The Sound of Syntax: Finetuning and Comprehensive Evaluation of Language Models for Speech Pathology

Anonymous ACL submission

Abstract

According to the U.S. National Institutes of Health, approximately 5%–9% of children experience speech disorders that require clinical intervention. However, the number of certified speech-language pathologists (SLPs) is roughly twenty times fewer than the number of affected children, highlighting a significant gap in care and a pressing need to automate aspects of SLP workflows. Existing AI approaches for supporting SLPs typically address individual tasks in isolation, resulting in inconsistent performance and high deployment costs. Moreover, the scarcity of annotated datasets further limits progress in this domain. Recent advances in multimodal large language models (LLMs), particularly speech LLMs, offer promising opportunities for automating key SLP tasks and generating high-quality datasets. Despite this potential, there has been limited exploration of speech LLMs in this context. In this work, we introduce the first unified and comprehensive benchmarking framework for five core SLP tasks: (1) disorder screening, (2) speech transcription, (3) disorder-type classification, (4) symptom identification, and (5) transcript-based classification. Furthermore, we develop a fine-tuning strategy based on cross-task knowledge transfer, which enhances model performance across multiple tasks. Our experiments with 15 state-of-the-art LLMs show that while base models perform adequately on coarse-grained tasks, finetuning on the transcription task can yield substantial improvements across a broader set of tasks, demonstrating up to more than 30% improvement over baseline approaches. We publicly release our datasets, models, and benchmark framework to support continued research in this area.

1 Introduction

Speech and language disorders in children can significantly impact communication, academic development, and long-term social outcomes (Hitchcock et al., 2015; Foster et al., 2023). Early detection and intervention by speech-language pathologists are critical to mitigating these adverse effects (Gibbard et al., 2004; Centers for Disease Control and Prevention, 2024). However, the availability of qualified clinicians is characterized by an uneven distribution across geographic and socioeconomic contexts, with only an expert for every 20 affected children, resulting in significant disparities in access to care and leading to "missing intervention" for many children who could benefit from timely support (U.S. National Institute on Deafness and Other Communication Disorders, 2025; Tucker and McKinnon, 2020). This gap underscores an urgent need for scalable and supportive technological solutions to assist clinicians by augmenting their capacity and extending the reach of vital interventions.

The shortage of qualified clinicians has led to significant gaps in diagnostic capacity, particularly in domains requiring specialized expertise such as speech-language pathology (SLP). Recent advancements in large language models (LLMs) present a promising opportunity to partially automate or augment diagnostic workflows (Lammert et al., 2025; Bhattacharya et al., 2024; Nagpal et al., 2025; Maqsood et al., 2024). Multimodal LLMs, including GPT-4[1] and Gemini[2], exhibit state-of-the-art capabilities in speech processing and contextual reasoning. Trained on diverse, large-scale datasets, these models are robust to the variability of clinical SLP data, making them well-suited for tasks such as transcribing atypical speech and supporting disorder screening and subtype classification.

Effective integration of LLMs into clinical SLP requires rigorous, domain-specific evaluation to establish their clinical validity and utility (Cordella et al., 2025). This process depends on large, high-quality datasets that capture the variability of pedi-

[1] https://openai.com/index/gpt-4/
[2] https://gemini.google.com

atric speech, especially disordered forms, and are annotated with clinically relevant features. Current progress is hindered by two key challenges: the scarcity of well-curated pediatric speech corpora and the lack of evaluation frameworks that address the unique acoustic-phonetic features of children's speech, which are often overlooked by general-purpose benchmarks (Suh et al., 2024).

In this work, we present a comprehensive approach to bridge these gaps. Our solution first involves developing systematic procedures for annotating child speech data, creating resources suitable for SLP-focused model evaluation, and fine-tuning. Second, leveraging these curated datasets, we extensively evaluate state-of-the-art speech-capable LLMs for tasks pertinent to SLP. Our evaluation utilizes a tailored benchmark built upon the HELM framework (Liang et al., 2023). The benchmark assesses models across five clinical scenarios covering a spectrum of tasks from foundational disorder detection to more granular symptoms, including Disorder Diagnosis, Transcription-Based Diagnosis, Transcription, Disorder Type Classification, and Symptom Classification. This structured investigation aims to quantify existing LLMs' current capabilities and limitations in SLP-relevant contexts and explore avenues for enhancing their performance through domain-specific adaptation. The systematic assessment of current multimodal models with this benchmark reveals sizeable performance gaps: macro-F1 scores routinely fall below clinically acceptable thresholds, especially on the more fine-grained tasks. Additionally, our work involves developing and evaluating fine-tuned speech LLMs designed to push the boundaries of current state-of-the-art results on these specialized tasks. Our contributions are stated as follows.

