MEDSONOR: LEARNING TO USE CLINICAL AUDIO IN MEDICAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical audio conveys medical information that transcripts may not preserve, but using it in medical responses requires interpreting it alongside clinical requests and dialogue context. Existing medical audio supervision largely targets classification or recognition, while general audio-language models lack specialized medical knowledge and direct feedback for open-ended responses. We present MedSonor, a medical dialogue model integrating medical knowledge, spoken dialogue, and clinical audio understanding. We pair authentic medical audio with clinical requests and combine these audio-integrated dialogues with collected medical audio and synthesized spoken dialogues for supervised fine-tuning, followed by two stages of rubric-guided reinforcement learning. We introduce MedVoice-Dial and MedVoice-AID to evaluate responses within supplied multi-turn medical contexts and to requests involving clinical audio, respectively. Across seven medical benchmarks, MedSonor outperforms the evaluated general-purpose open-weight audio-language baselines and ranks highest in aggregate physician preference evaluation, while broadly retaining general speech capabilities.

1 INTRODUCTION

Clinical assessment draws on patient reports and acoustic information that transcripts may omit. Respiratory and cardiac recordings support sound and murmur analysis Rocha et al. (2019); Oliveira et al. (2022), while voice research spans laryngeal, neurological, and mood-related conditions Bensoussan et al. (2024). While large language models (LLMs) show promise in medical reasoning and consultation Singhal et al. (2023); Chen et al. (2023); Tu et al. (2024), medical voice assistants must also interpret these acoustic cues alongside the clinical request and dialogue context.

However, existing audio datasets do not fully reflect real-world clinical tasks, limiting their ability to supervise useful medical responses. Bridging this gap also requires adapting medical knowledge to acoustic inputs. Further improving these responses remains difficult without task-specific feedback for medical audio dialogue.

We present MedSonor, an audio-conditioned medical dialogue model adapted from Qwen2.5-Omni-7B. We address these challenges through three components. **(1) Contextualized task construction:** We collect and standardize medical audio, synthesize spoken multi-turn dialogues across clinical specialties and user roles, and construct audio-integrated dialogues that pair authentic audio with clinical requests, connecting existing audio resources to medical response tasks. **(2) Adapting medical knowledge to audio-based tasks:** We jointly fine-tune the backbone on these complementary sources to develop medical knowledge, medical audio understanding, and context-sensitive response generation. **(3) Rubric-guided response optimization:** We define instance-specific rubrics covering required content, dialogue constraints, and prohibited errors, and train local item-level judges to provide feedback. These signals guide two successive reinforcement learning stages for medical dialogue and audio-integrated requests. Severity-based reward caps prevent high overall rewards when the judges detect major or critical violations.

We evaluate MedSonor using MedVoice-Dial for responses within supplied multi-turn medical contexts and MedVoice-AID for clinical requests containing medical audio, together with five public or speech-adapted medical benchmarks. MedSonor ranks first among the evaluated general-purpose open-weight audio-language baselines on all seven benchmarks and achieves the highest aggregate physician preference ranking in a study of 30 MedVoice-AID cases. Removing audio-integrated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dialogue supervision reduces the MedVoice-AID score from 0.7900 to 0.5957, supporting the value of explicitly pairing medical audio with clinical requests. Six general speech evaluations further show broad retention of the backbone's capabilities. Together, these findings support our approach to adapting general audio-language models to answer clinical questions.

Our contributions are: 1) We introduce MedSonor, combining multimodal medical SFT with two-stage RL guided by instance-specific rubrics and severity-based reward caps. 2) We construct medical audio and dialogue resources and introduce MedVoice-Dial and MedVoice-AID to evaluate medical dialogue responses informed by consultation history and responses to clinical audio. 3) MedSonor leads the evaluated general-purpose open-weight audio-language baselines across seven medical benchmarks and in aggregate physician preference on 30 cases. Ablations support explicit audio–request pairing, with general speech capabilities broadly retained.

2 Background

2.1 Clinical Role of Medical Audio

Medical audio serves two complementary roles: conveying spoken symptoms, histories, and requests, and providing acoustic evidence that transcripts may not preserve. Respiratory and cardiac sounds support auscultation, while voice quality, articulation, and prosody inform research on laryngeal function, neurological status, and emotional state Rocha et al. (2019); Oliveira et al. (2022); Bensoussan et al. (2024). Pain-related vocalizations can complement patient reports, while infant cries provide cues when verbal self-report is unavailable Dao et al. (2025); Bellieni et al. (2004). Table 1 summarizes representative categories and their clinical relevance. These roles make audio both a communication channel and a source of clinical evidence: effective assistance must understand what the user asks and determine what the available sounds can support.

Table 1: Clinical roles of medical audio, spanning physiological sounds, spoken information, and vocal cues. Acoustic findings require clinical context and do not establish diagnoses alone.

Audio category	**Key cues**	**Clinical relevance**
Lung sounds	Wheezes, crackles, breath patterns	Respiratory auscultation.
Snoring	Spectral and temporal patterns	Upper-airway obstruction research.
Cardiac audio	Heart sounds and murmurs	Murmur detection and characterization.
Cough	Cough events and acoustics	Symptom assessment and screening research.
Medical speech	Symptoms, history, concerns	History taking and clinical reasoning.
Spoken intent	Questions and requests	Consultation goals and follow-up.
Voice disorders	Voice quality and articulation	Voice and motor-speech assessment.
Pain-related speech	Speech changes and vocal cues	Complementing pain reports in research.
Infant cries	Pitch, intensity, timing	Pain assessment without self-report.
Affective speech	Prosody and emotion	Mood-related research and communication.

2.2 From Sound Interpretation to Clinical Responses

Representative models emphasize different aspects of medical audio and interaction (Table 2). HeAR focuses on health-acoustic representations, AMIE on text-based consultation, and Qwen2.5-Omni on general audio understanding and dialogue. SpeechMedAssist and StethoLM already connect medical sounds with spoken consultation or instruction-driven clinical tasks. Building on these directions, our focus is on supervision that connects acoustic evidence with responses across clinical requests and interaction contexts.

Recognizing a sound does not by itself determine how to answer a medical question. The same recording may prompt a request for a sound description, an explanation of its relevance, or guidance on what information is needed next. The appropriate response therefore depends on the recording, the current request, and any supplied dialogue history. Conversely, a fluent medical answer may misinterpret the sound or draw conclusions that the evidence cannot support. The learning objective must consequently extend beyond acoustic recognition to *using acoustic evidence in context-sensitive medical responses*.

Table 2: Reported scope of representative models. Entries describe task coverage, not equivalent performance or clinical validation.

Work	Audio modeling	Dialogue and response tasks
HeAR Baur et al. (2024)	Health-acoustic representations	Consultation outside reported scope
Qwen2.5-Omni Xu et al. (2025)	General audio understanding	General-purpose dialogue
AMIE Tu et al. (2024)	Clinical sound interpretation outside reported scope	Text-based clinical consultation
SpeechMedAssist Chen et al. (2026)	Cough and other vocal cues	Spoken medical consultation
StethoLM Wang et al. (2026)	Cardiopulmonary sound interpretation	Instruction-driven clinical tasks

Motivation for MedSonor Effective medical audio dialogue requires not only recognizing sounds, but learning how acoustic evidence should inform a response to a clinical request. Sound-centered datasets provide limited supervision for this connection, while spoken dialogues synthesized from text do not by themselves teach interpretation of authentic clinical recordings. Moreover, an accurate sound description may fail to address the clinical question, while a medically plausible answer may contradict the recording or dialogue history. These gaps motivate MedSonor, which combines collected medical audio, synthesized spoken dialogues, and audio-integrated dialogues to connect acoustic understanding with contextual medical responses. Beyond these demonstrations, two-stage rubric-guided RL provides instance-specific feedback on acoustic interpretation, clinical appropriateness, and dialogue consistency. The rubrics distinguish required content from prohibited errors, with severity-based reward caps preventing high rewards when major or critical violations are detected. Together, this supervision and feedback design targets both *understanding clinical audio* and *using it appropriately in medical responses*, as detailed in Section 3.

3 METHOD

3.1 OVERVIEW

MedSonor receives user audio and, when provided by the task, dialogue history and transcript context. Its Qwen2.5-Omni backbone supports text and speech generation through a Thinker–Talker architecture: the Thinker produces text, and the Talker conditions on its hidden representations and text to generate speech tokens Xu et al. (2025). For consistent evaluation of response content, we use the Thinker's text outputs.

3.2 DATA CONSTRUCTION

We construct three complementary sources that connect medical audio understanding with clinical response tasks, together with rubrics for training feedback (Figure 1).

Collected medical audio. To establish medical audio understanding, we curate physiological sounds and clinically relevant speech, including respiratory and cardiac sounds, medical speech with transcripts, recordings of emotional or disfluent speech, and recordings of voices affected by voice disorders. Examples contain audio, a task category, an available label or response, and a normalized clinical description. This stream includes 349 waveforms reconstructed from Bridge2AI-Voice v3.0.0 spectrograms, rather than original recordings (Appendix E). We remove empty or corrupted audio clips, standardize audio and metadata, and construct training and validation splits (Appendix E).

