RespLLM: Unifying Audio and Text with Multimodal LLMs for Generalized Respiratory Health Prediction

Yuwei Zhang[1], **Tong Xia**[1*], **Aaqib Saeed**[2], **Cecilia Mascolo**[1]

[1] **University of Cambridge, UK** [2] **Eindhoven University of Technology, The Netherlands**
*Corresponding author, {yz798, tx229}@cam.ac.uk

Abstract

The high incidence and mortality rates associated with respiratory diseases underscores the importance of early screening. Machine learning models can automate clinical consultations and auscultation, offering vital support in this area. However, the data involved, spanning demographics, medical history, symptoms, and respiratory audio, are heterogeneous and complex. Existing approaches are insufficient and lack generalizability, as they typically rely on limited training data, basic fusion techniques, and task-specific models. In this paper, we propose RespLLM, a novel multimodal large language model (LLM) framework that unifies text and audio representations for respiratory health prediction. RespLLM leverages the extensive prior knowledge of pretrained LLMs and enables effective audio-text fusion through cross-modal attentions. Instruction tuning is employed to integrate diverse data from multiple sources, ensuring generalizability and versatility of the model. Experiments on five real-world datasets demonstrate that RespLLM outperforms leading baselines by an average of 4.6% on trained tasks, 7.9% on unseen datasets, and facilitates zero-shot predictions for new tasks. Our work lays the foundation for multimodal models that can *perceive*, *listen to*, and *understand* heterogeneous data, paving the way for scalable respiratory health diagnosis.

1 Introduction

Respiratory diseases are the third leading cause of death worldwide, highlighting the critical need for early and accessible respiratory health screening (Labaki and Han, 2020). Clinical assessment of such diseases typically begins with gathering personal information (*consultation*), including demographics, medical history, symptoms, and other relevant details (hereafter collectively referred to as DMS). In addition, clinicians listen to respiratory sounds (*auscultation*) as a non-invasive method of screening, before proceeding to more invasive and costly examinations (Reyes et al., 2024). Consequently, automating both the consultation and auscultation processes using machine learning (ML), as illustrated in Figure 1, can significantly enhance early screening by increasing efficiency, accessibility, and affordability.

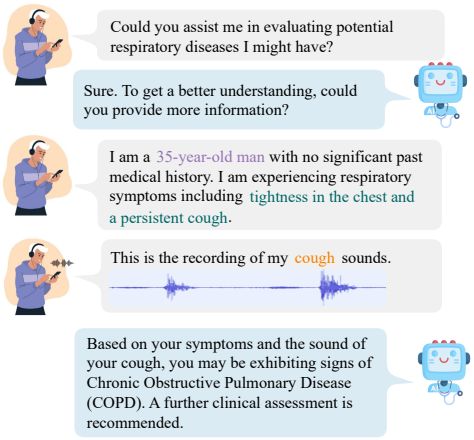


Figure 1: Automated consultation and auscultation for respiratory health screening.

Advancements In Medical Foundation Models (AIM-FM) Workshop @NeurIPS 2024.

Considering that the DMS and audio data are different modalities, presenting heterogeneous information, multimodal ML approaches that can effectively integrate them are needed. Early efforts have been made in this direction (Xia et al., 2023; Kim et al., 2024; Han et al., 2021); nevertheless, limitations hinder their application in real-world diagnostic scenarios. First, *these models are typically small-scale and trained on limited data*, restricting their ability to effectively learn from high-dimensional audio signals and unstructured DMS data. Second, *the fusion of DMS and audio remains inadequate, which may reduce model performance*. They commonly concatenate audio representations with DMS representations, encoded either into categorical vectors using a pre-defined mapping (Han et al., 2021) or into word embeddings (Kim et al., 2024). Such concatenation overlooks the differences in their embedding spaces and interrelationship between the two types of data.

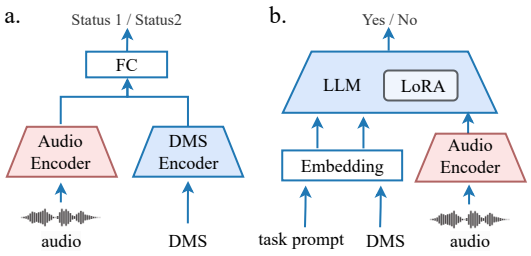


Figure 2: **Multimodal models for respiratory health prediction.** (a) Existing concatenation-based fusion method. (b) Our LLM-based fusion method.

More importantly, *existing models are task- and dataset-specific, which hinders their ability to generalize*. Traditional machine learning models rely on the IID (Independent and Identically Distributed) assumption, and when the data distribution shifts, their performance tends to degrade. However, respiratory health data for model training is often limited (Kim et al., 2024), and in real-world deployments, data ranging from DMS to audio, as well as the respiratory status included, can differ significantly from the training data. For example, a model trained to predict asthma may be required to predict COVID-19 status at the inference stage. Highly generalized models capable of handling these changes are necessary but currently lacking.