- We release four curated pediatric speech datasets comprising approximately 30,000 speech samples across English and French, encompassing both typical and disordered speech. These datasets provide a publicly available, high-quality resource to support reproducible benchmarking in SLP.
- We propose the first comprehensive evaluation framework for SLP, extending the HELM paradigm to unify five essential clinical tasks. This framework enables consistent, task-aligned evaluation and facilitates direct comparison of speech LLM performance under a standardized protocol.
- We introduce fine-tuned speech LLMs that achieve state-of-the-art performance across all evaluated SLP tasks, illustrating the efficacy of domain-specific adaptation in enhancing diagnostic and transcriptional capabilities.

2 Related Works

AI in Speech Language Pathology Assessment The use of artificial intelligence, particularly LLMs, in clinical speech-language assessment has gained increasing attention in recent years. Several recent studies have demonstrated the utility of LLMs in detecting and characterizing speech and language disorders. For instance, Bhattacharya *et al.* showed that pre-trained LLMs could effectively identify both the presence and type of aphasia, suggesting that these models can serve as viable tools for clinical screening and diagnosis of language disorders (Bhattacharya et al., 2024).

Beyond perception studies, a growing body of technical literature examines the use of speech and language features for automated assessment. Engelhardt *et al.* reviewed computational features used to assess cognitive and thought disorders, highlighting the relevance of acoustic and linguistic cues in differential diagnosis (Engelhardt et al., 2021). Similarly, (Heilmann et al., 2023) demonstrated that automatic language sample analysis tools can support clinical workflows, providing reliable linguistic metrics with reduced human effort.

Several efforts have focused on building automated tools for therapy and assessment. (Deka et al., 2025) systematically reviewed AI-based automated speech therapy tools for individuals with speech sound disorders, underscoring their potential and emphasizing the need for clinically validated benchmarks. (Themistocleous, 2024) introduced a framework for automatic language assessment using LLMs, proposing a scalable and adaptable approach for linguistic evaluation.

LLMs for Disordered Speech Analysis A recent survey of SLPs and graduate students revealed a combination of cautious optimism and skepticism regarding the integration of LLMs such as ChatGPT into diagnostic and therapeutic workflows (Schwartz et al., 2024). These practitioner attitudes highlight critical socio-technical barriers to the clinical adoption of AI-driven systems in speech-language pathology.

Recent research has explored the adaptation of LLMs and related models for disordered speech processing, with an emphasis on reinforcement learning. Zhang *et al.* employed reinforcement learning with human feedback (RLHF) to personalize automatic speech recognition (ASR) systems for disordered speech, demonstrating significant improvements in recognition accuracy through individual-level adaptation (Zhang et al., 2024). Sanguedolce *et al.* proposed a more generalized framework by fine-tuning Whisper on a dataset of stroke patients, resulting in a universal disordered-speech detection model. Their approach exhibited strong generalization across multiple neurological conditions, underscoring the potential of foundation models for broad-spectrum clinical speech applications (Sanguedolce et al., 2024).

Benchmarking Efforts Benchmarking has been instrumental in advancing speech-health research. The ADReSS Challenge (Luz et al., 2020) established a balanced benchmark for Alzheimer's detection from spontaneous speech, standardizing evaluation via F1 and MMSE-regression metrics. Similarly, the Children's ASR Benchmark (Fan et al., 2024) introduced standardized splits and Whisper/Wav2Vec baselines for speech recognition in children aged 6–14, highlighting age-specific acoustic challenges. Nonetheless, systematic benchmarking of speech LLMs in clinical contexts remains limited. To address this, we propose a unified evaluation framework for assessing speech LLMs across clinically relevant tasks, emphasizing both diagnostic performance and usability.