Synthesized spoken dialogue. To support contextual medical interaction, we combine 25 clinical specialties with 15 scenarios, forming 375 settings with specified topics, user roles, and objectives. GPT-5.4 generates multi-turn dialogues spanning professional, patient, and caregiver interactions. Utterance-level TTS preserves speaker and turn alignment. Scenario definitions and corpus statistics appear in Appendices A and B.

Audio-integrated dialogue. To connect audio with clinical requests, we assign each audio clip a department using its provenance and clinical category, then pair it with the same 15 scenarios. GPT-

5.4 uses the audio clip's description and annotation to generate a single-turn request and reference response. We synthesize the request with TTS and insert the authentic audio clip at the specified position, preserving segment boundaries and order.

Instance-specific rubrics. GPT-5.4 generates rubrics from the request, reference response, and relevant context: preceding user and reference-assistant turns for spoken dialogue, or the audio clip's description for audio-integrated requests. The criteria specify required content, contextual constraints, and prohibited errors, providing item-level feedback for reinforcement learning.

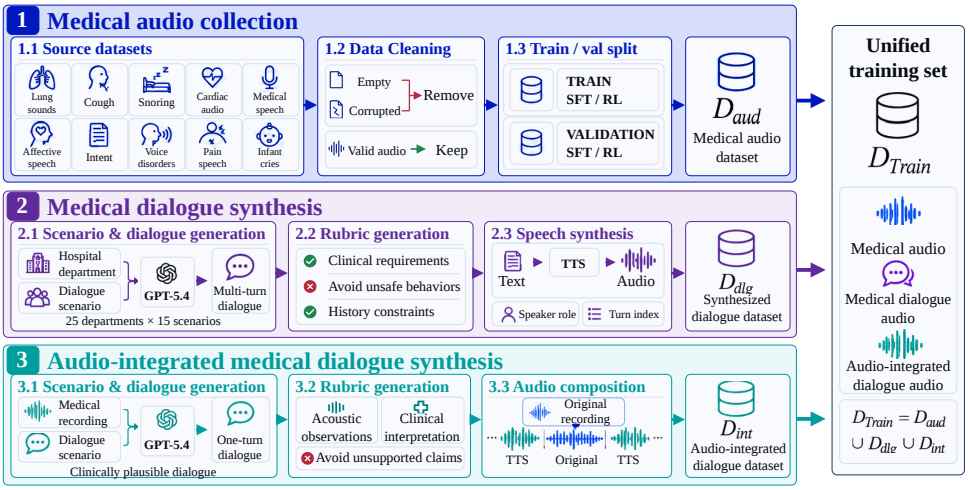

Figure 1: Data construction: standardized medical audio, synthesized spoken multi-turn dialogues with turn-level rubrics, and single-turn dialogues incorporating authentic medical audio.

3.3 TRAINING

As shown in Figure 2, training proceeds from medical SFT to medical dialogue RL and then audio-integrated dialogue RL, with both RL stages guided by severity-guarded rubric rewards.

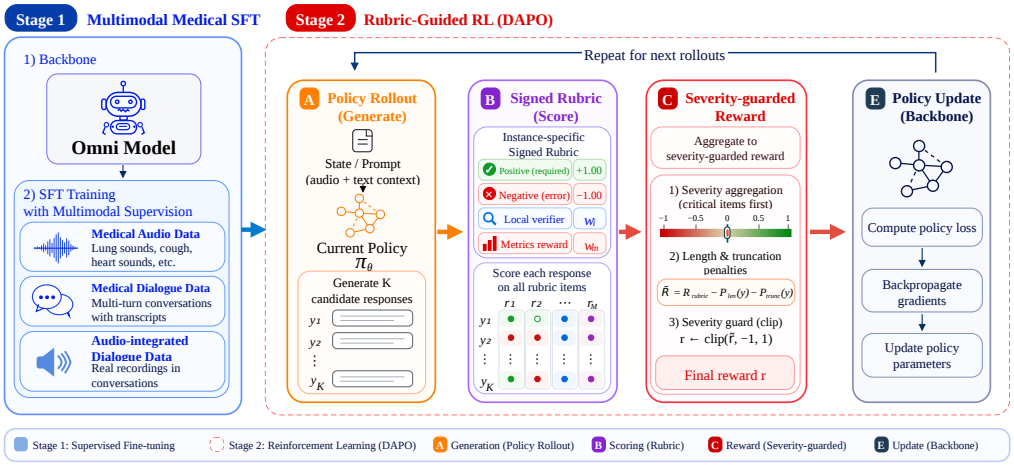

Figure 2: Post-training overview. Medical SFT initializes the policy; successive medical dialogue RL and audio-integrated dialogue RL stages optimize severity-guarded rubric rewards.

Multimodal medical SFT. Medical SFT adapts the general-purpose backbone to medical content and response requirements through collected medical audio, synthesized spoken dialogues, and audio-integrated dialogues. Inputs may contain text, medical audio, spoken requests, and dialogue history. Configuration details are reported in Section D.

Item-level judges. Instance-specific rubrics translate the requirements of each medical request into item-level feedback for optimizing open-ended responses. We train separate local Qwen3-4B-Instruct-2507 (Qwen3-4B-Instruct) judges for dialogue and audio-integrated responses. A rubric comprises positive requirements Γ^+, prohibited errors Γ^-, and history-derived constraints Γ^h. GPT-5.4 labels whether candidate responses satisfy positive or history items and trigger negative items. These binary labels supervise $z(H, u, y, r) \in \{0, 1\}$ for response y and rubric item r; we omit the conditioning context below for brevity. Training and validation details for the local judges appear in Section D.1.

Severity-guarded reward. Both RL stages use normalized weighted coverage of positive and history items, minus negative-item penalties and length/truncation penalties:

$$\widetilde{R}_\Gamma(y) = \frac{\sum_{r_i \in \Gamma^+ \cup \Gamma^h} w_i z(y, r_i)}{\sum_{r_i \in \Gamma^+ \cup \Gamma^h} w_i + \epsilon} - \frac{\sum_{r_j \in \Gamma^-} |w_j| z(y, r_j)}{\sum_{r_j \in \Gamma^-} |w_j| + \epsilon} - P_{\text{len}}(y) - P_{\text{trunc}}(y). \tag{1}$$

Here w_i is an item weight and ϵ stabilizes normalization. The terms $P_{\text{len}}(y)$ and $P_{\text{trunc}}(y)$ penalize excessive response length and truncated responses, respectively; their definitions and settings are provided in Section C. The final reward $R_\Gamma(y)$ caps this score at zero for a recognized major violation and at -0.5 for a critical violation, then clips it to $[-1, 1]$.

Sequential rubric-guided RL. Medical dialogue RL trains on multi-turn instances (H, u, Γ) from the synthesized dialogue corpus. For each input, the policy samples a group of responses, the dialogue judge supplies rubric rewards, and DAPO Yu et al. (2025) updates the policy using group-relative advantages. Both stages optimize reward with KL regularization against the policy that initializes the stage:

$$\max_\theta \; \mathbb{E}_{x \sim \mathcal{D}_{\text{RL}},\, y \sim \pi_\theta(\cdot|x)} \left[R_\Gamma(y) \right] - \beta \, D_{\text{KL}}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x)) \,. \tag{2}$$

Audio-integrated dialogue RL applies the same optimization to spoken requests combined with clinical audio recordings and their associated rubrics, using the separate audio-integrated judge. Data filtering and optimization settings are provided in Sections B and D.

4 Experiments

4.1 Evaluation Protocol

Medical response evaluation covers LiveMedBench Clean-TTS Yan et al. (2026), MMedFD-Lite Chen et al. (2025b), an ACI-Bench-derived response task Yim et al. (2023), HealthDial Hu et al. (2026), and the public ThReadMed-QA review sample Munnangi & Savage (2026). The adapted tasks use speech for the latest user turn, text for history, and hidden references. MedVoice-AID tests responses to medical audio; MedVoice-Dial tests medical dialogue across specialties. Both benchmarks and the three adapted tasks use GPT-5.4 signed instance-rubric scoring.

General speech retention is assessed across the backbone, SFT, and full model on LibriSpeech test-clean Panayotov et al. (2015) and AISHELL-1 Bu et al. (2017) (WER/CER), VoiceBench-IFEval Chen et al. (2024) (official score), and VoiceBench-CommonEval, OpenHermes-Audio test Wang et al. (2025), and a task-stratified WildSpeech-Bench subset Zhang et al. (2025) (mean of three independent GPT-5.5 judgments). Inputs, prompts, and greedy decoding are fixed across stages. We compare seven general-purpose open-weight baselines and three closed-source models (Tables 3 and 4), with matched task content and scoring within each setting. Baselines and input details appear in Appendix D.

4.2 Results

Human evaluation. We conducted a physician preference study on 30 MedVoice-AID cases across five clinical specialties. Two attending physicians and four physicians in standardized residency training compared MedSonor with six open-weight baselines. Given the audio and two anonymous responses, physicians selected a preferred response, a tie, or an unjudgeable comparison. Preferences are aggregated into Bradley–Terry scores with case-bootstrap uncertainty (Appendix F). Across 18 comparisons shared by all six physicians, pairwise reviewer agreement was 59.6% (95% item-bootstrap interval: 46.0–73.8%). Four comparisons received unanimous judgments from all six physicians. Figure 3 shows that MedSonor ranks above all six baselines, suggesting that our pipeline improves alignment with the participating physicians' preferences.