To overcome these limitations and progress towards the envisioned applications depicted in Figure 1, this paper puts forward a unique approach that harnesses the power of pre-trained LLMs to simultaneously interpret DMS and audio for respiratory health screening. The high-level concept of the proposed method is illustrated in Figure 2b. Unlike existing methods, which are constrained by limited data and model scale, our approach leverages LLMs extensively trained on large corpora, including medical materials (Goel et al., 2023), to extend model capacity beyond the available respiratory training data. For effective multimodal fusion, we generate sequences of audio representations from a pre-trained encoder and combine them with DMS text embeddings as a unified input to the LLM. This enables coherent integration of the two modalities through multi-layer and multi-head attention mechanisms. To enhance the model generalizability, we curate multiple data sources for training and create instructions applicable to a variety of tasks that combine DMS and audio. This approach equips the model with zero-shot inference capabilities for new datasets and unseen tasks.

Our contributions can be summarized as follows[1]:

1. To the best of our knowledge, this work presents, for the first time, the use of LLMs to jointly model DMS and audio data for respiratory health screening. The proposed multimodal LLM, RespLLM, can comprehensively *perceive*, *listen to*, *understand* heterogeneous inputs and then *diagnose* respiratory health.

2. We curate a large instruction-tuning set combining task prompts, DMS, and audio to optimize the proposed model. This approach ensures the model remains versatile (one model for multiple tasks) and generalized (performing well on new datasets or tasks).

3. We conduct extensive experiments on multiple open datasets. Results demonstrate the superiority of our model over existing methods, showing notable improvement in both trained and unseen tasks, along with the robustness of our approach in integrating different LLM models.

[1]Our code is available at `https://github.com/evelyn0414/RespLLM`

2 Related Work

2.1 ML for Respiratory Health

In clinical practice, respiratory health is assessed through various clinical examinations such as spirometry, auscultation, chest X-rays, plethysmography, and computed tomography scans (Reyes et al., 2024). Auscultation, combined with personal DMS information, is among the most comfortable and affordable approaches. Using an electronic stethoscope or a microphone, respiratory sounds, such as coughing and breathing, produced by airflow in the respiratory system can be easily recorded. These recordings contain valuable physiological information related to breathing difficulties, reduced oxygen saturation, and other conditions (Xia et al., 2022). Therefore, modeling respiratory audio and DMS data holds significant potential for ubiquitous respiratory health monitoring.

Traditionally, audio signal processing techniques were used to extract acoustic features that help distinguish between different respiratory conditions (Ma et al., 2022; Islam et al., 2018). Recently, deep learning (DL) has significantly advanced acoustic modeling by automatically capturing complex relationships from raw audio data or spectrograms. This advancement has led to high-performing applications, from detecting abnormal lung sounds to diagnosing conditions such as the flu and pulmonary diseases (Gairola et al., 2021; Fraiwan et al., 2022; Srivastava et al., 2021). When combined with additional information like DMS, DL-driven respiratory health prediction models demonstrate further performance improvements (Han et al., 2021; Xia et al., 2023; Kim et al., 2024; Moummad and Farrugia, 2023).

However, current methods to represent and fuse DMS and audio in the field of respiratory health remain simple and may fail to capture all the relevant information. DMS is typically encoded either by mapping variables into a uniform vector using a predefined dictionary (Figure 3a) (Han et al., 2021; Xia et al., 2023) or by extracting text embeddings from the unstructured data (Figure 3b) (Kim et al., 2024; Moummad and Farrugia, 2023). This representation is then concatenated with audio from a deep learning encoder, ignoring the differences and complex relationship between the two, limiting the potential of DL for health prediction. In the related field of chest X-ray modeling, more advanced multimodal techniques such as LSTM-based fusion (Hayat et al., 2022), cross-modal attention (Wang et al., 2018), and multimodal pre-training (Moon et al., 2022) have been explored. In this paper we explore how similar approaches could be beneficial to audio and DMS.

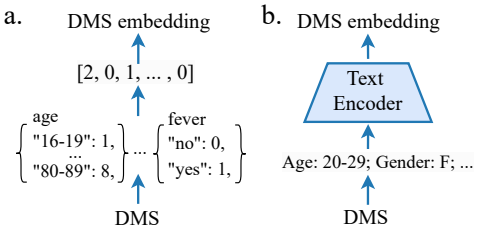


Figure 3: **Existing DMS encoding methods.** (a) Pre-defined mapping. (b) Text embedding.

2.2 LLMs for Health

Recently-emerged LLMs have demonstrated remarkable capabilities in various health diagnostic applications (Singhal et al., 2023; Liévin et al., 2024). This is primarily due to their pretraining on enormous and diverse datasets, including medical literature, clinical guidelines, research papers, and general knowledge (Goel et al., 2023). Such pretraining enables LLMs to understand medical terminology, concepts, and associations relevant to health diagnostics.

There is also a growing trend in extending LLMs, which are inherently language models, to handle multimodal data in a unified manner (Wu et al., 2023; Qiu et al., 2023). This capability is typically achieved by combining prompts, modality-specific encoders, and LLMs within a single framework (Moor et al., 2023; Yu et al., 2023; Liu et al., 2024). For example, Liu *et al.* Liu et al. (2024) leveraged LLMs to interpret electrocardiography signals and perform zero-shot diagnosis. To further enhance generalizability, instruction tuning has emerged as a promising approach for adapting LLMs to various tasks and domains (Aw et al., 2023). In this work, we make the first effort to leverage recent advancements in multimodal LLMs and curate an instruction-tuning dataset using diverse sources for generalized respiratory health prediction.

