline, using SenseVoiceSmall[1] for ASR and CosyVoice2[2] for TTS. We also consider ShizhenGPT (Chen et al., 2025), which takes multimodal input and generates text, focusing on traditional Chinese medicine (TCM). (2) **SpeechLMs**: As detailed in Section 2, 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 (OpenMoss, 2025), Qwen2-Audio (Chu et al., 2024), StepAudio2-mini (Wu et al., 2025). (3) **OmniLMs**: We also include the latest multimodal large 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).

5.2 Evaluation

To evaluate our model and compare it with other models, we construct a benchmark mainly covering four dimensions: medical knowledge, dialogue capability, robustness, and speech quality.

Single-turn Q&A To assess models' mastery of medical knowledge in both text and speech modalities, we introduce two medical multiple-choice evaluation sets, **CMB** (Wang et al., 2024) and **CMExam** (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. We also adopt MedSafetyBench (referred to as **Safety**) (Han et al., 2024) to evaluate the safety of model responses, with scores ranging from 1 to 5. Except for the multiple-choice tasks, which use a text-based interaction paradigm, all other evaluations are conducted through both text and speech interactions. The results of text-based evaluations are reported in Appendix B.

Multi-turn Conversation Speech interaction requires strong conversational abilities, and medical consultations demand that the model proactively engage with patients. To align evaluation with real-world scenarios, we build a virtual medical consultation environment similar to AIHospital (Fan et al., 2025), consisting of an LLM-driven patient, a chief examiner, and an intern doctor powered by the model under evaluation. The patient, provided with real doctor–patient dialogues (**MedDG** (Liu et al., 2022)) or real patient cases (**AIHospital**), consults the intern doctor through multi-turn interactions and can end the conversation once sufficient diagnostic advice and treatment have been obtained. The intern doctor cannot see the patient's information and must obtain relevant details from patient responses. Finally, the chief examiner as a judge (Zheng et al., 2023) evaluates the dialogue from six perspectives, which are detailed in Appendix C, and we report the average score. Both the patient and chief examiner in Table 2 are driven by Qwen2.5-72B-Instruct.

Wild To assess whether the models can handle real-world medical consultations, we collect 20 sets of patient questions recorded in real clinical environments. Unlike the synthesized speech used in testing, these real-world recordings contain significant background noise and disorganized speech.

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

Table 2: Evaluation results of various models on single-turn QA and multi-turn conversation metrics. '-' indicates the metric is not available for that model. **Bold** and underline indicate the highest and second highest performance, respectively.

Model	Single-turn Q&A				Multi-turn Conversation				Wild
	CMB ↑	CME ↑	Ency ↑	Safety ↓	MedDG ↑	AIHospital ↑	Resp.Len.	Conv.Num.	Vote ↑
Medical LLMs									
HuatuoGPT2	60.39	69.16	63.45	2.18	79.25	80.70	242.44	3.62	24
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
ShizhenGPT-Omni	74.58	71.95	53.72	2.18	76.06	76.51	1066.20	3.12	5
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	14
SpeechGPT2	35.57	35.57	56.65	2.48	82.36	80.28	114.28	3.54	5
LLaMA-Omni2	73.43	56.98	39.82	1.96	73.18	76.33	61.82	4.37	0
StepAudio2-mini	72.42	74.30	61.26	2.04	76.90	77.53	178.12	3.91	-
OmniLMs									
Qwen2.5-Omni	76.83	75.33	58.12	1.72	76.46	76.53	252.89	3.32	-
BaichuanOmni-1.5	64.15	70.48	62.16	1.90	80.28	80.63	148.60	3.80	-
MiniCPM-o 2.6	21.68	16.01	46.45	2.08	76.53	78.60	153.17	3.95	-
Ours									
SMA-stage1	**78.08**	74.45	44.17	1.56	72.81	70.68	52.77	5.26	
SMA-stage2-10k	77.71	**76.80**	58.14	**1.12**	81.81	81.16	51.44	4.91	**28**
SMA-stage2-198k	77.96	75.48	61.02	1.32	**83.26**	**83.40**	51.36	4.62	

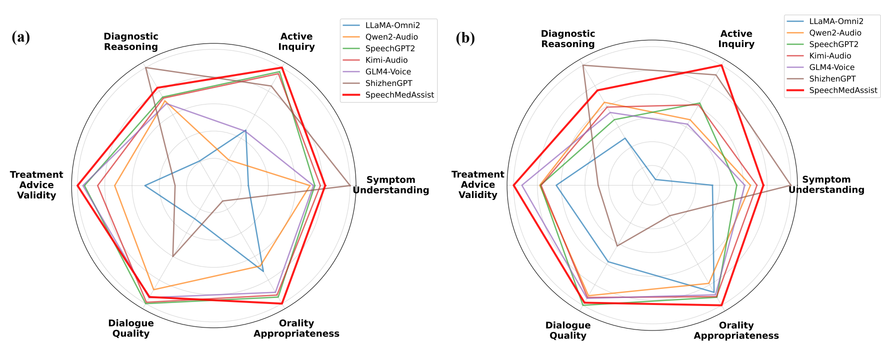

Figure 3: Comparison of our model with other SpeechLMs on multi-dimensional metrics of multi-turn conversations, based on consultation background from MedDG (a) and AIHospital (b). Apart from a few metrics that favor long-text responses, our model exhibits strong diagnostic capabilities.

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.

Speech Quality We evaluate the quality of models' speech responses from three dimensions. **(1) UTMOS**: We utilize a Mean Opinion Score (MOS) prediction model called UTMOS (Saeki et al., 2022) to measure the naturalness of the output speech. **(2) ASR-CER**: To assess the alignment between text responses and speech responses, we convert the speech output back to text using an ASR model and compare it with the original text response to calculate the Character Error Rate (CER). **(3) Latency**: We define the model's response latency as the time elapsed from the start of speech input into the speech encoder to the generation of the first chunk of speech by the model.

