Hearing Health in Home Healthcare: Leveraging LLMs for Illness Scoring and ALMs for Vocal Biomarker Extraction

Yu-Wen Chen[1*] **William Ho**[1*] **Sasha M. Vergez**[3] **Grace Flaherty**[3]
Pallavi Gupta[2] **Zhihong Zhang**[2] **Maryam Zolnoori**[2]
Margaret V. McDonald[3] **Maxim Topaz**[2] **Zoran Kostic**[1] **Julia Hirschberg**[1]
[1]The Fu Foundation School of Engineering and Applied Science, Columbia University
[2]School of Nursing, Columbia University
[3]Center for Home Care Policy & Research, VNS Health

Abstract

The growing demand for home healthcare calls for tools that can support care delivery. In this study, we explore automatic health assessment from voice using real-world home care visit data, leveraging the diverse patient information it contains. First, we utilize Large Language Models (LLMs) to integrate Subjective, Objective, Assessment, and Plan (SOAP) notes derived from unstructured audio transcripts and structured vital signs into a holistic illness score that reflects a patient's overall health. This compact representation facilitates cross-visit health status comparisons and downstream analysis. Next, we design a multi-stage preprocessing pipeline to extract short speech segments from target speakers in home care recordings for acoustic analysis. We then employ an Audio Language Model (ALM) to produce plain-language descriptions of vocal biomarkers and examine their association with individuals' health status. Our experimental results benchmark both commercial and open-source LLMs in estimating illness scores, demonstrating their alignment with actual clinical outcomes, and revealing that SOAP notes are substantially more informative than vital signs. Building on the illness scores, we provide the first evidence that ALMs can identify health-related acoustic patterns from home care recordings and present them in a human-readable form. Together, these findings highlight the potential of LLMs and ALMs to harness heterogeneous in-home visit data for better patient monitoring and care.

1 Introduction

The global demand for home healthcare highlights the urgent need for improved tools and interventions to enhance care delivery. Automatic health assessment systems have the potential to assist clinicians by identifying aspects that may be overlooked during visits and support individuals with limited access to clinical services. Recently, Large Language Models (LLMs) have undergone development at an unprecedented pace, demonstrating contextual understanding, reasoning, and task generalization across diverse domains [4, 13, 35]. These advancements are increasingly being leveraged in the field of healthcare, with LLMs being applied to tasks such as summarizing and extracting structured information from clinical notes, answering medical questions, and facilitating patient-provider communication [2, 31, 33, 36, 37]. However, to the best of our knowledge, no prior study has explored the use of LLMs to interpret and assess a patient's overall health condition from

*These authors contributed equally. {yu-wen.chen, william.ho}@columbia.edu.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

real-world home care visits. Such data is inherently heterogeneous and complex, comprising both structured and unstructured information [42]. For instance, Subjective, Objective, Assessment, and Plan (SOAP) notes [28] capture diverse aspects of a patient's condition in narrative form, while vital signs [29] consist of multiple numerical indicators reflecting distinct physiological states.

Recognizing the strength of LLMs in generating human-readable text and leveraging extensive textual knowledge, researchers have extended LLMs into Audio-Language Models (ALMs) [10, 15, 16, 32]. ALMs inherit LLMs' abilities to generate interpretable descriptions and knowledge of human voices, while also possessing the capability to extract meaningful information directly from audio signals. While ALMs have demonstrated strong performance in areas such as audio captioning and sound event classification [12], their applications in health assessment are largely underexplored – despite growing evidence that acoustic signals can provide valuable indicators of health conditions [3], such as respiratory symptoms [38] and depression [19].

In this study, we explore the performance of LLMs in assessing real-world home care visit data. We address the challenges of comparing heterogeneous data across visits by consolidating them into a single, holistic illness severity score (denoted as illness score) that reflects a patient's overall health status for each visit. This approach is motivated by evidence that LLMs encode extensive medical knowledge and can reason effectively about clinical tasks [2, 18, 21, 22, 30], and have the potential as automatic evaluators, generating pseudo-labels that offer a scalable and resource-efficient alternative to human assessment [7, 9, 20, 23, 24, 40]. While not a replacement for detailed clinical documentation, this unified score provides a more compact and comparable representation of health status, enabling downstream analyses such as investigating associations of vocal biomarkers with patient health.

We then apply ALMs to analyze acoustic signals from real-world clinician-patient conversations, investigating whether the differences in vocal biomarkers detected by ALMs correspond to patients' overall health conditions. Specifically, we perform multi-stage processing to extract short speech segments of the target speaker from the full conversation recording and then prompt an ALM to describe the input speech in terms of various vocal biomarkers, such as voice quality, pronunciation, and energy level. In contrast to commonly used speech features such as speech embeddings from acoustic models, which cannot be directly interpreted by humans, the ALM's ability to generate plain-language descriptions provides clinicians with more accessible insights into a patient's condition.

The contributions of this study are summarized as follows:

1. Using data from real-world home care visits, our study captures authentic patient–clinician interactions to examine whether, and how, characteristics of a person's voice are correlated with their overall health condition.
2. We propose using LLMs to integrate SOAP notes and vital signs into a compact numeric health score, serving as pseudo-labels that enable more effective and quantitatively driven downstream analyses. Experimental results demonstrate LLMs' capabilities to generate reliable illness scores, which show strong alignment with actual patient clinical outcomes.
3. To the best of our knowledge, we have conducted the first study employing ALMs to assess human health through speech signals. We show that ALMs capture acoustic patterns reflecting changes in health status and provide interpretable descriptions that facilitate clearer understanding for clinicians.

2 Analysis Framework

Our goal in this study is to investigate the correlation between people's voices and their health conditions using real-world home care visit data. First, we propose using an LLM to generate a numeric illness score to summarize the patient's overall health status based on both SOAP notes from clinician-patient conversation transcript and recorded vital signs. The numeric illness score functions as a pseudo-label, providing a quantitative representation of each visit's health status, making it easier to compare patient conditions over time or across individuals. Next, we extract short speech signals from a designated speaker in the conversation. We then apply an ALM to generate interpretable acoustic descriptions from the speech signals. Finally, we examine the relationship between these acoustic descriptions and the illness scores. Figure 1 outlines the analysis framework for this study.

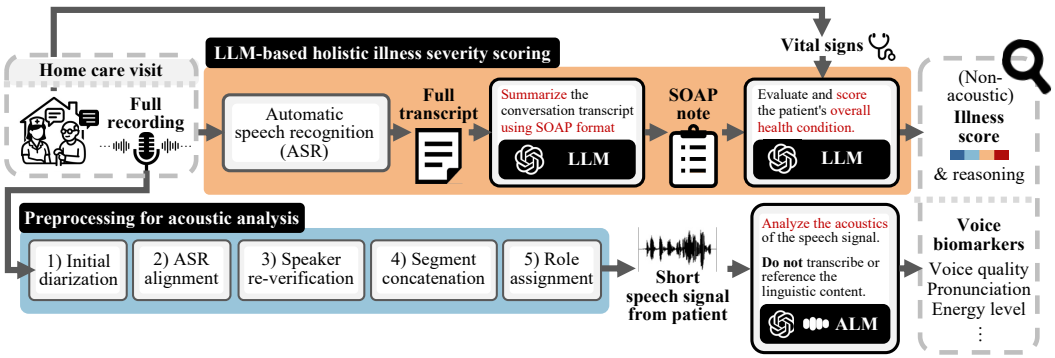


Figure 1: Overview of our analysis framework. Home care visit audio recordings are processed through two pipelines. **Top**: An LLM summarizes ASR transcripts into SOAP notes and combines them with vital signs to produce an illness score. **Bottom**: Preprocessed patient speech segments are analyzed by an ALM to generate acoustic descriptors, which are correlated with the illness score.