3 Method

3.1 Clinically-Informed Data Annotation

Existing datasets for children's speech-language pathology (SLP) research primarily focus on transcription and binary classification of speech as either disordered or typical (Benway et al., 2022; Eshky et al., 2018; Le Normand, 1997; Schneider et al., 2006). However, as discussed previously, critical SLP tasks involve finer-grained classification, including the identification of specific disorder types and associated symptoms—categories for which no large-scale publicly available datasets currently exist. Addressing this gap, we collaborated closely with certified SLP professionals to develop a detailed annotation schema that captures both disorder types and their characteristic symptoms. For each speech sample, we assign the most prominent disorder type and symptom, prioritizing the most salient diagnostic features when multiple conditions may co-occur. Speech samples exhibiting no observable signs of speech disorder are annotated as typical. Our annotation protocol and chosen taxonomy are informed by clinical guidelines from the U.S. National Institutes of Health (Simon and Rosenbaum, 2016) and SLP best practices (American Speech-Language-Hearing Association, 2016). After initial manual labeling, we conducted a verification phase in which all annotations were reviewed by certified speech-language pathologists to ensure consistency and clinical validity. This procedure resulted in a high-quality, expert-validated dataset suitable for training and evaluating models on clinically relevant SLP tasks.

3.2 Evaluation Pipelines

We evaluate five core tasks that collectively capture the essential stages of pediatric SLP, from initial screening to detailed diagnostic analysis: **(1) Disorder Diagnosis**, which assesses a model's ability to distinguish between typical and disordered speech—a critical early triage step for prioritizing clinical resources; **(2) Transcript-based Diagnosis,** which serves as a baseline for diagnostic accuracy by testing the assumption that speech from children with disorders deviates from expected utterances. This approach operates by matching model-generated transcripts to clinician prompts, which offers a minimal, interpretation-free method that could be readily deployed in clinical settings. By benchmarking against this heuristic, we quantify the value added by more sophisticated multimodal LLM reasoning; **(3) Transcription**, which measures the fidelity of automatic speech recognition (ASR) systems on child and disordered speech, a prerequisite for downstream diagnostic and documentation tasks; **(4) Disorder Type Classification**, which probes whether models can differentiate between *articulation disorders*—motor-based speech errors such as lisps or distortions—and *phonological disorders*, which involve rule-based sound pattern errors like consistent substitution of one phoneme for another (e.g., /k/ $\rightarrow$ /t/);**(5) Disorder Symptom Classification**, a fine-grained task, requires models to identify specific clinical symptoms, including *additions* (insertion of extra sounds), *substitutions* (replacing one sound with another), *omissions* (dropping expected sounds), and *stuttering* (disruptions in speech fluency). Figure 1 illustrates an overview of classification tasks in our pipeline. To assess model capacity under different prompting strategies, we evaluate performance using both zero-shot and five-shot prompting. Details of these prompts are presented in Appendix A. Evaluation metrics for classification tasks include Macro F1, Micro F1, and Exact Match Accuracy, while transcription performance is assessed using Word Error Rate (WER), Match Error Rate (MER), and Word Information Preserved (WIP).

We develop SLPHelm, an evaluation framework built upon the HELM benchmark (Liang et al., 2023), to enable standardized, systematic assessment across all tasks. By leveraging a unified pipeline, SLPHelm ensures consistent evaluation protocols and comparability across models and prompting strategies. To promote reproducibility, we publicly release all code, prompts, and configuration files associated with our framework.

3.3 Finetuning Methods

To investigate the impact of fine-tuning on model performance across multiple tasks, we explore two fine-tuning strategies. Prior work has shown that fine-tuning can facilitate cross-task and cross-lingual knowledge transfer, wherein a model fine-tuned on a simple task in a given language can exhibit improved performance on a range of downstream tasks in the same language (Ye, 2024; Egonmwan et al., 2019).

Our first strategy involves fine-tuning the model on a speech recognition task (Scenario 3, as described above), relying on the model's intrinsic ability to transfer knowledge to improve performance on related tasks. In this setup, both typical and disordered speech samples are labeled with the same expected transcriptions. However, assigning identical transcriptions to acoustically distinct inputs may introduce ambiguity and limit the model's ability to learn disorder-specific patterns. To mitigate this, our second strategy modifies the labeling of disordered speech by appending an asterisk to each word in its transcription. This lightweight labeling scheme serves to differentiate disordered speech from typical speech, thereby guiding the model to better recognize and transcribe disordered speech patterns without altering the overall task formulation. Details of fine-tuning prompts and hyperparameters are presented in Appendix B.