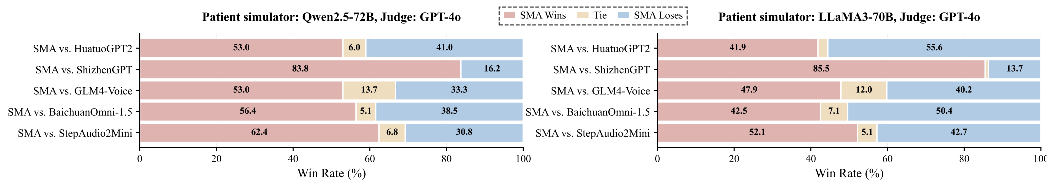

Figure 4: Win rates of our model against strong baselines, using Qwen2.5-72B and LLaMA3-70B as patient simulators and GPT-4o as the judge. Our model achieves higher win rates in most settings.

5.3 Main Results

Table 2 reports the evaluation results of medical LLMs, mainstream general-purpose SpeechLMs, and our model on single-turn Q&A, multi-turn conversation, and wild tasks. All metrics are assessed through text-based and speech-based interaction, except for CMB and CMExam only in text form. Results of the text-based evaluation are provided in the Appendix B. On most evaluation metrics, our model outperforms existing medical LLMs and general-purpose speech LLMs.

Knowledge Transfer and Alignment Evaluation on CMB and CMExam shows that our model acquires fundamental medical knowledge after the first stage of knowledge injection and maintains this knowledge without degradation after the second stage training. For medical encyclopedia Q&A and safety evaluation, model's performances show slight improvement after stage1 and increases significantly after stage2. This suggests that part of the medical knowledge acquired from text can be generalized to the speech modality, but the text-only training in stage1 weakens speech-text alignment, making the modality alignment in stage2 necessary. In later sections, we further demonstrate that effective alignment can be achieved with only a small amount of speech data.

Table 3: Evaluation results on the general speech QA benchmark VoiceBench, noise robustness on single-turn medical QA, and general knowledge retention measured by MMLU.

Model	Noise Robustness			VoiceBench				MMLU
	Noise=0	Noise=0.2	Noise=0.6	BBH	AdvBench	CommonEval	OpenBookQA	
Zhongjing+ASR+TTS	54.63	$53.49_{\downarrow 1.14}$	$50.95_{\downarrow 3.68}$	48.83	79.80	2.01	28.35	32.81
Qwen2-Audio	49.48	$46.34_{\downarrow 3.14}$	$43.85_{\downarrow 5.63}$	54.70	96.73	3.43	49.45	51.38
ShizhenGPT	53.72	$52.27_{\downarrow 1.45}$	$49.20_{\downarrow 4.52}$	46.51	53.46	1.28	37.80	66.36
GLM4-Voice	54.43	$53.60_{\downarrow 0.83}$	$48.25_{\downarrow 6.18}$	52.80	88.08	3.42	53.41	45.12
BaichuanOmni-1.5	62.16	$59.15_{\downarrow 3.01}$	$55.34_{\downarrow 6.82}$	62.70	97.31	4.050	74.51	66.25
LLaMA-Omni2	39.82	$30.47_{\downarrow 9.35}$	$29.78_{\downarrow 10.04}$	27.13	59.80	3.118	58.13	67.48
SMA-stage2-10k	58.14	$55.82_{\downarrow 2.32}$	$51.79_{\downarrow 6.35}$	55.81	79.80	2.028	59.80	69.49
SMA-stage2-198k	61.02	$58.99_{\downarrow 2.03}$	$58.67_{\downarrow 2.35}$	58.14	82.69	2.046	60.66	69.94

Diagnostic Performance Comparison In a simulated clinical setting, we evaluate models' diagnostic capabilities through dynamic multi-turn consultation. As shown in Table 2, under two background settings, our model consistently achieves the best performances while generating concise responses and maintaining a moderate average number of turns, which better aligns with real medical consultations. To better compare model capabilities, we visualize the performance of our model and other speechLMs across six dimensions in Figure 3. Overall, our model achieves superior results on most metrics. Notably, ShizhenGPT produces responses nearly 20 times longer than ours, which boosts its scores in reasoning and symptom understanding but reduces efficiency and practicality. In real-world settings, our model receives the most votes from medical professionals, highlighting its robustness and closer fidelity to actual clinical consultations.

To improve the reliability of our evaluation, we conduct pairwise comparisons between our model and several top-performing baselines. Specifically, we use Qwen2.5-72B and LLaMA3-70B independently as patient simulators, and employ GPT-4o as the judge to assess each matchup. We then compute the win rates across these settings. As shown in the Figure 4, our model consistently outperforms the others in most cases.

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 E. 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.

Table 4: Evaluation results of medical LLMs and general-purpose SpeechLMs on input/output capability and quality of output speech. Our model supports both text and speech input and output, and achieves a quality comparable to TTS model with very low latency. * means that the model is theoretically capable of streaming generation, but not implemented in the publicly released code.

Model	Input	Output	Stream?	Added Module	UTMOS ↑	ASR-CER ↓	Latency ↓
Medical LLMs	🗒	🗒	✗	ASR & TTS	**3.96**	6.77	3520ms
Qwen2-Audio	🎙, 🗒	🗒	✗	TTS	**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
SMA(Ours)	🎙, 🗒	🔊, 🗒	✓	-	3.75	7.71	**367ms**

5.4 Speech Input Robustness and Output Quality

Noise Robustness Real-world medical consultations involve diverse acoustic challenges such as dialects, pauses, speaker timbre, speaking rate, and environmental noise. To approximate these conditions, we have already used FishSpeech to synthesize patient speech with naturally varied characteristics in the previous evaluations. For noise robustness evaluation, we further add noise sampled from the MS-SNSD (Reddy et al., 2019) dataset (e.g., babble) to the synthesized speech in the single-turn setting and quantify its impact using CER. As shown in the table, noise levels of 0, 0.2, and 0.6 correspond to CER of 9.77%, 10.20%, and 12.19%, respectively. As shown in Table 3, while all models degrade under stronger noise, our model consistently maintains SOTA performance. Even at the highest noise level, our model remains competitive and surpasses most baselines, demonstrating strong robustness in challenging audio conditions.