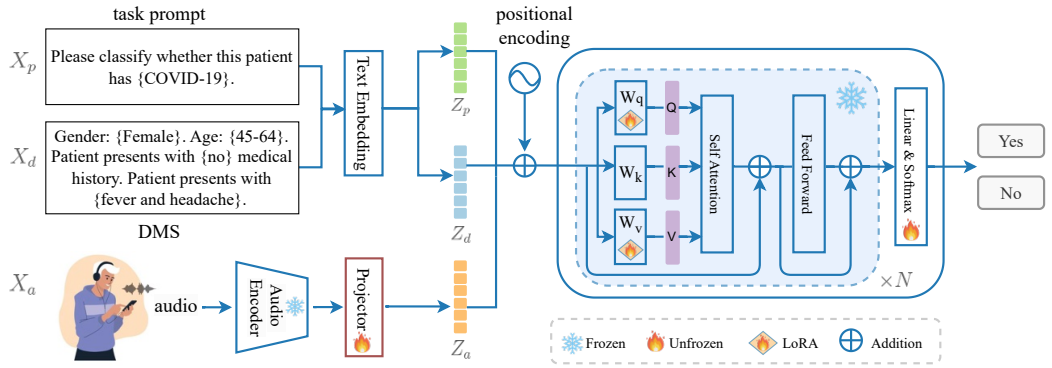


Figure 4: **The model architecture of RespLLM.** Text embeddings from task prompts and personal DMS, along with audio embeddings from respiratory sounds, are sequentialized as input for the LLM consisting of multiple transformer blocks.

3 Methodology

Figure 2b illustrates our proposed framework, a multimodal LLM that can model DMS and respiratory audio simultaneously. In this section, we begin by elaborating on the model architecture design. Then we delve into how we curate the instruction tuning dataset to train this model.

3.1 Model Architecture

Our overall model architecture is shown in Figure 4. Given the DMS X_d and the respiratory audio signal X_a, our goal is to provide a screening result/recommendation in response to the question in the prompt X_p. To achieve this, our model mainly consists of three modules: a text embedder that maps X_p and X_d into text token embeddings, an audio encoder with a projector to map X_a into audio embeddings, and an LLM to fuse all the given information for respiratory health screening. These modules are specified as follows.

Text embedding. The text embedding module will first split the given prompt X_p and DMS X_d into sequence of tokens using its tokenizer, and then map the words into a sequence of word embeddings, denoted by $Z_p \in \mathbb{R}^{L_p \times S}$ and $Z_d \in \mathbb{R}^{L_d \times S}$, where L_p and L_d are the lengths of the text and S is the dimension of the word embeddings. For consistency, we use the same tokenizer and word embeddings from the LLM that is used in the later stage. In this sense, S is also the dimension of the hidden state in the transformer blocks of the used LLM.

Audio Encoder with Projector. Given the high dimensionality and complexity of the audio data, we adapt a pre-trained audio encoder to obtain audio embeddings for X_a (Zhang et al., 2024). Each audio sample is first transformed into a spectrogram, which is then divided into small patches of equal size to derive embeddings. We feed the resulting sequence of L_a embeddings into the LLM, denoted by $z_a \in \mathbb{R}^{L_a \times A}$, where A is the dimension of the original audio embeddings. As the LLM has a different hidden embedding space of dimension S, we need to efficiently align the audio embeddings with word embeddings. Following insights from previous work (Ma et al., 2024), we use a simplistic linear layer as the projector $\mathcal{P}(\cdot)$. Then, we have the final audio embeddings $Z_a = \mathcal{P}(z_a)$, where $Z_a \in \mathbb{R}^{L_a \times S}$.

LLM and LoRA. For the three distinct embedding Z_p, Z_d, and Z_a, which correspond to task prompt, DMS, and audio information respectively, we first combine them into a longer sequence of embeddings. After this, we add positional embeddings to the resulting sequence, producing the final embedding $Z \in \mathbb{R}^{L \times S}$, $L = L_p + L_d + L_a$. Note that we use the same positional embedding approach as that employed by the chosen LLM model. This combined embedding Z is then fed into the LLM for further processing.

Since the LLM consists of multiple transformer blocks as shown by the blue shaded box in Figure 4, each containing several self-attention operations parameterized by W_q, W_k and W_v, the three types

of information are deeply fused. The final transformer block outputs a sequence of hidden states with a length of L, which are then flattened across the temporal dimension to generate a single vector representation. This vector is then passed through a linear layer with a *Softmax* function to produce the final output, yielding binary health predictions. To mitigate the risk of hallucinations in the original LLM output, we replace the original linear layer with a randomly initialized one containing two output nodes, representing the answer, either 'Yes' or 'No', to the question in the task prompt.

To balance between preserving the LLM's prior knowledge from large-scale pretraining and adapting it to respiratory health prediction tasks, we choose to update only part of the pretrained parameters. LoRA (Low-Rank Adaptation) (Hu et al., 2021) is a parameter-efficient fine-tuning method that reduces the computational cost of updating large models. As shown in Figure 4, we apply LoRA to the value (W_v) and query (W_q) mapping modules in the transformer blocks of the LLM, while keeping the rest of the parameters frozen.

3.2 Model Training

Data Curation. To increase the generality of our method, we propose to combine multiple data resources for training. Those data can differ in the audio modalities, DMS formats and the category of respiratory conditions. To unify them for model training, we design contextualized instructions containing task prompts, the description of DMS and the corresponding audio information. The templates of X_p and X_d are formulated as described below, with examples provided in Figure 5.