Our central hypothesis is that fine-tuning on a general task (e.g., speech recognition) alone is insufficient to yield improvements on specialized clinical tasks unless the fine-tuning data contains explicit information relevant to those tasks. This stems from the theoretical premise that general-purpose models primarily optimize for surface-level acoustic-linguistic alignment, which may not encode the deeper, disorder-specific features, such as atypical phonological patterns or motor-based distortions, necessary for clinical inference (Shor et al., 2019; Dorfner et al., 2024). We posit that enhanced task performance, particularly for disorder-specific tasks, requires either systematic cues (as in the second strategy) or explicit exposure to disorder-relevant data.

4 Experimental Resutls

4.1 Datasets, Models, and Configurations

Datasets In this study, we utilize four publicly available datasets: Ultrasuite(Eshky et al., 2018), ENNI(Schneider et al., 2006), LeNormand(Le Normand, 1997), and Percept-GFTA(Benway et al., 2022). These datasets encompass a range of child speech samples, both typical and disordered, and serve as the foundation for evaluating model performance across diagnostic tasks. Detailed dataset statistics are presented in Table 1. To ensure computational efficiency while maintaining representativeness, we randomly sample up to 1000 instances from each dataset for evaluation.

Table 1: Dataset statistics

Dataset	# Children	# Samples	Age Range
Ultrasuite	66	8338	5–13
ENNI	377	16546	4–9
LeNormand (French)	17	329	3–8
PERCEPT-GFTA	350	3664	6–17

Models We evaluate a total of 15 speech LLMs, encompassing both proprietary and open-source systems. Among the closed-source models, our evaluation includes the GPT-4 family (4o-audio, 4o-mini-audio, 4o-transcribe, and 4o-mini-transcribe), Whisper, and the Gemini 2.0 family (2.0-flash, 2.0-flash-lite). For open-source models, we consider multiple versions and sizes from the Qwen families (2.5-omni-7b, 2.5-omni-3b, 2-audio-7b, audio-chat), the Phi-4, and IBM Granite series (3.3-8b, 3.3-3b, 3.2-8b). Models are assessed across different parameter scales within each family to capture performance variability due to model capacity.

Inference pipelines We implement two distinct model inference pipelines within our evaluation framework. The first, referred to as the audio-to-

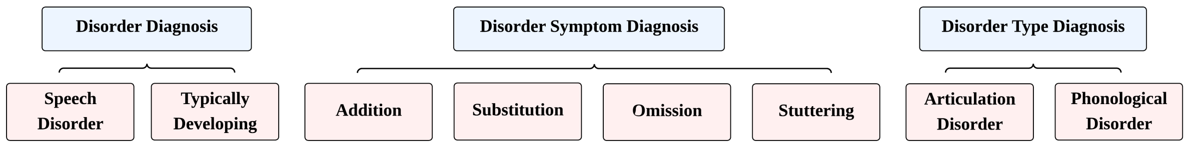

Figure 1: Taxonomy of classification tasks in SLPHelm. The benchmark includes three core diagnostic tasks: (i) disorder diagnosis, (ii) disorder symptom classification, and (iii) disorder type classification.

LLM prompting pipeline, is designed for models with native multimodal capabilities (e.g., GPT-4o-Audio, Gemini 2.0 Flash). In this setting, raw audio inputs are passed directly to the model alongside a task-specific prompt, enabling end-to-end processing of both acoustic and textual information. The second pipeline, termed transcription-based prompting, targets language-only models (denoted with the -transcribe suffix). Here, audio inputs are first transcribed using a base automatic speech recognition (ASR) model (e.g., Whisper or GPT-4o's internal ASR), and the resulting text is embedded into a structured prompt for downstream reasoning. This two-pronged architecture enables systematic comparison between models with native audio comprehension and those relying on cascaded ASR-to-LLM pipelines, providing insights into the trade-offs between direct speech understanding and transcription-mediated processing.