Knowledge Retention Ability Since both stages of our training pipeline rely exclusively on medical-domain data, there is a potential risk that domain adaptation may erode the model's original general-purpose knowledge. To assess whether such degradation occurs, we evaluate the model's text reasoning and generation ability using the MMLU dataset (Hendrycks et al., 2021), and its speech understanding quality using VoiceBench (Chen et al., 2024b). The results are summarized in Table 3. From the table, we observe that our model not only preserves but even improves performance on several QA tasks compared with the base model LLaMA-Omni2, with only minor drops on a few tasks. Notably, its performance on AdcBench (a benchmark for general speech safety), improves substantially, indicating that our training strategy strengthens safety without sacrificing capability. At the same time, the model maintains strong performance on general text-based QA. Overall, these results show that our training pipeline introduces minimal interference with the model's general knowledge and exhibits virtually no signs of catastrophic forgetting.

Output Speech Quality Except for diagnostic capability and interactivity, medical consultation also requires real-time responsiveness and accuracy. Table 4 shows the speech quality evaluation of cascaded models, general-purpose SpeechLMs, and our model. For medical LLMs, results are obtained by integrating same ASR and TTS modules with different models and then averaging their results. We observe that cascaded models generate speech directly from text, avoiding the noise introduced by unit-based synthesis in end-to-end models, which results in slightly better UTMOS and ASR-CER scores. However, they incur higher latency compared to end-to-end speech LLMs.

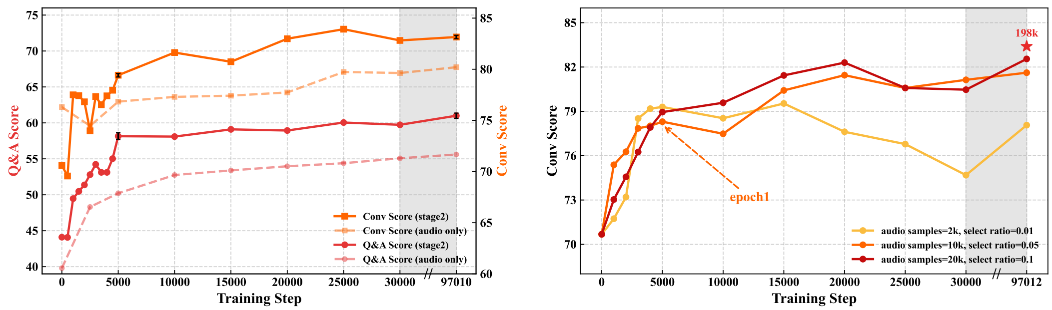

Figure 5: **Left**: Comparison of the performance between the model trained in stage2 and the model trained from scratch on speech data, for single-turn Q&A and multi-turn conversation evaluations across training steps. To ensure the reliability of our conclusions, we compute the variance at step 5k and 97k. **Right**: 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.

Overall, our model supports text–speech multimodal input and streaming output, achieving TTS-level speech quality and outperforming other general-purpose SpeechLMs.

5.5 Demand of Speech Data for Alignment

After Stage 1, the LLM core has already acquired relevant medical knowledge and diagnostic abilities, which can be proved by the results on text interactions in Appendix B. In Stage 2, we re-align speech and text modalities using only a small amount of speech data, enabling the model to transfer knowledge&capabilities learned from text to speech modality. To assess how much speech data is needed, we evaluate intermediate checkpoints with Ency and AIHospital, metrics used before.

As shown in the left of Figure 5, we observe that two speech-based metrics rise rapidly within the first 0–5k training steps, with growth rates **91 times** and **43 times** higher than those during the subsequent 5k–97k steps, respectively. We point out that modality alignment occurs primarily during the 0–5k training step, where medical knowledge and diagnostic abilities acquired in the first stage are quickly generalized to the speech modality. In the later steps of training, the model begins to acquire more fine-grained knowledge and abilities from speech-modality data, leading to a slower rise in the related metrics. We also omit Stage 1 and train the base model directly on speech data (audio only). The scores rise relatively slowly during training, indicating that acquiring knowledge and capabilities directly from complex speech modality is challenging for the LLM core.

The left figure shows that modality alignment mainly occurs within the first 0–5k steps, requiring only about 10k audio samples. To examine whether fewer samples can achieve alignment through more training epochs, we train models with 2k, 10k, and 20k audio samples and present how the conv scores evolve for each setting in the right of Figure 5, which shows the same rapid improvement within 0–5k steps. However, with continued training, the 2k-sample model overfits and its performance declines, while the 10k- and 20k-sample models continue to improve gradually, eventually approaching the performance of the model trained with 198k samples. These results indicate that too little audio data may lead to overfitting, whereas excessive data incurs high collection or synthesis costs. For our model, 10k audio samples are sufficient to achieve effective modality alignment.

6 Conclusion

In this work, we propose SpeechMedAssist, the first 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 the medical domain as well as other vertical domains that lack large-scale speech data , and has the potential to drive the adoption of SpeechLMs in vertical applications.

7 Ethical Considerations

All existing datasets used in this paper are publicly available for research purposes. The self-constructed TextMedDataset and SpeechMedDataset have been further processed to remove sensitive and private information, and they do not contain subjective bias or discriminatory content. It should be noted that although we impose safety constraints during data construction and training, and the evaluation results are satisfactory, the model may still generate harmful content especially in the absence of additional downstream control.

8 Reproducibility statement

Currently, only a limited number of medical LLMs or general-purpose SpeechLMs have open-sourced their training and evaluation code. To facilitate the adoption of SpeechLMs in vertical domains such as the medical domain, we release the code of the end-to-end SpeechLM, whose architexture is described in Section 2, and the training code, which corresponds to the two-stage training strategy introduced in Section 3. In addition, we provide the code for constructing medical text and speech datasets, described in Section 4. All evaluation-related code is also made available, including the reproduction code for the baseline models discussed in Section 5, which have been encapsulated into multiple classes with the interaction API. All of the corresponding code is currently hosted in an anonymous GitHub repository `https://anonymous.4open.science/r/SpeechMedAssist-Anonymous/`, and speech interaction examples are also included in this repository. After the review process, the camera-ready version will further include links to the datasets, model weights, and other resources.

