Adaptive Test-Time Scaling for Zero-Shot Respiratory Audio Classification

Tsai-Ning Wang T.N.WANG@TUE.NL
Eindhoven University of Technology, The Netherlands

Herman Teun den Dekker H.DENDEKKER@ERASMUSMC.NL
Erasmus University Medical Center, The Netherlands

Lin-Lin Chen L.CHEN@TUE.NL
Eindhoven University of Technology, The Netherlands

Neil Zeghidour NEIL@KYUTAI.ORG
Kyutai, France

Aaqib Saeed A.SAEED@TUE.NL
Eindhoven University of Technology, The Netherlands
Eindhoven Artificial Intelligence Systems Institute, The Netherlands

Abstract

Automated respiratory audio analysis promises scalable, non-invasive disease screening, yet progress is limited by scarce labeled data and costly expert annotation. Zero-shot inference eliminates task-specific supervision, but existing methods apply uniform computation to every input regardless of difficulty. We introduce TRIAGE, a tiered zero-shot framework that adaptively scales test-time compute by routing each audio sample through progressively richer reasoning stages: fast label-cosine scoring in a joint audio-text embedding space (Tier-L), structured matching with clinician-style descriptors (Tier-M), and retrieval-augmented large language model reasoning (Tier-H). A confidence-based router finalizes easy predictions early while allocating additional computation to ambiguous inputs, enabling nearly half of all samples to exit at the cheapest tier. Across nine respiratory classification tasks without task-specific training, TRIAGE achieves a mean AUROC of 0.744, outperforming prior zero-shot methods and matching or exceeding supervised baselines on multiple tasks. Our analysis shows that test-time scaling concentrates gains where they matter: uncertain cases see up to 19% relative improvement while confident predictions remain unchanged at minimal cost.

Data and Code Availability
We use only publicly available datasets; the dataset descriptions and corresponding citations are provided in the Experiments section. Our source code is provided as anonymized supplemental material during review and will be made publicly available on GitHub upon acceptance.

Institutional Review Board (IRB)
This research does not require IRB approval.

1. Introduction

Automated analysis of heart and lung sounds increasingly relies on audio–text embedding models that project recordings into a shared latent space, where classification reduces to similarity scoring against label prompts or lightweight classifiers (Zhang et al., 2024a,b). Yet current pipelines treat inference as a fixed-cost operation: a clean textbook wheeze and a faint crackle buried in motion artifact receive identical computation—one forward pass, one decision rule. This uniformity is mismatched to clinical reality. Auscultation recordings vary widely in signal quality, device characteristics, and diagnostic subtlety; some can be resolved from gross spectral features, while others require careful attention to timing within the respiratory cycle, the coexistence of multiple abnormalities, or acoustic cues that lie near the noise floor. When the same shallow inference is applied indiscriminately, difficult cases are under-served and overall robustness suffers.

Adapting models to each clinical environment through fine-tuning offers one remedy, but this path is often blocked: regulatory approval takes months, labeled data from the target site may be scarce or inac-

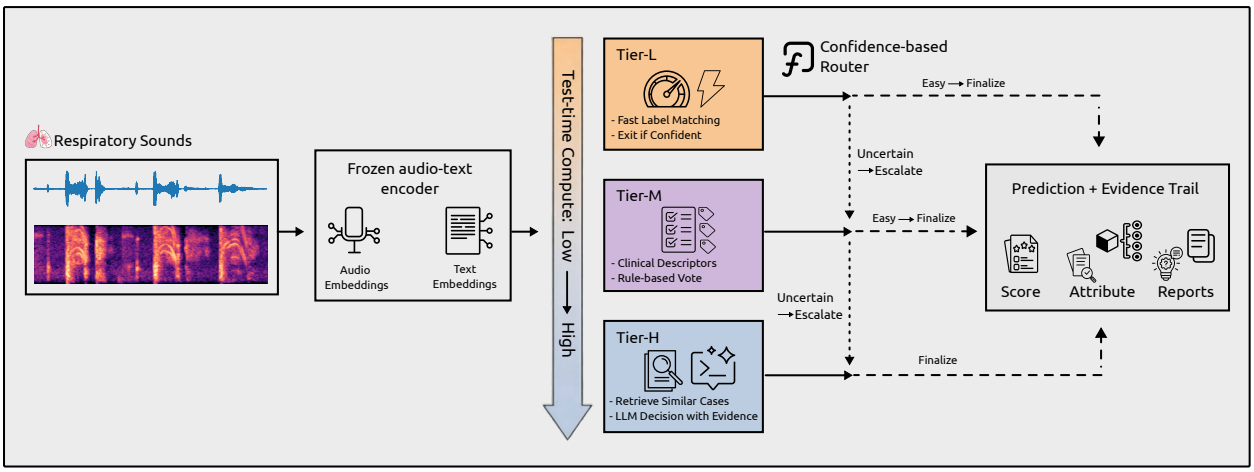

Figure 1: Overview of TRIAGE. A frozen audio–text encoder embeds recordings and medical text. **Tier-L** performs label cosine scoring with margin-based early exit; **Tier-M** matches clinical descriptors with rule voting; **Tier-H** retrieves nearest-neighbor reports and queries an off-the-shelf LLM, e.g., Gemini or GPT.

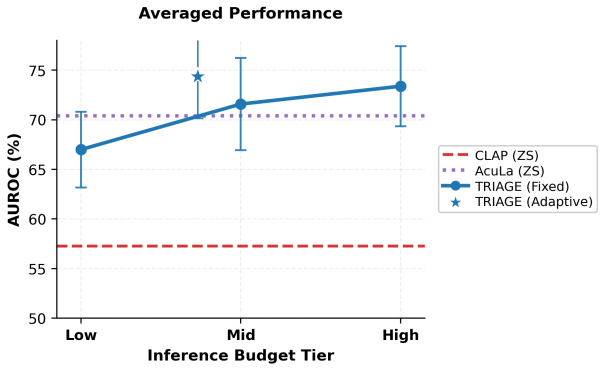

Figure 2: Tiered test-time inference improves zero-shot respiratory audio classification. Points show mean AUROC (%) averaged over 9 tasks versus budget tier. CLAP and AcuLa are zero-shot baselines (horizontal lines). TRIAGE (Fixed) evaluates three compute budgets by forcing all samples to run up to Tier-L/Tier-M/Tier-H (Low/Mid/High), while TRIAGE (Adaptive) reports the single operating point produced by confidence-based routing across tiers.

cessible, and computational resources are constrained. A more agile alternative is to improve inference itself—extracting more from a frozen model by spending additional computation selectively on recordings that need it.

This idea has gained traction in language and vision under the banner of *test-time scaling*: sampling multiple candidate outputs, applying self-verification, or routing inputs through progressively richer reasoning stages based on difficulty (Snell et al., 2024; Brown et al., 2024; Saad-Falcon et al., 2024). These strategies complement the familiar benefits of scaling models and data at training time (Hestness et al., 2017; Kaplan et al., 2020), yet they have been studied almost exclusively for autoregressive generators on open-domain benchmarks. For embedding-based systems—particularly in safety-critical medical applications—test-time compute remains underexplored as a resource that can be allocated differentially across inputs without any task-specific supervision.