4.2 Evaluation Results

Our findings indicate that current speech LLMs exhibit substantial potential in augmenting core SLP tasks. However, both existing proprietary and open-source models currently fall short of clinically acceptable performance thresholds. This limitation is likely attributable to the underrepresentation of disordered speech in training corpora, as such data is significantly less prevalent than typical speech samples available online. For reference, existing FDA-approved diagnostic systems typically achieve F1 scores in the range of 0.80 to 0.85 (Fanni et al., 2023; Abràmoff et al., 2018), which serves as a practical standard for clinical viability. Furthermore, model performance varies markedly across different task scenarios, highlighting the absence of a universally robust model capable of consistently addressing the diverse requirements of pediatric SLP applications. Figure 2 presents an overview of the performance of all models.

Scenario 1: Disorder Diagnosis In the disorder diagnosis task, performance remains limited, with no model exceeding a Macro F1 score of 0.71. The best result is achieved by Qwen 2.5-Omni-7B, outperforming GPT-4o-Mini-Transcribe (0.56). The fact that these models use different pipelines—audio-grounded vs. ASR+text—suggests no clear advantage of one approach over the other. Smaller variants within each family perform similarly, indicating diminishing returns from increased parameter count. Audio-grounded Granite models perform poorly (F1 < 0.1), likely due to their pretraining focus on speech-to-text and translation tasks (IBM Granite Team, 2025). While their WER is competitive with other audio models (e.g., Gemini 2.0 Flash (Saon et al., 2025)), they appear to miss prosodic and articulatory cues critical for disorder detection. Overall, even the strongest models misclassify nearly half of the cases, underscoring the challenge of this foundational diagnostic task.

Scenario 2: Transcription-based Diagnosis Substituting the audio-grounded prompts with a naïve transcribe-and-compare baseline precipitates a pronounced decline in performance. Macro-F1 scores fall by roughly an order of magnitude: the strongest system, GPT-4o-Mini-Transcribe, attains only 0.21, while most models approach zero. Error propagation from automatic-speech-recognition (ASR) output, compounded by brittle string-matching heuristics, underscores the necessity of end-to-end acoustic reasoning and establishes this baseline as a conservative lower bound.

Scenario 3: Transcription Word-error rates (WER) vary widely—from 8.3% to 66.4%. Gemini-2.0-Flash-Lite achieves the lowest WER (8.3%), closely followed by Gemini-2.0-Flash (9.4%). Importantly, transcription fidelity shows limited correlation with diagnostic accuracy: GPT-4o-Mini-Transcribe records a moderate WER of 15.4% yet ranks among the strongest classifiers in Scenario 1. These findings indicate that high-quality transcripts are neither necessary nor sufficient for dependable clinical reasoning.

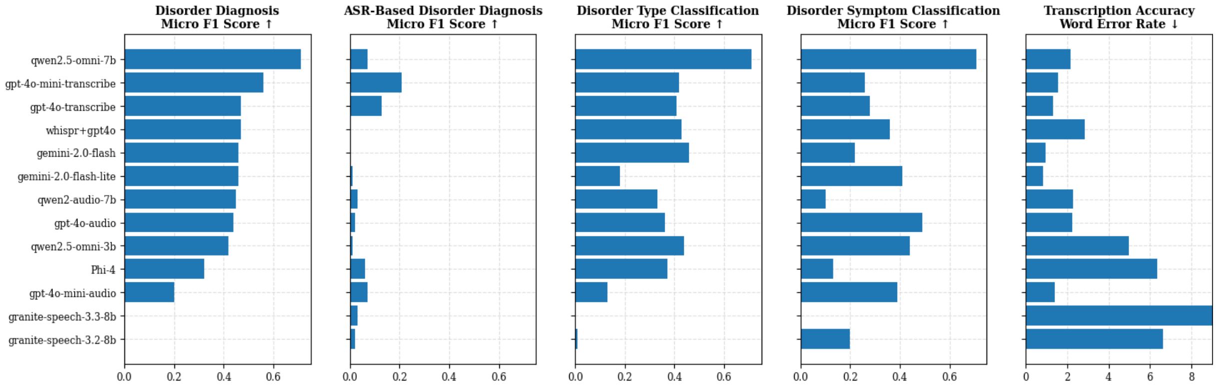

Figure 2: Metrics across all scenarios

Scenario 4: Disorder Type Classification In this scenario, improved accuracy has direct clinical implications: precise subtype identification supports more targeted and effective therapy plans, potentially reducing treatment duration and improving long-term speech outcomes. Closed-source multimodal models—especially Gemini-2.0-Flash—generally outperform open-source ones, suggesting benefits from broader or more diverse acoustic pretraining. However, Qwen2.5-7B is a notable exception, outperforming all models regardless of access or scale, hinting at architectural or pretraining advantages. The performance gap between audio-grounded and transcript-only variants is modest; for instance, GPT-4o-Mini-Transcribe lags its audio-capable version by just 6 Macro-F1 points. This indicates that LLMs can extract diagnostic signals from transcripts alone. Overall, ASR+LLM pipelines, while not yet optimal, offer a feasible alternative when audio is unavailable.