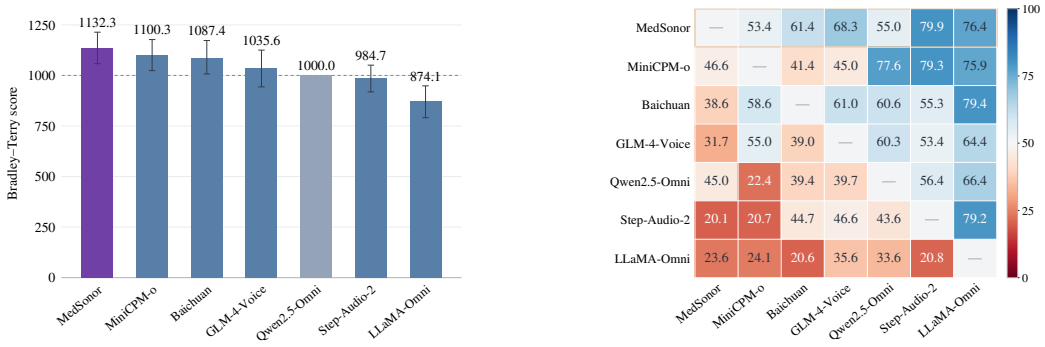

Figure 3: Physician preferences on 30 MedVoice-AID cases, based on final judgments from two attending and four resident physicians. Left: aggregate preference scores with 95% uncertainty intervals. Right: direct preference credits (%) for the row model against the column model, with matching model order. Colors span 0–100%, centered at 50%. Scoring and uncertainty estimation appear in Appendix F.

Benchmark performance. MedSonor ranks first among the evaluated open-weight systems on LiveMedBench Clean-TTS, MMedFD-Lite, MedVoice-AID, and MedVoice-Dial (Table 3). These evaluations cover spoken medical dialogue and responses to clinical requests containing medical audio. The consistent gains over the backbone suggest that our training pipeline strengthens medical response generation across both dialogue contexts and medical audio inputs.

Table 3: Results on medical dialogue benchmarks. MMedFD-Lite values are GPT-5.4 G-Eval scores; LiveMedBench Clean-TTS, MedVoice-AID and MedVoice-Dial values are GPT-5.4 rubric scores. Bold and underline indicate the best and second-best results within the closed-source and open-weight groups separately. The proposed model is included in the open-weight ranking.

Model	LiveMedBench Clean-TTS ↑	MMedFD-Lite ↑	MedVoice-AID ↑	MedVoice-Dial ↑
Closed-source models				
Gemini-2.5-Flash	<u>0.4079</u>	4.624	0.1676	0.1585
Gemini-3.7-Flash	**0.4536**	**6.872**	**0.5579**	**0.7126**
GPT-Audio (2025-08-28)	0.3798	<u>5.224</u>	<u>0.5319</u>	<u>0.6714</u>
Open-weight baselines				
Qwen2.5-Omni-7B	0.2593	5.532	0.3341	0.3237
MiniCPM-o-2.6	0.2283	5.440	<u>0.3903</u>	0.3307
Baichuan-Omni-1.5	<u>0.2803</u>	<u>5.692</u>	0.3490	<u>0.3558</u>
GLM-4-Voice-9B	0.2218	4.952	0.2839	0.2902
Step-Audio-2-mini	0.2358	5.604	0.3442	0.3367
LLaMA-3.1-8B-Omni	0.1122	2.556	0.1411	0.1772
SpeechGPT-2.0-preview	0.0688	4.272	-0.1349	-0.0106
Proposed model				
MedSonor	**0.2843**	**5.780**	**0.7900**	**0.5225**

On the adapted response evaluations using ACI-Bench, HealthDial, and ThReadMed-QA, MedSonor also ranks first among the evaluated open-weight systems (Table 4). It outperforms Gemini-2.5-Flash on HealthDial and ThReadMed-QA while narrowing the gap with stronger closed-source models. These results suggest that the gains extend to medical responses based on clinical encounters, knowledge-grounded dialogues, and patient–physician threads, supporting the model's ability to address diverse medical requests through spoken interaction.

Pairwise preference evaluation. We conducted an automated preference study on 100 MedVoice-AID examples, using Gemini-3.8-Flash to compare responses from MedSonor and six open-weight baselines. The evaluator received the same user audio and two responses without model identifiers, selecting the preferred response based on accuracy, grounding, usefulness, and clarity, with ties allowed. Each pair was evaluated in both presentation orders. We averaged these judgments and used the same ranking method as in the physician study, with uncertainty estimated by case-family bootstrap resampling. As shown in Figure 4, MedSonor ranks above all six baselines in the aggregate preference ranking. This result complements the rubric-based evaluation by showing that MedSonor's advantage is also reflected in judgments of complete medical responses under a different evaluator.

Table 4: Results on public medical dialogue datasets adapted for spoken response evaluation. ACI-Bench denotes our dialogue-response task, not its original note-generation benchmark. All values are GPT-5.4 signed instance-rubric scores. Bold and underline indicate the best and second-best results within the closed-source and open-weight groups separately. The proposed model is included in the open-weight ranking.

Model	ACI-Bench ↑	HealthDial ↑	ThReadMed-QA ↑
Closed-source models			
Gemini-2.5-Flash	0.2509	0.4661	<u>0.2973</u>
Gemini-3.7-Flash	**0.4026**	**0.5875**	**0.4915**
GPT-Audio (2025-08-28)	<u>0.3237</u>	<u>0.5569</u>	0.2858
Open-weight baselines			
Qwen2.5-Omni-7B	0.2002	<u>0.4489</u>	0.1800
MiniCPM-o-2.6	<u>0.2132</u>	0.4160	0.1835
Baichuan-Omni-1.5	0.1977	0.4109	<u>0.2100</u>
GLM-4-Voice-9B	0.1954	0.3678	0.1393
Step-Audio-2-mini	0.2080	0.3605	0.1912
LLaMA-3.1-8B-Omni	0.0989	0.3237	0.0845
SpeechGPT-2.0-preview	-0.0370	-0.0001	-0.0084
Proposed model			
MedSonor	**0.2263**	**0.5022**	**0.3403**

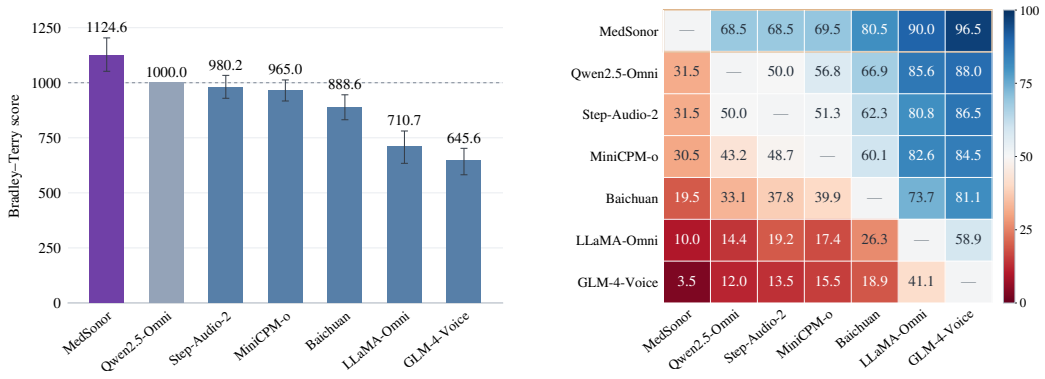

Figure 4: Arena-style preference evaluation on 100 MedVoice-AID examples from 47 case families, judged by Gemini-3.8-Flash. Left: Bradley–Terry scores with 95% intervals from 2,000 case-family bootstrap resamples; Qwen2.5-Omni-7B is fixed at 1000. Right: direct preference credits (%) for the row model against the column model, with model order matched across both panels. The heatmap colorbar spans 0–100%, centered at 50%; colors do not indicate statistical significance. Ties receive half credit, and the two presentation orders are averaged.

General speech capability retention. Table 5 summarizes performance on general speech tasks after medical post-training. MedSonor broadly retains the backbone's capabilities across speech recognition, spoken dialogue, and open-ended instruction following. Performance improves on several dialogue and instruction tasks, while modest declines occur on English ASR and OpenHermes-Audio. These results suggest that medical audio dialogue training strengthens domain-specific capabilities while largely preserving general speech performance.

4.3 Ablation Study

Contribution of multimodal medical SFT. We compare the backbone, SFT-only model, and full pipeline to examine how medical response capabilities develop through successive SFT and RL stages. An RL-only variant applies both RL stages directly to Qwen2.5-Omni-7B to assess the contribution of SFT initialization. As shown in Figure 5 and Table 19, SFT improves on the backbone on four of the five tasks, and the full pipeline further improves on SFT across all five. This pattern supports gains from medical SFT and subsequent RL. RL-only exceeds the backbone only on MedVoice-Dial and

Table 5: General speech capability retention after medical post-training. Libri and AISHELL report WER/CER (%; lower is better); IFEval reports %; CommonEval and OpenHermes report scores out of 5; WildSpeech reports a score out of 10. GPT-5.5 scores average three independent judgments.