We address this gap with **TRIAGE** (**T**iered **R**etrieval and **I**nference for **A**udio with **G**ated **E**scalation), a three-tier inference framework for zero-shot respiratory audio classification. TRIAGE leaves the underlying audio–text encoder frozen and instead modulates computation per recording through gated escalation. In the first tier, each recording is scored against natural-language label prompts via cosine similarity; cases with a confident margin are finalized immediately. Recordings that fall below this confidence threshold advance to a second tier, where the audio is matched against clinician-approved descriptor templates—structured queries probing sound timing, quality, and anatomical location. A rule-based aggregator distills these matches into an interpretable attribute profile. The most ambiguous cases proceed to a third tier: nearest-neighbor retrieval over an external corpus of audio–report pairs supplies supporting evidence, and a large language model synthesizes a final prediction grounded in retrieved context.

This architecture exposes an explicit compute–accuracy tradeoff. By tuning the confidence thresholds that govern tier transitions, practitioners can shift the operating point: resolve more cases cheaply when efficiency is paramount, or escalate aggressively when reliability matters most. Empirically, we evaluate TRIAGE on nine tasks spanning five public respiratory datasets. In a fully zero-shot setting—no fine-tuning, no labeled examples from the target data—TRIAGE improves over direct embedding similarity by a substantial margin on the majority of tasks, with the largest gains concentrated on recordings that require escalation. These results suggest that adaptive inference-time computation is a practical way for improving robustness in medical audio, independent of model or data scaling.

We make the following contributions:

- **Adaptive test-time scaling for medical audio embeddings.** We show that selectively increasing inference computation on uncertain recordings—without task-specific training—can substantially improve zero-shot classification, offering an alternative pathway to robustness when model adaptation is infeasible.
- **A compute-aware formulation of auscultation inference.** We cast respiratory audio classification as a staged decision process in which test-time computation is an explicit, tunable resource, enabling principled study of efficiency–accuracy tradeoffs across heterogeneous recording conditions.
- **An interpretable, clinically grounded inference pipeline.** TRIAGE decomposes classification into label scoring, descriptor-based attribute extraction, and retrieval-augmented reasoning. Each stage yields human-readable outputs—confidence scores, attribute profiles, retrieved evidence—providing transparency into why a recording was escalated and how the final decision was reached.

2. Related Work

Our work draws on three lines of research: domain-specific audio encoders for auscultation, audio–language models that enable zero-shot classification, and recent methods that treat inference-time computation as a tunable resource. We review each in turn, highlighting how TRIAGE builds on and departs from prior approaches.

2.1. Domain-Specific Audio Encoders for Auscultation

Early AI-aided auscultation used hand-crafted features or shallow CNNs trained on small, single-task datasets (e.g., murmur or wheeze detection, normal–abnormal screening). Such models were tightly coupled to their training distribution, and adapting to new devices, populations, or label spaces typically required retraining. Recent foundation-style encoders pretrained on large, heterogeneous respiratory audio improve transfer: OPERA aggregates multi-source cough and breath corpora and shows strong transfer to held-out tasks (Zhang et al., 2024a), while AcuLa aligns respiratory audio with a medical language model to inject clinical semantics and achieve state-of-the-art results across broad benchmarks (Wang et al., 2025b). In contrast, TRIAGE treats pretrained representations as fixed and improves performance through test-time inference, targeting recordings where a single forward pass and fixed decision rule (e.g., a linear head) are insufficient.

2.2. Audio–Language Models and Zero-Shot Classification

A complementary strategy learns joint audio–text embeddings for zero-shot transfer via natural-language prompts. CLIP popularized this paradigm in vision by training dual encoders on image–caption pairs and classifying by matching inputs to textual label descriptions (Radford et al., 2021). AudioCLIP and CLAP extend this idea to audio with contrastive audio–text alignment, enabling competitive zero-shot tagging and retrieval across diverse sound categories (Guzhov et al., 2021; Elizalde et al., 2022).

Beyond embedding models, larger audio–language systems combine acoustic front-ends with autoregressive decoders or instruction-following LLMs. In the clinical domain, RespLLM integrates respiratory audio with clinical text in a multimodal language model to improve robustness under dataset shift (Zhang et al., 2024b). CaReAQA couples a self-supervised auscultation encoder with an LLM for open-ended diagnostic QA over medical sounds and introduces a cardiorespiratory audio QA benchmark that also evaluates transfer to closed-ended classification (Wang et al., 2025a). These works show that shared audio–language representations can support flexible and interpretable zero-shot behavior. However, inference is typically

uniform across recordings: inputs receive the same procedure and compute budget, whether via embedding similarity or generation. TRIAGE retains zero-shot audio–text matching but adds a gated pipeline that allocates more computation to ambiguous cases by escalating them to progressively richer reasoning stages.

2.3. Test-Time Scaling and Adaptive Inference

Recent work treats inference-time computation as a tunable budget rather than a fixed cost, complementing classical scaling laws that relate performance to model size and training compute (Kaplan et al., 2020; Hestness et al., 2017). With the model held fixed, additional test-time effort can still improve outcomes. In language modeling, repeated sampling yields predictable accuracy gains as the number of candidate generations increases (Brown et al., 2024). Archon composes inference-time strategies (e.g., verification, critique, multi-agent debate) into pipelines that can outperform single-pass baselines under matched token budgets (Saad-Falcon et al., 2024), and recent theory begins to characterize when test-time scaling should succeed (Chen et al., 2025). For classification, TestNUC applies test-time consistency with nearest neighbors in embedding space (Zou et al., 2025). For dual encoders, decomposing inputs into fine-grained attribute comparisons can improve out-of-distribution retrieval without changing encoder weights (Xiao, 2024), and prompt engineering alone can unlock diverse behaviors from frozen encoders (Koukounas et al., 2024). TRIAGE brings these ideas to medical audio. We keep a pretrained auscultation encoder fixed and introduce a three-tier pipeline—label-name scoring, structured clinical-attribute matching, and retrieval-augmented LLM reasoning—that allocates more computation to ambiguous recordings. Our setting is embedding-based, safety-critical, and fully zero-shot: no gradient updates and no labeled examples from the target distribution.

3. Methodology

We present TRIAGE, a three-tier inference framework for zero-shot auscultation classification. We allocate test-time computation adaptively: easy recordings stop early, ambiguous ones escalate to richer reasoning. All tiers share a frozen audio–text encoder; no task-specific parameters are learned.

3.1. Problem Formulation

Let $\mathcal{X}$ denote auscultation recordings and $\mathcal{Y}_j$ the label set for task j. Given a test recording $x \in \mathcal{X}$, we predict $\hat{y} \in \mathcal{Y}_j$ in a fully zero-shot setting (no parameter updates and no labeled examples from the target task).

We assume access to three frozen resources:

1. a frozen audio–text embedding model with encoders f_{audio} and f_{text};
2. a clinician-defined descriptor system (groups + text templates) and task-specific rule tables;
3. a retrieval corpus $\mathcal{R}$ of audio–report pairs used only for Tier-H retrieval.

