Respiratory Audio Question Answering for Health Assessment Under Real-World Mobile Sensing Heterogeneity

Gaia A. Bertolino[1], Yuwei Zhang[1], Tong Xia[2], Domenico Talia[3] and Cecilia Mascolo[1]

[1]University of Cambridge, UK; [2]Tsinghua University, China; [3]University of Calabria, Italy

Introduction

Conversational multimodal AI is increasingly explored for health assessment, enabling systems that combine physiological signals and natural-language queries to generate clinically relevant answers. This is particularly promising in respiratory health, where non-invasive audio signals such as cough, breathing, and speech can support screening and diagnosis. However, respiratory audio is acquired through diverse sensing devices and recording settings, from smartphones and wearable microphones to digital stethoscopes, creating substantial variability in signal characteristics and quality. Moreover, the same recording may need to support different clinical queries, including wheeze or crackle detection, disease prediction, symptom severity assessment, and estimation of continuous physiological scores. Existing biomedical audio-language QA systems are not specifically designed or evaluated for respiratory heterogeneity in realistic sensing settings, leaving their robustness under real-world mobile sensing conditions unclear.

Methods

We present *RAMoEA-QA*, a hierarchical two-stage specialization model for respiratory-audio question answering designed to handle heterogeneous sensing devices, acquisition regimes, and output targets while preserving the efficiency of single-path inference. The model jointly specializes audio encoding and language adaptation: an Audio Mixture-of-Experts selects one encoder for each recording based on a shallow encoding of the audio and question, and a Language Mixture-of-Adapters selects one LoRA adapter on a shared frozen language model to match the query intent and answer format using the selected audio encoder's full embedding. By activating only one audio expert and one language adapter at test time, RAMoEA-QA avoids the cost of evaluating all experts for each sample, making conditional specialization more compatible with resource-constrained deployment. We evaluate the model on the *RA-QA* collection [1], a large-scale respiratory-audio QA benchmark spanning multiple recording devices, respiratory sound types, clinical attributes, and question formats.

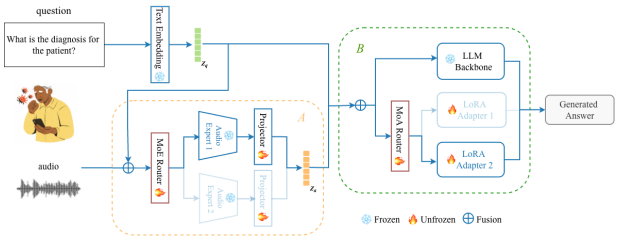

Figure 1: RAMoEA-QA architecture.

Results

General audio-language models transfer poorly to respiratory-audio QA, and monolithic models perform less robustly when faced with the device, modality, and acquisition variability typical of real-world respiratory sensing. Compared with the strongest baseline, *RAMoEA-QA* improves both classification and regression performance, reaching 0.72 accuracy and 2.29 MAE versus 0.67 and 2.61, respectively.

Under controlled modality, dataset, and task shifts, our model also shows greater robustness to changes in respiratory sound type, acquisition domain, and previously unseen tasks, achieving a 23% accuracy gain in the COPD modality-shift setting.

Routing analysis further shows that specialization reflects recording conditions: even within the same dataset, matched recordings from the same subject are often assigned to different experts (e.g., about 20% of matched deep- versus shallow-breath recordings in Coswara), indicating adaptation to within-dataset variation in signal characteristics. These results suggest that conditional routing improves adaptation to heterogeneous sensing regimes while retaining efficient single-path inference.

A fuller account of the benchmarks, results, and routing analyses is reported in the papers [1, 2].

References

[1] Gaia A. Bertolino, Yuwei Zhang, Tong Xia, Domenico Talia, and Cecilia Mascolo, "Ra-qa: A benchmarking system for respiratory audio question answering under real-world heterogeneity," 2026.

[2] Gaia A. Bertolino, Yuwei Zhang, Tong Xia, Domenico Talia, and Cecilia Mascolo, "Ramoea-qa: Hierarchical specialization for robust respiratory audio question answering," 2026.