	ASR		Dialogue		Open-ended instruction	
Model	**Libri**	**AISHELL**	**IFEval**	**CommonEval**	**OpenHermes**	**WildSpeech**
Qwen2.5-Omni-7B	2.26	2.93	35.55	1.77	3.49	3.35
MedSonor (SFT)	2.60	2.13	47.58	3.40	3.47	4.87
MedSonor	2.78	1.73	45.34	3.30	3.36	4.76

falls below the full pipeline on every task, supporting the value of medical SFT before reinforcement learning.

Contributions of spoken dialogue and collected audio. We remove synthesized spoken dialogue and collected medical audio separately to examine their contributions to medical response generation. As shown in Figure 5 and Table 19, removing spoken dialogue most strongly affects tasks requiring responses to conversation context, supporting its role in contextual medical interaction. Removing collected medical audio reduces performance across all evaluated tasks, suggesting that this supervision benefits medical response generation across different settings. These findings support combining spoken dialogue and collected audio in medical SFT.

Contribution of audio-integrated dialogue. We remove audio-integrated dialogue supervision to examine its contribution to answering clinical requests that involve medical audio. As shown in Figure 5 and Table 19, its removal produces the largest performance decline on MedVoice-AID, supporting the value of explicitly training the model to connect audio with clinical requests. Together with the spoken-dialogue ablation, this result suggests that complementary supervision formats contribute to the model's ability to handle both medical audio and dialogue context.

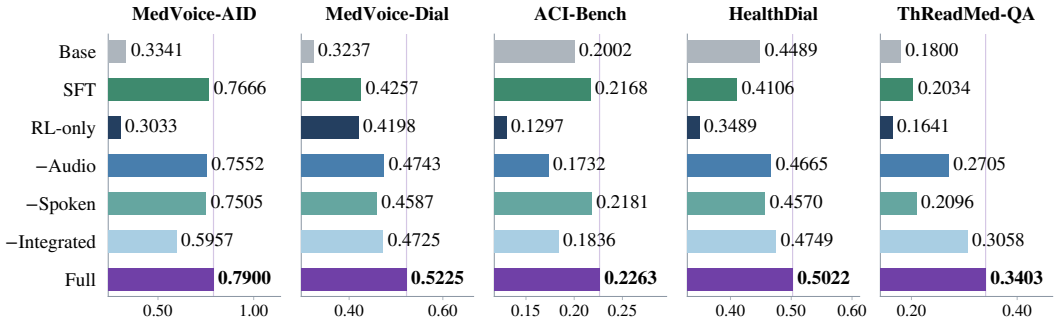

Figure 5: Ablations of medical SFT and training sources. Base is Qwen2.5-Omni-7B. SFT and RL-only omit RL and medical SFT, respectively. Source ablations remove one SFT source while retaining both RL stages. Panels report GPT-5.4 rubric scores (higher is better) with independently scaled, truncated axes. Purple lines indicate the full pipeline. Complete numerical results are reported in Appendix H, Table 19.

4.4 Robustness to Background Noise

Real-world clinical encounters often involve background noise, such as nearby conversations, that can interfere with spoken requests and medical audio. To evaluate robustness to this interference, we overlay babble noise from the MS-SNSD test set Reddy et al. (2019) on the complete input audio. We select one example from each of 100 independent audio sources in MedVoice-AID and use identical noise segments across models. After matching the sampling rate and duration, we mix the input x and zero-mean noise n as

$$x_\alpha = x + \alpha \frac{\mathrm{RMS}(x)}{\mathrm{RMS}(n)} n, \tag{3}$$

where $\alpha \in \{0.2, 0.6\}$ controls the noise amplitude, corresponding to approximately 14.0 and 4.4 dB whole-input signal-to-noise ratios, respectively. We evaluate medical responses with the same GPT-5.4 signed instance-rubric metric and greedy decoding.

As shown in Table 6, MedSonor achieves the highest medical response scores at both noise levels. Although all evaluated models score lower under strong noise than under moderate noise, MedSonor maintains relatively stable performance. These results demonstrate its robustness to background noise.

Table 6: Medical response quality under MS-SNSD babble noise (higher is better).

Model	**Clean**($\alpha = 0$)	**Moderate noise**($\alpha = 0.2$)	**Strong noise**($\alpha = 0.6$)
Qwen2.5-Omni-7B	0.3341	0.2554	0.2212
Baichuan-Omni-1.5	0.3490	0.3039	0.2140
GLM-4-Voice-9B	0.2839	0.2310	0.2076
Step-Audio-2-mini	0.3442	0.3173	0.2645
LLaMA-3.1-8B-Omni	0.1411	0.1858	0.1560
MedSonor	0.7900	0.7480	0.7282

5 Related Work

Medical models. Med-PaLM studies medical question answering, while MEDITRON adapts language models through pretraining on medical literature and guidelines Singhal et al. (2023); Chen et al. (2023). HuatuoGPT and DISC-MedLLM use medical dialogue supervision Zhang et al. (2023); Bao et al. (2023), and AMIE develops diagnostic dialogue through simulated consultations Tu et al. (2024). Med-PaLM M extends biomedical modeling to imaging and genomics Tu et al. (2023). These studies advance medical knowledge and interaction, but their reported tasks do not directly address clinical acoustic evidence in spoken consultation.

Medical audio. Qwen2-Audio and Qwen2.5-Omni support general audio understanding and interaction Chu et al. (2024); Xu et al. (2025), while HeAR and OPERA learn representations for health and respiratory acoustics Baur et al. (2024); Zhang et al. (2024). SpeechMedAssist connects spoken consultation with cough and other vocal cues Chen et al. (2026). StethoLM uses supervised and preference-based training for cardiopulmonary tasks, including reporting and reasoning Wang et al. (2026). We consider diverse clinical sounds and emphasize how their interpretation should inform responses to different requests and consultation contexts.

Rubric-based learning. HealthBench and MedDialogRubrics provide detailed criteria for evaluating medical responses and multi-turn consultations Arora et al. (2025); Gong et al. (2026). Rubrics as Rewards turns instance-specific criteria into feedback for open-ended reinforcement learning Gunjal et al. (2025), while AudioRubrics uses evolving, audio-grounded criteria for audio reasoning Yu et al. (2026). We jointly assess acoustic interpretation and clinical appropriateness with local item-level judges. Severity-based reward caps prevent other strengths from offsetting detected critical errors.

6 Conclusion

We presented MedSonor, which combines multimodal medical SFT with two-stage rubric-guided RL, and introduced MedVoice-Dial and MedVoice-AID for contextual medical dialogue and clinical audio evaluation. Across seven medical benchmarks, MedSonor outperforms the evaluated general-purpose open-weight audio-language baselines and achieves the highest aggregate physician preference ranking on 30 MedVoice-AID cases. Ablations support medical SFT before RL and combined audio–dialogue supervision, while evaluations show broadly retained speech capabilities and relative stability under babble noise. These results support adapting medical voice assistants through supervision that connects audio evidence to clinical requests.

Limitation Our work has several limitations. Our datasets are English-only, requiring extension to other languages and dialects and validation across regional clinical settings. Rubric-based evaluation relies on GPT-5.4, and larger-scale physician evaluation remains necessary despite supporting physician preference judgments. The collected-audio SFT stream includes reconstructed Bridge2AI-Voice waveforms, whose acoustic fidelity has not been clinically validated. MedSonor is for research use only and must not be used for clinical diagnosis without rigorous prospective validation and appropriate clinical oversight.

Ethics and Safety Considerations

Source datasets are distributed through public repositories or controlled-access portals and remain subject to their respective licenses and access conditions. Use requires qualified healthcare oversight, with professionals retaining responsibility for clinical decisions. Deployment requires audio quality and completeness checks, systematic output review for medical accuracy, unsupported claims, and harmful advice, plus ongoing monitoring and escalation of uncertain or high-risk cases to professionals.

AI Use Statement

We used LLMs to assist with language editing, data synthesis and code implementation. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. URL https://arxiv.org/abs/2505.08775.

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. *arXiv preprint arXiv:2308.14346*, 2023. URL https://arxiv.org/abs/2308.14346.

Sebastien Baur, Zaid Nabulsi, Wei-Hung Weng, Jake Garrison, Louis Blankemeier, Sam Fishman, Christina Chen, Sujay Kakarmath, Minyoi Maimbolwa, Nsala Sanjase, Brian Shuma, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Shruthi Prabhakara, Monde Muyoyeta, and Diego Ardila. HeAR – health acoustic representations. *arXiv preprint arXiv:2403.02522*, 2024. URL https://arxiv.org/abs/2403.02522.

Carlo V. Bellieni, Renata Sisto, Duccio M. Cordelli, and Giuseppe Buonocore. Cry features reflect pain intensity in term newborns: An alarm threshold. *Pediatric Research*, 55:142–146, 2004. doi: 10.1203/01.PDR.0000099793.99608.CB. URL https://www.nature.com/articles/pr200421.