3.2. Audio–Text Embedding Model

The backbone of TRIAGE is a frozen dual-encoder that embeds audio and text into a shared d-dimensional space. The audio encoder

$$f_{\text{audio}} : \mathcal{X} \to \mathbb{R}^d, \quad x \mapsto \mathbf{a} = f_{\text{audio}}(x),$$

maps a recording to a unit-normalized embedding, and the text encoder

$$f_{\text{text}} : \mathcal{T} \to \mathbb{R}^d, \quad t \mapsto \mathbf{t} = f_{\text{text}}(t),$$

maps a text string to a unit-normalized embedding. We score alignment by cosine similarity,

$$s(\mathbf{a}, \mathbf{t}) = \mathbf{a}^\top \mathbf{t},$$

and TRIAGE varies only the text queries and how similarities are aggregated.

3.3. Clinical Attribute System

Direct matching between audio embeddings and class-name embeddings can be brittle when class names are short or ambiguous. We therefore define a structured attribute system with K descriptor groups

$$\mathcal{G} = \{g_1, \ldots, g_K\}.$$

Each group g_k targets a clinically meaningful dimension (e.g., timing, anatomic location, sound quality, pitch, adventitious findings) and contains mutually exclusive options

$$\mathcal{O}_k = \{o_{k,1}, \ldots, o_{k,M_k}\}, \quad |\mathcal{O}_k| = M_k.$$

Each option $o_{k,m}$ is associated with a natural-language template $t_{k,m} \in \mathcal{T}$. The full descriptor set and templates (clinician-reviewed) are provided in Appendix A.

Given x with embedding $\mathbf{a} = f_{\text{audio}}(x)$, we score every option by cosine similarity:

$$s_{k,m}(x) = s\big(\mathbf{a},\, f_{\text{text}}(t_{k,m})\big), \quad k \in [K],\ m \in [M_k].$$

We select the top option per group,

$$m_k^*(x) = \arg\max_{m \in [M_k]} s_{k,m}(x),$$

yielding a descriptor profile

$$\mathbf{z}(x) = \big(o_{1,m_1^*}, \ldots, o_{K,m_K^*}\big) \in \mathcal{O}_1 \times \cdots \times \mathcal{O}_K.$$

Rule-based label mapping. For each task j with label set $\mathcal{Y}_j$, a fixed rule table

$$\Phi_j : \mathcal{O}_1 \times \cdots \times \mathcal{O}_K \to \mathbb{R}^{C_j}$$

maps $\mathbf{z}(x)$ to label scores, encoding which attribute configurations support each class. The rule tables require no learned parameters; an example is provided in Appendix A.

3.4. Three-Tier Inference Pipeline

TRIAGE processes each recording x through up to three stages (Figure 1): **Tier-L** (label cosine scoring), **Tier-M** (descriptor templates + rule voting), and **Tier-H** (kNN report retrieval + LLM decision).

3.4.1. Tier-L: Label-Similarity Scoring

For each class $y \in \mathcal{Y}_j$, we encode its name to obtain $\mathbf{t}_y = f_{\text{text}}(\text{name}(y))$. Given $\mathbf{a} = f_{\text{audio}}(x)$, we compute

$$s_y(x) = \mathbf{a}^\top \mathbf{t}_y, \quad y \in \mathcal{Y}_j,$$

and predict

$$\hat{y}^{(L)}(x) = \arg\max_{y \in \mathcal{Y}_j} s_y(x).$$

We use the top-two margin as confidence:

$$c_L(x) = s_{(1)}(x) - s_{(2)}(x).$$

3.4.2. Tier-M: Descriptor-Based Decision

Tier-M selects the highest-scoring template per descriptor group to form $\mathbf{z}(x) = \{m_k^*\}_{k=1}^K$. A task-specific rule table converts $\mathbf{z}(x)$ to label scores $\{r_y(x)\}_{y \in \mathcal{Y}_j}$, predicting

$$\hat{y}^{(M)}(x) = \arg\max_{y \in \mathcal{Y}_j} r_y(x),$$

with routing confidence

$$c_M(x) = r_{(1)}(x) - r_{(2)}(x).$$

3.4.3. Tier-H: Retrieval-Augmented LLM Reasoning

For remaining uncertain cases, we retrieve k nearest neighbors from a corpus $\mathcal{R}$ of audio embeddings paired with clinician-authored reports:

$$\mathcal{N}_k(x) = \text{top-k}_i\ s\big(\mathbf{a}, \mathbf{a}_i^{\mathcal{R}}\big).$$

We prompt the LLM with $\mathbf{z}(x)$, Tier-L scores $\{s_y(x)\}$, and retrieved report snippets, and parse the final prediction:

$$\hat{y}^{(H)}(x) = \textsc{Parse}\big(\text{LLM}(P(x))\big).$$

The prompt template is in Appendix B.

3.5. Gated Escalation Policy

The three tiers are composed via a gated escalation policy parameterized by thresholds τ_L and τ_M. Let $c_L(x)$ and $c_M(x)$ denote the confidence scores from Tier-L and Tier-M, respectively. The final prediction is:

$$\hat{y}(x) = \begin{cases} \hat{y}^{(L)}(x) & \text{if } c_L(x) \geq \tau_L, \\ \hat{y}^{(M)}(x) & \text{if } c_L(x) < \tau_L \text{ and } c_M(x) \geq \tau_M, \\ \hat{y}^{(H)}(x) & \text{otherwise.} \end{cases}$$

This policy ensures that computation scales with difficulty: high-confidence recordings terminate early, while ambiguous cases receive the full inference budget.

Threshold selection. τ_L and τ_M trade accuracy for compute: lower thresholds escalate more cases, while higher thresholds finalize earlier. We select (τ_L, τ_M) on the validation split and keep them fixed for test evaluation (Section 4.3).

Computational cost. Let T_L, T_M, T_H be per-recording costs and α_M, α_H the fractions reaching Tier-M and Tier-H. The expected cost is

$$\bar{T} = T_L + \alpha_M T_M + \alpha_H T_H.$$

Since $T_H \gg T_M > T_L$, controlling α_H is the main lever; the gate reserves Tier-H for genuinely uncertain cases.

4. Experiments

We evaluate TRIAGE on nine auscultation classification tasks spanning five public datasets. Our experiments assess zero-shot performance against supervised baselines, quantify the contribution of each tier, and ablate key design choices.

4.1. Tasks and Datasets

We evaluate nine auscultation classification tasks drawn from five public corpora of respiratory sounds (details in Appendix C). The tasks include COVID-19 detection from exhalation and cough (UK COVID-19; Coppock et al., 2023), COVID-19 and gender classification from crowdsourced coughs (CoughVID; Orlandic et al., 2021), COPD-versus-healthy screening on stethoscope recordings (ICBHI; Sun, 2023), smoking status and gender classification from cough audio (Coswara; Bhattacharya et al., 2023), obstructive-versus-healthy lung sound classification (KAUH; Fraiwan et al., 2021), and five-class COPD severity grading (Resp.@TR; Altan and Kutlu, 2020). For corpora that provide official subject-level splits, we adopt them directly; otherwise, we construct subject-disjoint train/validation/test partitions with approximate ratios 60/20/20. The proposed zero-shot pipeline is evaluated on the test splits only.