2.1 Real-World Home care Visit Data

The data utilized in this study was collected from real-world home care visits. Participants were enrolled in a home care service, either following hospitalization or through community referral, in which clinicians conducted home visits and recorded their interactions and conversations with the patients. All participants were instructed to maintain their usual interaction patterns and to disregard the presence of recording equipment to minimize behavior alterations. Data collected during each visit includes full audio recordings and patient's vital signs. Audio recordings were made using wireless microphones attached to both the clinician and the patient at the start of the visit. Each vital sign – including body mass index (BMI), blood pressure (diastolic and systolic), height, weight, pulse, respiration rate, temperature, oxygen saturation, blood sugar level, and pain level – may be measured once, multiple times, or not at all, depending on the clinician's judgment.

2.2 LLM for Holistic Illness Severity Scoring

We apply an automatic speech recognition (ASR) system to generate transcripts from the full audio recordings. The transcripts are then processed by an LLM to produce SOAP notes [8, 28]. The SOAP notes, combined with the patient's vital signs, are then fed to the LLM to generate an illness score along with a rationale for the score. Scoring ranges from 1 to 5, with 1 indicating good overall health and 5 indicating a critical condition for which hospitalization is recommended. We evaluate our illness scores by examining their alignment with real-world patient outcomes. Health conditions are inherently complex, making it challenging to obtain a definitive ground-truth score that represents the overall health status. One of the best available indicators of a home care patient's overall health is whether they were hospitalized or required an emergency department (ED) visit. Accordingly, we examine whether illness scores generated by LLMs are associated with patients' risk of being hospitalized or visiting the ED within a 60-day home care period. We categorize patients into two groups: 1) those with at least one ED visit or hospitalization during this period (denoted as ED/HOSP), and 2) those without (denoted as No ED/HOSP). For each patient, we calculate the average illness score, then compare these averages between the two groups. Experimental results (Section 4.1) show that patients who were hospitalized or visited the ED during the period had notably higher illness scores on average, indicating that the score can reflect overall health conditions. Therefore, we use these scores as health indicators in our acoustic analysis, serving as pseudo-labels to group patients by health severity and enabling comparisons of vocal characteristics across different health levels.

2.3 ALM for Vocal Biomarker Extraction

2.3.1 Preprocessing for Acoustic Analysis

To analyze each speaker's acoustic characteristics, we locate and extract their speech segments from raw multi-speaker recordings. Because our recordings were collected in real-world home

care environments, variability from third-party speakers and background noise often cause current diarization models to mis-assign speaker identities. We therefore design a multi-stage preprocessing pipeline to automatically exclude segments with ambiguous speaker labels or mixed voices.

Only the first 30 seconds of each processed audio recording are used for acoustic analysis, since the length of processed signals can vary depending on the duration and quality of the original recordings, a fixed short segment reduces variability and ensures more consistent and comparable analysis across visits. Because processing long inputs with ALMs is computationally expensive, we begin by evaluating shorter speech signals to determine if they can effectively capture useful acoustic patterns. Finally, limiting input duration constrains the available linguistic context, reducing the risk of ALMs relying on semantic information in ways that violate the instructions.

The multi-stage preprocessing pipeline includes:

1. **Initial diarization** assigns generic speaker labels to speech segments using a diarization model.
2. **ASR alignment** applies an ASR model to obtain word-level timestamps and map them to diarization results. Words are grouped into sentences based on end-of-sentence punctuation, and any sentences containing words with unassigned speakers are removed.
3. **Speaker re-verification** reduces possible speaker confusion by using speaker embeddings from speech-enhanced audio to confirm diarization labels. For each diarized (generic) speaker, we compute their average embedding from all associated segments. The cosine similarity is computed between each segment embedding and all speaker's average embeddings, assigning the most similar speaker to the segment. Segments whose newly inferred speaker deviates from the diarizer's label are discarded.
4. **Segment concatenation** merges verified speech segments into continuous speech with short silences inserted between them.
5. **Role assignment** leverages recognizable interaction patterns in clinician-patient conversations [41], e.g., clinicians asking patients condition-related questions, to classify each speaker as clinician, patient, or third party. An LLM performs the classification based on full transcripts, and manual inspection of a subset confirms high accuracy in role assignment.

2.3.2 Analysis of ALM Vocal Biomarkers and Health

For acoustic analysis, we evaluate the ALMs' ability to detect basic speaker traits, including gender and age, as baseline tasks to measure the model's capacity to extract speaker information not represented in the transcript. We then investigate whether the ALM's acoustic descriptions of *emotion*, *voice quality*, *pronunciation characteristics (fluency, prosody, articulation)*, *energy level*, and *potential signs of discomfort or fatigue* reflect variations in an individual's health condition. We then analyze speech signals from clinicians vs. patients based on the hypothesis that these two groups differ in overall health status. Finally, we explore how the ALM's acoustic descriptions of patient voices evolve with health status indicated by the illness scores.

3 Experimental Process

3.1 Data Collection and Statistics

The data used in this study were collected under Institutional Review Board (IRB) approval. Both patients and clinicians are compensated for their participation. All textual data were anonymized prior to our access. However, since audio recordings inherently reflect speaker identity information, all experiments were conducted within a secure, access-controlled computing environment to ensure data privacy and protection. Additionally, only open-source software or HIPAA-compliant commercial services (i.e., the HIPAA-compliant version of OpenAI's API) were used.

The dataset comprises single-channel (48 kHz) audio recordings from 160 home care visits (one recording per visit), ranging from approximately 6 to 70 minutes in duration, with an average length of about 30 minutes. At least one vital sign was recorded during each visit. The dataset includes 77 patients (23 male, 54 female, ages 20–99) and 9 clinicians, with no demographic information collected for clinicians. For each patient, we derived a binary outcome indicating whether an ED

visit or hospitalization occurred within 60 days of the start of home care – 20 patients experienced at least one such event. If a patient returned home after an ED visit or hospitalization during this period, clinicians could continue recording, and these post-event recordings were included. Although enrollment periods varied, all visit data was collected within the first 60 days after the start of care, spanning from June 16, 2024, to May 6, 2025.

3.2 Model settings

Illness scores and SOAP notes were generated independently for each visit. We compare the performance of *GPT-4.1* [1, 26] with default settings against open-source LLMs: *Qwen-3-8B* [39] and *Llama-3.1-8B-Instruct* [14], which serve as open-source alternatives to GPT-4.1, as well as *MediPhi-Instruct* [11] chosen for its specialized medical expertise. Detailed prompts are in Appendix A.