References

Scott J. Adams, Julián Nicolás Acosta, and Pranav Rajpurkar. How generative AI voice agents will transform medicine. *npj Digit. Medicine*, 8(1), 2025. doi: 10.1038/S41746-025-01776-Y. URL `https://doi.org/10.1038/s41746-025-01776-y`.

Sourav Banerjee, Ayushi Agarwal, and Promila Ghosh. High-precision medical speech recognition through synthetic data and semantic correction: UNITED-MEDASR. *CoRR*, abs/2412.00055, 2024. doi: 10.48550/ARXIV.2412.00055. URL `https://doi.org/10.48550/arXiv.2412.00055`.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. Disc-medllm: Bridging general large language models and real-world medical consultation. *CoRR*, abs/2308.14346, 2023. doi: 10.48550/ARXIV.2308.14346. URL `https://doi.org/10.48550/arXiv.2308.14346`.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann (eds.), *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*, pp. 137–144. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper/2006/hash/b1b0432ceafb0ce714426e9114852ac7-Abstract.html`.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T. Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F. Chen, and Stefan Winkler. MEDSAGE: enhancing robustness of medical dialogue summarization to ASR errors with llm-generated synthetic dialogues. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 23496–23504. AAAI Press, 2025. doi: 10.1609/AAAI.V39I22.34518. URL `https://doi.org/10.1609/aaai.v39i22.34518`.

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

Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Trans. Vis. Comput. Graph.*, 18 (12):2649–2658, 2012. doi: 10.1109/TVCG.2012.291. URL https://doi.org/10.1109/TVCG.2012.291.

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. Huatuogpt-ii, one-stage training for medical adaption of llms. *CoRR*, abs/2311.09774, 2023. doi: 10.48550/ARXIV.2311.09774. URL https://doi.org/10.48550/arXiv.2311.09774.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *CoRR*, abs/2412.18925, 2024a. doi: 10.48550/ARXIV.2412.18925. URL https://doi.org/10.48550/arXiv.2412.18925.

Junying Chen, Zhenyang Cai, Zhiheng Liu, Yunjin Yang, Rongsheng Wang, Qingying Xiao, Xiangyi Feng, Zhan Su, Jing Guo, Xiang Wan, et al. Shizhengpt: Towards multimodal llms for traditional chinese medicine. *arXiv preprint arXiv:2508.14706*, 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*, 2024b.

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. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024. doi: 10.48550/ARXIV.2407.10759. URL https://doi.org/10.48550/arXiv.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, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. Recent advances in speech language models: A survey. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *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*, pp. 13943–13970. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.682/.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. AISHELL-2: transforming mandarin ASR research into industrial scale. *CoRR*, abs/1808.10583, 2018. URL http://arxiv.org/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. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117, 2024. doi: 10.48550/ARXIV.2412.10117. URL https://doi.org/10.48550/arXiv.2412.10117.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, and Fei Huang. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. 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, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 10183–10213. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.680/.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 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, 2025a. URL https://openreview.net/forum?id=PYmrUQmMEw.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni 2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *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*, pp. 18617–18629. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.acl-long.912/.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. Medsafetybench: Evaluating and improving the medical safety of large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/3ac952d0264ef7a505393868a70a46b6-Abstract-Datasets_and_Benchmarks_Track.html.

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

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. Audiogpt: Understanding and generating speech, music, sound, and talking head. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *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*, pp. 23802–23804. AAAI Press, 2024. doi: 10.1609/AAAI.V38I21.30570. URL https://doi.org/10.1609/aaai.v38i21.30570.

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

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. Wavchat: A survey of spoken dialogue models. *CoRR*, abs/2411.13577, 2024. doi: 10.48550/ARXIV.2411.13577. URL https://doi.org/10.48550/arXiv.2411.13577.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report. *CoRR*, abs/2504.18425, 2025. doi: 10.48550/ARXIV.2504.18425. URL https://doi.org/10.48550/arXiv.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. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 11116–11141. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.ACL-LONG.599. URL https://doi.org/10.18653/v1/2024.acl-long.599.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W. Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In Amir Globersons, Lester Mackey, Danielle

Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/32b80425554e081204e5988ab1c97e9a-Abstract-Conference.html.

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

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *CoRR*, abs/2411.01156, 2024. doi: 10.48550/ARXIV.2411.01156. URL https://doi.org/10.48550/arXiv.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. Benchmarking large language models on cmexam - A comprehensive chinese medical exam dataset. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a48ad12d588c597f4725a8b84af647b5-Abstract-Datasets_and_Benchmarks.html.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In Wei Lu, Shujian Huang, Yu Hong, and Xiabing Zhou (eds.), *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*, pp. 447–459. Springer, 2022. doi: 10.1007/978-3-031-17120-8_35. URL https://doi.org/10.1007/978-3-031-17120-8_35.

Si-Ioi Ng, Lingfeng Xu, Ingo Siegert, Nicholas Cummins, Nina R. Benway, Julie Liss, and Visar Berisha. A tutorial on clinical speech AI development: From data collection to model validation. *CoRR*, abs/2410.21640, 2024. doi: 10.48550/ARXIV.2410.21640. URL https://doi.org/10.48550/arXiv.2410.21640.

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

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *CoRR*, abs/2502.19634, 2025. doi: 10.48550/ARXIV.2502.19634. URL https://doi.org/10.48550/arXiv.2502.19634.

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

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

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 4521–4525. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-439. URL https://doi.org/10.21437/Interspeech.2022-439.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 2840–2861. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.167. URL https://doi.org/10.18653/v1/2024.findings-acl.167.

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

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. CMB: A comprehensive medical benchmark in chinese. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *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*, pp. 6184–6205. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.343. URL https://doi.org/10.18653/v1/2024.naacl-long.343.