4.2. Embedding Model and Baselines

Our method assumes a frozen audio–text encoder that maps both auscultation audio and textual descriptions into a shared embedding space, as described in Section 3.2. In all experiments, we instantiate this backbone with AcuLa (Wang et al., 2025b), a recent medical audio–language encoder, to provide a strong and reproducible foundation and enable direct comparison to prior zero-shot methods. Concretely, we instantiate f_{audio} as a domain-specific encoder pretrained on a large multi-dataset collection of heart and lung sounds with self-supervised objectives, and f_{text} as a medical language encoder trained on clinical text. AcuLa leverages an alignment stage in which structured metadata from existing auscultation datasets are converted into synthetic clinical reports and used to align audio embeddings with text embeddings via a symmetric contrastive loss, while keeping the language encoder frozen. After this alignment stage, both encoders are frozen and used unchanged for all experiments.

4.3. Test-Time Inference Setup

We follow the pipeline described in Section 3.4 and specify here the test-time thresholds and hyperparameters used in evaluation.

Tier-L configuration. For binary tasks with label set $\mathcal{Y}_j = \{y^{(1)}, y^{(2)}\}$, we define the Tier-L confidence as the absolute margin:

$$c_L(x) = \left| s(\mathbf{a}, \mathbf{t}_{y^{(1)}}) - s(\mathbf{a}, \mathbf{t}_{y^{(2)}}) \right|.$$

For multi-class tasks, $c_L(x)$ is the margin between the top two similarities, as defined in Section 3.4.1. We finalize at Tier-L when $c_L(x) \geq \tau_L$ with $\tau_L = 0.20$; otherwise, the recording is routed to Tier-M.

Tier-M configuration. Tier-M produces rule-based label scores via the task-specific rule table Φ_j. We define $c_M(x)$ as the absolute difference between the top two label scores. The threshold τ_M is selected on the validation split: we sweep a small set of candidate values ($\tau_M \in \{0.04, 0.08, 0.12, 0.16, 0.20\}$), finalize recordings with $c_M(x) \geq \tau_M$, and choose the value that yields the best validation performance on the finalized subset. We fix τ_M per task and apply it unchanged on the test split.

Tier-H configuration. Tier-H is invoked only when $c_L(x) < \tau_L$ and $c_M(x) < \tau_M$. We retrieve the top $k = 3$ nearest neighbors from the retrieval corpus $\mathcal{R}$ using FAISS (Douze et al., 2025), concatenate the retrieved clinical reports into a context block, and query Gemini 3 Pro (Google DeepMind, 2025) with greedy decoding ($T = 0$) to obtain the final prediction. Appendix D reports an ablation over alternative LLM backends under identical retrieval and prompt settings.

4.4. Evaluation Protocol and Ablations

We report AUROC on the held out test set for all tasks. Each experiment is repeated with five random seeds for the supervised baselines and for stochastic components of the test time pipeline, and we report mean and standard deviation.

Tier isolation. To study the effect of adaptive inference, we compare four policies that share the same backbone and data splits: Tier-L only, Tier-M only, Tier-H only, and the full adaptive router that escalates recordings based on $c_L(x)$ and $c_M(x)$.

Table 1: Performance comparison across respiratory audio classification tasks. AUROC scores (↑) for supervised baselines (linear probing with frozen encoders), zero-shot methods, and our proposed zero-shot tiered inference policies. **Supervised baselines** require task-specific training data. **Zero-shot methods** operate without any task-specific training: CLAP uses text-audio similarity, while our tiered policies use Tier-L (label-score cosines), Tier-M (descriptor retrieval with rule voting), Tier-H (FAISS kNN report retrieval + LLM reasoning), and Adaptive (confidence-based hierarchical routing across all three tiers).

Method	UKCOV-EX-1	UKCOV-CO-1	CVID-CO-1	CVID-CO-2	ICBHI-LS-1	COSW-CO-1	COSW-CO-2	KAUH-LS-1	RESPTR-LS-1
Linear Probing Baselines — *(requires task-specific training)*									
VGGish	0.580±0.001	0.557±0.005	0.538±0.028	0.600±0.001	0.605±0.077	0.507±0.027	0.606±0.003	0.605±0.036	0.590±0.034
AudioMAE	0.549±0.001	0.616±0.001	0.554±0.004	0.628±0.001	0.886±0.017	0.549±0.022	0.724±0.001	0.616±0.041	0.510±0.021
CLAP	0.565±0.001	0.648±0.003	0.599±0.007	0.665±0.001	0.933±0.005	0.680±0.009	0.742±0.001	0.697±0.004	0.636±0.045
OGT	0.605±0.001	0.677±0.001	0.552±0.003	0.735±0.000	0.741±0.011	0.650±0.005	0.825±0.001	0.703±0.016	0.606±0.015
AcuLa	0.698±0.001	0.730±0.008	0.887±0.003	0.796±0.004	0.826±0.014	0.830±0.011	0.845±0.004	0.752±0.019	0.710±0.028
Zero-Shot Methods									
CLAP	0.528±0.000	0.542±0.000	0.540±0.000	0.574±0.000	0.687±0.000	0.556±0.000	0.608±0.000	0.566±0.000	0.552±0.000
AcuLa	0.602±0.000	0.665±0.000	0.768±0.000	0.683±0.000	0.789±0.000	0.755±0.000	0.714±0.000	0.702±0.000	0.656±0.000
TRIAGE (Ours)									
Tier-L	0.593±0.000	0.627±0.000	0.722±0.000	0.668±0.000	0.706±0.000	0.717±0.000	0.716±0.000	0.670±0.000	0.610±0.000
Tier-M	0.690±0.000	0.652±0.000	0.780±0.000	0.640±0.000	0.832±0.000	0.695±0.000	0.734±0.000	0.721±0.000	0.698±0.000
Tier-H	0.707±0.001	0.670±0.001	0.802±0.000	0.682±0.001	0.812±0.000	0.700±0.001	0.765±0.000	0.761±0.000	0.705±0.002
Adaptive	**0.703**±0.000	**0.672**±0.000	**0.810**±0.000	**0.700**±0.000	**0.835**±0.000	**0.728**±0.000	**0.766**±0.000	**0.768**±0.000	**0.710**±0.001
Δ vs CLAP (ZS)	+0.175	+0.130	+0.270	+0.126	+0.148	+0.172	+0.158	+0.202	+0.158
Δ vs AcuLa (ZS)	+0.101	+0.007	+0.042	+0.017	+0.046	-0.027	+0.052	+0.066	+0.054

Ablation studies. We conduct ablations along three axes: (i) descriptor coverage, by masking subsets of descriptor groups in Tier-M; (ii) retrieval depth, by varying the number of neighbors $k \in \{1, 3, 5, 10\}$ passed to the LLM in Tier-H; and (iii) LLM backend, by comparing Gemini 3 Pro, gpt-oss-20, Mistral-Small-3.2-24B-Instruct, and Kimi-K2-Instruct under identical retrieval and prompt conditions. All ablations hold other components fixed to isolate each factor's contribution.