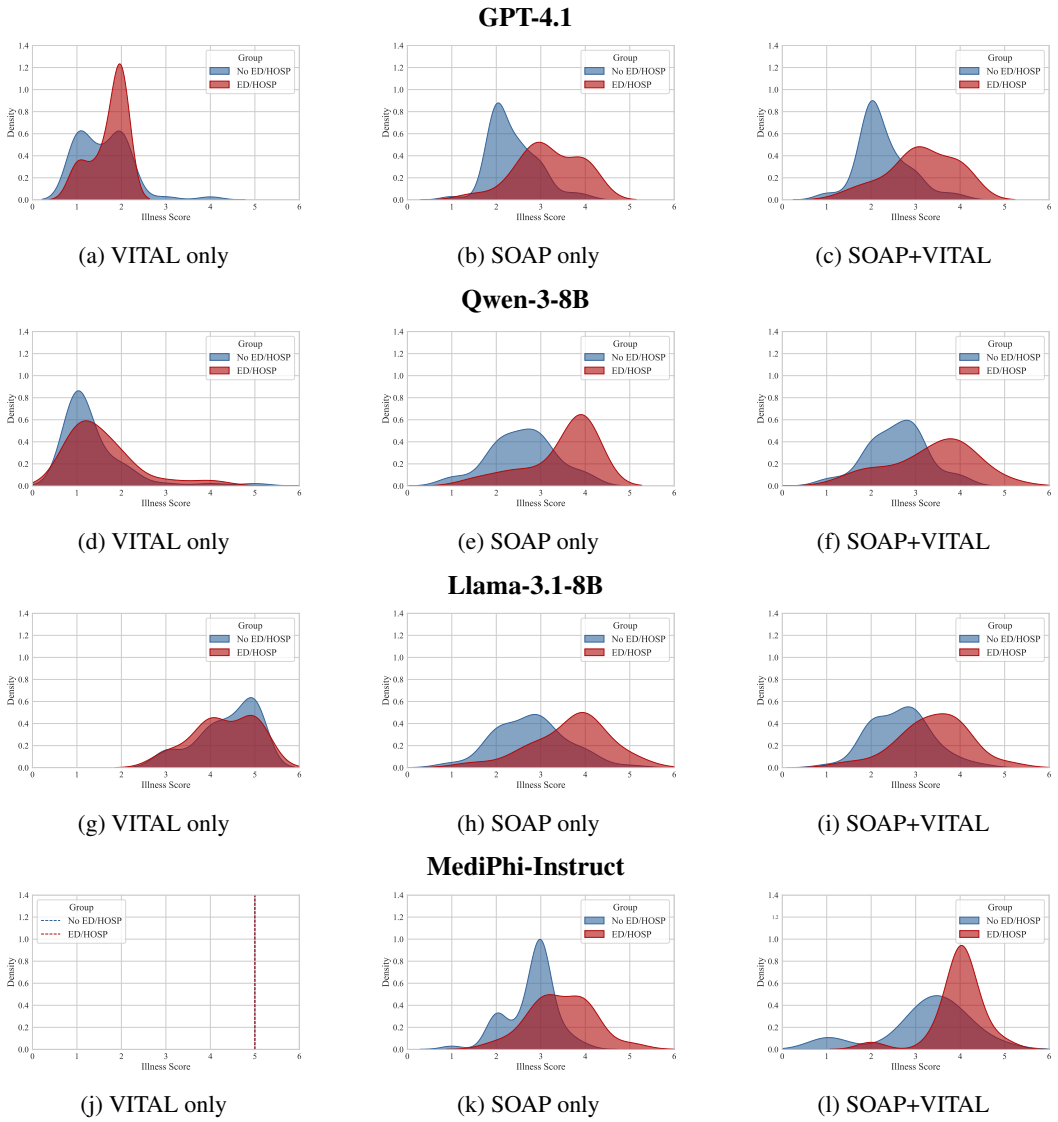
We used *nvidia/parakeet-tdt-0.6b-v2* [25] as the ASR model for both full recording transcription and the preprocessing for acoustic analysis. Because data primarily focuses on conversations between the patient and clinician, we selected *nvidia/diar_sortformer_4spk-v1* [27] as the initial diarization model due to its strong performance in few-speaker settings. However, since the full recordings are significantly longer than the training data, we split each recording into 250-second chunks with a 5-second overlap, and processed each chunk independently. Speakers were matched across chunks by extracting embeddings using the speaker verification model *nvidia/speakerverification_en_titanet_large* [17], which were then clustered into four groups via agglomerative hierarchical clustering. For speaker re-verification, SEMamba [5] with the *vd.pth* checkpoint was employed for speech enhancement, followed by verification using the same speaker verification model as above. Sentences that fail verification are discarded. Verified sentences longer than 0.5 seconds are concatenated with 0.5 seconds of silence inserted between them, and the first 30 seconds are retained for analysis. GPT-4.1 with default settings is used for role assignment in acoustic processing. The ASR, speech diarization, and speaker verification models used in this study are publicly available on Hugging Face. For ALMs, we utilized *gpt-4o-audio-preview* [16] configured with default settings and *Qwen2-Audio-7B-Instruct* [10].

4 Results

4.1 LLM-based Holistic Illness Severity Scoring

The performance of four LLMs was evaluated: GPT-4.1, Qwen3-8B, Llama3.1-8B, and MediPhi-Instruct. For each LLM, we conducted an ablation study using different input types (VITAL only, SOAP only, and SOAP+VITAL) to determine which information an LLM can most effectively use to assess overall health. Figure 2 shows Kernel Density Estimation (KDE) plots of resulting illness scores, with detailed summary statistics provided in the Appendix B.1. In all settings, ED/HOSP groups tend to have higher average illness scores compared to the No ED/HOSP group, except for MediPhi-Instruct and Llama3.1-8B with VITAL only. These results indicate that LLMs can effectively interpret SOAP notes, with some models also able to meaningfully incorporate vital signs, to generate scores aligning with the instruction that higher values correspond to more critical health conditions. We observe that illness scores derived from vital signs alone differ substantially from those generated with SOAP inputs (SOAP only and SOAP+VITAL). For VITAL only, the two groups tend to have similar distributions, reflecting that most patients' vital signs fall within normal ranges. The similarity between SOAP only and SOAP+VITAL suggests that the LLM's interpretation of SOAP notes is consistent, with vital signs providing subtle adjustments to the overall illness score. MediPhi-Instruct's results deviate more from the other LLMs. With VITAL only inputs, it assigns a uniform score of 5 to all cases, requiring additional information to generate a meaningful score.

To better compare the performance of each LLM, we calculate the Wasserstein distances (WD) [34] between the score distribution of the two groups (Figure 3). Higher WD indicates greater separation between the two groups and is therefore considered better performance. The results show that SOAP notes provide substantially more information for overall health assessment than vital signs, as VITAL only performed much lower across all LLMs. In addition, GPT-4.1 and MediPhi-Instruct benefit from including vital signs, with SOAP+VITAL inputs outperforming SOAP-only inputs. In contrast, Qwen3-8B and Llama3.1-8B perform worse when vital inputs are added, indicating that they cannot effectively integrate this information. Given that GPT-4.1 demonstrates the best performance on average and greater robustness with or without vital information, we selected its SOAP+VITAL results for the subsequent acoustic analysis.



(a) VITAL only (b) SOAP only (c) SOAP+VITAL

(d) VITAL only (e) SOAP only (f) SOAP+VITAL

(g) VITAL only (h) SOAP only (i) SOAP+VITAL

(j) VITAL only (k) SOAP only (l) SOAP+VITAL

Figure 2: KDE plots of generated illness score distributions of No ED/HOSP and ED/HOSP groups. Each row shows results for a specific LLM, while columns represent different inputs. Across most settings, ED/HOSP groups show higher average scores than No ED/HOSP groups, indicating alignment with clinical outcomes. SOAP-based inputs (SOAP only and SOAP+VITAL) show better group separation than VITAL only, with consistently higher scores for ED/HOSP patients.

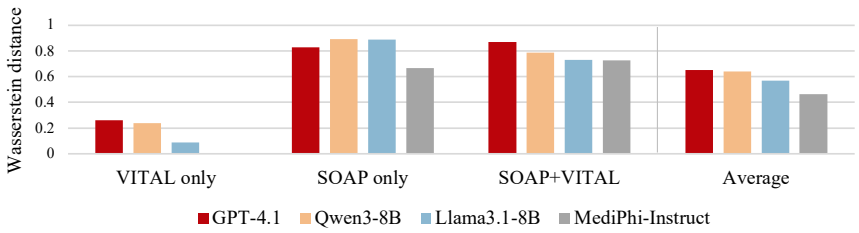


Figure 3: Wasserstein distances between illness score distributions of No ED/HOSP and ED/HOSP groups across four models. A higher Wasserstein distance indicates better separation between groups. SOAP notes provide substantially more discriminative information than vital signs alone. GPT-4.1 achieves the highest average performance and benefits from incorporating vital signs, while Qwen3-8B and Llama3.1-8B show degraded performance when vitals are added.