Yael Bensoussan, Satrajit Ghosh, Anais Rameau, Micah Boyer, Ruth Bahr, Stephanie Watts, Frank Rudzicz, Don Bolser, Jordan Lerner-Ellis, Shaheen Awan, Maria Powell, Jean-Christophe Belisle-Pipon, Vardit Ravitsky, Alistair Johnson, Alexandros Sigaras, Olivier Elemento, David Dorr, and Philip Payne. Developing multi-disorder voice protocols: A team science approach involving clinical expertise, bioethics, standards, and DEI. In *Interspeech 2024*, pp. 1445–1449, 2024. doi: 10.21437/Interspeech.2024-1926. URL https://www.isca-archive.org/interspeech_2024/bensoussan24_interspeech.html.

Yael Bensoussan, Alexandros Sigaras, Anais Rameau, Olivier Elemento, Maria Powell, David Dorr, Philip Payne, Vardit Ravitsky, Jean-Christophe Bélisle-Pipon, Ruth Bahr, et al. Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information. *PhysioNet*, 2025. doi: 10.13026/k81f-qr68. URL https://physionet.org/content/b2ai-voice/3.0.0/. Version 3.0.0.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017. doi: 10.1109/ICSDA.2017.8384449. URL https://doi.org/10.1109/ICSDA.2017.8384449.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4): 377–390, 2014. doi: 10.1109/taffc.2014.2336244. URL https://doi.org/10.1109/taffc.2014.2336244.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599, 2025a. doi: 10.18653/v1/2025.naacl-long.182. URL https://aclanthology.org/2025.naacl-long.182/.

Hongzhao Chen, Xiaoyang Wang, Jing Lan, Hexiao Ding, Yufeng Jiang, Minghui Yang, Danhui Xu, Jun Luo, Nga-Chun Ng, Gerald W. Y. Cheng, Yunlin Mao, and Jung Sun Yoo. Mmedfd: A real-world healthcare benchmark for multi-turn full-duplex automatic speech recognition. *arXiv preprint arXiv:2509.19817*, 2025b. URL https://arxiv.org/abs/2509.19817.

Sirry Chen, Jieyi Wang, Wei Chen, and Zhongyu Wei. SpeechMedAssist: Efficiently and effectively adapting speech language models for medical consultation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30914–30935, 2026. doi: 10.18653/v1/2026.acl-long.1428. URL https://aclanthology.org/2026.acl-long.1428/.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024. URL https://arxiv.org/abs/2410.17196.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023. URL https://arxiv.org/abs/2311.16079.

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. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Tu-Quyen Dao, Eike Schneiders, Jennifer Williams, John Robert Bautista, Tina Seabrooke, Ganesh Vigneswaran, Rishik Kolpekwar, Ritwik Vashistha, and Arya Farahi. TAME Pain: Trustworthy AssessMEnt of pain from speech and audio for the empowerment of patients. *PhysioNet*, 2025. doi: 10.13026/20e2-1g10. URL https://physionet.org/content/tame-pain/1.0.0/. Version 1.0.0.

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

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, Jinjie Gu, and Junwei Liu. MedDialogRubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*, 2026. URL https://arxiv.org/abs/2601.03023.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3123–3128, 2014. URL https://aclanthology.org/L14-1421/.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025. URL https://arxiv.org/abs/2507.17746.

Fu-Shun Hsu, Shang-Ran Huang, Chien-Wen Huang, Chao-Jung Huang, Yuan-Ren Cheng, Chun-Chieh Chen, Jack Hsiao, Chung-Wei Chen, Li-Chin Chen, Yen-Chun Lai, Bi-Fang Hsu, Nian-Jhen Lin, Wan-Ling Tsai, Yi-Lin Wu, Tzu-Ling Tseng, Ching-Ting Tseng, Yi-Tsun Chen, and Feipei Lai. Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—HF_Lung_V1. *PLOS ONE*, 16(7):e0254134, 2021. doi: 10.1371/journal.pone.0254134. URL https://doi.org/10.1371/journal.pone.0254134.

Songbo Hu, Yinhong Liu, Ej Zhou, Evgeniia Razumovskaia, Xiaobin Wang, Alexander Fraser, Ivan Vulić, and Anna Korhonen. Dial HEALTHDIAL for advice: A multilingual and multi-parallel spoken dialogue dataset for knowledge-grounded information seeking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25529–25552, 2026. doi: 10.18653/v1/2026.findings-acl.1275. URL https://aclanthology.org/2026.findings-acl.1275/.

Hagen Jaeger, Dhaval Trivedi, and Michael Stadtschnitzer. Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls, 2019. URL https://zenodo.org/records/2867216.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020. URL https://arxiv.org/abs/2009.13081.

Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P. Bigham. SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6798–6802, 2021. doi: 10.1109/icassp39728.2021.9413520. URL https://doi.org/10.1109/icassp39728.2021.9413520.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025. URL https://arxiv.org/abs/2501.15368.

Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, et al. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181–2213, 2016. doi: 10.1088/0967-3334/37/12/2181. URL https://doi.org/10.1088/0967-3334/37/12/2181.

Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391. URL https://doi.org/10.1371/journal.pone.0196391.

Monica Munnangi and Saiph Savage. ThReadMed-QA: A multi-turn medical dialogue benchmark from real patient questions. *arXiv preprint arXiv:2603.11281*, 2026. URL https://arxiv.org/abs/2603.11281.

Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alipio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, Andoni Elola, Ali Bahrami Rad, Reza Sameni, Gari D. Clifford, and Miguel T. Coimbra. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2524–2535, 2022. doi: 10.1109/JBHI.2021.3137048. URL https://doi.org/10.1109/JBHI.2021.3137048.

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

Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, 2021. doi: 10.1038/s41597-021-00937-4. URL https://doi.org/10.1038/s41597-021-00937-4.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390. URL https://doi.org/10.1145/2733373.2806390.

Chandan K. A. Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. A scalable noisy speech dataset and online subjective test framework. *arXiv preprint arXiv:1909.08050*, 2019. URL https://arxiv.org/abs/1909.08050.

Bruno M. Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P. Kahya, Nikša Jakovljevic, Tatjana L. Turukalo, Ioannis M. Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, Pantelis Natsiavas, Ana Oliveira, Cristina Jácome, Alda Marques, Nicos Maglaveras, Rui Pedro Paiva, Ioanna Chouvarda, and Paulo de Carvalho. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3):035001, 2019. doi: 10.1088/1361-6579/ab03ea. URL https://doi.org/10.1088/1361-6579/ab03ea.

Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4):523–541, 2012. doi: 10.1007/s10579-011-9145-0. URL https://doi.org/10.1007/s10579-011-9145-0.

Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy. Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *Interspeech 2020*, pp. 4811–4815, 2020. doi: 10.21437/interspeech.2020-2768. URL https://doi.org/10.21437/interspeech.2020-2768.

Ying Shen, Huiyu Yang, and Lin Lin. Automatic Depression Detection: an Emotional Audio-Textual Corpus and A GRU/BiLSTM-Based Model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6247–6251, 2022. doi: 10.1109/icassp43922.2022.9746569. URL https://doi.org/10.1109/icassp43922.2022.9746569.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *arXiv preprint arXiv:2307.14334*, 2023. URL https://arxiv.org/abs/2307.14334.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*, 2024. URL https://arxiv.org/abs/2401.05654.

Laura Verde and Giovanna Sannino. VOICED database. *PhysioNet*, 2018. doi: 10.13026/C25Q2N. URL https://physionet.org/content/voiced/1.0.0/. Version 1.0.0.

David Vilares and Carlos Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 960–966, 2019. doi: 10.18653/v1/P19-1092. URL https://aclanthology.org/P19-1092/.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025. doi: 10.18653/v1/2025.naacl-long.218. URL https://aclanthology.org/2025.naacl-long.218/.

Yishan Wang, Tsai-Ning Wang, Mathias Funk, and Aaqib Saeed. StethoLM: Audio language model for cardiopulmonary analysis across clinical tasks. *arXiv preprint arXiv:2603.00355*, 2026. URL https://arxiv.org/abs/2603.00355.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025. URL https://arxiv.org/abs/2507.16632.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. URL https://arxiv.org/abs/2503.20215.

Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. Livemedbench: A contamination-free medical benchmark for llms with automated rubric evaluation. *arXiv preprint arXiv:2602.10367*, 2026. URL https://arxiv.org/abs/2602.10367.

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. URL https://arxiv.org/abs/2408.01800.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. ACI-BENCH: A novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, 2023. doi: 10.1038/s41597-023-02487-3. URL https://doi.org/10.1038/s41597-023-02487-3.

Fangxu Yu, Tao Feng, Dehai Min, Zinan Lin, Weijia Xu, Michael Xu, Philip S. Yu, Ge Liu, and Tianyi Zhou. Reinforcement learning with evolving rubrics as rewards for audio reasoning. *arXiv preprint arXiv:2608.02831*, 2026. URL https://arxiv.org/abs/2608.02831.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

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

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. HuatuoGPT, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023. URL https://arxiv.org/abs/2305.15075.

Linhao Zhang, Jian Zhang, Bokai Lei, Chuhan Wu, Aiwei Liu, Wei Jia, and Xiao Zhou. WildSpeech-Bench: Benchmarking end-to-end SpeechLLMs in the wild, 2025. URL https://arxiv.org/abs/2506.21875.

