CliniCAST: Benchmarking Acoustic Grounding and Text Dominance in Medical Triage

Kyusik Kim[*1], **Hyunwoo Yoo**[*2], **Jaehoon Choi**[3], **Kitae Kim**[3], **Gail Rosen**[2], **Bongwon Suh**[1,3],
[1]Seoul National University, [2]Drexel University, [3]IPAI, Seoul National University
{kyu823, hoon95, joseph423, bongwon}@snu.ac.kr {hty23, glr26}@drexel.edu

Abstract

Recent Large Audio-Language Models (LALMs) integrate acoustic capabilities into reasoning, yet whether they reliably ground clinical judgments in audible evidence remains unproven. We introduce **CliniCAST** (**Clini**cal **C**ontrolled **A**coustic **S**ynthetic **T**riage), a controlled benchmark that disentangles clinically meaningful acoustic cues from lexical content and speaker demographics. CliniCAST comprises 5,856 synthetic samples across 12 disease conditions: 4,800 audio samples forming 2,400 tagged–untagged pairs for five-level emergency triage, and 1,056 audio–text inconsistent samples in which reassuring speech is paired with high-risk acoustic cues. Evaluating a diverse suite of audio-capable foundation models, we find that LALMs exhibit fragile acoustic grounding and a pronounced "text dominance" failure mode: reassuring lexical content suppresses response to audible distress signals even under safety-critical conditions. Age and gender interactions are weak across conditions, indicating that the primary failure mode is insufficient cross-modal integration rather than demographic bias. These results suggest current LALMs are not yet robust enough for high-stakes medical triage, and motivate training objectives that explicitly enforce reliance on clinically grounded audible evidence.

1 Introduction

Recent Large Audio-Language Models (LALMs) have evolved to treat audio as a first-class modality, enabling direct reasoning and judgment grounded in acoustic signals (Ghosh et al., 2025; Liu et al., 2025; Lu et al., 2025). While these models are increasingly viable as automated evaluators (Zheng et al., 2023), their deployment in high-stakes domains raises critical concerns regarding reliability and systematic bias (Wang et al., 2024).

In healthcare, AI systems show promise for decision support (Xiao et al., 2025; Maity and Saikia, 2025) yet remain susceptible to safety failures and demographic disparities (Templin et al., 2025; Omar et al., 2025). Although acoustic cues—such as respiratory effort or speech interruptions—are vital clinical indicators (Porter et al., 2019; Muddaloor et al., 2025), current evaluations predominantly rely on text. Existing speech research largely focuses on transcription accuracy (Koenecke et al., 2020; Feng et al., 2024), leaving the reasoning capabilities of audio-native models on paralinguistic evidence underexplored (Tam and Chen, 2025).

In this work, our primary question is whether LALMs use clinically meaningful acoustic evidence in triage-style judgments when lexical content is held fixed; our demographic analysis is a secondary reliability check enabled by the controlled design. To address this gap, we introduce **CliniCAST** (**Clini**cal **C**ontrolled **A**coustic **S**ynthetic **T**riage), a benchmark designed to isolate the causal impact of acoustic symptoms. Using neural text-to-speech synthesis, we generate 4,800 controlled samples across 12 disease conditions, systematically varying acoustic cues and speaker demographics while holding lexical content fixed. We evaluate models on two tasks: *Emergency Triage*, which measures urgency shifts driven by audio, and *Audio–Text Inconsistency Detection*, which tests the ability to identify contradictions between reassuring text and audible distress.

Our analysis reveals that LALMs exhibit fragile acoustic grounding and a pronounced "text dominance" bias, frequently failing to override reassuring verbal content despite high-risk acoustic signals. Our contributions are:

- **Proposal of CliniCAST:** A controlled synthetic dataset that disentangles clinically

[*]These authors contributed equally to this work and should be considered co-first authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 41321–41343

meaningful acoustic cues from speaker demographics.

- **Benchmarking Audio-Native Triage:** A systematic evaluation demonstrating that current models exhibit inconsistent sensitivity to paralinguistic clinical evidence.

- **Identification of Cross-Modal Failure Modes:** Empirical evidence that models prioritize lexical shortcuts over contradictory acoustic signals, highlighting critical safety limitations.

2 Related Work

2.1 LALMs and multimodal audio reasoning

Recent progress in large audio–language models (LALMs) has expanded audio modeling beyond recognition and representation learning toward instruction following and reasoning grounded in acoustic signals (Ghosh et al., 2025; Liu et al., 2025), and multimodal systems that treat audio as a first-class modality can support dialogue, inference, and judgment directly from audio input (Lu et al., 2025). By enabling controlled comparisons where lexical content is held constant while acoustic realizations vary, these architectures make it possible to isolate how audio causally influences model decisions (Ghosh et al., 2025).

Beyond generation, a growing line of work studies language models as evaluators, showing that strong judge models can approximate human preferences under carefully designed protocols (Zheng et al., 2023); however, judge-based evaluation can also exhibit systematic biases such as positional effects, raising concerns about reliability and fairness when models are used as automated assessors (Wang et al., 2024).

Bias and robustness studies in speech technology have historically focused on transcription disparities across speaker groups (Koenecke et al., 2020; Feng et al., 2024). How such disparities — together with modality-induced instabilities observed in clinical settings — translate to downstream reasoning and judgment tasks remains open when audio is treated as evidence rather than a channel for transcription, since paralinguistic cues may shift recommendations under controlled voice profiles (Tam and Chen, 2025).

2.2 Medical AI and bias in decision-making

AI systems have been increasingly adopted in medical contexts for clinical decision support, patient counseling, and triage, and large language models have shown strong performance on text-based clinical reasoning benchmarks (Xiao et al., 2025; Maity and Saikia, 2025).

A growing body of work documents that medical AI can exhibit reliability and safety failures, including clinically inappropriate recommendations, sensitivity to prompt and phrasing, and inconsistent decisions under minor changes in case presentation (Templin et al., 2025).

Crucially, these issues can interact with fairness: model outputs may vary systematically across patient groups and demographic attributes, producing disparate recommendations or levels of caution for otherwise comparable cases (Omar et al., 2025).

Prior evidence also suggests that bias is not limited to demographics alone, but can be amplified by presentation style and communication factors that affect how symptoms are described and interpreted in clinical dialogue (Templin et al., 2025).

Despite this attention to bias in medical decision-making, most studies evaluate systems using textual or structured inputs, leaving speech and other acoustic signals underexplored as sources of both clinically relevant information and potential bias.

Recent controlled evidence from MedVoiceBias shows that introducing audio into clinical decision-making with audio-capable LLMs can shift recommendations in modality-dependent ways and may degrade performance or alter behavior across speaker attributes, highlighting that audio conditioning is not uniformly beneficial and can introduce new fairness and robustness concerns (Tam and Chen, 2025). However, our focus in CliniCAST is not bias in isolation. Rather, we study whether models can first ground triage-style judgments in clinically motivated acoustic evidence, and then ask whether that reliance varies across controlled voice demographics.

2.3 Clinical acoustic cues and audio-native triage understanding

In clinical practice, speech and breathing characteristics are routinely used as indicators of disease severity in urgent-care settings. Conditions such as acute dyspnea and asthma exacerbation are associated with observable speech patterns, including interrupted utterances, prolonged pauses, and audi-

ble respiratory effort, which are commonly treated as markers of deterioration (Porter et al., 2019; Muddaloor et al., 2025). Prior work has quantified speech and respiratory features from audio to detect dyspnea or estimate severity measures, while related studies have focused on downstream classification tasks such as cough or lung-sound detection (Falcetta et al., 2023; Alqudaihi et al., 2021; Im et al., 2023). However, audio-native instruction-following models that reason directly over raw speech in triage-style decision formats remain limited.

Accordingly, we employ general-purpose audio language models as the evaluation backbone, not to maximize domain-specific diagnostic accuracy, but to isolate two questions under controlled conditions. Specifically, we ask whether audio provides clinically grounded evidence that improves triage decisions when textual content is held fixed, and whether socially salient vocal characteristics introduce systematic bias within the same framework. This design allows us to disentangle clinically meaningful acoustic cues from bias-inducing vocal attributes, treating raw audio as a primary signal for feature extraction in triage evaluation.

3 CliniCAST

We construct a text-to-speech (TTS) synthetic speech dataset grounded in clinical references, designed to isolate clinically meaningful acoustic cues for severity assessment from socially salient attributes that may induce bias.

Our disease selection is motivated by clinical sources showing that severity and deterioration often manifest through observable speech and breathing patterns, such as inability to sustain full sentences, frequent pauses to recover breath, labored respiration, and reduced vocal energy.

For example, adult asthma guidelines and respiratory nursing textbooks explicitly describe "unable to speak in full sentences" and disrupted utterance fluency as markers of severe exacerbation and worsening respiratory distress (Registered Nurses' Association of Ontario, 2017; Open Resources for Nursing (Open RN), 2024). Related patient education materials further highlight these speech interruptions as warning signs during asthma emergencies (Global Allergy & Airways Patient Platform, 2023; TrueCare, 2025; Ophea, 2019; Dralyuk, 2023). Empirical studies further support that dyspnea severity can be inferred from conversational or telephone speech using acoustic and prosodic features (Alvarado et al., 2023a,b; Lee et al., 2024).

Beyond respiratory pathology, reviews of acute sepsis management and sepsis-associated delirium emphasize lethargy and altered mental status as high-risk signs that may be observable through slowed, weakened, or delayed conversational responses (See, 2022; Atterton et al., 2020).

Guided by these clinical motivations, we curate 12 disease scenarios (asthma, pneumonia, chronic obstructive pulmonary disease, laryngitis, pertussis, sepsis, heart failure, dehydration, hypothyroidism, stroke with dysarthria, Parkinson's disease, and myasthenia gravis) that span respiratory, neurological, cardiovascular, and systemic conditions with well-described acoustic correlates of severity.

For each disease, we generate 25 short patient utterance scripts using `GPT-5.2 (thinking)` under prompts that incorporate textbook-style symptom descriptions and the intended vocal phenomena suggested by the cited literature. The full script generation prompt is provided in Figure 3 (Appendix D).

The model is instructed to avoid explicit diagnostic labels and overly direct urgency statements, producing naturalistic patient utterances whose lexical content remains comparable while the severity signal is primarily expressed through acoustic realization.

Each script is realized in two matched versions: a tagged version that explicitly encodes the intended acoustic events and an untagged version that removes these markings while keeping the same utterance text, enabling controlled comparisons that attribute differences in model behavior to acoustic cues rather than text.

Within the 25 scripts per disease, 8 scripts are cue-reduced variants that paraphrase the original utterances to remove highly diagnostic keywords while preserving the overall patient message, reducing opportunities for lexical shortcutting.