5. Results

5.1. Overall performance across tasks

Table 1 reports AUROC on nine respiratory audio classification tasks, comparing supervised linear-probing baselines (frozen encoders with task-specific training), prior zero-shot methods, and our zero-shot tiered inference policies. All TRIAGE tiers use the frozen AcuLa audio–text encoder; CLAP (ZS) is evaluated as an external baseline with its own pretrained encoder. Among zero-shot approaches, AcuLa (ZS) is a strong baseline, while CLAP (ZS), which relies on text–audio similarity scoring, is consistently weaker across tasks. Starting from label-cosine scoring (Tier-L), adding external evidence via Tier-M and Tier-H further improves performance. Overall, the Adaptive router achieves the best zero-shot results (mean AUROC 0.744), substantially improving over CLAP (ZS) and outperforming AcuLa (ZS) on 8 of 9 tasks (the exception is COSW-CO-1). Within our methods, accuracy increases with richer test-time information: Tier-M and Tier-H both improve over Tier-L, and Adaptive performs best on most tasks. Despite using no task-specific training data, Adaptive also compares favorably to supervised linear probing, outperforming VGGish across tasks and matching or exceed-

Table 2: AUROC (↑) on nine respiratory audio classification tasks. We compare **supervised** linear-probing baselines (frozen encoders with task-specific training) and **zero-shot** methods (no task-specific training). CLAP (ZS) predicts via audio–text similarity with its own pretrained encoder. **TRIAGE** uses a frozen AcuLa audio–text encoder and allocates test-time compute across three tiers: **Tier-L** (label-name cosine scoring), **Tier-M** (clinician-approved descriptor template matching with rule voting), and **Tier-H** (FAISS kNN report retrieval + LLM reasoning). **Adaptive** routes examples hierarchically based on confidence. Δ rows report gains of Adaptive over CLAP (ZS) and AcuLa (ZS).

	TL-Finalized (High Conf.)			**TM-Finalized (Medium Conf.)**				**TH-Escalated (Low Conf.)**			
Task	**%**	**TL**	**Adapt.**	**%**	**TL**	**Adapt.**	**Rel.↑**	**%**	**TL**	**Adapt.**	**Rel.↑**
UKCOV-EX-1	47	.645	.645	38	.562	**.690**	23%	15	.520	**.705**	36%
UKCOV-CO-1	62	.680	.680	25	.602	**.662**	10%	13	.566	**.658**	16%
CVID-CO-1	52	.779	.779	31	.711	**.804**	13%	17	.660	**.783**	19%
CVID-CO-2	43	.707	.707	37	.640	**.695**	9%	20	.638	**.678**	6%
ICBHI-LS-1	47	.761	.761	36	.679	**.797**	17%	17	.633	**.802**	27%
COSW-CO-1	67	.712	.712	21	.690	**.720**	4%	12	.639	**.722**	13%
COSW-CO-2	36	.745	.745	43	.728	**.776**	7%	21	.680	**.750**	10%
KAUH-LS-1	33	.703	.703	40	.645	**.728**	13%	27	.605	**.740**	22%
RESPTR-LS-1	32	.677	.677	41	.590	**.718**	22%	27	.545	**.731**	34%
Mean	**46**	**.712**	**.712**	**35**	**.646**	**.732**	**13%**	**19**	**.621**	**.741**	**19%**

Note: Rel.↑ = Relative improvement over TL baseline. Color intensity proportional to improvement magnitude. TH-Escalated shows highest mean relative gain (19%), validating LLM reasoning value on complex cases. The TL-Finalized block omits Rel.↑ because Adaptive and TL are identical for examples finalized at Tier-L.

Table 3: Impact of random descriptor masking on Tier-M retrieval performance. Each task evaluated at three masking levels: 0% (baseline), 20%, and 50% random descriptor removal. Performance degradation (Δ) color-coded: minimal ($|\Delta| < 0.02$), moderate ($0.02 \leq |\Delta| < 0.05$), severe ($|\Delta| \geq 0.05$). Results based on 5 repeated maskings per rate.

Task ID	**Classification Task**	**Baseline**	**20% Masking**		**50% Masking**		**Sensitivity**
		AUROC	**AUROC**	**Δ**	**AUROC**	**Δ**	**Rank**
ICBHI-LS-1	COPD (Lung sounds)	0.832	0.794	-0.038	0.739	**-0.093**	High
UKCOV-EX-1	COVID-19 (Exhalation)	0.690	0.656	-0.034	0.607	**-0.083**	High
RESPTR-LS-1	COPD Severity (Lung)	0.698	0.662	-0.036	0.660	-0.038	High
CVID-CO-1	COVID-19 (Cough)	0.780	0.763	-0.017	0.734	-0.046	Medium
KAUH-LS-1	Obstructive Disease (Lung)	0.721	0.703	-0.018	0.690	-0.031	Medium
COSW-CO-1	Smoking Status (Cough)	0.695	0.688	-0.007	0.673	-0.022	Low
UKCOV-CO-1	COVID-19 (Cough)	0.652	0.642	-0.010	0.635	-0.017	Low
COSW-CO-2	Gender (Cough)	0.734	0.727	-0.007	0.726	-0.008	Low
CVID-CO-2	Gender (Cough)	0.640	0.637	-0.003	0.635	-0.005	Low
Mean Performance		**0.716**	**0.697**	**-0.019**	**0.678**	**-0.038**	–
Std. Deviation		**0.058**	**0.053**	**0.014**	**0.047**	**0.032**	–

Note: Sensitivity ranking based on performance drop at 50% masking: High ($\Delta \leq -0.05$), Medium ($-0.05 < \Delta \leq -0.03$), Low ($\Delta > -0.03$). Lung sound tasks show higher sensitivity to descriptor removal than cough-based tasks, suggesting stronger reliance on semantic retrieval.

Table 4: Tier-H LLM performance versus retrieval context depth (single call, b=1). Each task evaluated with 1, 3, 5, and 8 retrieved documents. Optimal context size (highest AUROC) highlighted in bold. Gain metrics show improvement from minimal context (d=1) to optimal depth. Results demonstrate diminishing returns beyond d=3 for most tasks.

Task ID	Classification Task	d=1 (baseline)	d=3	d=5	d=8	Optimal d	Gain Δ
High Context Sensitivity (Gain > 0.035)							
RESPTR-LS-1	COPD Severity (Lung)	0.667	0.705	0.700	0.698	**3**	+0.038
CVID-CO-2	Gender (Cough)	0.646	0.682	**0.682**	0.680	**3/5**	+0.036
ICBHI-LS-1	COPD (Lung sounds)	0.785	0.812	**0.818**	0.817	**5**	+0.033
Medium Context Sensitivity (0.020 < Gain ≤ 0.035)							
UKCOV-EX-1	COVID-19 (Exhalation)	0.680	0.707	0.705	**0.709**	**8**	+0.029
CVID-CO-1	COVID-19 (Cough)	0.784	0.802	**0.810**	0.805	**5**	+0.026
UKCOV-CO-1	COVID-19 (Cough)	0.645	**0.670**	0.669	0.666	**3**	+0.025
COSW-CO-1	Smoking Status (Cough)	0.681	0.700	**0.705**	0.702	**5**	+0.024
KAUH-LS-1	Obstructive Disease (Lung)	0.738	**0.761**	0.760	0.757	**3**	+0.023
COSW-CO-2	Gender (Cough)	0.746	0.765	**0.769**	0.768	**5**	+0.023
Mean Performance		**0.708**	**0.734**	**0.735**	**0.734**	–	**+0.028**
Optimal Context Distribution		0 tasks	4 tasks	4 tasks	1 task	**Mode: 3–5**	–

Note: Mean AUROC plateaus at d=3 (0.734), with negligible improvement at d=5 (+0.001) and d=8 (-0.001). Most tasks (8/9) achieve optimal performance with 3–5 documents. Only UKCOV-EX-1 benefits from extended context (d=8). Gain magnitude correlates with task difficulty: lowest baseline performers show largest context-driven improvements.