4.2 ALM for Speaker Age and Gender Classification

Figure 4 presents the results for GPT-4o-audio. For gender prediction, the model achieves an accuracy of 0.8625, with higher precision for female voices than for male. For age prediction, the model classifies speakers into categories: young adult, adult, middle-aged, middle-aged to older, and older adult. The predicted age categories largely correspond to the speakers' true ages. These results indicate that GPT-4o-audio can reliably capture speaker characteristics from audio signals, supporting more complex acoustic analyses, such as health assessment. Compared with GPT-4o-audio, GPT-4o-mini-audio shows a reduced ability in vocal biomarker extraction, e.g., achieving only 0.7484 gender prediction accuracy. Furthermore, the open-source Qwen2-Audio-7B-Instruct demonstrates intermediate performance, with an accuracy of 0.8064 on gender prediction. Given its stronger performance in extracting basic speaker traits, the following ALM analyses are based on GPT-4o-audio, while results for GPT-4o-mini and Qwen2-Audio-7B-Instruct are provided in Appendix C.

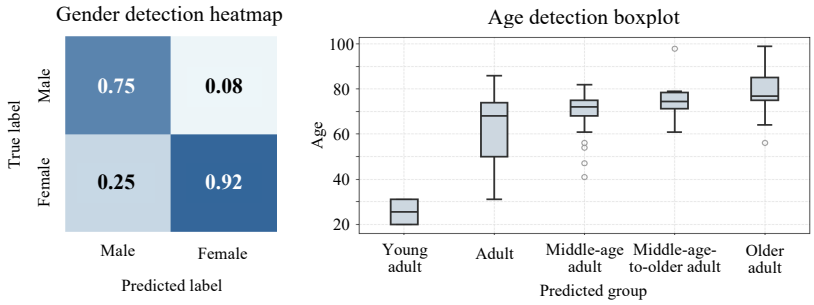


Figure 4: GPT-4o-audio performance on a baseline speaker classification tasks. **Left**: Confusion matrix showing GPT-4o-audio achieves 0.8625 accuracy on gender detection. **Right**: Age prediction boxplot showing predicted age categories that align well with true ages.

4.3 ALM Analysis of Clinician and Patient Voices

Figure 5 presents wordclouds of ALM-generated phrases describing various aspects of the speech signals from each group. The results show that the most commonly identified emotions for both clinicians and patients were neutral and calm in home care conversations. However, clinicians were more frequently described using positive emotions whereas patients' speech was more often associated with negative emotions like frustration. When examining voice quality, clinicians' voices were typically characterized as smooth and clear, consistent with professional speech patterns. Conversely, patients' voices were more likely to be described as raspy, breathy, or nasal, which may be indicative of underlying health conditions affecting their vocal quality. For pronunciation, both groups showed moderate pitch variation in prosody, but the patients' speech tended to be more monotone. For fluency, clinicians generally spoke in a consistent and smooth manner and patients tended to hesitate more often. For energy level, clinician speech is most often described as consistent or moderate. While these descriptors are also common for patients, their speech is also frequently characterized as low energy and soft. Lastly, indicators of discomfort and fatigue were more prominent in patients' speech. Overall, the ALM's descriptions of clinician voices align more closely with characteristics typically associated with healthier individuals, whereas descriptions of patient voices reflect traits indicative of poorer health. These results demonstrate that the ALM-generated vocal descriptions are interpretable and capable of distinguishing individuals with differing health conditions based on their speech signals.

4.4 ALM Analysis of Patient Voices Across Health Conditions

We used illness scores generated by the GPT-4.1 (SOAP+VITAL) to group samples accordingly and analyze the corresponding changes in the acoustic descriptions. Table 1 presents the top three important phrases identified using Term Frequency–Inverse Document Frequency (TF-IDF) for each illness score group. The observed trend is consistent with the clinician–patient comparison discussed above. Across all groups, emotional tone is predominantly neutral or calm. As health status

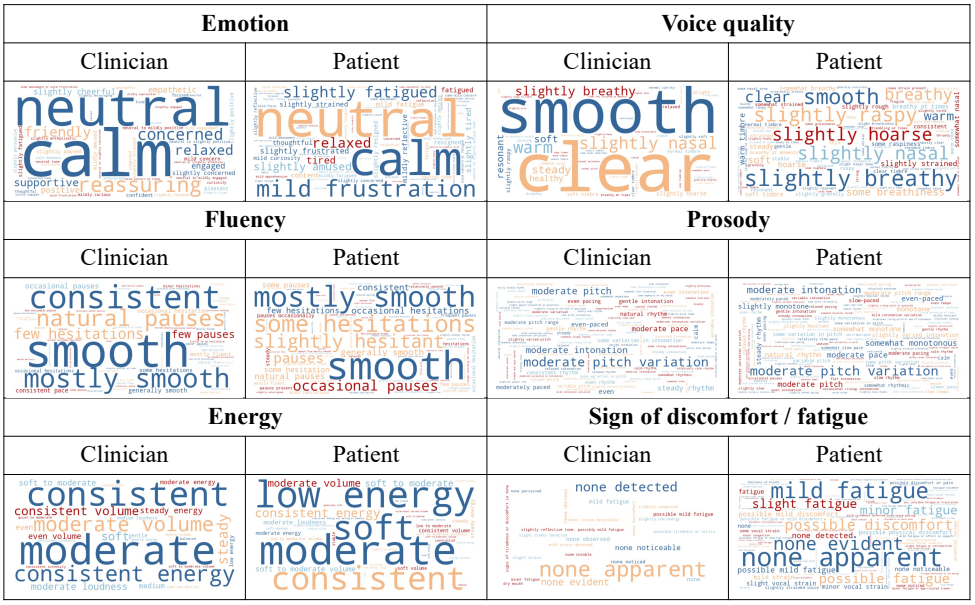


Figure 5: Wordcloud of GPT-4o-audio generated phrase frequencies for clinician and patient voices.

Table 1: Top-3 acoustic descriptions for patients per health condition group using TF-IDF analysis. The phrases within each group are ranked from most to least frequently generated by GPT-4o-audio.

Acoustic Feature	Illness Score			
	1 (Good health)	**2**	**3**	**4 (Poor health)**
Emotion	Neutral Calm Some mild concern	Neutral Calm Slightly fatigued	Neutral Calm Mild frustration	Neutral Calm Resigned
Voice Quality	Smooth Soft Slightly breathy	Smooth Slightly breathy Slightly raspy	Slightly breathy Smooth Slightly nasal	Slightly breathy Slightly raspy Slightly hoarse
Prosody	Moderate pitch variation Even Steady rhythm	Moderate intonation Moderate pitch variation Moderate pace	Moderate pitch variation Paced Moderate intonation	Somewhat monotonous Monotone Moderate pitch variation
Fluency	Mostly smooth Smooth Occasional pauses	Smooth Mostly smooth Occasional pauses	Occasional hesitations Slightly hesitant Mostly smooth	Slightly hesitant Some hesitations Pauses frequently

declines from relatively healthy to worse, voice quality tends to shift from smooth to increasingly breathy, raspy, or hoarse. Fluency changes from smoother delivery to more hesitant speech, while prosody changes from moderate pitch variation to a more monotone pattern. For features with a continuous scale, such as energy level and potential signs of fatigue or discomfort, we converted the descriptive phrases into numerical values to enable quantitative analysis. For energy, we mapped the output description phrases to pre-defined categories by calculating sentence similarity [6] between the original phrase and the following reference categories: *low/soft*, *soft to moderate*, *moderate,* and *consistent/steady*, corresponding to energy levels 1 to 4. For example, phrases most similar to *low/soft* were assigned an energy score of 1, whereas those most similar to *consistent/steady* received a score of 4. Similarly, for possible signs of discomfort and fatigue, we defined anchor categories as follows: *no evident/no apparent*, *mild discomfort/fatigue*, *moderate discomfort/fatigue*, and *noticeable discomfort/fatigue*. Output phrases were assigned numeric scores based on sentence similarity to these anchors, allowing quantification of severity on a scale from 1 (no discomfort/fatigue) to 4 (noticeable discomfort/fatigue). Results shown in Figure 6 show that both lower energy levels and increased signs of fatigue or discomfort detected by the ALM are indicative of poorer health status.