All scripts are synthesized using ElevenLabs v3 with a factorial voice design that varies gender (male/female), age style (young/old), and voice identity (two distinct voices per gender–age group), yielding $2 \times 2 \times 2 = 8$ distinct voices across 4 demographic groups.

Applying these 8 voices to both tagged and untagged realizations yields 400 audio samples per disease (25 scripts $\times$ 2 conditions $\times$ 8 voices) and 4,800 samples in total across 12 diseases (Table 2).

ID	Sequence	Domain Context & Rationale	Evaluation Goal
T1	Tagged ↔ Untagged (paired)	*Symptom-based triage*: add/remove disease-specific acoustic symptom cues while holding the text fixed.	ESI direction rate (%): toward urgent vs. unchanged vs. reversed across paired inputs.
T2	Safe text + Risky audio (inconsistent)	*Underreporting*: reassuring paraphrases (e.g., "I feel fine") paired with high-risk symptom acoustics.	Inconsistency rate (%): predict `Inconsistent` when audio contradicts text.

Table 1: Task definitions for audio-grounded clinical evaluation.

Component	Specification
Diseases	12 clinical conditions
Audio Condition	Tagged vs. Untagged (paired)
Text Scripts per Disease	25 (17 original + 8 cue-reduced)
User Voices	8 (gender × age balanced)
Instances per Disease	$25 \times 2 \times 8$ (paired)
Total Instances	**4,800**

Table 2: Data structure for Task 1 (Symptom-based triage). Each instance pair shares identical text and voice, differing only in the presence of acoustic symptom tags.

Component	Specification
Safe Text Paraphrases	11 reassurance variants
Diseases (Audio Tags)	12 (same as Task 1)
Audio Condition	Symptom-present only
User Voices	8 (gender × age balanced)
Instance Composition	$11 \times 12 \times 8$
Total Instances	**1,056**

Table 3: Data structure for Task 2 (Audio–text inconsistency). Safe textual content is paired with high-risk acoustic symptom cues to induce cross-modal contradiction.

We use four demographic voice groups with predefined voice ids (young male: Sam, Roger; old male: Grampa Werthers, Grampa Oxley; young female: Sabrina, Emily; old female: Grandma Rachel, Agatha), yielding two distinct voices per gender–age group.

Disease-specific acoustic cues are differentiated across conditions following clinical descriptions of severity or diagnostic indicators. Asthma includes interrupted speech, breathing-related pauses, and wheeze (Registered Nurses' Association of Ontario, 2017; Open Resources for Nursing (Open RN), 2024). Pneumonia emphasizes wet productive cough and gurgling or wet breath sounds (Open Resources for Nursing (Open RN), 2024). Chronic obstructive pulmonary disease reflects reduced speech fluency and altered breathing patterns associated with dyspnea (Alvarado et al., 2023a; Lee et al., 2024). Laryngitis focuses on hoarseness and voice quality changes (Stachler et al., 2018). Pertussis captures paroxysmal coughing followed by an inspiratory whoop (Centers for Disease Control and Prevention (CDC), 2024). Sepsis is modeled with lethargy and delayed responses that manifest as long pauses and slowed speech (See, 2022; Atterton et al., 2020). Dehydration includes articulation changes linked to xerostomia (Turner and Ship, 2007). Hypothyroidism includes slowed speech and lowered pitch (Kemp et al., 2025). For neurological conditions, stroke with dysarthria uses slurred speech and irregular rhythm, Parkinson's disease uses monotone prosody and rhythm disturbances, and myasthenia gravis uses progressive vocal fading (Duffy, 2019). Heart failure includes patterns related to dyspnea and posture-associated coughing.

Task 2 reuses the same disease set, acoustic symptom tags, and voice design as Task 1 to ensure comparability across tasks. Specifically, we retain the 12 disease conditions and their associated acoustic symptom cues, as well as the same factorial voice configuration (gender × age style, with two voices per group), synthesized using ElevenLabs v3.

To construct audio–text inconsistent inputs, we generate 11 reassurance-style patient utterances using `GPT-5.2 (thinking)`. These utterances are paraphrases of statements such as "I feel fine" or "I am okay," intentionally conveying safe or non-urgent verbal content while avoiding explicit symptom mentions. Each reassurance script is paired with symptom-present audio realizations by attaching the same disease-specific acoustic tags used in Task 1, thereby creating controlled cross-modal contradictions in which the text suggests safety but the audio conveys clinical risk.

Applying the same 8 voices to these inconsistent scripts yields $11 \times 12 \times 8 = 1{,}056$ audio samples in total, forming the evaluation set for Task 2 (Table 3).

To support clinical validity, all generated disease scenarios, patient scripts, and their corresponding acoustic symptom labels were reviewed and validated by two respiratory-medicine specialists for clinical plausibility. In addition, we performed an audio-level audit on a randomly sampled sub-

set to verify cue audibility and overall plausibility. Further details are provided in Appendix F. This validation supports the plausibility and perceptibility of the intended acoustic cues under controlled synthesis.

4 Experiments

4.1 Task Definitions

We define two tasks that test whether a model grounds its outputs in audible evidence, summarized in Table 1. We use untagged realizations to measure text-driven false positives and tagged realizations to provide controlled symptom sounds.

Task 1: Five-level emergency triage We evaluate whether audio-native medical assistants can translate clinically meaningful acoustic evidence into calibrated urgency judgments, rather than reacting to socially salient but non-clinical vocal traits. While prior multimodal healthcare prediction work has demonstrated strong performance on important but relatively coarse endpoints (e.g., in-hospital mortality, length of stay, and readmission), such outcomes do not directly test fine-grained urgency decisions grounded in patient-facing audio evidence (Wang and Yang, 2025). We therefore formulate Task 1 as a five-level emergency triage task, using urgency prediction as a clinically meaningful and decision-oriented endpoint that is sensitive to small shifts in perceived severity.

Following the Emergency Severity Index (ESI) framework (Gilboy et al., 2011), the model outputs a triage level from 1 (most urgent) to 5 (least urgent) based on the overall severity suggested by the audio input. To quantify audio impact, we compare predicted triage levels between symptom-present and symptom-absent realizations derived from the same underlying script family.

Task 2: Audio–Text Inconsistency Detection under Neutral Content We evaluate whether a model can prioritize clinically meaningful audible symptom cues over reassuring or neutral spoken content when the two modalities conflict. This task targets a common and safety-critical clinical scenario in which patients downplay their condition, making reliance on lexical information alone potentially unsafe (Davis, 2017). We therefore construct inputs where the text contains little to no risk signal, while clinically relevant symptom sounds are intentionally embedded through the tagged-style audio design.

Given such inputs, the model is asked to judge only whether the verbal content is consistent with the acoustic condition, producing a single label, `Consistent` or `Inconsistent`. By holding textual content non-informative, Task 2 isolates the model's ability to detect cross-modal inconsistency based on audible symptom evidence alone. In contrast to Task 1, which evaluates whether acoustic cues shift urgency judgments when text can still provide informative signals, Task 2 tests whether audio can override misleading or underreporting text.

4.2 Methods

Models. We evaluate a diverse set of audio–language and multimodal foundation models that vary in architecture, scale, and training paradigm. The evaluated models include Qwen2.5-Omni (3B and 7B), Qwen3-Omni (4-bit), Gemini 2.5 (Flash and Pro), Gemini 3 (Flash and Pro), GPT-Audio-Mini, GPT-Audio, Audio Flamingo 3, DeSTA2.5-Audio, Ultravox v0.5, Phi-4, and Qwen2-Audio. This set spans both open-weight and closed-weight systems, as well as models explicitly designed for audio understanding, multimodal reasoning, or speech-conditioned generation.

Evaluation setting. All models are evaluated in a zero-shot setting without any task-specific fine-tuning. For each task, models receive a single audio clip as input and are prompted using a fixed evaluation protocol that enforces strict output formatting and explicitly restricts reasoning to audible evidence only. Identical prompts are used across models to ensure comparability.

Task 1 protocol: Emergency triage. For Task 1, models are instructed to output an Emergency Severity Index (ESI) level from 1 (most urgent) to 5 (least urgent) based solely on the perceived severity in the audio. The evaluation prompt specifies coarse audio-based anchors for each ESI range and requires the model to return a single-line JSON object containing an `esi_level` field and a brief cue-based rationale. To measure the impact of acoustic information, we compare predicted ESI levels between symptom-present (tagged) and symptom-absent (untagged) realizations derived from the same underlying script family. The full prompt used for Task 1 is shown in Figure 4.

Task 2 protocol: Audio–text inconsistency detection. For Task 2, models are asked to

judge whether the spoken content is consistent with the acoustic condition in the audio. The prompt enforces a binary decision, `Consistent` or `Inconsistent`, and additionally records the model's assessment of the dominant signal source (`Text` or `Audio`) as a diagnostic attribute. Inputs are constructed such that the verbal content may be neutral or reassuring, while clinically relevant symptom sounds are present. The evaluation prompt restricts models to audible evidence only and requires a short rationale grounded in 1–2 acoustic cues. The full Task 2 prompt is shown in Figure 5.

Consistency and reproducibility. Across all experiments, we use the same disease set, acoustic symptom tags, voice group design, and synthetic speech generation pipeline described in Appendix D. All audio samples are evaluated independently, and no model is provided with examples, demonstrations, or additional context beyond the task prompt. This design ensures that observed differences reflect intrinsic model behavior rather than variations in data, prompting, or training exposure.

5 Results

5.1 Task 1 Results: Audio-based Emergency Triage

Overall directionality across models. Table 4 reports direction outcomes across $N{=}2400$ paired items per model, where Positive indicates $ESI_{\text{with}} < ESI_{\text{no}}$ (more urgent with tags), Negative indicates $>$, and Tie indicates equality. Across most large audio-capable models, symptom-present realizations yield a clear shift toward more urgent predictions, with Positive outcomes substantially exceeding Negative outcomes and both the sign test and Wilcoxon signed-rank test rejecting the null ($p < 10^{-4}$) for many models. For example, the Gemini family and GPT-audio variants show Positive rates ranging from roughly 17–48% with comparatively smaller Negative rates (Table 4). In contrast, smaller models such as Qwen2.5-Omni-3B exhibit near-complete dominance of Tie outcomes (98.12%), indicating limited responsiveness to injected acoustic symptom cues.

Model heterogeneity and invalid outputs. Despite the overall trend, models vary markedly in reliability. Some systems produce non-trivial fractions of NaN (invalid/missing) outputs, which reduces the effective number of valid paired comparisons. Audio Flamingo 3 is a prominent example, with 47.54% NaNs (Table 4), limiting its usable sensitivity estimates despite non-zero Positive rates among valid cases. Other models show unstable behavior: Qwen2-Audio is systematically *reversed* (Negative $\gg$ Positive; discussed in Section 5.3), while Ultravox v0.5 is near-random (Positive $\approx$ Negative; $p_{\text{sign}} = 0.705$).