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. Huatuo-26m, a large-scale chinese medical QA dataset. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 3828–3848. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.211. URL https://doi.org/10.18653/v1/2025.findings-naacl.211.

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, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Nie Hao, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, and Yuxiang Zhang. Step-audio 2 technical report. *CoRR*, abs/2507.16632, 2025. doi: 10.48550/ARXIV.2507.16632. URL https://doi.org/10.48550/arXiv.2507.16632.

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.

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

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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *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*, pp. 19368–19376. AAAI Press, 2024b. doi: 10.1609/AAAI.V38I17.29907. URL https://doi.org/10.1609/aaai.v38i17.29907.

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.

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. *CoRR*, abs/2412.02612, 2024. doi: 10.48550/ARXIV.2412.02612. URL https://doi.org/10.48550/arXiv.2412.02612.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=3tukjsVyrE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 15757–15773. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.FINDINGS-EMNLP.1055. URL https://doi.org/10.18653/v1/2023.findings-emnlp.1055.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. *CoRR*, abs/2401.13527, 2024. doi: 10.48550/ARXIV.2401.13527. URL https://doi.org/10.48550/arXiv.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. Huatuogpt, towards taming language model to be a doctor. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 10859–10885. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.FINDINGS-EMNLP.725. URL https://doi.org/10.18653/v1/2023.findings-emnlp.725.

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

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. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. Large model driven radiology report generation with clinical quality reinforcement learning. *CoRR*, abs/2403.06728, 2024. doi: 10.48550/ARXIV.2403.06728. URL https://doi.org/10.48550/arXiv.2403.06728.

A Text Embedding changes in the training process

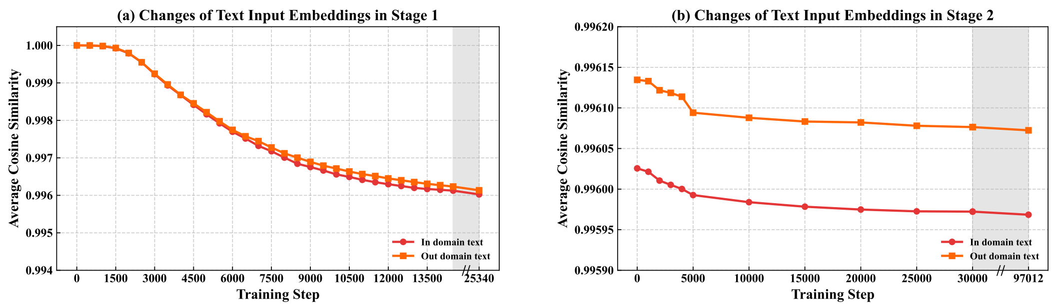

Figure 6: Average cosine similarity between the text input embeddings of the original model and those of the model at each training step.

Since medical consultation is a subset of dialogue tasks, and general-purpose speech LLMs are already trained on large-scale text and speech dialogues, further training on medical text dialogues 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 6, with the left panel illustrating changes during stage 1 training and the right panel illustrating stage 2 training. 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.

B The results of Text-based Multi-turn Conversation Evaluation

Table 5: Evaluation results of various models on text-based multi-turn conversation using real-world patient-doctor conversations as background from MedDG dataset.

Model	Symptom Understanding	Active Inquiry	Diagnostic Reasoning	Treatment Advice Validity	Dialogue Quality	Orality Appropriateness	Average
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-stage1	7.95	8.01	7.45	8.47	8.58	9.08	8.26
SMA-stage2	8.03	8.02	7.51	8.53	8.67	9.15	8.32

Table 6: Evaluation results of various models on text-based multi-turn conversation using patient info as background from AIHospital dataset.

Model	Symptom Understanding	Active Inquiry	Diagnostic Reasoning	Treatment Advice Validity	Dialogue Quality	Orality Appropriateness	Average
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-stage1	8.49	7.55	8.29	8.54	9.01	9.41	8.55
SMA-stage2	8.44	7.57	8.21	8.58	8.96	9.52	8.55

C DEFINITION OF SIX DIMENSIONS FOR MULTI-TURN DIALOGUE EVALUATION

We formulate the evaluation metrics based on publicly available medical guidelines and physicians' ethical standards, assessing doctors' mastery of professional knowledge and dialogue skills with patients 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).

D Overview of Training Data and Model Parameters

We construct the first medical speech interaction dataset, SpeechMedAssist, consisting of 198k samples totaling 600 hours. Existing large speech models are pretrained on millions of hours of audio, which is nearly impossible for specialized domains. In contrast, we require only about 10k samples (30 hours of audio) to adapt a general-purpose SpeechLM into a medical SpeechLM.

Table 7: Summary of training data and model parameters for different SpeechLMs.

Attribute	GLM4-Voice	Kimi-Audio	Qwen2-Audio	LLaMA-Omni2	SpeechMedAssist
Training Audio Data (Hour / Count)	700k hour	16m hour	500k hour	200k samples	600 hour / 198k samples
Base model Parameters	9B	7B	7B	0.5B-32B	7B
Speech encoder/tokenizer	Zeng et al. (2025)	Zeng et al. (2024)&Whisper		the encoder of Whisper-large-v3	

E Conversation cases of different models

We present in this section examples of interactions between different models, act as doctors, and one same virtual patient provided with the patient profile from AIHospital dataset. The entire interactions are conducted in Chinese speech, which we perform ASR and translation to provide a bilingual text dialogue examples.