I. The task prompt X_p is a diagnostic query with respect to the condition that can be predicted from the given audio and DMS. It is formulated as:

> "***Dataset description****: This data comes from the {D}.* ***Task description****: classify whether the participant has {C} given the following information and audio of the person's {T} sounds. Please output 1 for {C1}, and 0 for {C2}. "*

Here, D distinguishes the data resource, T presents the sound type, and C denotes the condition to be predicted from C1 and C2 restricts the output space.

II. For the text input of DMS X_d, we use the following template:

> "***Gender****: {G}.* ***Age****: {A}. Patient presents with {M}* ***medical history*** *conditions. Patient presents with the following* ***respiratory symptoms****: {S}. Recorded location: {L}. "*

Here, G denotes the gender, A represents age, M specifies medical history, and S is the list of symptoms. L represents the location where the audio was recorded for lung sounds. For any missing or non-applicable data field, the corresponding description is omitted.

Instruction Tuning. Since various data resources have been unified into instructions, we can now shuffle these instructions from multiple sources to create batches for model training. To make the most of the pre-trained knowledge in the audio encoder and the LLM, we will only train the projector, the LoRA parameters, and the final fully connected layer for the LLM in our model, as shown in Figure 4. For the objective function, we use the cross-entropy loss, comparing the output of the LLM with the actual answer to the diagnostic question in the prompt.

Zero-shot Prediction. As mentioned earlier, since the diagnostic task and personal DMS are formulated in text, our model can easily extend to new data and unseen respiratory conditions. This allows for zero-shot inference without requiring any parameter changes when deploying to a new domain.

4 Experiments

In this section, we conduct extensive experiments with real-world data to answer the following questions:

- **RQ1:** How does our model perform compared to the state-of-the-art baselines for respiratory health prediction?
- **RQ2:** How well does our model generalize to new data and unseen tasks?
- **RQ3:** How do the model design and the choice of LLMs impact the performance of our method?

Table 1: **Summary of *source* and *target* datasets and tasks used in this study.** The five datasets are UK COVID, COVID-19 Sounds, ICBHI, Coswara, and KAUH. For task IDs, S1−S7 refer to the source tasks, and T1T6 refer to the target tasks. In audio types, 's' is short for shallow, 'h' for heavy, and 'd' for deep.

Data	ID	Label	Audio Type	#Train/Test
1	S1	Covid	Exhalation	1500/1000
1	S2	Covid	Cough	1500/1000
2	S3	Covid	Breath	1162/324
2	S4	Covid	Cough	1162/324
2	S5	Smoker	Breath	2570/1419
2	S6	Smoker	Cough	2570/1419
3	S7	COPD	Lung sounds	462/366
4	T1	Covid	Cough-s	-/100
4	T2	Covid	Cough-h	-/100
4	T3	Covid	Breath-s	-/40
4	T4	Covid	Breath-d	-/40
5	T5	COPD	Lung sounds	-/38
5	T6	Asthma	Lung sounds	-/116

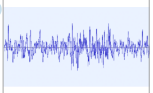
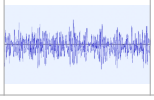
Task	Text	Audio	Answer
S1 (Training)	**Task prompt**: Dataset description: This data comes from the UK COVID-19 Vocal Audio Dataset. Task description: classify whether the participant has COVID-19 given the following information and audio of the person's exhalation sounds. Please output 1 for COVID19, and 0 for non-COVID19. **DMS text**: Gender: Female. Age: 45-64. Patient presents with the following medical history conditions: asthma. Patient presents with the following respiratory symptoms: cough, fatigue, headache.		1
S6 (Training)	**Task prompt**: Dataset description: This data comes from the COVID-19 Sounds dataset. Task description: classify whether the person is a smoker or not given the following information and audio of the person's cough sounds. Please output 1 for smoker, and 0 for non-smoker. **DMS text**: Gender: Female. Age: 50-59. Patient presents with no medical history conditions. Patient presents with no obvious respiratory symptoms.		0
S7 (Training)	**Task prompt**: Dataset description: This data comes from the ICBHI Respiratory Sound Database Dataset. Task description: classify whether the person has Chronic obstructive pulmonary disease (COPD) given the following information and audio of the person's lung sounds. Please output 1 for COPD, and 0 for healthy. **DMS text**: Gender: M. Age: 65. Record location: right posterior chest.		1
T4 (Testing)	**Task prompt**: This data comes from the Coswara Covid-19 dataset. Task description: classify whether the participant has COVID-19 given the following information and audio of the person's breathing-deep sounds. Please output 1 for COVID19, and 0 for non-COVID19. **DMS text**: Gender: male. Age: 35. Patient presents with the following respiratory symptoms: cold.		0
T6 (Testing)	**Task prompt**: Dataset description: This data comes from the KAUH lung sound dataset, containing lung sounds recorded from the chest wall using an electronic stethoscope. Task description: classify whether the person has asthma given the following information and audio of the person's lung sounds. Please output 1 for asthma, and 0 for healthy. **DMS text**: Gender: F. Record location: posterior right upper.		1

Figure 5: **Examples of instructions used in our work.** The variables that differ across samples and datasets are highlighted. For any missing data in a field, the corresponding description is omitted.