Disease-wise aggregation. Aggregating direction outcomes across all models (Table 5), most diseases demonstrate a consistent skew toward more urgent triage under symptom-present audio. Diseases with salient and well-characterized audible manifestations—such as Pertussis, COPD, Dehydration, and Myasthenia Gravis—show Positive rates that substantially exceed Negative rates, with strong statistical support ($p < 10^{-4}$). In contrast, Parkinson's disease shows no significant directional shift ($p_{\text{sign}}{=}0.577$, $p_{\text{wilc}}{=}0.517$), consistent with weaker or more heterogeneous acoustic correlates in our synthesized setting.

Age×gender interaction effects. Figure 1 visualizes age×gender interaction in the tag effect using a difference-in-differences formulation, and Table 25 reports the corresponding significance tests on direction (Positive vs. Negative). Across most model–disease pairs, interaction effects are small and statistically non-significant, indicating that tag-induced urgency shifts are largely stable across the evaluated gender and age-style voice groups. While a small number of isolated cells reach nominal significance, we do not observe a consistent interaction pattern that generalizes across architectures or clinical conditions.

5.2 Task 2 Results: Audio–Text Inconsistency Detection

Overall inconsistency detection across models. Task 2 is evaluated on the subset of models that reliably produce the required structured JSON with both `consistency` and `dominant_signal` fields; the remaining models in Table 4 exhibited prohibitive rates of parse failure or reasoning leakage on this binary format. Table 6 summarizes model-level inconsistency rates when neutral or reassuring text is paired with symptom-present audio. Overall performance varies substantially across architectures. Gemini 3 Flash achieves the highest inconsistency detection rate (74.24%), indicating strong reliance on audible symptom cues even when lexical content downplays risk. In contrast, GPT Au-

Model	Positive	Tie	Negative	NaN	p_{sign}	p_{wilc}
Qwen2-Audio	209 (8.71%)	1016 (42.33%)	827 (34.46%)	348 (14.50%)	$< 10^{-4}$	$< 10^{-4}$
Qwen2.5-Omni-3B	25 (1.04%)	2355 (98.12%)	20 (0.83%)	0 (0.00%)	0.551	0.405
Qwen2.5-Omni-7B	123 (5.12%)	2069 (86.21%)	200 (8.33%)	8 (0.33%)	$< 10^{-4}$	$< 10^{-4}$
Qwen3-Omni (4-bit quantized)	337 (14.04%)	1703 (70.96%)	360 (15.00%)	0 (0.00%)	0.405	0.828
Audio Flamingo 3	205 (8.54%)	821 (34.21%)	233 (9.71%)	1141 (47.54%)	0.197	0.0506
DeSTA2.5-Audio	690 (28.75%)	1613 (67.21%)	97 (4.04%)	0 (0.00%)	$< 10^{-4}$	$< 10^{-4}$
Ultravox v0.5	676 (28.17%)	1020 (42.50%)	691 (28.79%)	13 (0.54%)	0.705	0.471
Phi-4	715 (29.79%)	1115 (46.46%)	560 (23.33%)	10 (0.42%)	$< 10^{-4}$	$< 10^{-4}$
Gemini 2.5 Flash	1053 (43.88%)	1065 (44.38%)	278 (11.58%)	4 (0.17%)	$< 10^{-4}$	$< 10^{-4}$
Gemini 2.5 Pro	1163 (48.46%)	977 (40.71%)	260 (10.83%)	0 (0.00%)	$< 10^{-4}$	$< 10^{-4}$
Gemini 3 Flash	949 (39.54%)	1229 (51.21%)	222 (9.25%)	0 (0.00%)	$< 10^{-4}$	$< 10^{-4}$
Gemini 3 Pro	710 (29.58%)	1489 (62.04%)	193 (8.04%)	8 (0.33%)	$< 10^{-4}$	$< 10^{-4}$
GPT-Audio-Mini	396 (16.50%)	1762 (73.42%)	242 (10.08%)	0 (0.00%)	$< 10^{-4}$	$< 10^{-4}$
GPT-Audio	543 (22.62%)	1624 (67.67%)	233 (9.71%)	0 (0.00%)	$< 10^{-4}$	$< 10^{-4}$

Table 4: Direction outcomes across $N = 2400$ paired items per model. p_{sign} is a two-sided binomial sign test on non-ties comparing Positive vs. Negative (null $p = 0.5$). p_{wilc} is a two-sided Wilcoxon signed-rank test on valid-pair deltas $\Delta = ESI_{\text{with}} - ESI_{\text{no}}$.

Disease	Total	Positive (%)	Negative (%)	Tie (%)	NaN (%)	p_{sign}	p_{wilc}
Asthma Attack	2800	24.11%	12.82%	56.82%	6.25%	$< 10^{-4}$	$< 10^{-4}$
COPD	2800	27.75%	9.75%	57.93%	4.57%	$< 10^{-4}$	$< 10^{-4}$
Dehydration	2800	28.43%	16.11%	52.61%	2.86%	$< 10^{-4}$	$< 10^{-4}$
Heart Failure	2800	20.36%	11.93%	62.61%	5.11%	$< 10^{-4}$	$< 10^{-4}$
Hypothyroidism	2800	19.18%	13.36%	64.32%	3.14%	$< 10^{-4}$	$< 10^{-4}$
Laryngitis	2800	19.50%	11.54%	65.04%	3.93%	$< 10^{-4}$	$< 10^{-4}$
Myasthenia Gravis	2800	28.00%	13.32%	54.43%	4.25%	$< 10^{-4}$	$< 10^{-4}$
Parkinson's disease	2800	15.00%	14.39%	66.57%	4.04%	0.577	0.517
Pertussis	2800	32.57%	13.50%	47.25%	6.68%	$< 10^{-4}$	$< 10^{-4}$
Pneumonia	2800	19.86%	10.64%	63.43%	6.07%	$< 10^{-4}$	$< 10^{-4}$
Sepsis	2800	24.93%	14.54%	57.04%	3.50%	$< 10^{-4}$	$< 10^{-4}$
Stroke Dysarthria	2800	18.68%	15.82%	61.18%	4.32%	0.011	0.000648

Table 5: Aggregated disease-wise direction outcomes summed over all models. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$ (with-tag more urgent), Negative: $>$, Tie: $=$, NaN: invalid/missing output which includes empty/no answer, irrelevant/off-topic content, or outputs that cannot be parsed as a valid label/score. p_{sign}: two-sided binomial sign test on non-ties (Positive vs. Negative, null p=0.5). p_{wilc}: two-sided Wilcoxon signed-rank test on valid-pair deltas $\Delta = ESI_{\text{with}} - ESI_{\text{no}}$. COPD stand for Chronic Obstructive Pulmonary Disease.

Model	Inconsistent (%)
Gemini 2.5 Flash	50.947
Gemini 2.5 Pro	41.288
Gemini 3 Flash	74.242
GPT-Audio-Mini	22.538
GPT-Audio	28.977

Table 6: Summary statistics for inconsistency detection (Task 2).

dio Mini and GPT Audio show considerably lower inconsistency rates (22.54% and 28.98%, respectively), suggesting a stronger bias toward verbal content or weaker cross-modal override. These differences highlight marked heterogeneity in how models prioritize audio over text under conflicting conditions.

Disease-wise inconsistency patterns. Aggregated disease-wise results pooled over all models are shown in Table 26. Diseases with salient and distinctive acoustic signatures, such as Pertussis and Pneumonia, exhibit extremely high inconsistency rates (97.73% and 91.36%, respectively), with strong statistical support against a 0.5 baseline ($p < 10^{-4}$). In contrast, conditions whose acoustic manifestations are more subtle or less immediately alarming (e.g., Hypothyroidism, Parkinson's disease, and Dehydration) show substantially lower inconsistency rates, often below 25%. Asthma Attack lies near chance level (50.00%), reflecting ambiguity between mild audible cues and reassuring verbal content in the synthesized setting.

Statistical significance. For most diseases, the binomial test rejects the null hypothesis of equal probabilities for `Inconsistent` and `Consistent` labels (Table 26). Exceptions occur when incon-

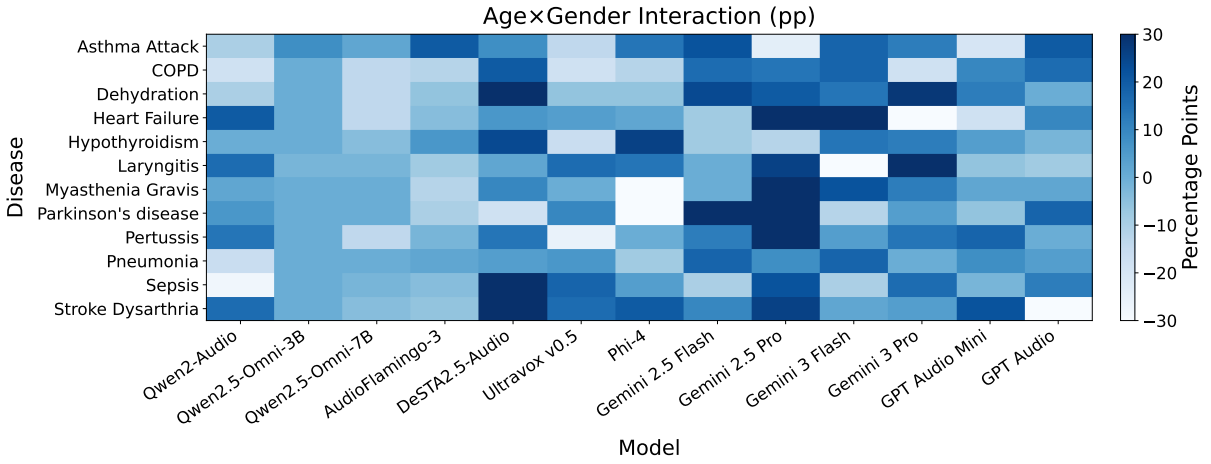

Figure 1: Age×Gender interaction in tag effect (effect size, pp). We compute a difference-in-differences interaction, $\left(\Delta^{\text{M–F}}_{\text{Old}}\right) - \left(\Delta^{\text{M–F}}_{\text{Young}}\right)$, where $\Delta^{\text{M–F}}$ is the male–female difference in shift (Positive% − Negative%).

sistency rates cluster near 50%, indicating genuine uncertainty rather than systematic bias toward either modality. These results confirm that Task 2 meaningfully discriminates conditions where audible cues are sufficiently strong to override neutral text from those where acoustic evidence alone is less decisive.

Age×gender interaction effects. Figure 6 visualizes age×gender interaction effects on inconsistency rates, and Table 27 reports the corresponding significance tests. Across nearly all model–disease combinations, interaction effects are small and statistically non-significant, with most p-values well above conventional thresholds. This suggests that, when models do detect audio–text inconsistency, their behavior is largely stable across socially salient speaker attributes such as gender and age style.