Scenario 5: Disorder Symptom Classification Accurate identification of these symptoms directly informs treatment goals and therapy design in speech-language pathology. Qwen2.5-7B once again leads in performance but still falls well short of clinically actionable accuracy. Moreover, transcription-first models underperform across all metrics, underscoring that critical acoustic cues needed for symptom detection are often lost or degraded during transcription. Three consistent trends emerge across tasks. First, audio grounding becomes increasingly vital as tasks grow more granular: while the performance gap between audio-first and transcript-only models is small for binary screening (Scenario 1), it widens significantly for symptom-level tagging (Scenario 5). Second, model scale is not the sole determinant of performance—smaller, well-aligned models such as Gemini-Flash-Lite achieve strong transcription results. Third, high transcription accuracy does not imply clinical accuracy, as evidenced by a weak correlation between performance on ASR and diagnostic tasks.

Finetuning results Fine-tuning large models can significantly enhance their performance on downstream tasks. In our setting, fine-tuning solely on automatic speech recognition (ASR) data, regardless of whether disordered speech is explicitly marked, leads to noticeable improvements in ASR-based tasks (Scenarios 2 and 3). However, not differentiating between typical and disordered speech introduces ambiguity in the input-label mapping, which in turn results in degraded performance. Incorporating a simple asterisk mitigates this issue, yielding more stable performance.

Cross Language Analysis Figure 4 shows a consistent pattern: macro-F1 is higher in French than in English, yet WER is markedly worse. A plausible explanation lies in the way these systems were pre-trained. Their ASR components are heavily optimized on English text-to-speech pairs, so lexical recognition degrades when confronted with French phonotactics, inflating WER. By contrast, the diagnostic classifiers operate on higher-level acoustic embeddings learned during large-scale audio pretraining that is largely language-agnostic (Klempíř and Krupička, 2024). Those embeddings could still capture phonological and articulatory cues relevant to speech-disorder detection, so classification accuracy can rise even as word-level transcription falters. In short, limited French supervision hurts the ASR stage but leaves the downstream pathology signal largely intact, highlighting that transcript fidelity and clinical utility depend on different slices