4.1 Datasets and Tasks

We use five open respiratory audio datasets for our experiments, featuring recordings of coughing, breathing, and lung sounds related to respiratory health statuses like smoking, COVID-19, and other respiratory diseases. These datasets also contain rich DMS information including age, gender, medical histories, symptoms, and recording locations for lung sounds. Using these datasets, we define 13 respiratory health tasks, as shown in Table 1. Among these, only the source tasks are used for model training, while the others are reserved for testing. Examples of the instructions we generated by combining task prompts, DMS, and audio recordings are illustrated in Figure 5.

Table 2: **Performance when training and testing on the same data sources.** Baselines are task-specific (trained and tested on each single task), while RespLLM is trained collectively with all tasks. Best results are bold and the second best are underlined.

	Task	**S1 → S1**	**S2→S2**	**S3→S3**	**S4→S4**	**S5→S5**	**S6→S6**	**S7→S7**	***Avg.***
Single-modal	Audio	0.6025	0.6729	0.5828	0.6260	0.5517	0.6247	0.9575	0.6597
	DMS - Hard	0.7626	0.7521	0.6427	0.6427	0.5485	0.5485	0.8341	0.6759
	DMS - Soft	0.9126	0.8900	0.7406	0.7406	0.5594	0.5594	0.9938	0.7709
Multimodal	Fusion - Hard	0.5936	0.6905	0.6171	0.6747	0.5714	0.6250	0.9845	0.6795
	Fusion - Soft	0.8668	0.8954	0.6997	0.7390	0.5692	**0.6336**	0.9981	0.7717
	RespLLM (Ours)	**0.9244**	**0.9002**	**0.7958**	**0.7840**	**0.6189**	0.6274	**1.0000**	**0.8072**

Table 3: **Performance of zero-shot prediction on new datasets.** For baselines, Sx presents the models used for testing, i.e., S2&4→T1, S2&4→T2, S1&3→T3, S1&3→T4, S7→T5. The average performance is reported when multiple models can be transferred. For RespLLM, Sx refers to our trained model with all source task data.

	Task	**Sx→T1**	**Sx→T2**	**Sx→T3**	**Sx→T4**	**Sx→T5**	***Avg.***	**Sx→T6**
Single-modal	Audio	0.6076	0.4940	0.5963	0.4875	0.5823	0.5535	-
	DMS - Hard	0.4956	0.4956	0.6312	0.6312	0.5375	0.5582	-
	DMS - Soft	0.5834	0.5834	0.5525	0.5525	0.5312	0.5606	-
Multimodal	Fusion - Hard	0.5528	0.5276	0.6288	0.5550	0.5708	0.5670	-
	Fusion - Soft	0.6190	0.5928	0.6288	0.6400	0.5542	0.6070	-
	RespLLM (Ours)	**0.6424**	**0.6284**	**0.6525**	**0.6750**	**0.6750**	**0.6547**	0.5865

4.2 Experimental Setup

For comparison, we implement both single-modal and multimodal baselines. Regarding single-modal methods, we compare to ***Audio***, which fine-tunes the pre-trained audio encoder alongside a linear classifier for respiratory condition prediction (Xia et al., 2021). For DMS-only methods, we consider to use the hard encoding in Figure 3a and soft text embedding in Figure 3b to fit a linear model, namely ***DMS-hard*** and ***DMS-soft***, respectively. Based on these two methods for DMS, we compare to the multimodal method as illustrated in Figure 2a, and name them ***Fusion-hard*** (Han et al., 2021) and ***Fusion-soft*** (Kim et al., 2024), as our multimodal baselines.

The audio encoder used in both the baselines and our method is the pre-trained OPERA-CT model (Zhang et al., 2024), a hierarchical token-semantic audio transformer. It processes an 8-second audio input (padded or cropped) into a spectrogram of size 256×64 and output embeddings of 64 patches, each with a dimension of 768. The LLM model that we modify is OpenBioLLM-8B[2] which is an open-source LLM designed for the biomedical domain. The instruction tuning is completed on a single A-100 GPU. For all tasks, we use AUROC as the metric to report the health condition prediction performance.

4.3 Results

Health Prediction Performance (RQ1). To answer RQ1, we first examine the performance of our model and the baselines when testing on training datasets (held-out testing set). Since the baselines are task-specific by design, they are trained and tested on the same task, whereas our model utilizes all data resources, resulting in a single RespLLM capable of performing well on multiple tasks. The results are summarized in Table 2. Among the seven evaluated tasks, our model outperforms the state-of-the-art baselines on six tasks, with the average AUROC across all seven tasks surpassing the best baseline by 4.6% (0.8072 vs. 0.7717). It can also be observed that the fusion baselines compared cannot consistently outperform their single-modal counterparts, and their average AUROCs are very close. This suggests that the fusion methods are insufficient. In contrast, our model demonstrates superiority by effectively fusing DMS and audio information via the LLM for respiratory health prediction.

[2] https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B

Generalizability (RQ2). To demonstrate our model's generalizability and address RQ2, we evaluate its performance not only on in-distribution data but also on new, unseen datasets and tasks. Specifically, we train the models on source task data and test them on target tasks. As shown in Table 1 and Figure 5, both the types of sounds and the information from DMS vary between source and target tasks. Our model can be directly tested, while for the baselines, we report cross-task transfer performance. Since no fine-tuning is applied, this constitutes zero-shot prediction, with the results summarized in Table 3.