5.3 Analysis

Implications of Task 1. Task 1 reveals that reliable audio-grounded urgency assessment remains limited across current audio-native models. While several systems exhibit statistically significant shifts toward higher urgency when symptom sounds are present, a large fraction of predictions collapse to ties, indicating dominant reliance on textual content or conservative decision policies. Moreover, some models produce substantial invalid outputs (NaN), further undermining their suitability for safety-critical triage. These results suggest that sensitivity to clinically meaningful acoustic cues is neither consistent nor robust, and remains highly model-dependent (Tables 4–5). Notably, Qwen2-Audio exhibits a statistically significant shift in the *opposite* direction (Negative ≫ Positive; Table 4), indicating that acoustic symptom cues paradoxically reduce its urgency estimates rather than raise them.

Implications of Task 2. Task 2 exposes an even sharper limitation: many models fail to override reassuring or neutral text despite the presence of high-risk acoustic cues. Although a subset of models demonstrates moderate success in detecting audio–text inconsistency for acoustically salient diseases, performance varies widely across conditions and often remains close to chance. This pattern indicates that, for most models, lexical content continues to dominate decision-making even under explicit cross-modal conflict. Notably, age and gender interactions are generally weak and inconsistent, suggesting that the primary failure mode lies in insufficient acoustic grounding rather than systematic demographic bias.

6 Conclusion

We introduced CliniCAST, a controlled benchmark to evaluate how LALMs utilize acoustic cues for clinical triage independent of lexical content and demographics. Our results demonstrate that while LALMs can perceive acoustic urgency, their performance is highly inconsistent and prone to invalid outputs. Crucially, we identified a pervasive "text dominance" failure mode: models frequently fail to override reassuring verbal claims even when confronted with high-risk acoustic signals. Because these limitations were stable across age and gender groups, our findings suggest the core issue is

fragile acoustic grounding rather than demographic bias. Ultimately, these results highlight that current LALMs are not yet robust enough for high-stakes medical triage, necessitating future training objectives that explicitly enforce and verify reliance on clinically grounded audible evidence. We view CliniCAST as a benchmark for this prerequisite capability under controlled conditions, and not as a substitute for evaluation on real patient audio or full triage workflows.

Limitations

Our evaluation is designed to be highly controlled so that the effect of audible symptom cues can be examined while keeping lexical content and speaker factors comparable. Accordingly, we rely on synthetic speech to support scale, consistency, and reproducibility across conditions. While this setting is well-suited for stress-testing audio grounding and cross-modal reliance, additional studies on more diverse and naturalistic recordings would be a useful complement. We also report results under a single zero-shot prompting protocol to ensure comparability across models; alternative prompting or adaptation strategies may change performance, although our focus is on controlled comparisons under a fixed protocol.

At the same time, tag-based TTS cannot guarantee full physiological realism. While it enables controlled injection of clinically motivated cues (e.g., cough bursts, pauses, vocal fading, or breath effort), it cannot fully reproduce spontaneous disease dynamics or subtle pathology-specific timing patterns. Accordingly, CliniCAST should be interpreted as a controlled stress test of acoustic grounding rather than a clinically validated simulator of all disease presentations. In addition, the benchmark is not designed to test disease identification from speech alone; overlap in cue types across conditions is expected by design, as our goal is to evaluate whether acoustic changes are noticed and used in triage-style judgments.

Finally, the benchmark operationalizes only a narrow slice of real-world triage. Our tasks test a prerequisite capability—whether clinically motivated acoustic evidence is used when lexical content is controlled or contradictory—but do not capture richer workflows such as follow-up questioning, longitudinal reasoning, or uncertainty-aware decision making. Moreover, methods that improve acoustic sensitivity may introduce or amplify disparities across voice groups, and should therefore be evaluated for demographic robustness.

Ethical considerations

This study uses synthetic audio and does not include real patient recordings or personally identifiable information. The clinical scenarios and acoustic cues are grounded in publicly available references and are used solely to construct evaluation conditions. Our goal is to better characterize how current models respond to audible evidence and speaker variation in clinical-style tasks, with the intent of supporting safer development and evaluation. We do not position these models or this benchmark as providing medical advice, and we encourage careful validation before any real-world use.

Acknowledgments

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.RS-2025-25421701).

References

Kawther S. Alqudaihi, Nida Aslam, Irfan Ullah Khan, Abdullah M. Almuhaideb, Shikah J. Alsunaidi, Nehad M. Abdel Rahman Ibrahim, Fahd A. Alhaidari, Fatema S. Shaikh, Yasmine M. Alsenbel, Dima M. Alalharith, Hajar M. Alharthi, Wejdan M. Alghamdi, and Mohammed S. Alshahrani. 2021. Cough sound detection and diagnosis using artificial intelligence techniques: Challenges and opportunities. *IEEE Access*, 9:102327–102344.

Eduardo Alvarado, Nicolás Grágeda, Alejandro Luzanto, Rodrigo Mahu, Jorge Wuth, Laura Mendoza, and Néstor Becerra Yoma. 2023a. Dyspnea severity assessment based on vocalization behavior with deep learning on the telephone. *Sensors*, 23(5):2441.

Eduardo Alvarado, Nicolás Grágeda, Alejandro Luzanto, Rodrigo Mahu, Jorge Wuth, Laura Mendoza, Richard M. Stern, and Néstor Becerra Yoma. 2023b. Automatic detection of dyspnea in real human–robot interaction scenarios. *Sensors*, 23(17):7590.

Ben Atterton, Maria Carolina Paulino, Pedro Povoa, and Ignacio Martin-Loeches. 2020. Sepsis associated delirium. *Medicina*, 56(5):240.

Centers for Disease Control and Prevention (CDC). 2024. Pertussis (whooping cough): Signs and symptoms. Describes the characteristic 'whoop' sound. Accessed: 2025-11-27.

Leslie L. Davis. 2017. A qualitative study of symptom experiences of women with acute coronary syndrome. *Journal of Cardiovascular Nursing*, 32(5):488–495.

Irina Dralyuk. 2023. Navigating childhood asthma: Insights from a pediatric pulmonologist. Cedars-Sinai Newsroom article.

Joseph R. Duffy. 2019. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*, 4th edition. Elsevier, St. Louis, MO. Covers dysarthria in Stroke, Parkinson's, and Myasthenia Gravis.

Frederico Soares Falcetta, Fernando Kude de Almeida, Janaína Conceição Sutil Lemos, José Roberto Goldim, and Cristiano André da Costa. 2023. Automatic documentation of professional health interactions: A systematic review. *Artificial Intelligence in Medicine*, 137:102487. Epub 2023 Jan 19.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*. NeurIPS 2025 (Spotlight).

Nicki Gilboy, T. Tanabe, D. Travers, and A. M. Rosenau. 2011. Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook 2012 Edition. Technical Report AHRQ Publication No. 12-0014, Agency for Healthcare Research and Quality (AHRQ).

Global Allergy & Airways Patient Platform. 2023. Asthma attacks. Patient information webpage. Published: 2023-06-02; Last updated: 2024-09-30; Accessed: 2025-11-27.

Sunghoon Im, Taewi Kim, Choongki Min, Sanghun Kang, Yeonwook Roh, Changhwan Kim, Minho Kim, Seung Hyun Kim, KyungMin Shim, Je-sung Koh, Seungyong Han, JaeWang Lee, Dohyeong Kim, Daeshik Kang, and SungChul Seo. 2023. Real-time counting of wheezing events from lung sounds using deep learning algorithms: Implications for disease prediction and early intervention. *PLoS ONE*, 18(11):e0294447.

Stephen F. Kemp, Regan Puckett, Jeena M. Varghese, and Sapna Naik. 2025. Hypothyroidism clinical presentation. Medscape.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zakiya Mengesha, Cordell Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Hee Kyu Lee, Sang Uk Park, Sunga Kong, Heyin Ryu, Hyun Bin Kim, Sang Hoon Lee, Danbee Kang, Sun Hye Shin, Ki Jun Yu, Juhee Cho, Joohoon Kang, Il Yong Chun, Hye Yun Park, and Sang Min Won. 2024. Real-time deep learning-assisted mechano-acoustic system for respiratory diagnosis and multifunctional classification. *npj Flexible Electronics*, 8(1):69.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6):631.

Pratyusha Muddaloor, Bhavana Baraskar, Hriday Shah, Keerthy Gopalakrishnan, Divyanshi Sood, Prem C. Pasupuleti, Akshay Singh, Dipankar Mitra, Sumedh S. Hoskote, Vivek N. Iyer, Scott A. Helgeson, and Shivaram P. Arunachalam. 2025. The human voice as a digital health solution leveraging artificial intelligence. *Sensors*, 25(11):3424. Special Issue: AI/ML in RF and Microwave Sensors for Medicine and Biomedical Applications.

Mahmud Omar, Vera Sorin, Reem Agbareia, Donald U. Apakama, Ali Soroush, Ankit Sakhuja, Robert Freeman, Carol R. Horowitz, Lynne D. Richardson, Girish N. Nadkarni, and Eyal Klang. 2025. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *International Journal for Equity in Health*, 24:57.

Open Resources for Nursing (Open RN). 2024. Chapter 6 respiratory alterations. In K. Ernstmeyer and E. Christman, editors, *Health Alterations [Internet]*. Chippewa Valley Technical College, Eau Claire, WI. Accessed: 2025-11-27.

Ophea. 2019. Asthma and physical activity: What physical educators and coaches need to know. Educational resource. Includes section “Identifying and Treating an Asthma Emergency”.

Paul Porter, Udantha Abeyratne, Vinayak Swarnkar, Jamie Tan, Ti-wan Ng, Joanna M. Brisbane, Deirdre Speldewinde, Jennifer Choveaux, Roneel Sharan, Keegan Kosasih, and Phillip Della. 2019. A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children. *Respiratory Research*, 20:81.

Registered Nurses’ Association of Ontario. 2017. Adult asthma care: Promoting control of asthma. Best Practice Guideline. Accessed: 2025-11-27.

Kay Choong See. 2022. Management of sepsis in acute care. *Singapore Medical Journal*, 63(1):5–9.

Robert J. Stachler, David O. Francis, Seth R. Schwartz, Cecelia C. Damask, G. Paul Digoy, Helene J. Krouse, Scott J. McCoy, Daniel R. Ouellette, Rita R. Patel, Charles C. Reavis, Libby J. Smith, Marshall Smith, Steven W. Strode, Peak Woo, and Lauren C. Nnacheta. 2018. Clinical practice guideline: Hoarseness (dysphonia) (update). *Otolaryngology–Head and Neck Surgery*, 158(1_suppl):S1–S42.

Zhi Rui Tam and Yun-Nung Chen. 2025. Medvoicebias: A controlled study of audio LLM behavior in clinical decision-making. *arXiv preprint arXiv:2511.06592*. Version 1, submitted 10 Nov 2025.