Overall, our results suggest that speech signals can exhibit distinct patterns associated with differences in health conditions. The ALM extracts interpretable acoustic descriptions that help characterize these patterns. While acoustic signals alone should not be used for diagnosis, such vocal characteristics could serve as supplementary cues in healthcare monitoring systems.

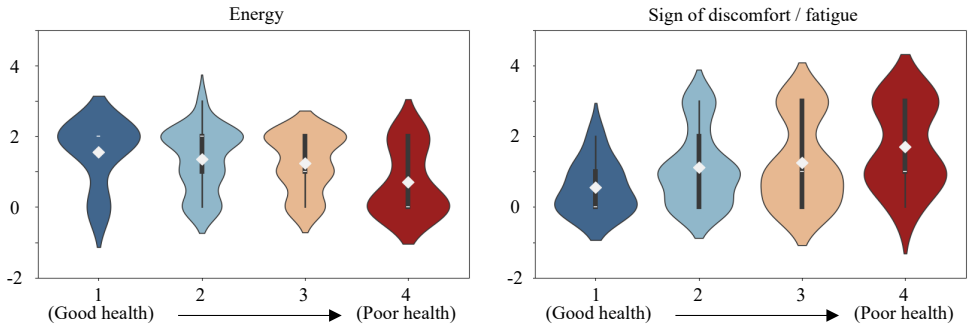


Figure 6: Violin plots showing the relationship between GPT-4o-audio extracted acoustic features and GPT-4.1 generated illness scores. Diamond markers indicate group means, and the width represents the density of observations at that value. **Left**: Energy levels decline as health worsens. **Right**: Signs of discomfort/fatigue increase with declining health status.

5 Limitations

Our study analyzes audio from a real-world home care setting, which presents challenges such as background noise, variable recording conditions, and frequent overlapping speech. While our automated acoustic processing pipeline substantially mitigates their impact on audio analysis, it remains imperfect and occasionally fails to fully isolate speakers, resulting in leakage from others and potentially blurring speaker-specific signals. Also, we examine only the first 30 seconds of the processed patient speech for acoustic analysis. This may omit important acoustic cues that occur later, including changes in the speaker's health condition.

Although we explicitly instruct the ALM to avoid transcribing the audio and to disregard linguistic content, it may not truly adhere to this, relying on contextual information to support its outputs. Further work is needed to assess the degree to which the ALM can disentangle acoustic from linguistic information. In addition, GPT-Audio declines to provide direct clinical judgments based solely on acoustic signals, while accepting the analysis of various acoustic features. The choice of prompt influences the ALM's outputs, but this effect was not examined in depth in the present study.

Moreover, this study focuses on analyzing acoustic signals in relation to overall health conditions, but in real-world healthcare contexts, voice alone should not serve as the sole basis for diagnosis. Instead, it should be regarded as a complementary signal, useful for capturing conditions not reflected in textual or numerical records, or in situations where such records are incomplete or inconsistent.

Finally, although experimental results demonstrate that our proposed LLM-generated illness scores align with clinical outcomes, we did not compare these scores directly with clinician annotations. This is because clinical judgments for overall health assessment are highly subjective and require extensive expertise, making such annotation both challenging and costly. Given the complexity and resource demands of this task, comparing LLM labels with clinician labels is beyond the scope of the present study.

6 Conclusion

This study investigates the use of LLMs to automatically assess patient health by generating holistic illness scores in a home care setting. We find that LLMs can effectively interpret SOAP notes extracted from visit transcripts to produce scores that align with clinical outcomes, whereas vital signs are less informative. To our knowledge, this is the first work to apply ALMs to assess patient health directly from speech, showing that they can capture clinically relevant acoustic patterns that, when quantified, correlate strongly with our LLM-based illness scores. Furthermore, we demonstrate ALMs' ability to describe acoustic features in interpretable plain language, enabling clinicians to incorporate speech as a meaningful health biomarker, adding another dimension of patient monitoring. Although our approach faces several challenges, it demonstrates the potential of leveraging generative AI for multimodal health assessment in real-world care and highlights the need to further explore the opportunities that ALMs offer for healthcare applications.

Acknowledgments

This work was supported by the National Institutes of Health under Grant No. 1R01AG081928-01, in addition to Grant No. P30AG073105 and R00AG076808 under the National Institute of Aging.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

[3] Sebastien Baur, Zaid Nabulsi, Wei-Hung Weng, Jake Garrison, Louis Blankemeier, Sam Fishman, Christina Chen, Sujay Kakarmath, Minyoi Maimbolwa, Nsala Sanjase, et al. HeAR–health acoustic representations. *arXiv preprint arXiv:2403.02522*, 2024.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.

[5] Rong Chao, Wen-Huang Cheng, Moreno La Quatra, Sabato Marco Siniscalchi, Chao-Han Huck Yang, Szu-Wei Fu, and Yu Tsao. An investigation of incorporating Mamba for speech enhancement. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 302–308. IEEE, 2024.

[6] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[7] Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025.

[8] Yu-Wen Chen and Julia Hirschberg. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain SOAP notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 1–9, 2024.

[9] Cheng-Han Chiang and Hung-yi Lee. A closer look into automatic evaluation using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, 2023.

[10] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[11] Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A modular approach for clinical slms driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, 2025.

[12] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36: 18090–18108, 2023.

[13] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128. Association for Computational Linguistics, 2024.

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[15] Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio Flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*, 2025.

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. TitaNet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8102–8106. IEEE, 2022.

[18] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

[19] Samara Soares Leal, Stavros Ntalampiras, and Roberto Sassi. Speech-based depression assessment: A comprehensive survey. *IEEE Transactions on Affective Computing*, 2024.

[20] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: a comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

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

[22] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. Large language models are few-shot health learners. *arXiv preprint arXiv:2305.15525*, 2023.

[23] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[24] Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023.

[25] Nvidia. Parakeet tdt 0.6b v2 (en). `https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2`. Accessed: 2025-08-12.

[26] OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, 2025.

[27] Taejin Park, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, Nithin Rao Koluguri, Krishna C Puvvada, Jagadeesh Balam, and Boris Ginsburg. Sortformer: A novel approach for permutation-resolved speaker supervision in speech-to-text systems. In *Forty-second International Conference on Machine Learning*, 2025.

[28] Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. Soap notes. In *StatPearls [internet]*. StatPearls Publishing, Treasure Island, FL, 2020.

[29] Amit Sapra, Ahmad Malik, and Priyanka Bhandari. Vital sign assessment. In *StatPearls [internet]*. StatPearls Publishing, Treasure Island, FL, 2023.

[30] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[31] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[32] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[34] Cédric Villani. The wasserstein distances. In *Optimal transport: old and new*, pages 93–111. Springer, 2009.

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, volume 35, pages 24824–24837, 2022.