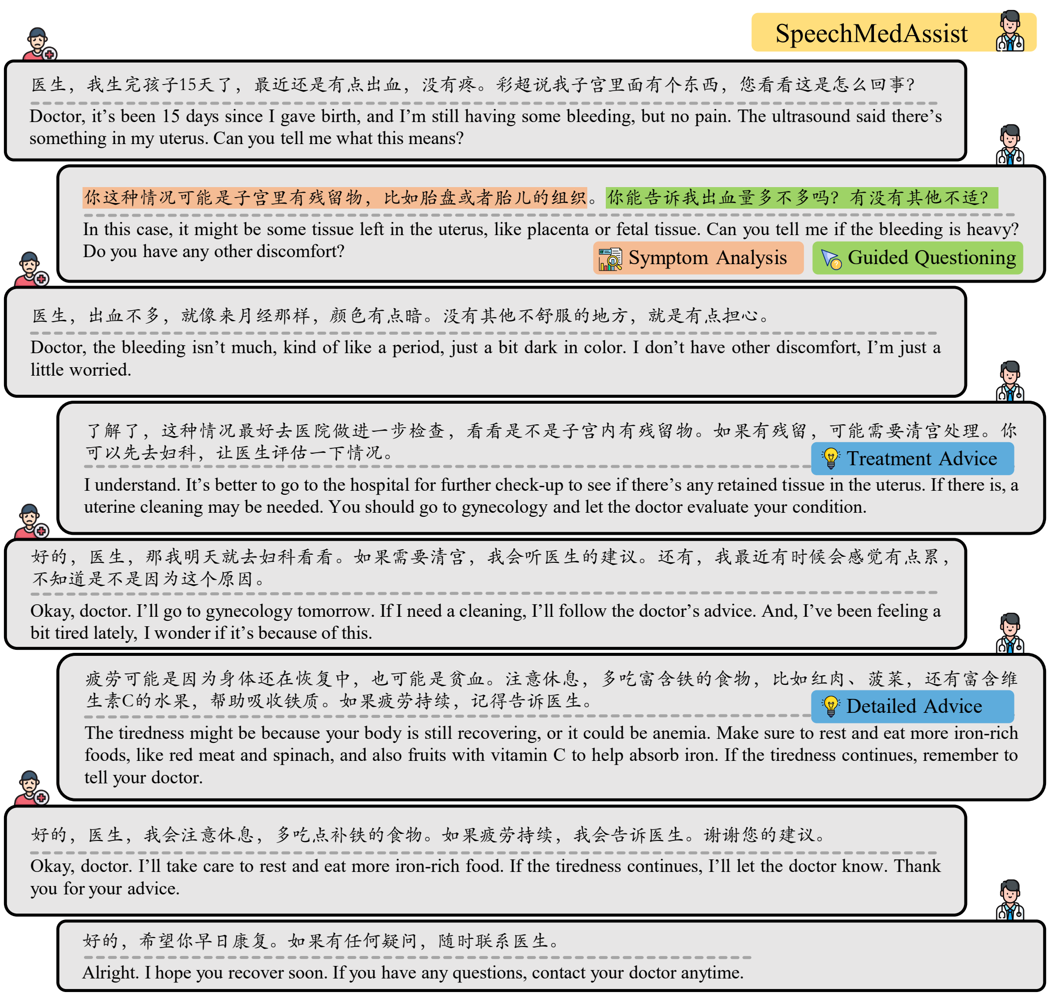

Figure 7: Dialogue between SpeechMedAssist as a consultation assistant and a virtual patient.

ShizhenGPT

医生，我生完孩子已经15天了，但是最近还是有点出血，没有肚子疼。昨天做了个彩超，说子宫里面有个东西，是不是胎盘没排干净啊?

Doctor, it's been 15 days since I gave birth, but I'm still having some bleeding. There's no stomach pain. I had an ultrasound yesterday, and it showed something inside my uterus. Could it be that the placenta hasn't come out completely?

根据您的描述，产后15天仍有出血且彩超提示子宫内有异常情况，需要高度重视并及时就医排查原因。以下是可能的情况分析及建议：\n\n---\n\n### **一、可能的原因**\n1. **胎盘/胎膜残留** \n - 最常见于产后出血的原因之一，彩超显示的“子宫内东西”可能是残留的胎盘组织或蜕膜组织。残留物会干扰子宫收缩，导致持续出血。\n - 伴随症状：出血量时多时少（可能夹血块）、轻微腹痛（但您未提及）。\n\n2. **子宫复旧不全** \n - 子宫未能正常收缩恢复到孕前状态，影响止血功能。\n - 可能与贫血、感染或激素水平异常有关。\n\n3. **产褥感染** \n - 若合并感染，可能出现发热、分泌物异味、下腹压痛等症状，需结合检查判断。\n\n4. **凝血功能障碍** \n - 罕见但需警惕，尤其有妊娠期高血压、糖尿病等高危因素者。\n\n---\n\n### **二、急需处理措施**\n1. **立即就医复查** \n - 建议尽快到妇产科就诊，完善以下检查： \n ✅ **血常规**（查是否贫血或感染） \n ✅ **凝血功能**（排除凝血异常） \n ✅ **HCG检测**（排除滋养细胞疾病） \n ✅ **阴道分泌物培养**（排查感染）\n\n2. **针对性治疗** \n - **药物治疗**：医生可能开具缩宫素、益母草等促进子宫收缩；若存在感染，需用抗生素。 \n - **清宫术**：若残留物较大或出血量多，可能需要手术清除。 \n - **中药调理**：可配合活血化瘀类中药（如生化汤加减），但需在医生指导下使用。\n\n---\n\n### **三、日常注意事项**\n1. **观察出血情况** \n - 记录出血量（如卫生巾使用数量）、颜色（鲜红/暗红/褐色）、是否含血块。 \n - 若出现头晕、心慌、乏力（贫血加重表现），需立即就诊。\n\n2. **避免剧烈活动** \n - 适当休息，避免提重物或久站，促进子宫恢复。\n\n3. **保持清洁** \n - 每日温水清洗外阴，勤换卫生巾，预防感染。\n\n---\n\n### **四、中医辅助调理建议**\n- **食疗方**： \n 黄芪当归炖乌鸡（补气养血）、红豆薏米粥（健脾利湿）。 \n- **穴位按摩**： \n 足三里、三阴交穴，每日按揉5分钟，辅助改善气血循环。 \n- **禁忌**： \n 避免生冷、辛辣食物，忌盲目服用活血类中药（如红花、桃仁）。\n\n---\n\n### **总结** \n产后出血超过2周属于异常情况，需优先排除胎盘残留和感染。请务必在24小时内到医院妇科就诊，明确病因后规范治疗，避免延误导致贫血或感染加重。祝您早日康复！