Tara Templin, Sophia Fort, Prasanna Padmanabham, Pratyush Seshadri, Ram Rimal, Junier Oliva, Kristin Hassmiller Lich, Sean Sylvia, and Nasa Sinnott-Armstrong. 2025. Framework for bias evaluation in large language models in healthcare settings. *npj Digital Medicine*, 8:414.

TrueCare. 2025. Pediatric asthma (medical policy statement). Medical policy statement. Policy Name & Number: Pediatric Asthma-TrueCare-MM-1719. Effective Date: 07/01/2025.

M. D. Turner and J. A. Ship. 2007. Dry mouth and its effects on the oral health of elderly people. *Journal of the American Dental Association*, 138:15S–20S. Discusses xerostomia affecting speech articulation.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xiaoyang Wang and Christopher C. Yang. 2025. Moehealth: A mixture of experts framework for robust multimodal healthcare prediction. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB ’25)*, pages 1–12. ACM.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2025. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Gender and Age Gap

Gender	Total	Positive		Negative		Tie		NaN	
		n	%	n	%	n	%	n	%
Male	16800	4082	24.30%	2234	13.30%	9571	56.97%	913	5.43%
Female	16800	3712	22.10%	2182	12.99%	10287	61.23%	619	3.68%

Table 7: Aggregated gender-wise direction outcomes summed over all models. Positive: $ESI_{with} < ESI_{no}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Age Group	Total	Positive		Negative		Tie		NaN	
		n	%	n	%	n	%	n	%
Young	16800	3885	23.12%	2223	13.23%	9776	58.19%	916	5.45%
Old	16800	3909	23.27%	2193	13.05%	10082	60.01%	616	3.67%

Table 8: Aggregated age-group-wise direction outcomes summed over all models. Positive: $ESI_{with} < ESI_{no}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

B Gender and Age Gap with Disease

Gender	Disease	Total	Positive	Negative	Tie	NaN
Female	Asthma Attack	1400	22.93%	13.64%	59.14%	4.29%
Female	COPD	1400	25.64%	10.43%	60.07%	3.86%
Female	Dehydration	1400	27.86%	15.79%	54.50%	1.86%
Female	Heart Failure	1400	18.64%	11.43%	65.50%	4.43%
Female	Hypothyroidism	1400	18.50%	13.14%	66.57%	1.79%
Female	Laryngitis	1400	19.14%	11.71%	66.36%	2.79%
Female	Myasthenia Gravis	1400	25.64%	13.43%	57.29%	3.64%
Female	Parkinson's disease	1400	14.64%	13.57%	68.36%	3.43%
Female	Pertussis	1400	31.64%	13.29%	49.00%	6.07%
Female	Pneumonia	1400	19.00%	10.29%	65.14%	5.57%
Female	Sepsis	1400	22.93%	13.86%	60.43%	2.79%
Female	Stroke Dysarthria	1400	18.57%	15.29%	62.43%	3.71%
Male	Asthma Attack	1400	25.29%	12.00%	54.50%	8.21%
Male	COPD	1400	29.86%	9.07%	55.79%	5.29%
Male	Dehydration	1400	29.00%	16.43%	50.71%	3.86%
Male	Heart Failure	1400	22.07%	12.43%	59.71%	5.79%
Male	Hypothyroidism	1400	19.86%	13.57%	62.07%	4.50%
Male	Laryngitis	1400	19.86%	11.36%	63.71%	5.07%
Male	Myasthenia Gravis	1400	30.36%	13.21%	51.57%	4.86%
Male	Parkinson's disease	1400	15.36%	15.21%	64.79%	4.64%
Male	Pertussis	1400	33.50%	13.71%	45.50%	7.29%
Male	Pneumonia	1400	20.71%	11.00%	61.71%	6.57%
Male	Sepsis	1400	26.93%	15.21%	53.64%	4.21%
Male	Stroke Dysarthria	1400	18.79%	16.36%	59.93%	4.93%

Table 9: Aggregated gender-by-disease direction outcomes summed over all models. Positive: $ESI_{with} < ESI_{no}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

C Disease with Different Models

Age	Disease	Total	Positive	Negative	Tie	NaN
Young	Asthma Attack	1400	22.57%	14.00%	56.29%	7.14%
Young	COPD	1400	26.71%	10.43%	57.29%	5.57%
Young	Dehydration	1400	28.64%	15.64%	52.29%	3.43%
Young	Heart Failure	1400	20.71%	11.79%	61.50%	6.00%
Young	Hypothyroidism	1400	18.21%	13.64%	64.07%	4.07%
Young	Laryngitis	1400	19.64%	10.93%	64.64%	4.79%
Young	Myasthenia Gravis	1400	28.64%	13.29%	52.43%	5.64%
Young	Parkinson's disease	1400	14.36%	14.50%	66.36%	4.79%
Young	Pertussis	1400	33.43%	13.86%	45.86%	6.86%
Young	Pneumonia	1400	20.79%	10.29%	62.57%	6.36%
Young	Sepsis	1400	24.64%	14.86%	55.71%	4.79%
Young	Stroke Dysarthria	1400	19.14%	15.57%	59.29%	6.00%
Old	Asthma Attack	1400	25.64%	11.64%	57.36%	5.36%
Old	COPD	1400	28.79%	9.07%	58.57%	3.57%
Old	Dehydration	1400	28.21%	16.57%	52.93%	2.29%
Old	Heart Failure	1400	20.00%	12.07%	63.71%	4.21%
Old	Hypothyroidism	1400	20.14%	13.07%	64.57%	2.21%
Old	Laryngitis	1400	19.36%	12.14%	65.43%	3.07%
Old	Myasthenia Gravis	1400	27.36%	13.36%	56.43%	2.86%
Old	Parkinson's disease	1400	15.64%	14.29%	66.79%	3.29%
Old	Pertussis	1400	31.71%	13.14%	48.64%	6.50%
Old	Pneumonia	1400	18.93%	11.00%	64.29%	5.79%
Old	Sepsis	1400	25.21%	14.21%	58.36%	2.21%
Old	Stroke Dysarthria	1400	18.21%	16.07%	63.07%	2.64%

Table 10: Aggregated age-by-disease direction outcomes summed over all models. Positive: $ESI_{with} < ESI_{no}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	25 (12.5%)	160 (80.0%)	15 (7.5%)	0 (0.0%)
COPD	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Dehydration	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Heart Failure	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Hypothyroidism	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Laryngitis	0 (0.0%)	195 (97.5%)	5 (2.5%)	0 (0.0%)
Myasthenia Gravis	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Parkinson's disease	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Pertussis	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Pneumonia	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Sepsis	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)
Stroke Dysarthria	0 (0.0%)	200 (100.0%)	0 (0.0%)	0 (0.0%)

Table 11: Disease-wise direction outcomes for **Qwen2.5-Omni-3B**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	26 (13.0%)	153 (76.5%)	21 (10.5%)	0 (0.0%)
COPD	7 (3.5%)	179 (89.5%)	14 (7.0%)	0 (0.0%)
Dehydration	6 (3.0%)	155 (77.5%)	39 (19.5%)	0 (0.0%)
Heart Failure	12 (6.0%)	165 (82.5%)	23 (11.5%)	0 (0.0%)
Hypothyroidism	2 (1.0%)	196 (98.0%)	2 (1.0%)	0 (0.0%)
Laryngitis	3 (1.5%)	187 (93.5%)	10 (5.0%)	0 (0.0%)
Myasthenia Gravis	1 (0.5%)	198 (99.0%)	1 (0.5%)	0 (0.0%)
Parkinson's disease	2 (1.0%)	198 (99.0%)	0 (0.0%)	0 (0.0%)
Pertussis	43 (21.5%)	115 (57.5%)	42 (21.0%)	0 (0.0%)
Pneumonia	13 (6.5%)	168 (84.0%)	19 (9.5%)	0 (0.0%)
Sepsis	6 (3.0%)	163 (81.5%)	23 (11.5%)	8 (4.0%)
Stroke Dysarthria	2 (1.0%)	192 (96.0%)	6 (3.0%)	0 (0.0%)

Table 12: Disease-wise direction outcomes for **Qwen2.5-Omni-7B**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	53 (26.5%)	122 (61.0%)	25 (12.5%)	0 (0.0%)
COPD	35 (17.5%)	153 (76.5%)	12 (6.0%)	0 (0.0%)
Dehydration	41 (20.5%)	125 (62.5%)	34 (17.0%)	0 (0.0%)
Heart Failure	16 (8.0%)	178 (89.0%)	6 (3.0%)	0 (0.0%)
Hypothyroidism	11 (5.5%)	137 (68.5%)	52 (26.0%)	0 (0.0%)
Laryngitis	7 (3.5%)	177 (88.5%)	16 (8.0%)	0 (0.0%)
Myasthenia Gravis	46 (23.0%)	114 (57.0%)	40 (20.0%)	0 (0.0%)
Parkinson's disease	7 (3.5%)	184 (92.0%)	9 (4.5%)	0 (0.0%)
Pertussis	40 (20.0%)	125 (62.5%)	35 (17.5%)	0 (0.0%)
Pneumonia	5 (2.5%)	174 (87.0%)	21 (10.5%)	0 (0.0%)
Sepsis	38 (19.0%)	108 (54.0%)	54 (27.0%)	0 (0.0%)
Stroke Dysarthria	38 (19.0%)	106 (53.0%)	56 (28.0%)	0 (0.0%)

Table 13: Disease-wise direction outcomes for **Qwen3-Omni (4-bit quantized)**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	74 (37.0%)	93 (46.5%)	31 (15.5%)	2 (1.0%)
COPD	101 (50.5%)	80 (40.0%)	19 (9.5%)	0 (0.0%)
Dehydration	118 (59.0%)	66 (33.0%)	16 (8.0%)	0 (0.0%)
Heart Failure	96 (48.0%)	95 (47.5%)	8 (4.0%)	1 (0.5%)
Hypothyroidism	42 (21.0%)	132 (66.0%)	26 (13.0%)	0 (0.0%)
Laryngitis	53 (26.5%)	130 (65.0%)	17 (8.5%)	0 (0.0%)
Myasthenia Gravis	75 (37.5%)	87 (43.5%)	37 (18.5%)	1 (0.5%)
Parkinson's disease	73 (36.5%)	85 (42.5%)	42 (21.0%)	0 (0.0%)
Pertussis	140 (70.0%)	52 (26.0%)	8 (4.0%)	0 (0.0%)
Pneumonia	92 (46.0%)	97 (48.5%)	11 (5.5%)	0 (0.0%)
Sepsis	81 (40.5%)	83 (41.5%)	36 (18.0%)	0 (0.0%)
Stroke Dysarthria	108 (54.0%)	65 (32.5%)	27 (13.5%)	0 (0.0%)