Zero-shot transfer prediction shows a degraded performance compared to Table 2, due to changes in data sources, audio types, and DMS information. Despite this challenge, our model consistently outperforms all compared baselines, with the average AUROC surpassing the best baseline by 7.9% (0.6547 vs. 0.6070). This demonstrates the stronger generalizability of our method over the baselines. Notably, in T6, where asthma is a new class not included in the training data, none of the baselines can predict this condition (e.g., a model trained to distinguish COVID/non-COVID in S1 cannot differentiate asthma from healthy cases). In contrast, our model achieves an AUROC of 0.5865, comparable to the baselines' average performance on T1-5. This capability largely stems from our instruction-tuning approach, which effectively retrieves relevant knowledge from the pretrained LLM for zero-shot generalization.

Effect of Training and Model Design (RQ3). To further validate the superiority of our framework with cross-data training, we perform several ablation studies. We combine S1-7 into a multi-label task and use all data to train the multimodal baselines for direct comparison of different fusion methods: concatenation fusion as used in the baseline, add-on fusion from (Blandfort et al., 2019), and cross-attention fusion from (Wang et al., 2022). The results for normal testing on source tasks and zero-shot prediction on target tasks are shown in Table 4 and Table 5. Concatenation outperforms addition, as the audio and text embeddings are in very different spaces, and simply adding them may confuse the model. Concatenation also outperforms cross-attention fusion, likely because attention introduces additional parameters to train, which increases the data demand. Our model outperforms all these ablations due to the use of more complex architectures with pretrained parameters and knowledge.

Table 4: Performance of different fusion methods in our framework when testing on source datasets.

Task	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**Avg.**
Fusion - Soft	0.9065	0.8927	0.7436	0.7396	0.5884	0.5833	0.9543	0.7726
Fusion - Add	0.7525	0.7828	0.7289	0.7223	0.5653	0.5930	0.6941	0.6913
Fusion - CrossAttn	0.8131	0.8369	0.7870	0.7805	0.5754	0.5872	0.7942	0.7392
RespLLM (Ours)	**0.9244**	**0.9002**	**0.7958**	**0.7840**	**0.6189**	**0.6274**	**1.0000**	**0.8072**

Table 5: Performance of different fusion methods in our framework when zero-shot testing on test datasets.

Task	**T1**	**T2**	**T3**	**T4**	**T5**	**Avg.**	**T6**
Fusion - Soft	0.6284	0.6504	0.6550	0.6375	0.6458	0.6434	-
Fusion - Add	0.6552	0.6396	0.5725	0.5350	0.5500	0.5905	-
Fusion - CrossAttn	0.7272	0.7112	0.5500	0.6125	0.5292	0.6260	-
RespLLM (Ours)	0.6424	0.6284	0.6525	0.6750	0.6750	**0.6547**	0.5865

We also compare different open-source LLMs within our framework, with their performance summarized in Table 6 and Table 7. The four LLMs show similar AUROCs across tasks, demonstrating the robustness of our training approach. Notably, OpenBioLLM achieves a higher AUROC in the zero-shot setting on the target tasks, likely due to its specialized pre-training on medical corpora, enhancing its diagnostic knowledge for generalized health screening.

Table 6: Performance of different LLMs in our framework when testing on source datasets.

Task	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**Avg.**
Gemma2 (2B)	0.9221	0.8927	0.7555	0.7202	0.5840	0.5709	0.9953	0.7772
Phi-3.5(4B)	0.9250	0.8989	0.7909	0.7886	0.5964	0.6050	1.0000	0.8007
Mistral (7B)	0.9236	0.9006	0.7889	0.7765	0.6040	0.6096	1.0000	0.8005
LLaMA (7B)	0.9225	0.9055	0.7899	0.7934	0.5986	0.6010	1.0000	0.8016
LLaMA3 (8B)	0.9269	0.9061	0.8048	0.7988	0.6131	0.6171	1.0000	**0.8095**
OpenBioLLM	0.9244	0.9002	0.7958	0.7840	0.6189	0.6274	1.0000	0.8072

Table 7: Performance of different LLMs in our framework when zero-shot testing on test datasets.

Task	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**Avg.**
Gemma2 (2B)	0.6456	0.6256	0.6500	0.5850	0.6833	0.5514	0.6255
Phi (4B)	0.6232	0.6200	0.5975	0.6375	0.6583	0.5039	0.6097
Mistral (7B)	0.6264	0.6068	0.6425	0.6575	0.6958	0.5826	0.6368
LLaMA (7B)	0.6368	0.6340	0.6400	0.6050	0.7083	0.5565	0.6322
LLaMA3 (8B)	0.6388	0.6152	0.6425	0.6625	0.6750	0.5797	0.6372
OpenBioLLM (8B)	0.6424	0.6284	0.6525	0.6750	0.6750	0.5865	**0.6449**

5 Discussion

In this work, we introduced RespLLM, the first audio-text multimodal LLM for respiratory health prediction. The model not only outperforms state-of-the-art baselines in typical in-distribution testing but also demonstrates stronger generalizability in zero-shot predictions on new datasets and tasks that it was not exposed to during training.

We anticipate that the rise of multimodal LLMs will create exciting opportunities for modality fusion (via Transformers) and for grounding models in heterogeneous data sources (via instruction tuning). Thus, our work serves as a foundational step toward more generalist medical AI models.