Qing Zhang, Jing Zhang, Jiajun Yuan, Huajie Huang, Yuhang Zhang, Baoqin Zhang, Gaomei Lv, Shuzhu Lin, Na Wang, Xin Liu, Mingyu Tang, Yahua Wang, Hui Ma, Lu Liu, Shuhua Yuan, Hongyuan Zhou, Jian Zhao, Yongfu Li, Yong Yin, Liebin Zhao, Guoxing Wang, and Yong Lian. SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database. *IEEE Transactions on Biomedical Circuits and Systems*, 16(5):867–881, 2022. doi: 10.1109/tbcas.2022.3204910. URL https://doi.org/10.1109/tbcas.2022.3204910.

Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. *arXiv preprint arXiv:2406.16148*, 2024. URL https://arxiv.org/abs/2406.16148.

A Data Synthesis Details

A.1 Department–Scenario Design

The audio-conditioned medical dialogue dataset is synthesized from a controlled department–scenario design. We define 25 medical departments and 15 clinical interaction scenarios, yielding 375 department–scenario combinations. This design allows the generated dialogues to cover diverse specialties, user roles, clinical environments, and urgency levels.

Table 7: Medical departments used for synthesis.

Group	Departments
Internal medicine	Neurology; Cardiology; Respiratory Medicine; Gastroenterology; Endocrinology; Nephrology; Hematology; Rheumatology and Immunology; Infectious Disease; Medical Oncology.
Surgery	General Surgery; Orthopedics; Neurosurgery; Cardiothoracic Surgery; Urology.
Women and children	Gynecology; Obstetrics; Pediatrics.
Specialty care	Ophthalmology; ENT; Dermatology; Psychiatry; Rehabilitation Medicine.
Acute and critical care	Emergency Medicine; ICU.

Table 8: Controlled synthesis space.

Item	Value
Medical departments	25
Interaction scenarios	15
Department–scenario pairs	375
Topics per pair	50
Full topic pool	18,750
Sampled topics for dialogue synthesis	5,000
Samples per pair (approx.)	13–14

B Data Statistics and Filtering

Filtering retains examples with baseline reward strictly between 0 and 0.9.

C Reward and Severity Details

When severity labels are explicitly available, we use them directly. Otherwise, severity is inferred from item weights and safety-sensitive medical keywords. Negative items with weight magnitude $|w_j| \geq 8$ or containing high-risk concepts such as contraindication, unsafe dosage, stroke, sepsis, anaphylaxis, airway compromise, respiratory failure, bleeding, overdose, urgent escalation, or emergency care are treated as critical. Otherwise, negative items with weight magnitude $|w_j| \geq 5$ are treated as major. The remaining negative items are treated as minor.

Table 9: Clinical interaction scenarios used for dialogue synthesis.

ID	User role	Scenario
1	Doctor	Doctor and patient conversing in a consultation room, where the doctor seeks medical AI assistance midway.
2	Patient	Patient at home in a bedroom, directly having a voice consultation dialogue with medical AI.
3	Nurse	Nurse at an emergency triage desk receiving patients and consulting a medical AI system.
4	Elderly patient	Elderly person at a community health center communicating with medical AI through a device.
5	Paramedic	Ambulance doctor in a moving vehicle urgently seeking medical AI support.
6	Pharmacist	Pharmacist at a hospital pharmacy window consulting medical AI when encountering medication-related questions.
7	Researcher	Researcher in a university office having a medical professional dialogue with medical AI.
8	Family member	Patient family member outside an ICU ward confiding in and consulting medical AI.
9	Resident doctor	Resident trainee doing ward rounds in an inpatient department and asking medical AI via tablet.
10	Surgeon	Surgeon in a sterile operating room interacting with medical AI through voice.
11	Rehabilitation therapist	Rehabilitation therapist in a physical therapy room activating medical AI for guidance assistance.
12	Pregnant woman	Pregnant woman at home in a living room communicating with medical AI via smartphone.
13	Laboratory doctor	Laboratory doctor in a hospital lab seeking medical AI opinion when facing test reports.
14	Dietitian	Dietitian in a hospital meal preparation room seeking medical AI advice.
15	Outpatient visitor	Patient at a hospital outpatient lobby talking with medical AI through a guidance kiosk.

Table 10: Dialogue rubrics before filtering.

Statistic	Value
Dialogue turns with rubrics	8,000
Unique conversations	2,000
Mean number of positive items	6.32
Range of positive items	5–10
Mean number of negative items	3.30
Range of negative items	2–5

Table 11: Baseline-based filtering statistics.

Statistic	Value
Input samples	8,000
Retained samples	7,127
Removed high-reward samples	556
Removed non-positive-reward samples	317
Samples with missing scores	0

Length and truncation penalties. For response length L in policy-tokenizer tokens, the length penalty increases linearly from zero at f to $p_{\max}$ at h:

$$P_{\text{len}}(y) = p_{\max} \min\left(1, \max\left(0, \frac{L-f}{h-f}\right)\right), \qquad P_{\text{trunc}}(y) = p_{\text{EOS}}\, \mathbf{1}\{y \text{ is nonempty and lacks EOS}\}. \tag{4}$$

Medical dialogue RL and audio-integrated dialogue RL use $p_{\max} = p_{\text{EOS}} = 0.15$ and a 256-token completion budget. The free/hard thresholds are 1.10/1.50 times the reference length, rounded up and adjusted to satisfy $1 \le f < h \le 256$; without a reference, they are 120/200 tokens. EOS denotes <|im_end|> and is excluded from length counting.

D Training Configuration

Training setup. We initialize from Qwen2.5-Omni-7B and sequentially perform medical SFT, medical dialogue RL and audio-integrated dialogue RL. Table 12 summarizes the data sizes and hyperparameters.

Table 12: Training data sizes and key hyperparameters for SFT and the two RL stages.

Setting	SFT	Medical dialogue RL	Audio-integrated dialogue RL
Training examples	18,657	6,927	4,546
Learning rate	8×10^{-6}	5×10^{-6}	5×10^{-7}
Per-device batch size	8	4	4
Completions per prompt	–	10	10

All stages use fused AdamW, BF16, gradient checkpointing, and seed 42. SFT uses a cosine learning-rate schedule, while both RL stages use constant learning rates after warmup. RL uses DAPO with asymmetric clipping thresholds of 0.20/0.28 and KL regularization against each stage's initial policy. SFT freezes the visual encoder and trains the language model and aligner; the visual branch receives no image or video inputs during RL.

D.1 Local judge training and validation

Both judges fine-tune Qwen3-4B-Instruct for one epoch using LoRA (rank 32, $\alpha = 64$, dropout 0.05) and GPT-5.4 binary labels for individual response–criterion pairs.

To include responses spanning a range of quality levels, the dialogue data contain responses from GPT-4.1, GPT-4o-mini, and GPT-4.1-mini, with disjoint training and validation conversations. Audio training combines current-policy responses with dialogue replay; audio splits are disjoint by source-audio family, and replay uses only dialogue training data. Audio evidence is supplied to the judges as reference text.

Table 13: Local judge data sizes and agreement with held-out GPT-5.4 labels. Counts are response–criterion pairs; metrics are percentages.

Setting / metric	Dialogue judge	Audio judge
Training examples	112,514	22,872
Validation examples	2,900	3,771
Parse success	100.00	100.00
Overall accuracy	94.76	89.90
Positive-item accuracy	92.37	92.31
Negative-item accuracy	99.39	85.77

Baselines include seven general-purpose open-weight models: Qwen2.5-Omni-7B Xu et al. (2025), MiniCPM-o-2.6 Yao et al. (2024), Baichuan-Omni-1.5 Li et al. (2025), GLM-4-Voice-9B Zeng et al. (2024), Step-Audio-2-mini Wu et al. (2025), LLaMA-3.1-8B-Omni Fang et al. (2024), and SpeechGPT-2.0-preview Open-Moss (2025); closed-source references are Gemini-2.5-Flash, Gemini-3.7-Flash, and GPT-Audio (2025-08-28).

ACI-Bench adaptation. The original ACI-Bench task generates visit notes from encounter transcripts Yim et al. (2023). We instead extract patient–doctor exchanges and hold out a doctor utterance as the reference response. The preceding patient utterance is converted to speech, and earlier turns are supplied as text. The target is the original doctor utterance from the transcript, not the visit note. Consequently, our rubric scores measure dialogue-response quality under this adaptation and are not the original ACI-Bench note-generation scores.

Evaluation inputs. Table 14 summarizes the inputs available to each policy. Compared models receive matched task content in their respective input formats. ASR tasks additionally include a transcription instruction. MedVoice-Dial provides user speech with its original TTS source text and reference-assistant history.

E Audio Data

Table 15 gives the three-source composition of the final SFT training manifest; the 6,000 audio-integrated examples comprise two 3,000-example sources. The 1,400 validation examples are counted separately.

Table 14: Policy-visible inputs within each evaluation setting. Compared models receive the same task content, encoded using their respective model/API formats. Current-target references and scoring rubrics remain judge-only.