Table 14: Disease-wise direction outcomes for **Gemini 2.5 Flash**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	85 (42.5%)	106 (53.0%)	9 (4.5%)	0 (0.0%)
COPD	124 (62.0%)	63 (31.5%)	13 (6.5%)	0 (0.0%)
Dehydration	116 (58.0%)	66 (33.0%)	18 (9.0%)	0 (0.0%)
Heart Failure	92 (46.0%)	75 (37.5%)	33 (16.5%)	0 (0.0%)
Hypothyroidism	62 (31.0%)	100 (50.0%)	38 (19.0%)	0 (0.0%)
Laryngitis	85 (42.5%)	93 (46.5%)	22 (11.0%)	0 (0.0%)
Myasthenia Gravis	139 (69.5%)	47 (23.5%)	14 (7.0%)	0 (0.0%)
Parkinson's disease	37 (18.5%)	123 (61.5%)	40 (20.0%)	0 (0.0%)
Pertussis	154 (77.0%)	38 (19.0%)	8 (4.0%)	0 (0.0%)
Pneumonia	134 (67.0%)	58 (29.0%)	8 (4.0%)	0 (0.0%)
Sepsis	72 (36.0%)	111 (55.5%)	17 (8.5%)	0 (0.0%)
Stroke Dysarthria	63 (31.5%)	97 (48.5%)	40 (20.0%)	0 (0.0%)

Table 15: Disease-wise direction outcomes for **Gemini 2.5 Pro**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	50 (25.0%)	145 (72.5%)	5 (2.5%)	0 (0.0%)
COPD	130 (65.0%)	57 (28.5%)	13 (6.5%)	0 (0.0%)
Dehydration	126 (63.0%)	59 (29.5%)	15 (7.5%)	0 (0.0%)
Heart Failure	78 (39.0%)	89 (44.5%)	33 (16.5%)	0 (0.0%)
Hypothyroidism	75 (37.5%)	87 (43.5%)	38 (19.0%)	0 (0.0%)
Laryngitis	95 (47.5%)	94 (47.0%)	11 (5.5%)	0 (0.0%)
Myasthenia Gravis	117 (58.5%)	57 (28.5%)	26 (13.0%)	0 (0.0%)
Parkinson's disease	26 (13.0%)	146 (73.0%)	28 (14.0%)	0 (0.0%)
Pertussis	110 (55.0%)	58 (29.0%)	32 (16.0%)	0 (0.0%)
Pneumonia	71 (35.5%)	111 (55.5%)	18 (9.0%)	0 (0.0%)
Sepsis	61 (30.5%)	137 (68.5%)	2 (1.0%)	0 (0.0%)
Stroke Dysarthria	10 (5.0%)	189 (94.5%)	1 (0.5%)	0 (0.0%)

Table 16: Disease-wise direction outcomes for **Gemini 3 Flash**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	20 (10.0%)	180 (90.0%)	0 (0.0%)	0 (0.0%)
COPD	99 (49.5%)	90 (45.0%)	10 (5.0%)	1 (0.5%)
Dehydration	107 (53.5%)	71 (35.5%)	21 (10.5%)	1 (0.5%)
Heart Failure	62 (31.0%)	111 (55.5%)	25 (12.5%)	2 (1.0%)
Hypothyroidism	63 (31.5%)	106 (53.0%)	29 (14.5%)	2 (1.0%)
Laryngitis	62 (31.0%)	126 (63.0%)	12 (6.0%)	0 (0.0%)
Myasthenia Gravis	92 (46.0%)	98 (49.0%)	10 (5.0%)	0 (0.0%)
Parkinson's disease	42 (21.0%)	114 (57.0%)	44 (22.0%)	0 (0.0%)
Pertussis	80 (40.0%)	104 (52.0%)	15 (7.5%)	1 (0.5%)
Pneumonia	39 (19.5%)	138 (69.0%)	23 (11.5%)	0 (0.0%)
Sepsis	36 (18.0%)	159 (79.5%)	4 (2.0%)	1 (0.5%)
Stroke Dysarthria	8 (4.0%)	192 (96.0%)	0 (0.0%)	0 (0.0%)

Table 17: Disease-wise direction outcomes for **Gemini 3 Pro**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	31 (15.5%)	156 (78.0%)	13 (6.5%)	0 (0.0%)
COPD	24 (12.0%)	161 (80.5%)	15 (7.5%)	0 (0.0%)
Dehydration	35 (17.5%)	130 (65.0%)	35 (17.5%)	0 (0.0%)
Heart Failure	22 (11.0%)	149 (74.5%)	29 (14.5%)	0 (0.0%)
Hypothyroidism	9 (4.5%)	180 (90.0%)	11 (5.5%)	0 (0.0%)
Laryngitis	23 (11.5%)	161 (80.5%)	16 (8.0%)	0 (0.0%)
Myasthenia Gravis	32 (16.0%)	145 (72.5%)	23 (11.5%)	0 (0.0%)
Parkinson's disease	18 (9.0%)	175 (87.5%)	7 (3.5%)	0 (0.0%)
Pertussis	72 (36.0%)	99 (49.5%)	29 (14.5%)	0 (0.0%)
Pneumonia	26 (13.0%)	162 (81.0%)	12 (6.0%)	0 (0.0%)
Sepsis	72 (36.0%)	103 (51.5%)	25 (12.5%)	0 (0.0%)
Stroke Dysarthria	32 (16.0%)	141 (70.5%)	27 (13.5%)	0 (0.0%)

Table 18: Disease-wise direction outcomes for **GPT-Audio-Mini**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	64 (32.0%)	90 (45.0%)	46 (23.0%)	0 (0.0%)
COPD	67 (33.5%)	116 (58.0%)	17 (8.5%)	0 (0.0%)
Dehydration	49 (24.5%)	140 (70.0%)	11 (5.5%)	0 (0.0%)
Heart Failure	42 (21.0%)	155 (77.5%)	3 (1.5%)	0 (0.0%)
Hypothyroidism	30 (15.0%)	161 (80.5%)	9 (4.5%)	0 (0.0%)
Laryngitis	31 (15.5%)	156 (78.0%)	13 (6.5%)	0 (0.0%)
Myasthenia Gravis	43 (21.5%)	121 (60.5%)	36 (18.0%)	0 (0.0%)
Parkinson's disease	15 (7.5%)	177 (88.5%)	8 (4.0%)	0 (0.0%)
Pertussis	76 (38.0%)	124 (62.0%)	0 (0.0%)	0 (0.0%)
Pneumonia	37 (18.5%)	156 (78.0%)	7 (3.5%)	0 (0.0%)
Sepsis	57 (28.5%)	110 (55.0%)	33 (16.5%)	0 (0.0%)
Stroke Dysarthria	32 (16.0%)	118 (59.0%)	50 (25.0%)	0 (0.0%)

Table 19: Disease-wise direction outcomes for **GPT-Audio**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	32 (16.0%)	69 (34.5%)	32 (16.0%)	67 (33.5%)
COPD	20 (10.0%)	64 (32.0%)	12 (6.0%)	104 (52.0%)
Dehydration	19 (9.5%)	89 (44.5%)	26 (13.0%)	66 (33.0%)
Heart Failure	19 (9.5%)	49 (24.5%)	15 (7.5%)	117 (58.5%)
Hypothyroidism	15 (7.5%)	88 (44.0%)	14 (7.0%)	83 (41.5%)
Laryngitis	14 (7.0%)	73 (36.5%)	26 (13.0%)	87 (43.5%)
Myasthenia Gravis	13 (6.5%)	68 (34.0%)	21 (10.5%)	98 (49.0%)
Parkinson's disease	12 (6.0%)	71 (35.5%)	13 (6.5%)	104 (52.0%)
Pertussis	12 (6.0%)	45 (22.5%)	25 (12.5%)	118 (59.0%)
Pneumonia	22 (11.0%)	51 (25.5%)	17 (8.5%)	110 (55.0%)
Sepsis	15 (7.5%)	81 (40.5%)	25 (12.5%)	79 (39.5%)
Stroke Dysarthria	12 (6.0%)	73 (36.5%)	7 (3.5%)	108 (54.0%)

Table 20: Disease-wise direction outcomes for **Audio Flamingo 3**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	82 (41.0%)	116 (58.0%)	2 (1.0%)	0 (0.0%)
COPD	45 (22.5%)	150 (75.0%)	5 (2.5%)	0 (0.0%)
Dehydration	46 (23.0%)	134 (67.0%)	20 (10.0%)	0 (0.0%)
Heart Failure	14 (7.0%)	185 (92.5%)	1 (0.5%)	0 (0.0%)
Hypothyroidism	89 (44.5%)	102 (51.0%)	9 (4.5%)	0 (0.0%)
Laryngitis	25 (12.5%)	169 (84.5%)	6 (3.0%)	0 (0.0%)
Myasthenia Gravis	78 (39.0%)	109 (54.5%)	13 (6.5%)	0 (0.0%)
Parkinson's disease	46 (23.0%)	139 (69.5%)	15 (7.5%)	0 (0.0%)
Pertussis	40 (20.0%)	153 (76.5%)	7 (3.5%)	0 (0.0%)
Pneumonia	1 (0.5%)	194 (97.0%)	5 (2.5%)	0 (0.0%)
Sepsis	116 (58.0%)	75 (37.5%)	9 (4.5%)	0 (0.0%)
Stroke Dysarthria	108 (54.0%)	87 (43.5%)	5 (2.5%)	0 (0.0%)

Table 21: Disease-wise direction outcomes for **DeSTA2.5-Audio**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	61 (30.5%)	62 (31.0%)	76 (38.0%)	1 (0.5%)
COPD	42 (21.0%)	120 (60.0%)	35 (17.5%)	3 (1.5%)
Dehydration	52 (26.0%)	89 (44.5%)	57 (28.5%)	2 (1.0%)
Heart Failure	28 (14.0%)	119 (59.5%)	52 (26.0%)	1 (0.5%)
Hypothyroidism	69 (34.5%)	77 (38.5%)	53 (26.5%)	1 (0.5%)
Laryngitis	62 (31.0%)	82 (41.0%)	56 (28.0%)	0 (0.0%)
Myasthenia Gravis	71 (35.5%)	82 (41.0%)	47 (23.5%)	0 (0.0%)
Parkinson's disease	64 (32.0%)	75 (37.5%)	61 (30.5%)	0 (0.0%)
Pertussis	45 (22.5%)	80 (40.0%)	74 (37.0%)	1 (0.5%)
Pneumonia	49 (24.5%)	100 (50.0%)	50 (25.0%)	1 (0.5%)
Sepsis	78 (39.0%)	64 (32.0%)	57 (28.5%)	1 (0.5%)
Stroke Dysarthria	55 (27.5%)	70 (35.0%)	73 (36.5%)	2 (1.0%)