[36] Bo Wen, Raquel Norel, Julia Liu, Thaddeus Stappenbeck, Farhana Zulkernine, and Huamin Chen. Leveraging large language models for patient engagement: The power of conversational ai in digital health. In *2024 IEEE International Conference on Digital Health (ICDH)*, pages 104–113. IEEE, 2024.

[37] Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K Meyer, Radhika Juglan, Zunamys I Carrero, Daniel Paech, Jens Kleesiek, Matthias P Ebert, et al. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digital Medicine*, 7(1):257, 2024.

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

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[40] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025.

[41] Maryam Zolnoori, Sasha Vergez, Sridevi Sridharan, Ali Zolnour, Kathryn Bowles, Zoran Kostic, and Maxim Topaz. Is the patient speaking or the nurse? automatic speaker type identification in patient–nurse audio recordings. *Journal of the American Medical Informatics Association*, 30 (10):1673–1683, 2023.

[42] Maryam Zolnoori, Ali Zolnour, Sasha Vergez, Sridevi Sridharan, Ian Spens, Maxim Topaz, James M Noble, Suzanne Bakken, Julia Hirschberg, Kathryn Bowles, et al. Beyond electronic health record data: leveraging natural language processing and machine learning to uncover cognitive insights from patient-nurse verbal communications. *Journal of the American Medical Informatics Association*, 32(2):328–340, 2025.

A Prompts

Prompt for generating SOAP note (Figure 7)

```
You are a professional healthcare provider. The following is a conversation between you and a homecare patient.
Please summarize the conversation using the SOAP (Subjective, Objective, Assessment, Plan) note format.

  Input format:
  A transcript of the conversation as a single string.

  Output:
  A JSON object in the following format:{
    "S": "Subjective findings — the patient's reported symptoms, concerns, or experiences.",
    "O": "Objective findings — observable, measurable facts such as vital signs, physical exam findings, or clinician
observations.",
    "A": "Assessment — your clinical judgment or diagnosis based on the subjective and objective information.",
    "P": "Plan — recommended next steps, treatments, referrals, or follow-up care."
  }

  Note: Do not include any other text other than the json object.
  Input:
```

Figure 7: Prompt template used for LLM-based SOAP note summarization from conversation transcripts.

Prompt for LLM-based holistic illness severity scoring (Figure 8)

Prompt for ALM acoustic analysis (Figure 9): We ask the ALM to generate a list of phrases rather than full sentences or fixed categories. This approach is motivated by several considerations: full sentences are difficult to evaluate because the same description can be expressed in many different ways, even though the underlying meaning is the same. Fixed categories, on the other hand, would constrain the model's output.

```
You are a professional healthcare provider reviewing SOAP (Subjective, Objective, Assessment, Plan) notes and
VITAL information from homecare visits. Your task is to evaluate and score the patient's overall health condition.

   Instructions:
   - Lower numbers to indicate better health (e.g., 1 = Good overall health, 5 = Critical condition, hospital referral
recommended.)
   - Provide a brief reason for each ranking.

   Input format:{
      "SOAP_note": {
       "S": "<subjective note>",
       "O": "<objective note>",
       "A": "<assessment>",
       "P": "<plan>"
       },
      "VITAL":{ "BMI": <List of BMI values recorded during the visit>,
                "DIASTOLICBLOODPRESSURE": <List of diastolic blood pressure values recorded during the
visit>,
                "HEIGHT": <List of height measurements recorded during the visit>,
                "PULSE": <List of pulse measurements recorded during the visit>,
                "RESPIRATIONS": <List of respiration rates recorded during the visit>,
                "SYSTOLICBLOODPRESSURE": <List of systolic blood pressure values recorded during the visit>,
                "TEMPERATURE": <List of body temperature readings recorded during the visit>,
                "WEIGHTLBS": <List of body weight in pounds recorded during the visit>,
                "O2SATURATION": <List of oxygen saturation levels recorded during the visit>,
                "BLOODSUGAR": <List of blood sugar levels recorded during the visit>,
                "PAIN": <List of self-reported pain levels recorded during the visit>,
                "WEIGHTKGS": <List of body weight recorded during the visit>
            }
   }

   Note:
   - Each vital sign may be measured multiple times or not at all during a visit.
   - The presence or absence of a measurement may reflect clinical judgment or specific concerns.

   Output:
   A dict object in the following format:{
     "health_condition_score": <score_assigned_to_visit>,
     "reason": "<explanation justifying the assigned score>"
   }

   Note: Do not include any other text other than the dict.
   Input:
```

Figure 8: Prompt template used for LLM-based illness scoring from SOAP notes and vital signs.

```
You are a specialist in speech and speaker analysis.

  You are given an audio sample consisting of several concatenated segments supposedly spoken by the same
individual. (but might mixed with other's voice) Your task is to analyze the **acoustic and prosodic characteristics**
of the speaker's voice—**not** the content or meaning of the words.
  Provide **specific, interpretable observations** as concise phrases or words in the following categories:
  - Speaker traits (e.g., perceived age group, gender)
  - Prosody
  - Fluency
  - Articulation
  - Emotional state inferred from vocal tone
  - Loudness and energy
  - Voice quality
  - Signs of fatigue, discomfort, or illness: Inferred **solely from voice** if applicable.

  **Do not** transcribe or reference the linguistic content. Analyze only how the speaker sounds.
  Return observations as a list of short descriptive phrases

  **Output format (JSON only):**
  ```json{
   "speaker_traits": ["<age>", "<gender>"],
   "prosody": ["", "", ...],
   "fluency": ["", "", ...],
   "articulation":["", "", ...],
   "emotion": ["", "", ...],
   "loudness_energy": ["", "", ...],
   "voice_quality": ["", "", ...]
   "signs_of_fatigue_discomfort":[""]
  }

  Note:
  - only output the JSON object. Do not include any additional text, explanation, or formatting.
  Input:
```

Figure 9: Prompt template used for ALM acoustic analysis instructing the model to analyze acoustic characteristics without transcribing linguistic content.

B Additional Results

B.1 Summary statistics for LLM-based holistic illness severity scoring (Table 2)

Table 2: Summary statistics for LLM-generated illness scores across different inputs types.

Model	Input	No ED/HOSP			ED/HOSP		
		Min	Max	Mean	Min	Max	Mean
GPT-4.1	VITAL only	1.0	4.0	1.60	1.0	2.0	1.73
	SOAP only	1.0	4.0	2.38	1.5	4.0	3.18
	SOAP+VITAL	1.0	4.0	2.27	1.5	4.0	3.14
Llama3.1-8B	VITAL only	3.0	5.0	4.38	3.0	5.0	4.30
	SOAP only	1.0	5.0	2.78	1.5	5.0	3.67
	SOAP+VITAL	1.0	4.5	2.67	1.5	5.0	3.39
MediPhi-Instruct	VITAL only	5.0	5.0	5.00	5.0	5.0	5.00
	SOAP only	1.0	4.0	2.79	2.0	5.0	3.46
	SOAP+VITAL	1.0	5.0	3.25	2.0	5.0	3.97
Qwen3-8B	VITAL only	1.0	5.0	1.37	1.0	4.0	1.57
	SOAP only	1.0	4.0	2.60	1.5	4.0	3.49
	SOAP+VITAL	1.0	4.0	2.59	1.5	5.0	3.38
```

B.2 Examples of LLM's rationale for illness scores

1 (Good health): The patient demonstrates excellent recovery after the procedure, with normal vital signs, good functional mobility, independence with daily activities and exercise, and no current pain or acute issues. Minor symptoms (occasional dizziness, mild intermittent pain) are self-managed and non-limiting. Discharge from home care is appropriate; following up with specialists is planned for further optimization.