Benchmark	Audio	Current-turn transcript	Supplied history
LiveMedBench Clean-TTS	Spoken query	None	None
MMedFD-Lite	Spoken query	None	None
MedVoice-AID	Composed audio request	None	None
MedVoice-Dial	Current and preceding user speech	Original TTS source text	User audio and source text; reference-assistant text
ACI-Bench	Final patient turn	None	Text dialogue
HealthDial	Current user turn	None	Text dialogue
ThReadMed-QA	Current user turn	None	Text dialogue
LibriSpeech	Speech audio	None	None
AISHELL-1	Speech audio	None	None
VoiceBench	Spoken query	None	None
OpenHermes-Audio	Spoken query	None	None
WildSpeech-Bench	Spoken query	None	None

Table 15: Composition of the final multimodal medical SFT training manifest following selection and assembly. Counts are training examples.

Supervision source	**Training examples**
Collected medical audio	7,569
Synthesized spoken dialogue	5,088
Audio-integrated dialogue	6,000
Total	18,657

Fine-grained composition and source references. Table 16 partitions the 7,569 selected training examples in the collected medical-audio stream by source domain. Table 17 lists the corresponding source references and selected counts.

Table 16: Source-domain breakdown of the collected medical-audio SFT stream in the final training manifest.

Source domain	Training examples
Lung and respiratory sounds	2,697
Pediatric respiratory sounds	386
Heart sounds	365
Cough and respiratory-health audio	590
Environmental health-related sounds	34
Voice and motor-speech disorders	414
Stuttering	199
Clinical voice features reconstructed as audio	349
Emotional speech	339
Depression-related speech	100
Pain-related vocalizations	167
Infant cries	88
Medical speech and transcription	738
Spoken medical intent	163
Medical QA converted to speech	940
Total	7,569

Dataset references. Published descriptions accompany the respiratory sources HF Lung V1 Hsu et al. (2021), ICBHI Rocha et al. (2019), SPRSound Zhang et al. (2022), Coswara Sharma et al. (2020), and COUGHVID Orlandic et al. (2021); the cardiac sources PhysioNet/CinC 2016 Liu et al. (2016) and CirCor DigiScope Oliveira et al. (2022); and the speech and voice sources TORGO Rudzicz et al. (2012), MDVR-KCL Jaeger et al. (2019), VOICED Verde & Sannino (2018), and SEP-28k Lea et al.

Table 17: Audio-data provenance and references. Each linked dataset name points to its source repository or dataset record. Counts are selected SFT training examples; they do not represent the full size of the source datasets.

Source (linked reference)	Count	Source (linked reference)	Count
HF Lung V1	2,575	CREMA-D	200
ICBHI 2017	122	RAVDESS	139
SPRSound	386	DAIC-WOZ	27
PhysioNet/CinC 2016	202	EATD-Corpus	73
CirCor DigiScope	163	TAME Pain	167
Coswara	146	Donate-a-Cry	88
COUGHVID	444	United-Syn-Med	476
ESC-50	34	Medical ASR Recording	99
TORGO	172	MedDialog-Audio	79
VOC-ALS	90	Medical Audio Dataset	3
MDVR-KCL	7	YouTube Medical Audio	81
Speech Disorders Voice	110	Medical Intent Audio	163
VOICED	35	MedQA	715
SEP-28k	199	HEAD-QA	151
Bridge2AI-Voice (reconstructed)	349	MedBullets	74

(2021). Additional source descriptions cover ESC-50 Piczak (2015), CREMA-D Cao et al. (2014), RAVDESS Livingstone & Russo (2018), the DAIC interview corpus underlying DAIC-WOZ Gratch et al. (2014), EATD-Corpus Shen et al. (2022), MedQA Jin et al. (2020), HEAD-QA Vilares & Gómez-Rodríguez (2019), and MedBullets Chen et al. (2025a). The links in Table 17 identify the dataset records and community releases used in this study.

Reconstructed Bridge2AI-Voice audio. The 349 Bridge2AI-Voice training examples use waveforms reconstructed from the spectrogram features in the v3.0.0 release Bensoussan et al. (2025). These are derived audio, not the original clinical recordings, which are not included in that PhysioNet release. The preprocessing pipeline uses Griffin–Lim reconstruction and audio post-processing. Reconstruction may alter clinically relevant acoustic details, so these examples should not be interpreted as evidence of diagnostic performance on original Bridge2AI-Voice recordings.

F Human Evaluation Details

Cohort and assignment. The physician evaluation compares MedSonor with Qwen2.5-Omni-7B, MiniCPM-o-2.6, Baichuan-Omni-1.5, GLM-4-Voice-9B, Step-Audio-2-mini, and LLaMA-3.1-8B-Omni. The 30 cases cover respiratory medicine, pediatrics, cardiology, neurology, and otorhinolaryngology. Each of the six physicians completed 120 comparisons, comprising 18 comparisons shared by all reviewers and 102 additional assignments unique to that reviewer. This yields 630 unique comparisons and 720 judgments. Twelve judgments were unjudgeable; the remaining 708 judgments cover 621 comparisons with at least one valid vote. The interface omitted model identifiers and randomized question order and response position. Each physician spent approximately 1.5 hours completing the evaluation and received RMB 1,000 in compensation.

Preference aggregation and uncertainty. We use the final submitted judgments and exclude unjudgeable votes. For each comparison, a win, tie, and loss receive credits of 1, 0.5, and 0, respectively. Direct pairwise preference credits average these values across comparisons after averaging valid votes within each comparison. Valid reviewer credits are averaged before fitting the Bradley–Terry model, so a shared comparison does not outweigh a comparison assigned to one reviewer. Scores use the same logistic fit and scale as the automated Arena analysis, with Qwen2.5-Omni-7B anchored at 1000. We resample source-case clusters 2,000 times to obtain percentile intervals, retaining all comparisons from a sampled case together.

Shared-item agreement. Responses are aligned by model identity rather than displayed A/B position; ties form a separate category and unjudgeable votes are treated as missing. The 18 shared comparisons contain 105 valid judgments. There are 152 agreeing pairs among 255 comparable reviewer pairs, giving 59.6% agreement (95% item-bootstrap interval: 46.0–73.8%). Four comparisons receive the

same choice from all six physicians. Agreement is 52.9% between the two attending physicians, 59.8% among resident physicians, and 60.3% across the two groups.

G Prompt Templates and Evaluation Protocols

The prompt templates in this section make the data construction and evaluation protocols reproducible. Double braces denote fields substituted from the locked manifest. The code blocks show the core templates used in the reported runs; minor dataset-specific metadata is passed through the same JSON fields.

Table 18: Prompt inventory for data construction, training, and evaluation.

ID	Use	Output contract
P1	Topic generation	One concise clinical topic per department–scenario pair
P2	Spoken dialogue synthesis	JSON alternating user and assistant turns
P3	Audio-integrated case synthesis	JSON case, audio script, reference, and signed rubric
P4	Policy input	Multimodal messages with audio markers and transcript fields
P5	Rubric generation and repair	Instance-specific signed rubric JSON
P6	Local item judge	Binary labels for every positive and negative item
P7	Medical benchmark scoring	Criterion-level JSON or official G-Eval JSON
P8	General speech scoring	ASR text, IFEval score, or GPT-5.5 scalar score

G.1 Data synthesis prompts

Topic and spoken-dialogue synthesis. Topic generation receives one controlled department–scenario pair. Dialogue synthesis then receives the selected topic and the same scenario metadata. Both stages require natural spoken English, role-consistent language, and structured JSON output.

```
[system]
You are a senior clinical data curator and medical dialogue writer.
Return exactly one JSON object and no markdown. Keep all user-facing
utterances natural, clinically grounded, and suitable for text-to-speech.
[input]
department: {{department}}
scenario: {{scenario}}
user_role: {{user_role}}
topic: {{topic}}
[requirements]
Generate an alternating user--assistant conversation with four user turns
and four assistant turns. Match the vocabulary and authority to the role,
make the turns progress logically, avoid bullet points and annotation language,
and do not reveal hidden labels, filenames, or grading criteria.
[output]
{"dialogues":[{"role":"user","content":"..."},
{"role":"assistant","content":"..."}]}
```

Audio-integrated case synthesis. The audio case builder receives the source-audio analysis and must separate observable acoustic evidence from hidden source metadata. It produces the spoken request, reference response, audio construction instructions, and instance-specific rubric in one JSON object.

```
[system]
You are a senior clinical data curator creating single-turn medical-audio data.
Return exactly one JSON object and no markdown. Ground every statement in the
supplied audio evidence; never infer a diagnosis from a recording alone.
[input]
source_modality: {{source_modality}}
source_analysis: {{source_analysis}}
department: {{department}}
scenario: {{scenario}}
user_role: {{user_role}}
[requirements]
Create one natural current user decision, an audio construction script, a
reference assistant response, and a signed rubric. State recording/device
context when relevant, preserve the real source recording, use proportionate
safety guidance, and keep unsupported diagnoses and metadata out of speech.
```

G.2 TRAINING INPUT AND REWARD PROMPTS

The canonical policy system message is shared by the medical dialogue SFT and RL inputs. Transcript availability in evaluation depends on the benchmark (Table 14). The current target reference response, rubric, and reward are not visible to the policy during rollout; preceding reference-assistant turns form the supplied dialogue history.

```
[system]
You are a professional medical voice AI assistant. Your role is to:
- Provide accurate, empathetic medical information and support.
- Help users understand symptoms, treatments, and medical conditions.
- Give medication, post-operative, chronic-care, and emergency guidance.
- Explain medical information and test results clearly.
Always remain professional, supportive, and patient-centered.
[optional single-turn audio suffix]
Answer the current question first, use only supported audible evidence, match
the user's role, and give the next step and uncertainty when relevant.
[user]
<audio>
{{current_user_transcript}}
[assistant history]
{{causal_dialogue_history}}
```