Limitations. This work presents a proof-of-concept. As such, RespLLM is *not* intended for clinical use and should not be considered safe for such applications. The experiments conducted in this study are limited to respiratory conditions such as COVID-19, COPD, and asthma. We have not tested the model performance on other conditions, such as the flu, due to the limited data available at the moment. However, we hope that such data will become more available in the future, enabling further research.

Future Work To mitigate the hallucinations that frequently occur in large language models, we replaced the final linear layer in the original LLM with a custom linear layer that only outputs 'Yes' or 'No' for a given condition. An exciting direction for future work would be to explore the use of the full language model for more comprehensive diagnostics and reasoning in respiratory conditions while maintaining trustworthiness. Additionally, we plan to integrate more biosignal modalities, such as photoplethysmography signals and body temperature dynamics, which could provide a more holistic approach to respiratory health screening.

Acknowledgments and Disclosure of Funding

This work was supported by ERC Project 833296 (EAR), EPSRC Project RELOAD, and Nokia Bell Labs through a donation. Y. Z. is additionally supported by the Cambridge Trust Scholarship. A. S. is financially supported by the NGF AiNed Fellowship Grant.

References

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. In *First Conference on Language Modeling*.

Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C Chandrakiran, Sahiti Nori, KK Suhail, Sadhana Gonuguntla, et al. 2023. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. *Scientific Data* 10, 1 (2023), 397.

Philipp Blandfort, Tushar Karayil, Federico Raue, Jörn Hees, and Andreas Dengel. 2019. Fusion strategies for learning user embeddings with neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

Harry Coppock, George Nicholson, Ivan Kiskin, Vasiliki Koutra, Kieran Baker, Jobie Budd, Richard Payne, Emma Karoune, David Hurley, Alexander Titcomb, et al. 2024. Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers. *Nature Machine Intelligence* (2024), 1–14.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Mohammad Fraiwan, Luay Fraiwan, Mohanad Alkhodari, and Omnia Hassanin. 2022. Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. *Journal of Ambient Intelligence and Humanized Computing* (2022), 1–13.

Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. 2021. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief* 35 (2021), 106913.

Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. 2021. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 527–530.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*. PMLR, 82–100.

Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2021. Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8328–8332.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. 2022. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Machine Learning for Healthcare Conference*. PMLR, 479–503.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Md Ariful Islam, Irin Bandyopadhyaya, Parthasarathi Bhattacharyya, and Goutam Saha. 2018. Multichannel lung sound analysis for asthma detection. *Computer methods and programs in biomedicine* 159 (2018), 111–123.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

June-Woo Kim, Miika Toikkanen, Yera Choi, Seoung-Eun Moon, and Ho-Young Jung. 2024. BTS: Bridging Text and Sound Modalities for Metadata-Aided Respiratory Sound Classification. *arXiv preprint arXiv:2406.06786* (2024).

Wassim W Labaki and MeiLan K Han. 2020. Chronic respiratory diseases: a global view. *The Lancet Respiratory Medicine* 8, 6 (2020), 531–533.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns* 5, 3 (2024).

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024. Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement. In *Forty-first International Conference on Machine Learning*.

Zhizhong Ma, Yuanhang Qiu, Feng Hou, Ruili Wang, Joanna Ting Wai Chu, and Christopher Bullen. 2022. Determining the best acoustic features for smoker identification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8177–8181.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. *arXiv preprint arXiv:2402.08846* (2024).

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics* 26, 12 (2022), 6070–6080.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*. PMLR, 353–367.

Ilyass Moummad and Nicolas Farrugia. 2023. Pretraining respiratory sound representations using metadata and contrastive learning. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 1–5.

Jielin Qiu, Jiacheng Zhu, Shiqi Liu, William Han, Jingqi Zhang, Chaojing Duan, Michael A Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. 2023. Automated Cardiovascular Record Retrieval by Multimodal Learning between Electrocardiogram and Clinical Report. In *Machine Learning for Health (ML4H)*. PMLR, 480–497.

Francisco M Reyes, Piyush Modi, and John K Le. 2024. *Lung Exam* (updated 2024 may 1 ed.). StatPearls Publishing, Treasure Island (FL). https://www.ncbi.nlm.nih.gov/books/NBK459253/ Available from: https://www.ncbi.nlm.nih.gov/books/NBK459253/.

Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. 2019. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement* 40, 3 (2019), 035001.

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. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

Arpan Srivastava, Sonakshi Jain, Ryan Miranda, Shruti Patil, Sharnil Pandya, and Ketan Kotecha. 2021. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Computer Science* 7 (2021), e369.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9049–9058.

Yangtao Wang, Yanzhao Xie, Jiangfeng Zeng, Hanpin Wang, Lisheng Fan, and Yufan Song. 2022. Cross-modal fusion for multi-label image classification with attention mechanism. *Computers and Electrical Engineering* 101 (2022), 108002.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519* (2023).

Tong Xia, Jing Han, Abhirup Ghosh, and Cecilia Mascolo. 2023. Cross-device federated learning for mobile health diagnostics: A first study on COVID-19 detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

Tong Xia, Jing Han, and Cecilia Mascolo. 2022. Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine* 247, 22 (2022), 2053–2061.