Table 5: **Effect of Tier-L confidence cutoff τ_1 on TRIAGE performance and compute.** AUROC and tier usage (% of samples handled by Tier-L / Tier-M / Tier-H) for $\tau_1 \in \{0.30, 0.45, 0.60\}$. The Tier-M→Tier-H escalation threshold τ_2 is fixed across all settings, and Tier-H uses the same retrieval+LLM backend throughout.

ID	Task	τ_1=0.30				τ_1=0.45				τ_1=0.60			
		AUROC	%T-L	%T-M	%T-H	AUROC	%T-L	%T-M	%T-H	AUROC	%T-L	%T-M	%T-H
UKCOV-EX-1	COVID (Exh.)	0.708	41	43	16	0.713	34	48	18	0.711	27	54	19
UKCOV-CO-1	COVID (Cgh.)	0.675	57	29	14	0.678	47	37	16	0.677	39	43	18
CVID-CO-1	COVID (Cgh.)	0.817	46	36	18	0.819	38	42	20	0.818	30	48	22
CVID-CO-2	Gender (Cgh.)	0.702	36	43	21	0.705	31	47	22	0.703	23	54	23
ICBHI-LS-1	COPD (Lung)	0.844	43	39	18	0.844	33	47	20	0.846	27	52	21
COSW-CO-1	Smoker (Cgh.)	0.727	63	24	13	0.730	53	32	15	0.728	43	40	17
COSW-CO-2	Gender (Cgh.)	0.770	28	50	22	0.772	24	53	23	0.771	20	56	24
KAUH-LS-1	Obstr. (Lung)	0.775	25	46	29	0.777	20	50	30	0.779	16	53	31
RESPTR-LS-1	COPD Sev.	0.717	24	47	29	0.720	19	51	30	0.721	17	52	31
Mean		0.748	40.3	39.7	20.0	0.751	33.2	45.2	21.6	0.750	26.9	50.2	22.9
Avg. Esc. Rate			59.7%				66.8%				73.1%		
Tasks w/ Best AUROC			0 tasks				6 tasks				3 tasks		

ing the strongest linear-probing baseline (AcuLa) on several datasets. Additional linear-probing baselines (OPERA-CT/OPERA-CE) on the same splits are reported in Appendix E.

5.2. Where adaptive routing helps: gains concentrate on uncertain cases

Table 2 stratifies test examples by the tier where Adaptive stops: *TL-Finalized* (resolved at Tier-L), *TM-Finalized* (resolved at Tier-M), and *TH-Escalated* (es-

calated to retrieval-augmented LLM reasoning). For each bucket, we report its share, Tier-L AUROC on that subset, and Adaptive AUROC on the same subset.

Across tasks, **46%** of samples are finalized at Tier-L, where Adaptive matches Tier-L (**.712** vs. **.712** mean AUROC), indicating routing leaves high-confidence decisions unchanged. Gains come from harder cases: for *TM-Finalized* samples (**35%**), mean AUROC rises from **.646** to **.732** (**+13%** relative); for *TH-Escalated* samples (**19%**), it increases from **.621** to **.741** (**+19%** relative). Thus, adaptive computation concentrates on ambiguous inputs, with the largest gains typically in *TH-Escalated* (e.g., UKCOV-EX-1, RESPTR-LS-1), where retrieval-augmented LLM reasoning is most beneficial.

5.3. Tier-M robustness: descriptor masking ablation

Tier-M predicts labels from clinician-reviewed descriptors using sentence-bank cosine matching and a task-specific rule table. We test sensitivity to missing descriptors by randomly masking descriptor groups at inference time (0%, 20%, 50%; Table 3). Mean AUROC declines from **0.716** (0%) to **0.697** (20%, $\Delta = -0.019$) and **0.678** (50%, $\Delta = -0.038$). The largest 50% drops occur for lung-sound/exhalation tasks (ICBHI-LS-1: $-$**0.093**; UKCOV-EX-1: $-$**0.083**), while cough-based tasks are comparatively stable (e.g., COSW-CO-2: $-$**0.008**; CVID-CO-2: $-$**0.005**). This supports escalating to Tier-H when descriptor evidence is missing or unreliable.

5.4. Tier-H context scaling: retrieval depth ablation

We study how Tier-H performance depends on retrieval context size by varying the number of retrieved documents included in the LLM prompt ($d \in \{1, 3, 5, 8\}$) while keeping a single LLM call ($b = 1$) and all other settings fixed (Table 4). Increasing context from $d = 1$ to a small number of documents yields consistent gains: mean AUROC rises from **0.708** at $d = 1$ to **0.734** at $d = 3$. Additional context provides limited benefit on average (**0.735** at $d = 5$ and **0.734** at $d = 8$). Most tasks achieve their best performance with **3**–**5** documents (8/9 tasks), with UKCOV-EX-1 as the only case that benefits from longer context ($d = 8$). Overall, these results show that Tier-H benefits from moderate retrieval context, and that prompt length beyond a few documents typically yields diminishing returns. We provide qualitative examples of retrieved report snippets and their alignment with query audio in Appendix F.

5.5. Compute–performance tradeoff: Tier-L threshold and tier distribution

We examine the compute–performance tradeoff by sweeping the Tier-L cutoff τ_1, which controls how often examples are finalized at Tier-L versus escalated to higher tiers (Table 5). We evaluate three settings, $\tau_1 \in \{0.30, 0.45, 0.60\}$, while keeping the Tier-M→Tier-H threshold fixed.

Increasing τ_1 shifts more examples to Tier-M and Tier-H. The mean fraction finalized at Tier-L drops from **40.3%** at $\tau_1 = 0.30$ to **26.9%** at $\tau_1 = 0.60$, and the overall escalation rate increases from **59.7%** to **73.1%**. Mean AUROC changes only slightly across the sweep (**0.748**, **0.751**, **0.750** for $\tau_1 = 0.30, 0.45, 0.60$, respectively). The setting $\tau_1 = 0.45$ achieves the best AUROC on the majority of tasks (6/9) with a lower escalation rate than $\tau_1 = 0.60$, illustrating that modest increases in Tier-H usage yield limited additional gains beyond an intermediate cutoff.