Instance-specific rubric generation. The rubric writer sees the full preceding conversation, the final request, and the reference answer. It returns 8–12 atomic binary criteria, with positive items in the majority and signed weights reflecting clinical importance.

```
[system]
You are a senior clinical rubric writer designing reward criteria for
medical-AI reinforcement learning. Create a self-contained rubric for a new
candidate response to the final user request.
[requirements]
Use the complete dialogue history. Reward specific, clinically grounded,
useful content; detect hallucinations, unsafe advice, false reassurance,
contradictions, and dangerous overclaims. Do not require reference wording.
Every item must be independently judgeable with a binary answer. Use positive
weights +1..+10 and negative weights -1..-10; reserve magnitude 8..10 for
clinically consequential decisions.
[output]
{"rubric_version":"instance-signed-v1",
 "clinical_focus":"...",
 "historical_constraints":["..."],
 "positive_items":[{"id":"P1","scope":"current_turn",
   "criterion":"Does the response ...?","weight":8,"rationale":"..."}],
 "negative_items":[{"id":"N1","scope":"dialogue_consistency",
   "criterion":"Does the response incorrectly ...?","weight":-8,
   "rationale":"..."}]}
```

Rubric review and repair. A second reviewer prompt removes redundant or logically inverted items before the rubric enters the training manifest.

```
[system]
You are a strict reviewer of a signed medical RL rubric. Repair the candidate
rubric while preserving its schema and clinical intent. Merge redundant items,
remove negative items that only invert a positive item, keep omission penalties
only when the omission creates immediate safety risk, preserve supported
historical constraints, keep 8--12 total items with positive items in the
majority, and renumber IDs consecutively. Return only repaired JSON.
[user]
{{candidate_rubric}}
```

Local item-level judge. The local Qwen3-4B-Instruct LoRA judge receives one candidate response and one rubric criterion per query. It returns `{"answer": 0}` or `{"answer": 1}`; a label of 1 indicates that the response satisfies a positive criterion or exhibits the prohibited behavior described by a negative criterion. The following compact template preserves this single-item interface.

```
[system]
Judge exactly one criterion using the causal dialogue and judge-only evidence.
Return only {"answer": 0} or {"answer": 1}. Answer 1 when the criterion is true
of the current candidate. Negative items require affirmative prohibited claims
or actions, not quotation, denial, warnings, or marking a claim unsupported.
```

```
Treat Gold Audio Evidence as authoritative; do not infer findings from an
<audio> placeholder or from the candidate response.
[user]
## Complete Dialogue Before Candidate Response
{{history}}
## Gold Audio Evidence (Judge-Only Reference)
{{audio_evidence_summary}}
## Text Constraints (Judge-Only Reference)
{{text_constraints}}
## Historical Constraints
{{historical_constraints}}
## Candidate Assistant Response
{{candidate}}
## Single Rubric Criterion
Scope: {{scope}}
Criterion: {{criterion}}
Return ONLY {"answer": 0} or {"answer": 1}.
```

G.3 MEDICAL BENCHMARK JUDGE PROMPTS

The external medical scorers are separate from the local RL judges. LiveMedBench Clean-TTS uses one criterion at a time, MMedFD-Lite uses the released batch G-Eval format, and ACI-Bench, HealthDial, ThReadMed-QA, MedVoice-AID, and MedVoice-Dial use signed instance-rubric scoring with GPT-5.4 at low reasoning effort.

```
[system]
You are a strict medical AI evaluator. Judge one candidate response against
one supplied criterion using only the patient narrative and current request.
For a negatively weighted criterion, mark it true only when the undesirable
behavior is present. Return JSON only.
[user]
Patient narrative: {{narrative}}
Patient request: {{request}}
Candidate response: {{candidate}}
Rubric criterion: {{criterion}}
{"explanation":"brief evidence-based reason",
 "criteria_met":true}
```

```
[MMedFD-Lite G-Eval]
Given a user query and a model reply, evaluate the response quality. Identify
repetition, off-topic content, role confusion, contradictions, and unsafe advice.
Return one JSON object containing the item id, an integer score from 1 to 10,
a short rationale, and the binary failure flags.
```

For signed instance-rubric evaluation, the judge also receives the preceding dialogue, constraints established in earlier turns, all positive and negative rubric items, and a reference response available only to the judge. Before computing the signed score, the scorer verifies that each rubric item has been assessed exactly once. Arena evaluation uses a separate blinded pairwise prompt and reports Bradley–Terry scores, with model identities and training stages hidden from the judge.

```
[Arena pairwise judge]
[system]
You are an independent blinded evaluator of medical assistant responses. Use
the causal history and current user audio. Prefer medically accurate, safe,
grounded, directly useful, and clear answers. Penalize invented patient facts,
unsupported audio findings, dangerous advice, and irrelevant content. Do not
favor any model or reward verbosity. Return JSON only.
[user]
Causal text history: {{history}}
Current user audio: {{audio}}
Candidate A: {{candidate_A}}
Candidate B: {{candidate_B}}
{"winner":"A|B|tie|unjudgeable", "both_bad":false,
 "decisive_dimension":"...", "reason":"...",
 "audio_understood":true}
```

G.4 GENERAL SPEECH EVALUATION PROMPTS

All models are evaluated on the same fixed examples using greedy decoding. The prompt shown to the policy differs by benchmark, while the text query, reference answer, or checklist used by a judge is kept outside the policy input.

```
[LibriSpeech ASR]
system: You are a helpful assistant.
user: <audio>
      Transcribe the spoken content exactly. Output only the transcription.
[AISHELL-1 ASR]
system: You are a careful audio understanding assistant.
user: <audio>
      Please accurately transcribe the Chinese speech in the audio and output only the
      ↪  transcription.
[VoiceBench-IFEval and VoiceBench-CommonEval]
system: You are a helpful assistant.
user: <audio>
```

```
[VoiceBench-CommonEval judge]
system: You are a helpful assistant who tries to help answer the user's question.
user:
I need your help to evaluate the performance of several models in the speech
interaction scenario. The models receive a speech input from the user and
need to understand and respond to it. Rate the response using the user-input
transcription [Instruction] and the model-output transcription [Response].
Please evaluate the response on a scale of 1 to 5:
1 point: largely irrelevant, incorrect, or fails to address the query.
2 points: somewhat relevant but inaccurate or incomplete.
3 points: relevant and mostly accurate but not concise.
4 points: relevant, accurate, and concise.
5 points: exceptionally relevant, accurate, and to the point.
### [Instruction]: {{query}}
### [Response]: {{candidate}}
Output the score only without explanations.
[OpenHermes-Audio judge]
[Reference Answer]
{{reference}}
[Model Answer]
{{candidate}}
[Question]
{{question}}
Rate accuracy and relevance from 0 to 5. A refusal or completely irrelevant
answer receives 0; 3 means generally aligned but incomplete; 5 means highly
accurate, relevant, and detailed. Return:
Explanation: ...
Rating: <integer>
[WildSpeech-Bench judge]
User query: {{query}}
AI response: {{candidate}}
Checklist: {{checklist}}
Rate spoken-response quality from 1 to 10, considering instruction adherence,
correctness, relevance, repetition, and the characteristics of voice dialogue.
Return:
Strengths: ...
Weaknesses: ...
Score: <integer>
```

GPT-5.5 evaluates CommonEval, OpenHermes-Audio, and WildSpeech-Bench with three independent calls per response. The requests use a 1,024-token completion limit, top-$p = 0.95$, and temperatures 0.5, 0.7, and 0.5 in the same order as those three benchmarks. ASR uses WER/CER after the fixed normalization rules; paired intervals use 20,000 bootstrap replicates.

Prompt provenance. The reported tables use the prompts above and the corresponding frozen script versions.

H Complete Ablation Results

Table 19 provides the complete numerical results for Figure 5.

Table 19: Ablation of multimodal medical supervised fine-tuning and its training sources. SFT omits both RL stages. RL-only starts directly from Qwen2.5-Omni-7B, while the source-removal variants omit the named SFT source before applying the same medical dialogue RL followed by audio-integrated dialogue RL. The full pipeline applies SFT followed by these two RL stages. All values are GPT-5.4 signed instance-rubric scores under the same evaluation protocol.

Training pipeline	**MedVoice-AID**	**MedVoice-Dial**	**ACI-Bench**	**HealthDial**	**ThReadMed-QA**
Qwen2.5-Omni-7B	0.3341	0.3237	0.2002	0.4489	0.1800
SFT	0.7666	0.4257	0.2168	0.4106	0.2034
RL-only	0.3033	0.4198	0.1297	0.3489	0.1641
w/o collected medical audio	0.7552	0.4743	0.1732	0.4665	0.2705
w/o synthesized spoken dialogue	0.7505	0.4587	0.2181	0.4570	0.2096
w/o audio-integrated dialogue	0.5957	0.4725	0.1836	0.4749	0.3058
MedSonor	**0.7900**	**0.5225**	**0.2263**	**0.5022**	**0.3403**