2: Patient is making good post-operative progress with stable vital signs (BP 120/80, pulse 69, respirations 13, temp 98.1). Incision is healing well with no signs of infection, pain is mild and well controlled, and patient is able to perform activities of daily living with minimal assistance. There are no reports of serious symptoms or complications. Continued monitoring and gradual reduction in opioid use is appropriate for this recovery stage. Only mild issues such as minor discomfort and adhesive residue are present.

3: The patient is stable with no acute distress, pain, or new skin issues, and displays improved mobility and motivation. However, there is persistent right-sided paralysis with significant ongoing mobility limitations, elevated blood glucose consistently above target indicating suboptimal diabetes control, and safety concerns regarding stairs. Edema is improving and vital signs are otherwise stable. The condition is chronic and requires ongoing management but does not meet criteria for hospital referral at this time.

4 (Poor health): Patient is elderly, recently hospitalized, and requires ongoing oxygen therapy for chronic respiratory issues with worsening dyspnea. The patient is unable to stand or transfer independently and needs significant physical assistance and equipment for mobility. There is concern for hypoxemia requiring further respiratory management (possible BIPAP), and care is complicated by coordination of multiple services and progressive de-conditioning. Although vitals are stable, the overall functional and respiratory status is poor, indicating high risk but not immediate critical instability.

B.3 Analysis of articulation

Figure 10 and Table 3 present the ALM analysis results for articulation with GPT-4o-audio. In home care visits, articulation is a less informative indicator of health condition, as participants generally speak in their familiar language and therefore articulate their words clearly.

Table 3: Top-3 articulation descriptions for patients per health condition group using TF-IDF analysis. The phrases within each group are ranked from most to least frequently generated by GPT-4o-audio.

Acoustic Feature	**Illness Score**			
	1 (Good health)	**2**	**3**	**4 (Poor health)**
Articulation	Clear Well Pronounced	Clear Well Pronounced	Clear Well Enunciated	Clear Slightly unclear Deliberate

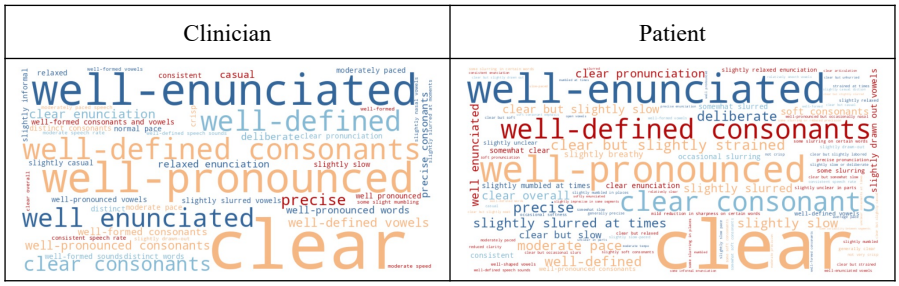


Figure 10: Wordcloud of GPT-4o-audio generated phrase frequencies describing articulation in clinician and patient voices.

C Experimental results of GPT-4o-mini-audio and Qwen2-Audio-7B-Instruct

C.1 Age and Gender Classification with GPT-4o-mini-audio (Figure 11)

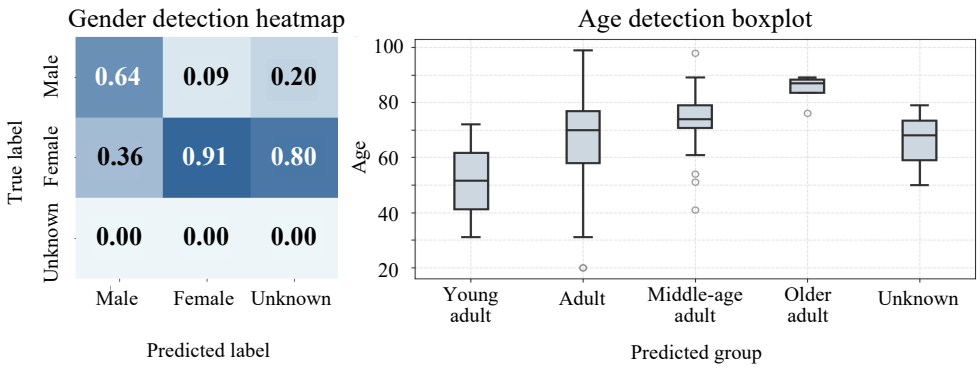


Figure 11: GPT-4o-mini performance in gender and age classification.

C.2 Age and Gender Classification with Qwen2-Audio-7B-Instruct (Figure 12)

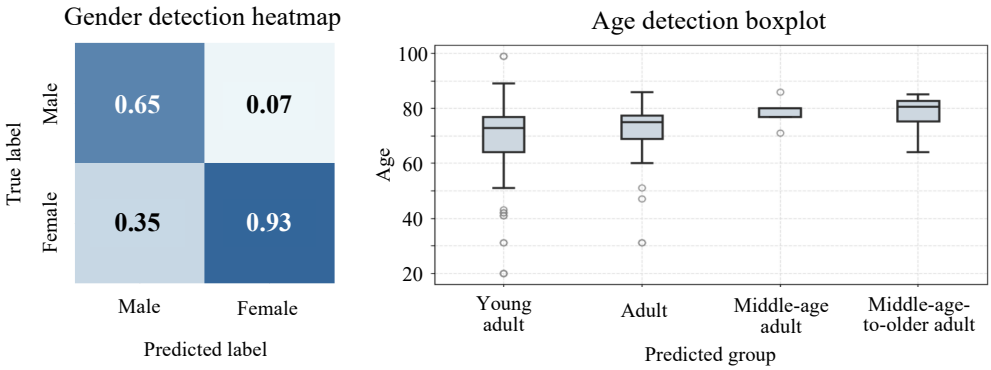


Figure 12: Qwen2-Audio-7B-Instruct performance in gender and age classification.

C.3 Analysis of Clinician and Patient Voices with GPT-4o-mini-audio (Figure 13)

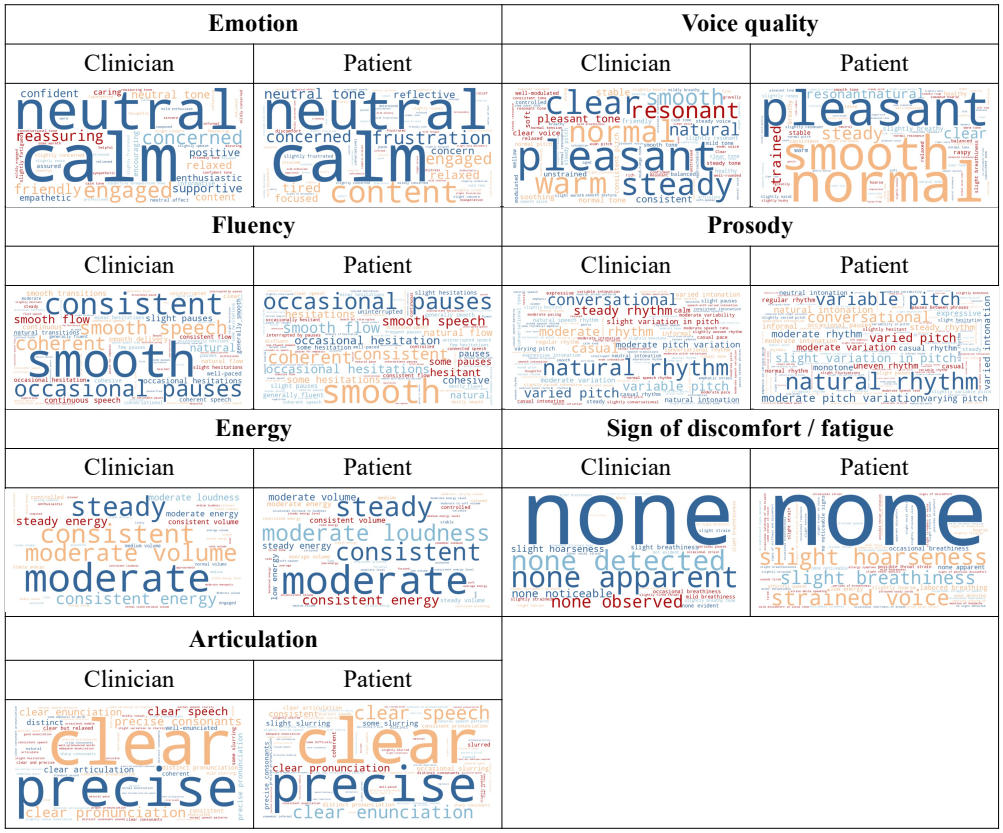


Figure 13: Wordcloud of GPT-4o-mini-audio-generated phrase frequencies for clinician and patient voices.