6. Conclusion

We introduced TRIAGE, a tiered zero-shot inference framework for respiratory audio classification that routes each test example through progressively richer reasoning stages based on prediction confidence, from lightweight label-cosine scoring (Tier-L), through descriptor-based matching (Tier-M), to retrieval-augmented LLM reasoning (Tier-H). Across nine diverse tasks spanning lung sounds, coughs, and exhalation recordings, our Adaptive router achieves a mean AUROC of 0.744 without any task-specific training, outperforming prior zero-shot baselines by substantial margins and matching or exceeding supervised linear probing on several benchmarks. Crucially, the adaptive routing mechanism concentrates additional computation on ambiguous inputs: high-confidence predictions are finalized early with no loss in accuracy, while the hardest cases benefit through retrieval-augmented reasoning. These findings demonstrate that structured, confidence-aware inference can unlock strong zero-shot performance in medical audio analysis, reducing reliance on costly labeled data

while providing interpretable, tiered decision pathways suitable for clinical integration.

Acknowledgments

This work was supported by the NWO AiNed Fellowship Grant of A.S., and in part by Google.org and the Google Cloud Research Credits program through the Gemini Academic Program. We also acknowledge the use of the Dutch National Supercomputer Snellius for essential computational tasks.

References

Gokhan Altan and Yakup Kutlu. Respiratory-database@tr (copd severity analysis), 2020.

D. Bhattacharya, N. K. Sharma, D. Dutta, et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific Data*, 10:397, 2023. doi: 10.1038/s41597-023-02266-0.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. Provable scaling laws for the test-time compute of large language models, 2025. URL https://arxiv.org/abs/2411.19477.

Harry Coppock, Jobie Budd, Emma Karoune, Chris Holmes, Kieran Baker, Davide Pigoli, George Nicholson, Richard Payne, Ivan Kiskin, Josef Packham, Ana Tendero Cañadas, Selina Patel, Sabrina Egglestone, Alexander Titcomb, David Hurley, Lorraine Butler, Tracey Thornley, Jonathon Mellor, Stephen Roberts, Steven Gilmour, Björn Schuller, Vasiliki Koutra, Radka Jersakova, Peter Diggle, Sylvia Richardson, UK Health Security Agency, and The Alan Turing Institute. The uk covid-19 vocal audio dataset, October 2023. URL https://zenodo.org/records/10043978.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025. URL https://arxiv.org/abs/2401.08281.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022. URL https://arxiv.org/abs/2206.04769.

Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope, 2021.

Google DeepMind. Gemini 3 Pro: External model card. Technical report, Google DeepMind, 2025. URL https://deepmind.google/models/model-cards/gemini-3-pro.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. URL https://arxiv.org/abs/2106.13043.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL https://arxiv.org/abs/1712.00409.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. Jina clip: Your clip model is also your text retriever, 2024. URL https://arxiv.org/abs/2405.20204.

Mistral AI. Mistral-small-3.2-24b-instruct-2506, June 2025. URL https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506. Hugging Face model card/repository.

Moonshot AI. Kimi-k2-instruct. Hugging Face model repository, July 2025. URL https://huggingface.co/moonshotai/Kimi-K2-Instruct. Initial repo commit Jul 11, 2025. Accessed 2026-01-20.

OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms, February 2021. URL https://zenodo.org/records/7024894.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Jon Saad-Falcon, Adrian Gamarra Lafuente, Shlok Natarajan, Nahum Maru, Hristo Todorov, Etash Guha, E. Kelly Buchanan, Mayee Chen, Neel Guha, Christopher Ré, and Azalia Mirhoseini. Archon: An architecture search framework for inference-time techniques. In *International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2409.15254. URL https://arxiv.org/abs/2409.15254.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Zhiqiang Sun. ICBHI 2017 challenge, 2023.

Tsai-Ning Wang, Lin-Lin Chen, Neil Zeghidour, and Aaqib Saeed. Careaqa: A cardiac and respiratory audio question answering model for open-ended diagnostic reasoning, 2025a. URL https://arxiv.org/abs/2505.01199.

Tsai-Ning Wang, Lin-Lin Chen, Neil Zeghidour, and Aaqib Saeed. Language models as semantic teachers: Post-training alignment for medical audio understanding, 2025b. URL https://arxiv.org/abs/2512.04847.

Han Xiao. Scaling test-time compute for embedding models. https://jina.ai/news/scaling-test-time-compute-for-embedding-models/, December 2024. Jina AI Blog.

Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu, Mohammed Mosuily, J Ch, and Cecilia Mascolo. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. *Advances in Neural Information Processing Systems*, 37:27024–27055, 2024a.

Yuwei Zhang, Tong Xia, Aaqib Saeed, and Cecilia Mascolo. Respllm: Unifying audio and text with multimodal llms for generalized respiratory health prediction. *arXiv preprint arXiv:2410.05361*, 2024b.

Henry Peng Zou, Zhengyao Gu, Yue Zhou, Yankai Chen, Weizhi Zhang, Liancheng Fang, Yibo Wang, Yangning Li, Kay Liu, and Philip S. Yu. Test-NUC: Enhancing test-time computing approaches and scaling through neighboring unlabeled data consistency. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30750–30762, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1486. URL https://aclanthology.org/2025.acl-long.1486/.

Appendix A. Tier-M Descriptor Taxonomy and Class Prototypes

In Table 6, we provide a concrete example of the Tier-M descriptor system for ICBHI-LS-1 (COPD vs. Healthy). For each descriptor group, we list the full set of clinician-approved templates used at inference time and the prototypical descriptor choice associated with each class in our rule table (`task_rules`). During Tier-M, a recording is first mapped to one selected template per group via cosine matching, and the resulting descriptor profile is compared against these class prototypes to produce the label decision.

Table 6: Tier-2 descriptor groups and prototypical configurations for COPD vs Healthy lung sounds in task ICBHI-LS-1. For each descriptor group we list the full option set used in our system, and the specific option selected by the COPD and Healthy prototypes in `task_rules`.

Descriptor group	**System option set**	**COPD prototype**	**Healthy prototype**
Breath sound character	normal vesicular breathing; diminished or distant breath sounds; bronchial breathing with high pitch; absent breath sounds; amphoric hollow breathing; cavernous breathing sounds; harsh breathing with increased intensity	diminished or distant breath sounds	normal vesicular breathing
Wheeze presence	no wheeze detected; mild expiratory wheeze; moderate expiratory wheeze; severe expiratory wheeze; inspiratory wheeze present; biphasic wheeze (inspiratory and expiratory); polyphonic multiple wheeze; monophonic single wheeze	moderate expiratory wheeze	no wheeze detected
Respiratory phase timing	normal inspiratory to expiratory ratio 1:2; prolonged expiratory phase ratio 1:3 or greater; shortened expiratory phase ratio 1:1; prolonged inspiratory phase; equal inspiratory and expiratory phases; irregular variable phase timing; rapid shallow breathing pattern	prolonged expiratory phase ratio 1:3 or greater	normal inspiratory to expiratory ratio 1:2
Crackle characteristics	no crackles present; fine high-pitched crackles early inspiratory; fine crackles late inspiratory; coarse low-pitched crackles early inspiratory; coarse crackles throughout inspiration; bibasilar crackles at lung bases; diffuse crackles throughout lung fields; velcro-like crackles	coarse low-pitched crackles early inspiratory	no crackles present
Respiratory effort	normal and effortless; mildly increased effort; moderately labored; severely labored with accessory muscle use; paradoxical breathing pattern; shallow with minimal effort; gasping or air hunger pattern	moderately labored	normal and effortless
Spectral frequency profile	normal frequency distribution 100–1000 Hz; low frequency dominance below 400 Hz; high frequency dominance above 800 Hz; broadband frequency distribution; narrow band frequency concentration; bimodal frequency peaks; irregular frequency scatter	low frequency dominance below 400 Hz	normal frequency distribution 100–1000 Hz

Appendix B. Prompt Example for Tier-H LLM Decision

At Tier-H, we query Gemini with retrieval-augmented clinical evidence to produce the final binary decision for diagnostically uncertain cases. The prompt supplies the top-k retrieved report snippets as context and constrains the model to select a diagnosis from a predefined label set. To make outputs consistent and directly usable for evaluation, we enforce a strict JSON format containing only the predicted label and a brief justification, preventing extraneous explanations.