Tong Xia, Dimitris Spathis, J Ch, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Erika Bondareva, Ting Dang, Andres Floto, Pietro Cicuta, et al. 2021. COVID-19 sounds: a large-scale audio dataset for digital respiratory screening. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

Han Yu, Peikun Guo, and Akane Sano. 2023. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. In *Machine Learning for Health (ML4H)*. PMLR, 650–663.

Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, Jagmohan Chauhan, and Cecilia Mascolo. 2024. Towards Open Respiratory Acoustic Foundation Models: Pretraining and Benchmarking. *arXiv preprint arXiv:2406.16148* (2024).

A Data description

COVID-19 Sounds (Xia et al., 2021) . The COVID-19 Sounds dataset consists of 53,449 audio samples (over 552 hours in total) crowd-sourced from 36,116 participants through the COVID-19 Sounds app. This dataset is comprehensive in terms of demographics and spectrum of health conditions. It also provides participants' self-reported COVID-19 testing status with 2,106 samples tested positive. It consists of three modalities including breathing, cough, and voice recordings. Only breathing and cough modalities are used in this paper.

UK COVID-19 (Coppock et al., 2024). The UK COVID-19 Vocal Audio Dataset is designed for the training and evaluation of machine learning models that classify SARS-CoV-2 infection status or associated respiratory symptoms using vocal audio. The UK Health Security Agency recruited voluntary participants through the national Test and Trace programme and the REACT-1 survey in England from March 2021 to March 2022, during dominant transmission of the Alpha and Delta SARS-CoV-2 variants and some Omicron variant sublineages. Audio recordings of volitional coughs, exhalations, and speech (speech not included in open access version, nor used in this paper) were collected in the 'Speak up to help beat coronavirus' digital survey alongside demographic, self-reported symptom and respiratory condition data, and linked to SARS-CoV-2 test results.

ICBHI (Rocha et al., 2019). The ICBHI Respiratory Sound Database contains audio samples, collected independently by two research teams in two different countries, over several years. Ethical approval was obtained from the ethics committees of the appropriate institutions.

Most of the database consists of audio samples recorded by the School of Health Sciences, University of Aveiro (ESSUA) research team at the Respiratory Research and Rehabilitation Laboratory (Lab3R), ESSUA and at Hospital Infante D. Pedro, Aveiro, Portugal. The second research team, from the Aristotle University of Thessaloniki (AUTH) and the University of Coimbra (UC), acquired respiratory sounds at the Papanikolaou General Hospital, Thessaloniki and at the General Hospital of Imathia (Health Unit of Naousa), Greece. The database consists of a total of 5.5 hours of recordings in 920 annotated audio samples from 126 subjects.

Coswara (Bhattacharya et al., 2023). The Coswara dataset contains respiratory sounds recorded between April 2020 and February 2022 from 2635 individuals (1819 SARS- CoV-2 negative, 674 positive, and 142 recovered subjects). The respiratory sounds contained nine sound categories associated with variants of breathing, cough and speech. The metadata contains demographic information associated with age, gender and geographic location, as well as the health information relating to the symptoms, pre-existing respiratory ailments, comorbidity and SaRS-CoV-2 test status.

KAUH (Fraiwan et al., 2021). The KAUH dataset includes sounds from seven ailments (i.e., asthma, heart failure, pneumonia, bronchitis, pleural effusion, lung fibrosis, and chronic obstructive pulmonary disease (COPD) as well as normal breathing sounds. The dataset contains the audio recordings from the examination of the chest wall at various vantage points using an electronic stethoscope. The

stethoscope placement on the subject was determined by the specialist physician performing the diagnosis. Each recording was replicated three times corresponding to various frequency filters that emphasize certain bodily sounds. The dataset can be used for the development of automated methods that detect pulmonary diseases from lung sounds or identify the correct type of lung sound.

B Implementation Details

B.1 RespLLM

Audio encoder. The audio encoder that we adopt is the pre-trained OPERA-CT model (Zhang et al., 2024). It is a hierarchical token-semantic audio transformer (HTS-AT) model trained with a contrastive learning objective of instance discrimination on respiratory sounds. All audio recordings are padded or cropped to 8 seconds, resampled to 16 kHz and merged into a mono channel. They are then transformed into spectrograms using 64 Mel filter banks with a 64 ms Hann window that shifts every 32 ms, resulting in a spectrogram of 126×64 dimension. It output patch embeddings of 64 patches, which is input into the LLM as 64 tokens after the alignment module.

LLM and LoRA. We use the OpenBioLLM model, which has 8B parameters and uses a LLaMA3 architecture. It was developed by Saama AI Lab and released in May 2024 and achieves state-of-the-art performance across various biomedical tasks. To efficiently adapt the LLM model to our tasks, we employ a LoRA module of rank $r = 16$ and $\alpha = 32$.

For the ablation study, we also explored LLaMA-7B (Touvron et al., 2023), LLaMA3-8B[3], Mistral (Jiang et al., 2023), Gemma-2(2B)[4] and Phi-3.5[5].

B.2 Baselines

We use the pre-traiend BERT (Devlin, 2018) for the wording embeddings in the soft fusion baselines, which are of the same dimension of the audio embeddings.

[3] https://ai.meta.com/blog/meta-llama-3/

[4] https://huggingface.co/google/gemma-2-2b

[5] https://huggingface.co/microsoft/Phi-3.5-mini-instruct