C.4 Analysis of Clinician and Patient Voices with Qwen2-Audio-7B-Instruct (Figure 14)

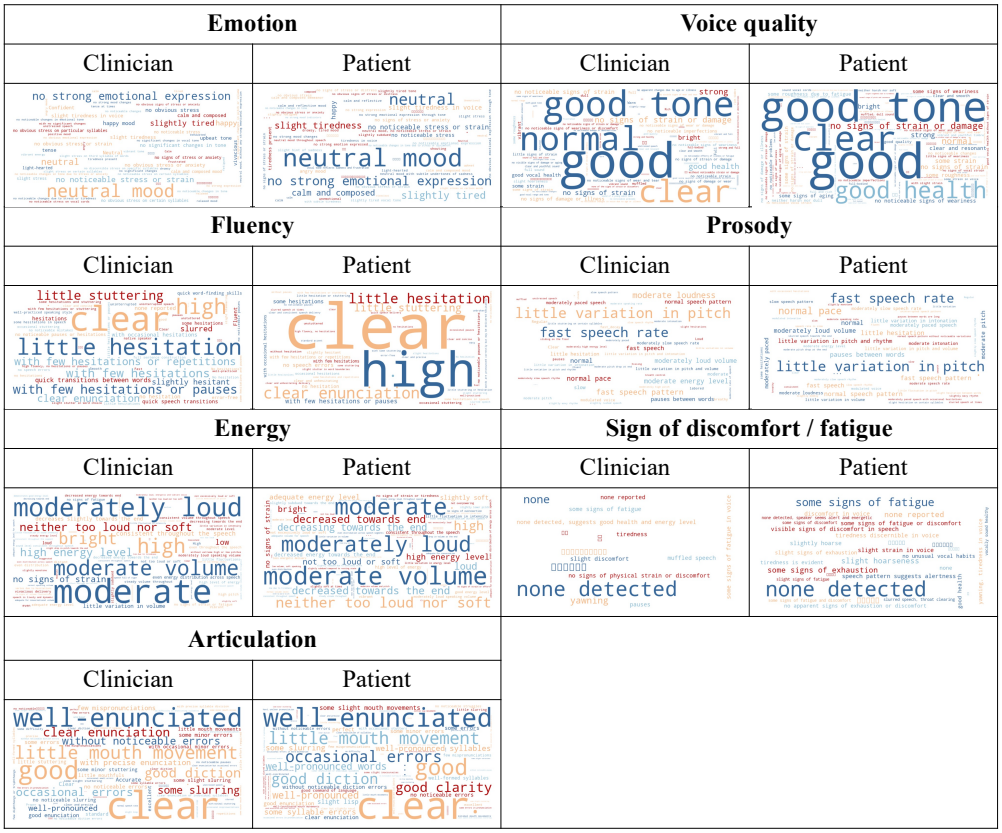


Figure 14: Wordcloud of Qwen2-Audio-7B-Instruct-generated phrase frequencies for clinician and patient voices.

C.5 Analysis of Patient Voices Across Health Conditions with GPT-4o-mini-audio (Table 4)

Table 4: Top-3 GPT-4o-mini-audio-generated acoustic descriptions for patients per health condition group using TF-IDF analysis. The phrases within each group are ranked from most frequent to least frequent.

Acoustic features	**Illness score**			
	1(Good health)	**2**	**3**	**4 (Poor health)**
Emotion	- Calm - Content - Neutral tone	- Calm - Neutral - Engaged	- Calm - Neutral - Frustration	- Calm - Neutral - Tired
Voice quality	- Pleasant - Warm - Slightly breathy	- Smooth - Normal - Pleasant	- Pleasant - Smooth - Strained	- Natural - Slightly raspy - Hoarse
Prosody	- Natural rhythm - Conversational - Even rhythm	- Natural rhythm - Variable pitch - Conversational	- Natural rhythm - Moderate pitch variation - Conversational	- Natural rhythm - Varied pitch - Moderate rhythm
Fluency	- Smooth - Smooth flow - Coherent	- Smooth - Occasional pauses - Smooth speech	- Smooth - Occasional pauses - Hesitations	- Smooth - Occasional pauses - Smooth flow
Articulation	- Clear - Precise - Clear pronunciation	- Clear - Precise - Consistent	- Clear - Precise - Clear enunciation	- Clear - Precise - Slurred
Energy	- Moderate - Consistent energy - Moderate loudness	- Moderate - Consistent - Steady	- Moderate - Steady - Low energy	- Moderate - Consistent volume - Low
Sign of discomfort / fatigue	- None - Mild discomfort in vocal tone - None detected	- None - Slight hoarseness - Occasional breathiness	- None - Slight hoarseness - Labored breathing	- No noticeable signs - Low energy - Strained voice

C.6 Analysis of Patient Voices Across Health Conditions with Qwen2-Audio-7B-Instruct (Table 5)

Table 5: Top-3 Qwen2-Audio-7B-Instruct-generated acoustic descriptions for patients per health condition groups using TF-IDF analysis. The phrases within each group are ranked from most frequent to least frequent.

Acoustic features	**Illness score**			
	1(Good health)	**2**	**3**	**4 (Poor health)**
Emotion	- Neutral mood - No noticeable stress or strain - Calm and reflective	- Neutral mood - Neutral - No strong emotional expression	- Neutral mood - No strong emotional expression - Slight tiredness	- Neutral mood - Neutral - Slightly tired
Voice quality	- Good - Good tone - Normal range of pitch and resonance	- Good - Good tone - Clear	- Good tone - Good - Clear	- Clear - Normal - Some strain
Prosody	- Fast speech rate - Little variation in pitch - Moderately paced speech	- Little variation in pitch - Fast speech rate - Normal pace	- Little variation in pitch - Fast speech rate - Normal pace	- Little variation in pitch - Fast speech rate - Normal
Fluency	- Clear - High - Little hesitation	- Clear - Clear enunciation - High	- Clear - Clear enunciation - High	- Clear - Little hesitation - High
Articulation	- Enunciated - Well - Some slight mouth movements	- Well - Enunciated - Clear	- Well - Enunciated - Clear	- Good - Clear - Clear speech with occasional stuttering
Energy	- Moderate volume - Moderately loud - No signs of fatigue	- Moderate volume - Neither too loud nor soft - Moderate	- Moderate volume - Moderate - Moderately loud	- Moderate - Decreased towards end - Moderately loud
Sign of discomfort / fatigue	- None detected - Slight strain in voice - Throat clearing	- None detected - Some signs of fatigue - No unusual vocal habits	- None detected - Slight hoarseness - Some signs of fatigue	- None detected - Some signs of fatigue - Throat clearing