Gemini Prompt for Tier-H

You are a highly experienced cardiopulmonary doctor. Given the following reports, select the most likely/probable diagnosis from the given classes below and write very few words justification.

Reports: - The presence of expiratory wheezes in the posterior lower lung fields in this 62-year-old male with COPD suggests airway obstruction typically associated with this condition.
- In a 58-year-old female with COPD, expiratory wheezes are noted in the posterior right lower lung field, indicating likely airway narrowing or obstruction in this region.
- The respiratory examination of this 59-year-old male with asthma reveals expiratory wheezes located at the posterior right lower lung field. These findings may indicate airway obstruction or constriction in this region.

Classes: Obstructive, Healthy

Your output should be JSON of the following structure: {"result": ..., "justification": ...}. Do not provide any other explanation.

Example Gemini Response (JSON)

{"result":"Obstructive", "justification":"Expiratory wheezes with COPD/asthma indicate airway obstruction."}

Appendix C. Downstream Tasks and Datasets

This section summarizes the downstream respiratory audio classification benchmarks used in our evaluation. Table 7 lists each task ID, label space, audio modality (cough, exhalation, or lung sounds), and dataset statistics (sample counts and class distributions). Tasks span three categories: respiratory disease detection, demographic classification, and COVID-19 detection.

Table 7: Downstream evaluation tasks for respiratory audio classification. All tasks use binary or multi-class classification on audio recordings from real-world clinical and crowdsourced datasets.

Task ID	Classification Task	Audio Type	Samples	Class Distribution
COVID-19 Detection				
UKCOV-EX-1	Covid vs. Non-covid	Exhalation	2,500	840 / 1,660
UKCOV-CO-1	Covid vs. Non-covid	Cough	2,500	840 / 1,660
CVID-CO-1	Covid vs. Non-covid	Cough	6,175	547 / 5,628
Demographic Classification				
CVID-CO-2	Female vs. Male	Cough	7,263	2,468 / 4,795
COSW-CO-2	Female vs. Male	Cough	2,496	759 / 1,737
Respiratory Disease Detection				
ICBHI-LS-1	COPD vs. Healthy	Lung sounds	828	793 / 35
COSW-CO-1	Smoker vs. Non-smoker	Cough	948	201 / 747
KAUH-LS-1	Obstructive vs. Healthy	Lung sounds	234	129 / 105
RESPTR-LS-1	COPD Severity (5-class)	Lung sounds	504	72 / 60 / 84 / 84 / 204

Appendix D. Tier-H backend choice: LLM ablation

We replace the LLM used in Tier-H while keeping the encoder, retrieval database, and prompt fixed (d=3, same budget b; Table 8). Gemini 3 Pro achieves the best AUROC on all nine tasks, with a mean AUROC of **0.734**. Kimi-K2 is second (**0.711**), followed by gpt-oss (**0.695**) and Mistral-Small (**0.689**). Backend choice therefore affects absolute Tier-H performance under matched inference cost, and we use Gemini 3 Pro as the default Tier-H backend in the remaining experiments.

Table 8: **Tier-H LLM backend ablation.** Per-task AUROC when swapping the Tier-H LLM while keeping retrieval and prompting fixed (d=3, same budget b). Backends: Gemini 3 Pro (Google DeepMind, 2025), gpt-oss-20b (OpenAI, 2025), Mistral-Small-3.2-24B-Instruct (Mistral AI, 2025), and Kimi-K2-Instruct (Moonshot AI, 2025).

ID	Task	AUROC (↑)			
		Gemini 3 Pro	gpt-oss	Mistral-Small	Kimi-K2
UKCOV-EX-1	Covid (Exhale)	0.707	0.686	0.660	0.683
UKCOV-CO-1	Covid (Cough)	0.670	0.643	0.638	0.659
CVID-CO-1	Covid (Cough)	0.802	0.756	0.752	0.767
CVID-CO-2	Gender (Cough)	0.682	0.654	0.648	0.668
ICBHI-LS-1	COPD (Lung)	0.812	0.755	0.757	0.776
COSW-CO-1	Smoker (Cough)	0.700	0.664	0.663	0.685
COSW-CO-2	Gender (Cough)	0.765	0.716	0.718	0.742
KAUH-LS-1	Obstructive (Lung)	0.761	0.713	0.710	0.739
RESPTR-LS-1	COPD severity (Lung)	0.705	0.666	0.659	0.676

Appendix E. Additional linear-probing baselines

Table 9 reports the per-task linear-probing AUROC of OPERA-CT and OPERA-CE (Zhang et al., 2024a) on the same train/test splits as Table 1. For completeness, we also include the corresponding tiered inference results (TL–TH and Adaptive) on the same tasks.

Table 9: Appendix: AUROC (↑) for the omitted linear-probing baselines OPERA-CT and OPERA-CE, alongside our three-tier inference policies (same splits as Table 1).

ID	OCT	OCE	TL-only	TM-only	TH-only	Adaptive
UKCOV-EX-1	0.586	0.551	0.593	0.690	0.707	0.703
UKCOV-CO-1	0.701	0.629	0.627	0.652	0.670	0.672
CVID-CO-1	0.578	0.566	0.722	0.780	0.802	0.810
CVID-CO-2	0.795	0.721	0.668	0.640	0.682	0.700
ICBHI-LS-1	0.855	0.872	0.706	0.832	0.812	0.835
COSW-CO-1	0.685	0.674	0.717	0.695	0.700	0.728
COSW-CO-2	0.874	0.801	0.716	0.734	0.765	0.766
KAUH-LS-1	0.722	0.741	0.670	0.721	0.761	0.768
RESPTR-LS-1	0.625	0.683	0.610	0.698	0.705	0.710

Appendix F. Qualitative Retrieval Examples

This section presents qualitative retrieval results from the Tier-H stage. For each query auscultation clip, FAISS retrieves the top-3 nearest clinical reports in the shared audio–text embedding space. The retrieved reports serve as supporting context for the Tier-H decision module, and they also provide an interpretable view of what the retrieval stage considers most similar to the query.

Figure 3: Top-3 clinical reports retrieved for auscultation clips (Tier-H). Left: query spectrogram and its reference (or generated) report. Right: the three closest reports returned by FAISS in the shared embedding space.