Non-interactive

Based on your description, having bleeding 15 days postpartum along with an ultrasound showing something abnormal in the uterus needs to be taken seriously and checked right away. Here are some possible causes and suggestions:1. Possible causesRetained placenta/membranesOne of the most common causes of postpartum bleeding. What the ultrasound showed might be retained placental tissue or decidual tissue. This can interfere with uterine contraction and cause ongoing bleeding.Possible symptoms: irregular bleeding (sometimes with clots), mild abdominal pain (though you haven't mentioned it).Subinvolution of the uterusThe uterus hasn't contracted and returned to its pre-pregnancy state, affecting its ability to stop bleeding.May be related to anemia, infection, or hormonal imbalance.Puerperal infectionIf infection is present, symptoms may include fever, foul-smelling discharge, or lower abdominal tenderness, which need further examination.Coagulation disordersRare but should be considered, especially if you had pregnancy complications like hypertension or diabetes.2. What needs to be done urgentlySee a doctor as soon as possibleGo to the obstetrics/gynecology department for the following tests: ✅ Blood test (to check for anemia or infection) ✅ Coagulation profile (to rule out clotting problems) ✅ HCG test (to rule out gestational trophoblastic disease) ✅ Vaginal secretion culture (to check for infection)Targeted treatmentMedications: Doctors may prescribe oxytocin, motherwort, or other drugs to help uterine contraction; antibiotics if infection is present.Uterine evacuation (D&C): If retained tissue is significant or bleeding is heavy, surgical removal may be necessary.Traditional Chinese medicine: May use blood-activating herbal formulas (e.g., modified Shenghua Decoction) but only under a doctor's guidance.3. Daily precautionsMonitor bleedingKeep track of amount (e.g., number of pads used), color (bright red/dark red/brown), and presence of clots.If you experience dizziness, palpitations, or fatigue (signs of worsening anemia), seek medical help right away.Avoid strenuous activityRest well, avoid lifting heavy objects or standing too long to support uterine recovery.Maintain hygieneWash the perineal area daily with warm water, change pads frequently, prevent infection.4. TCM supportive careDiet therapy:Astragalus and Angelica stewed with black chicken (to tonify qi and blood); red bean and coix seed porridge (to strengthen the spleen, drain dampness).Acupoint massage:Zusanli (ST36), Sanyinjiao (SP6), massage 5 minutes daily to improve blood circulation.Avoid:Cold, spicy food; and don't take blood-activating herbs (e.g., safflower, peach kernel) without guidance.SummaryBleeding more than 2 weeks postpartum is abnormal. Retained placenta and infection need to be ruled out first. Please go to the hospital's gynecology department within 24 hours for proper diagnosis and treatment to avoid worsening anemia or infection. Wishing you a speedy recovery!

医生，你说的我都明白了。我现在就去妇产科做那些检查。如果检查结果出来有什么问题，我会再联系您的。谢谢医生！

Doctor, I understand everything you said. I'll go to the obstetrics and gynecology department now for those tests. If the results show any problems, I'll contact you again. Thank you, doctor!

Figure 8: Dialogue between ShizhenGPT as a consultation assistant and a virtual patient.

SpeechGPT2

医生，我生完孩子15天了，最近还是有点出血，但不疼。今天做了个彩超，说是子宫里面有个东西，是不是没排干净？这严重吗？

Doctor, it's been 15 days since I gave birth. I'm still having some bleeding, but no pain. I had an ultrasound today, and they said there's something in my uterus. Could it be that it hasn't come out completely? Is this serious?

听你这么说，可能是有点残留，不过呢，很多时候身体会自己吸收这些东西。你看啊，彩超显示有东西，但不一定就是大问题。要是没出血或者不疼，通常问题不大。不过呢，还是建议你去医院再查查，医生会给你更具体的建议。你平时有没有注意休息和饮食呢？

Unclear

From what you said, it might be some tissue left inside, but often the body can absorb it on its own. You see, the ultrasound shows something, but it's not always a big problem. If there's no heavy bleeding or pain, usually it's not serious. But still, I suggest you go back to the hospital for another check. By the way, have you been paying attention to rest and diet?

医生，我这几天都在家里休息，吃得也还可以。就是有点担心这个子宫里面的东西，会不会影响恢复啊？需要怎么处理呢？

Doctor, I've been resting at home these days, and eating pretty well. I'm just worried about this thing in my uterus. Will it affect recovery? How should it be treated?

你在家休息和注意饮食是对的，挺好的。至于那个子宫里的东西，如果医生说是残留，可能会建议你再观察一段时间。身体有时候会自己处理这些问题。要是一直不消失，或者有其他症状，就得再去医院看看。你有没有感觉到其他不舒服的地方呢？

Lack of medical evidence

It's good that you've been resting and eating well. As for that thing in the uterus, if the doctor says it's retained tissue, they may suggest observing for a while. Sometimes the body can take care of it by itself. If it doesn't go away or you develop other symptoms, then you'll need to go back to the hospital. Have you noticed any other discomfort?

医生，我没有别的不舒服，就是偶尔会有点担心。那我接下来应该怎么办呢？需要吃药或者做些什么吗？

Doctor, I don't feel any other discomfort, just sometimes a bit worried. So what should I do next? Do I need to take medicine or do anything else?

既然没有其他不舒服，那就先不用太担心。一般来说，保持良好的生活习惯就好。你可以多喝水，注意饮食清淡，别太劳累。至于药物，最好还是听医生的建议，别自己乱吃。要是有任何不适，及时去医院复查。对了，你平时有做什么运动吗？

Lack of Information

Since you don't have other discomfort, you don't need to worry too much for now. Generally, just keep healthy habits. Drink more water, eat light, and don't get too tired. As for medicine, it's better to follow the doctor's advice and not take anything on your own. If you feel unwell, go back to the hospital for a check. By the way, do you usually do any exercise?