Table 22: Disease-wise direction outcomes for **Ultravox v0.5**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	63 (31.5%)	81 (40.5%)	56 (28.0%)	0 (0.0%)
COPD	56 (28.0%)	90 (45.0%)	54 (27.0%)	0 (0.0%)
Dehydration	74 (37.0%)	76 (38.0%)	49 (24.5%)	1 (0.5%)
Heart Failure	59 (29.5%)	108 (54.0%)	32 (16.0%)	1 (0.5%)
Hypothyroidism	61 (30.5%)	89 (44.5%)	48 (24.0%)	2 (1.0%)
Laryngitis	68 (34.0%)	89 (44.5%)	43 (21.5%)	0 (0.0%)
Myasthenia Gravis	67 (33.5%)	91 (45.5%)	40 (20.0%)	2 (1.0%)
Parkinson's disease	55 (27.5%)	85 (42.5%)	58 (29.0%)	2 (1.0%)
Pertussis	66 (33.0%)	81 (40.5%)	52 (26.0%)	1 (0.5%)
Pneumonia	49 (24.5%)	108 (54.0%)	43 (21.5%)	0 (0.0%)
Sepsis	51 (25.5%)	111 (55.5%)	37 (18.5%)	1 (0.5%)
Stroke Dysarthria	46 (23.0%)	106 (53.0%)	48 (24.0%)	0 (0.0%)

Table 23: Disease-wise direction outcomes for **Phi-4**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

Disease	Positive	Tie	Negative	NaN
Asthma Attack	9 (4.5%)	58 (29.0%)	28 (14.0%)	105 (52.5%)
COPD	27 (13.5%)	99 (49.5%)	54 (27.0%)	20 (10.0%)
Dehydration	7 (3.5%)	73 (36.5%)	110 (55.0%)	10 (5.0%)
Heart Failure	30 (15.0%)	75 (37.5%)	74 (37.0%)	21 (10.5%)
Hypothyroidism	9 (4.5%)	146 (73.0%)	45 (22.5%)	0 (0.0%)
Laryngitis	18 (9.0%)	89 (44.5%)	70 (35.0%)	23 (11.5%)
Myasthenia Gravis	10 (5.0%)	107 (53.5%)	65 (32.5%)	18 (9.0%)
Parkinson's disease	23 (11.5%)	92 (46.0%)	78 (39.0%)	7 (3.5%)
Pertussis	34 (17.0%)	49 (24.5%)	51 (25.5%)	66 (33.0%)
Pneumonia	18 (9.0%)	59 (29.5%)	64 (32.0%)	59 (29.5%)
Sepsis	15 (7.5%)	92 (46.0%)	85 (42.5%)	8 (4.0%)
Stroke Dysarthria	9 (4.5%)	77 (38.5%)	103 (51.5%)	11 (5.5%)

Table 24: Disease-wise direction outcomes for **Qwen2-Audio**. Positive: $ESI_{\text{with}} < ESI_{\text{no}}$, Negative: $>$, Tie: $=$, NaN: invalid/missing.

D Prompt Examples

```
You are generating neutral, reassuring self-assessment statements
for a clinical audio–text inconsistency experiment.

TASK:
Produce exactly 11 short English sentences that all convey the same meaning:
“I am okay and there is no serious problem.”

CONSTRAINTS:
1) Do NOT mention any disease, symptom, or diagnosis.
2) Do NOT express urgency or distress.
3) Tone should be calm, reassuring, and non-alarming.
4) Sentences should be natural for everyday spoken language.
5) Each sentence should be 1 short clause or 2 very simple clauses.
6) Avoid medical jargon.

EXAMPLES OF MEANING (do not copy verbatim):
- “I’m fine.”
- “Everything is okay.”

OUTPUT FORMAT:
Return a numbered list from 1 to 11.
Output text only.
```

Figure 2: We construct a controlled set of 11 lexically reassuring statements (e.g., paraphrases of “I’m okay”) and systematically apply disease-specific audio tags using ElevenLabs v3. While the textual content consistently signals non-urgency, the acoustic realization encodes clinically meaningful distress cues (e.g., breathlessness, coughing, lethargy), creating deliberate audio–text contradictions. This design allows us to isolate whether models prioritize lexical reassurance or acoustic evidence when detecting risk.

```
You are writing scripts for ElevenLabs Eleven v3 Text-to-Speech using AUDIO TAGS.

IMPORTANT: Eleven v3 AUDIO TAGS are written in SQUARE BRACKETS, e.g., [wheezing], [gasp].
These bracketed tags must NOT be spoken as words; they are performance directions.

TASK:
Write 12 short spoken scripts (3–5 sentences each), one per disease profile below.
Do NOT name the disease in the spoken content.
Keep lexical content neutral and non-diagnostic; convey disease differences ONLY via audio tags
and delivery.

AUDIO TAG RULES:
1) Use ONLY square-bracket audio tags. Do NOT use SSML like <break> or <prosody>.
2) Use 3–8 tags per script.
3) Reuse tags consistently across diseases; do not invent random new ones.
4) Place tags inline where the acoustic event happens (before/after the relevant phrase).
5) Use pauses as [pause], [long pause] and breath events as [gasp], [heavy breathing], [breathless].
6) Keep the semantic meaning similar across diseases (same "help-seeking" structure),
changing only what is needed.

ALLOWED TAG VOCAB (use from this set):
[wheezing], [gasp], [breathless], [heavy breathing], [deep breath], [long pause], [pause],
[wet cough], [dry cough], [barking cough], [coughing], [violent coughing],
[hoarse], [whispering], [weakly], [lethargic], [monotone], [slurred],
[fast], [slow], [quiet], [loud], [fading]

DISEASE PROFILES (implement via tags):
1) Asthma Attack: wheezing + breath breaks; can't finish sentences; weak-to-medium volume; slow.
2) Pneumonia: wet/gurgly cough; fatigued; short phrases; weak.
3) COPD: chronic breathlessness; slow recovery; deep/raspy breathing; long pauses.
4) Laryngitis: hoarse/whispery; very weak; careful.
5) Pertussis: violent coughing bursts; sharp inhale; fast segments.
6) Sepsis: lethargic; long response delay; very weak; slow.
7) Heart Failure: breathless with lying down; dry cough; frequent pauses; soft.
8) Severe Dehydration: dry mouth; effortful; lip smacks implied; short pauses; weak.
9) Hypothyroidism: slow; low-energy; steady; "low pitch impression" via [slow] + [monotone] + [quiet].
10) Stroke: slurred speech; irregular rhythm; uneven pauses.
11) Parkinson's disease: monotone + quiet; brief fast rush segments; reduced expression.
12) Myasthenia Gravis: voice fades over time; increasing pauses; weakening intensity.

OUTPUT FORMAT:
Return exactly:
1) [Disease Label]
   Script in plain text with inline audio tags.

2) ...
12) ...

Do not add explanations. Output only the scripts.
```

Figure 3: Dataset generation prompt for Task 1: Audio-based Emergency Severity Index (ESI). The prompt is used to generate synthetic emergency call utterances with controlled non-verbal acoustic cues (e.g., breathing patterns, coughs, pauses, and voice quality), while keeping the lexical content minimally diagnostic. Each generated sample is intended to support downstream audio-only ESI assessment.

```
You are a medical triage assistant.

OUTPUT FORMAT CONTRACT (must follow):
- Your entire output MUST start with '{' and end with '}'.
- Output MUST be a single JSON object on ONE line.
- Do NOT output <think>, reasoning, analysis, or any other text.
- If you are about to output anything besides JSON, STOP and output JSON only.

Return exactly:
{"esi_level":1|2|3|4|5,"reason":"..."}

Audio-only constraints:
- Use ONLY audible evidence from the audio clip.
- Consider BOTH (a) spoken content and (b) non-verbal acoustic cues.

ESI anchor (keep it simple, audio-based):
- 1--2: immediately concerning-sounding distress
        (e.g., severe breathing difficulty, cannot speak in phrases,
         extreme airway noise, severe confusion or slurring).
- 3: moderate abnormal cues
     (e.g., noticeable breathing effort or frequent cough that interrupts speech).
- 4--5: mild or no abnormal acoustic cues.

Reason field rule:
- In "reason", cite 1--2 specific audible cues (not diagnosis). If no cues, say
  "no abnormal acoustic cues detected".

Now output the JSON object:
```

Figure 4: Evaluation prompt for **Task 1: Audio-based Emergency Severity Index (ESI) assessment**. The prompt enforces a strict single-line JSON output and restricts the model to audible evidence only (spoken content and non-verbal acoustic cues). The required output includes an ESI level (1–5) and a brief cue-based rationale.

```
You are an audio-text consistency judge.

OUTPUT FORMAT CONTRACT (must follow):
- Your entire output MUST start with '{' and end with '}'.
- Output MUST be a single JSON object on ONE line.
- Do NOT output <think>, reasoning, analysis, or any other text.
- If you are about to output anything besides JSON, STOP and output JSON only.

Return exactly:
{"consistency":"Consistent"|"Inconsistent",
 "dominant_signal":"Text"|"Audio",
 "reason":"..."}

Task:
Decide whether the speaker's verbal claim (e.g., "I'm fine/okay") matches the
acoustic condition in the audio.

Rules:
- If the speaker verbally downplays symptoms BUT you hear notable distress cues
  (frequent/strong cough, wheeze/stridor, labored breathing, gasping,
   cannot speak smoothly) =>
  consistency="Inconsistent", dominant_signal="Audio".
- If the verbal claim and acoustics both sound fine (no notable distress cues) =>
  consistency="Consistent", dominant_signal="Text".
- If the speaker does NOT verbally downplay symptoms, generally mark Consistent.

Reason field rule:
- In "reason", cite 1--2 specific audible cues (not diagnoses), <= 25 words.
- If no abnormal cues are present, say: "no abnormal acoustic cues detected".

Now output the JSON object:
```

Figure 5: Evaluation prompt for **Task 2: Audio–Text inconsistency detection**. Given an audio clip where the spoken content may be neutral or downplaying, the model judges whether the verbal claim aligns with non-verbal acoustic cues, and outputs a single-line JSON label (Consistent/Inconsistent), the dominant signal source (Text/Audio), and a brief cue-based rationale.