医生，我不怎么运动，就是在家干点轻活。那我接下来就多注意休息，多喝水，观察一段时间。要是有什么不对劲，我就再去医院复查。谢谢医生。

Doctor, I don't really exercise, just some light chores at home. So I'll pay more attention to resting, drink more water, and observe for a while. If anything feels wrong, I'll go back to the hospital. Thank you, doctor.

不客气，你这样做就很对了。多休息，多喝水，保持心情愉快，对身体恢复都很有帮助。偶尔可以做点轻松的活动，比如散步，别太剧烈就行。要是有啥问题，随时去医院。你看，生活中有时候也需要放松心情，对吧？

You're doing the right thing. Rest more, drink water, keep a good mood—it all helps recovery. You can do some light activities, like walking, but nothing too heavy. And if you notice anything wrong, go to the hospital right away. You see, sometimes in life it's also important to relax, right?

Figure 9: Dialogue between SpeechGPT2 as a consultation assistant and a virtual patient.

HuatuoGPT2

医生，我生完孩子已经15天了，但最近还是有点出血，没有疼痛。昨天做了个彩超，说子宫里面有个东西，是不是没排干净的胎盘？这严重吗？

Doctor, it's been 15 days since I gave birth, but I'm still having some bleeding, without pain. Yesterday I had an ultrasound, and it showed something inside my uterus. Could it be placenta that hasn't come out completely? Is this serious?

您好，首先恭喜您顺利生产。产后出血是正常的生理现象，一般会持续一周左右，但如果超过这个时间还有出血，就需要引起注意。这可能是由于子宫内膜恢复不全或者残留的胎盘组织等原因引起的。\n\n如果超声检查显示子宫内有异物，可能是未排出的胎盘碎片或者其他物质。这种情况如果不及时处理，可能会引发感染，甚至可能导致子宫收缩不良，影响身体的恢复。\n\n对于这种情况，通常需要进行清宫手术，也就是清除子宫内的异物。这是一种相对简单的手术，一般在全麻或者局麻下进行，手术后需要服用一些抗生素来预防感染。\n\n但是，具体的治疗方案还需要根据您的具体情况和医生的建议来确定。如果您有任何不适，比如发热、腹痛等症状，应立即就医。同时，保持良好的个人卫生，避免剧烈运动，多吃营养丰富的食物，有助于身体的恢复。\n\n总的来说，虽然这种情况需要引起重视，但只要及时治疗，一般不会对身体健康造成太大影响。希望这些信息对您有所帮助，祝您早日康复。

Non-interactive

First of all, congratulations on your delivery. Postpartum bleeding is normal and usually lasts about a week, but if it continues longer, it needs attention. It may be due to incomplete recovery of the uterine lining or retained placental tissue.If the ultrasound shows something inside the uterus, it could be retained placental fragments or other material. If not treated in time, this could cause infection or poor uterine contraction, affecting recovery.In such cases, a uterine evacuation (D&C) is often needed to remove what's inside. It's a relatively simple procedure, usually done under anesthesia, and antibiotics are given afterward to prevent infection.The exact treatment, however, depends on your individual condition and your doctor's advice. If you have symptoms like fever or abdominal pain, you should seek care immediately. At the same time, keep good personal hygiene, avoid strenuous activity, and eat nutritious food to support recovery.In short, while this situation needs attention, with timely treatment it usually doesn't have a big impact on health. I hope this helps—wishing you a smooth recovery.

医生，我明白了。那我需要再做一些检查吗？比如血常规或者其他的？还有，手术大概需要住院多久？

Doctor, I understand. Do I need more tests, like blood work or others? Also, how long would I need to stay in the hospital after surgery?

是的，你可能需要进行更多的检查以评估你的整体健康状况以及手术的风险。这可能包括血液检查，以确保你的身体准备好接受手术，以及其他可能的检查，如心电图或胸部X光。\n至于手术后的住院时间，这也取决于手术的复杂性和你的身体恢复情况。一般来说，清宫术后你需要在医院住一两天，以便医生可以密切监测你的恢复情况。然后，你可能会回家继续恢复。然而，如果你的恢复情况不佳或有并发症，你可能需要更长时间的住院治疗。\n请记住，每个人的情况都是独特的，所以最好是与你的医生讨论你的具体情况，以便他们可以提供最适合你的建议。

Yes, you'll likely need more tests to evaluate your health and the surgical risk. This may include blood tests, to make sure your body is ready, and possibly an ECG or chest X-ray.As for hospital stay, it depends on how complex the procedure is and how you recover. Usually, after a D&C, you may stay in the hospital for one or two days so doctors can monitor you. After that, you can go home to continue recovery. If recovery is slower or complications occur, you may need to stay longer.Since everyone's situation is unique, it's best to discuss directly with your doctor so they can give you the most suitable advice.

Confused

医生，我明白了。那我明天就去做这些检查。如果需要手术，我会听从您的建议。手术后住一两天应该没问题。谢谢您。

Doctor, I understand. I'll go for those tests tomorrow. If surgery is needed, I'll follow your advice. Staying in the hospital one or two days should be fine. Thank you.

不客气！我希望你能好好休息，并在手术前后一切顺利。如果你有任何其他问题或疑虑，请随时联系。祝你好运！😊

You're very welcome! I hope you can rest well and that everything goes smoothly before and after surgery. If you have any other questions or concerns, feel free to reach out. Best of luck! 😊

Figure 10: Dialogue between HuatuoGPT2 as a consultation assistant and a virtual patient.

F 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 11. 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 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.

Figure 11: Two examples with relatively low scores in MedSafetyBench.

G Prompt

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

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

H THE USE OF LARGE LANGUAGE MODELS (LLMS)

In order to reduce typos during the writing process and to optimize complex sentence structures so that the article becomes simpler and easier to read, we use mainstream large language models to refine certain paragraphs. For example, we use prompts such as "Help me correct the typos and grammatical errors in the above text, and streamline the logic to make it clear and easy to understand."