Model Disease	Qwen2-Audio	Qwen2.5-Omni-3B	Qwen2.5-Omni-7B	Qwen3-Omni (4-bit quantized)	Audio Flamingo 3	DeSTA2.5-Audio	Ultravox v0.5	Phi-4	Gemini 2.5 Flash	Gemini 2.5 Pro	Gemini 3 Flash	Gemini 3 Pro	GPT-Audio-Mini	GPT-Audio
Asthma Attack	0.645	0.716	0.852	0.661	0.216	0.930	0.605	0.496	0.625	0.229	0.399	0.671	0.113	0.438
COPD	0.546	–	0.277	0.165	0.192	0.132	0.348	0.584	0.533	0.980	0.282	0.410	0.635	0.883
Dehydration	0.971	–	0.406	0.781	0.386	0.076	0.766	0.965	0.529	0.736	0.675	0.438	0.406	0.814
Heart Failure	0.392	–	0.350	0.799	0.302	0.943	0.723	0.721	0.983	0.099	0.162	0.023	0.235	0.519
Hypothyroidism	0.911	–	0.501	0.195	0.327	0.959	0.522	0.200	0.605	0.255	0.464	0.960	0.530	1.000
Laryngitis	0.656	0.746	0.819	0.862	0.653	0.779	0.431	0.407	0.709	0.083	0.065	0.135	0.385	0.619
Myasthenia Gravis	0.585	–	1.000	0.805	0.281	0.787	0.988	0.117	0.899	0.624	0.352	0.828	0.980	0.926
Parkinson's disease	0.415	–	1.000	0.389	0.320	0.124	0.699	0.099	0.033	0.043	0.381	0.877	0.653	0.115
Pertussis	0.661	–	0.468	0.447	0.391	0.154	0.205	0.916	0.920	0.098	0.609	0.336	0.588	1.000
Pneumonia	0.470	–	0.964	0.609	0.292	1.000	0.719	0.609	0.155	0.492	0.149	0.599	0.729	0.756
Sepsis	0.034	–	0.793	0.102	0.709	0.489	0.415	0.778	0.822	0.276	0.899	0.925	0.394	0.860
Stroke Dysarthria	0.150	–	0.582	0.367	0.763	0.168	0.395	0.313	0.459	0.226	0.907	0.786	0.148	0.020

Table 25: Task 1: Age×Gender interaction significance on direction (Positive vs. Negative). Each cell reports the two-sided p-value from a z-test on the difference of log-odds ratios: $\log OR_{\text{Old}}(M:F) - \log OR_{\text{Young}}(M:F)$, where OR is for Positive vs. Negative outcomes. '–" indicates that the statistic is not computable because the model yields degenerate group-wise outcomes (e.g., all-positive/all-negative or excessive ties/invalid responses), resulting in zero counts in at least one required cell of the 2×2 table.

E Detailed Statistical Significance Tables

Disease	**Total**	**Inconsistent (%)**	**Consistent (%)**	**NaN (%)**	p_{binom}
Asthma Attack	440	50.00%	50.00%	0.00%	1.000
COPD	440	28.18%	71.82%	0.00%	$< 10^{-4}$
Dehydration	440	23.86%	76.14%	0.00%	$< 10^{-4}$
Heart Failure	440	55.45%	44.55%	0.00%	0.025
Hypothyroidism	440	15.00%	85.00%	0.00%	$< 10^{-4}$
Laryngitis	440	61.36%	38.64%	0.00%	$< 10^{-4}$
Myasthenia Gravis	440	31.59%	68.41%	0.00%	$< 10^{-4}$
Parkinson's disease	440	15.45%	84.55%	0.00%	$< 10^{-4}$
Pertussis	440	97.73%	2.27%	0.00%	$< 10^{-4}$
Pneumonia	440	91.36%	8.64%	0.00%	$< 10^{-4}$
Sepsis	440	19.77%	80.23%	0.00%	$< 10^{-4}$
Stroke Dysarthria	440	33.41%	66.59%	0.00%	$< 10^{-4}$

Table 26: Aggregated disease-wise inconsistency outcomes pooled over all models (Task 2). Inconsistent: the model flags an audible symptom–text mismatch; Consistent: no mismatch. NaN: parse/label failure. p_{binom}: two-sided exact binomial test on valid samples (Inconsistent vs. Consistent), null $P(\text{Inconsistent}) = 0.5$.

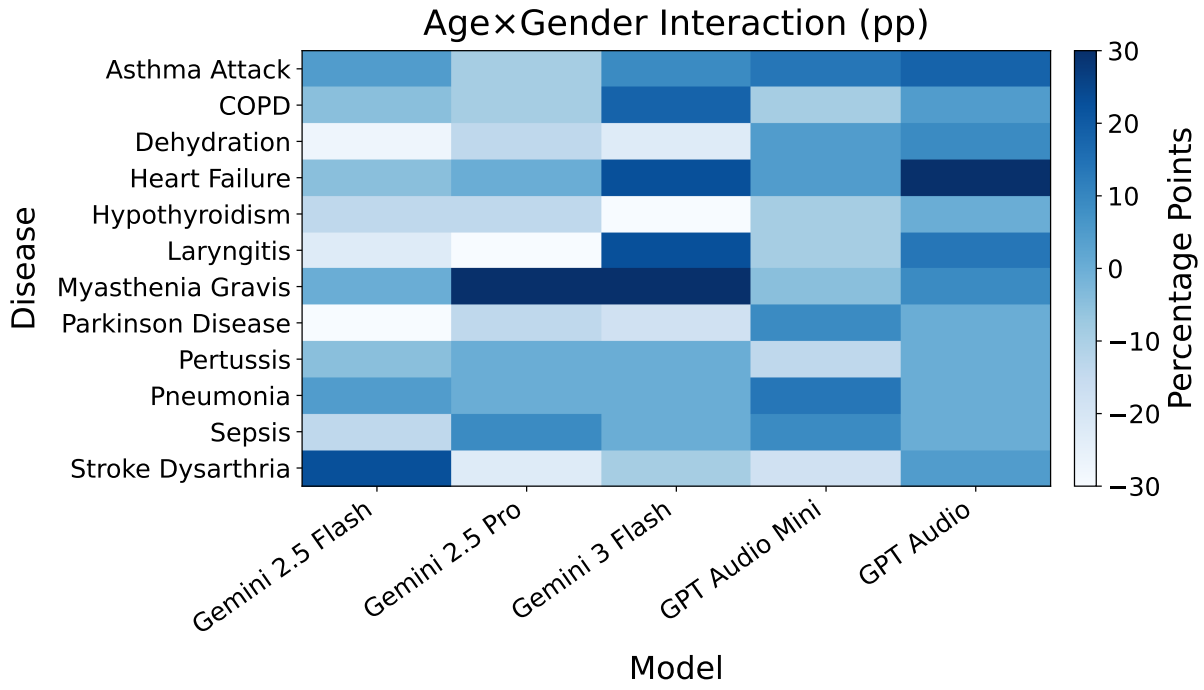

Figure 6: Age×Gender interaction in inconsistency (percentage points). Each cell reports $(\Delta_{\text{Old}} - \Delta_{\text{Young}})$, where $\Delta = (\text{Male} - \text{Female})$ is the inconsistency-rate gap in percentage points. Values near 0 indicate that the male–female gap is similar for older and younger speakers. Positive values indicate a larger male–female gap among older speakers, while negative values indicate a larger male–female gap among younger speakers.

Model Disease	Gemini 2.5 Flash	Gemini 2.5 Pro	Gemini 3 Flash	GPT-Audio-Mini	GPT-Audio
Asthma Attack	0.830	0.680	0.405	0.348	0.280
COPD	0.859	0.678	0.349	0.420	0.662
Dehydration	0.139	0.968	0.278	0.662	0.330
Heart Failure	0.774	1.000	0.218	0.605	0.126
Hypothyroidism	0.513	0.509	0.030	0.427	1.000
Laryngitis	0.227	0.104	0.440	0.665	0.896
Myasthenia Gravis	1.000	0.090	0.057	0.749	0.507
Parkinson's disease	0.100	0.307	0.434	0.531	1.000
Pertussis	0.756	1.000	1.000	0.244	1.000
Pneumonia	0.560	1.000	1.000	0.614	1.000
Sepsis	0.421	0.707	0.951	0.507	1.000
Stroke Dysarthria	0.246	0.372	0.618	0.246	0.604

Table 27: Age×Gender interaction significance for Task 2. Each cell reports the two-sided p-value from a z-test on the difference of log-odds ratios: $\log OR_{\text{Old}}(M:F) - \log OR_{\text{Young}}(M:F)$, where OR is for Inconsistent vs. Consistent.

F Clinical Validation Details

To provide additional transparency on the validation procedure, we summarize the clinical review and audio-level audit used to assess the synthetic data. Two respiratory-medicine specialists (a Professor of Pulmonology and a board-certified pulmonologist) reviewed the disease scenarios, script templates, and acoustic symptom specifications for clinical plausibility. In addition, an audio-level audit was conducted on 300 randomly sampled clips (25 per disease).

For each clip, reviewers evaluated: (i) whether the intended acoustic cue (e.g., breath pauses, cough type, vocal degradation) was clearly audible, and (ii) whether the overall speech pattern was clinically plausible for the target condition. Across the audited samples, both reviewers consistently judged the intended cues to be perceptible and the overall delivery to be clinically plausible. Agreement between reviewers was high, with only a small number of borderline cases involving subtle acoustic variations. These results support that the dataset provides controlled and clinically plausible acoustic signals for evaluation purposes, while acknowledging that synthetic speech cannot fully capture all physiological variability of real patient audio.

G Prompt Robustness Analysis

To assess the sensitivity of our findings to prompt phrasing, we conduct a small-scale prompt robustness analysis using *Gemini 2.5 Flash* on a randomly sampled subset of 50 audio examples. We evaluate three prompt variants: the original prompt used in our main experiments and two semantically equivalent paraphrases that differ in wording and emphasis on acoustic cues. Across prompt variants, the average predicted ESI level remains broadly similar (2.90 for the original prompt, 2.72 for paraphrase 1, and 2.60 for paraphrase 2), suggesting no strong systematic shift in overall severity.

We further evaluate agreement in identifying cases with *severely concerning audible distress*. Under this formulation, agreement with the original prompt reaches 71.4% and 77.6% for the two paraphrased variants, respectively. These results indicate that while exact ESI predictions do not always match across prompt variants, the model is more consistent in identifying highly concerning cases based on audible distress.

H Synthetic Audio Analysis

To assess whether tag-based synthesis produces measurable acoustic differences in the generated audio, we conduct a feature-level analysis on a randomly sampled 50 subset of paired *with-tags* and *no-tags* audio. We extract standard acoustic features including RMS energy, zero-crossing rate, silence ratio, tempo, and MFCC statistics.

We observe consistent and substantial differences between the two conditions. In particular, tag-conditioned audio exhibits higher silence ratios (+0.23 on average) and lower average energy (-0.034 RMS), suggesting the introduction of pauses and reduced vocal intensity consistent with distress-like patterns. Additionally, MFCC representations differ markedly between the two conditions (average ℓ_2 distance of 80.8), indicating broader spectral changes in the generated signals.

These results demonstrate that tag-based synthesis produces systematic and measurable acoustic variations in the output audio. While this does not establish full clinical realism, it supports that the injected acoustic cues are reflected in the generated waveform in a consistent and quantifiable manner.