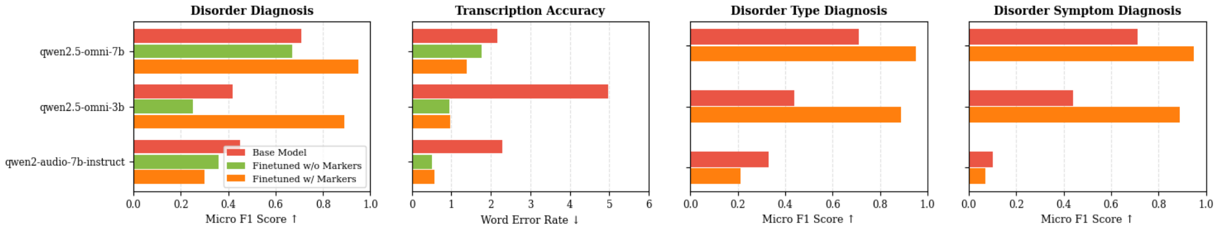

Figure 3: Model performance after finetuning

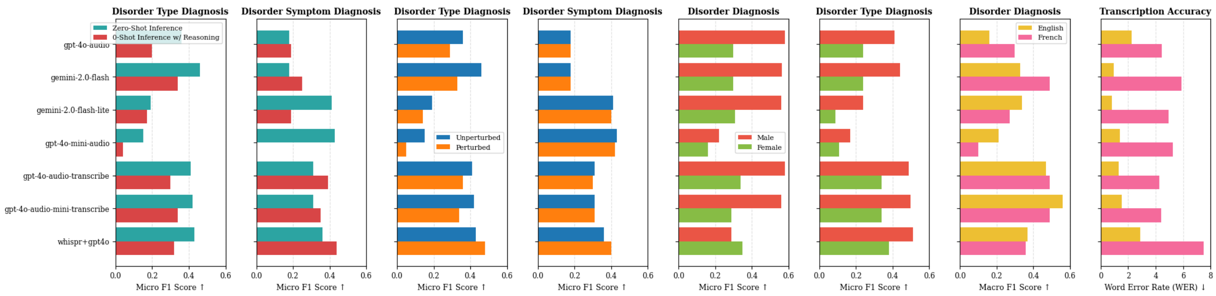

Figure 4: Compares model performance with reasoning, robustness under noisy conditions, across gender and languages

of the model's pre-training pipeline. This once again highlights the divergence and lack of correlation between the diagnostic capabilities of a model and its performance under transcription tasks

Cross Gender Analysis Figure 4 showcases that across two diagnostic tasks, the models exhibit a systematic gender performance gap that favors male speech. For both tasks, we run model evaluation on 1000 utterances for each gender on the UltraSuite dataset since it makes demographic identifiers available through its metadata. For the binary disorder-screening scenario, macro-F1 for male speakers averages 0.59 versus 0.40 for female speakers. The disparity persists, though it narrows in disorder-type classification. This pattern is remarkably consistent: almost every architecture in the screening task posts a positive male-female differential, and even that baseline reverses to a male advantage once finer-grained labels are required. Notably, the gap is not confined to a particular modeling strategy; it appears in fully audio-grounded systems (e.g., GPT-4o-audio, Gemini-2.0-Flash) as well as in transcript-conditioned variants, indicating that either the upstream acoustic encoders or the training data itself encode gender-skewed priors. The magnitude of the divergence suggests practical consequences for clinical deployment, as female speech receives both lower sensitivity and lower precision across disorder categories. Taken together, the results underscore the need for targeted auditing and, potentially, gender-balanced fine-tuning to ensure equitable diagnostic performance across child speakers.

Robustness Analysis We analyse model robustness by having the models evaluate audio recordings with artificial 3 different perturbations added to them - road noise, classroom noise, and office noise; and aggregating the results to comprehensively model performance under these conditions. We added 20 dB of that background noise to approximately simulate conditions that a given LLM might face in the clinical SLP setting. Figure 4 shows that the bulk of the metric degradation is concentrated in the disorder type diagnosis, while performance for symptom diagnosis remains virtually unchanged. These observations suggest that noise resilience is not strictly a function of architecture class - audio-grounded, and transcript-conditioned pipelines each appear at both ends of the robustness spectrum—but is instead tied to model-specific design, scale, and potentially, training data. Critically, the disproportionate degradation in Disorder Type diagnosis indicates that intermediate-level labels rely on acoustic cues most vulnerable to the injected perturbations, whereas symptom-level tagging may benefit from label sparsity that cushions small score shifts.

Impact of Reasoning Introducing an explicit chain-of-thought (CoT) prompt ("Let's think step by step . . . ") systematically depressed F1 scores on the intermediate Disorder-Type task but produced a mixed picture on the more fine-grained Symptom task. The pattern aligns with recent evidence that CoT can hamper tasks where the optimal decision boundary is compact or where answer formatting is unforgiving, because the additional reasoning tokens introduce distraction or bleed into the predicted label (Liu et al., 2024). Conversely, when the label space becomes larger and more conceptually diffuse, as in symptom diagnosis, CoT can help larger models articulate latent acoustic cues, echoing earlier results that larger LLMs benefit from self-generated rationales on complex problems (Kojima et al., 2022). Taken together, these findings caution against the blanket adoption of CoT in clinical speech pipelines: its utility is contingent on task granularity and model capacity, and careless deployment can hamper accuracy in resource-constrained systems.

Impact of Fewshot examples The results of the GPT-4 family across the first three scenarios under few-shot prompting indicate that few-shot examples do not consistently enhance the model's intrinsic capabilities; the benefits of prompting are not uniformly evident. For instance, while few-shot prompting significantly improves the performance of GPT-4o-Mini-Transcribe and GPT-4o-Transcribe in Scenario 1, it leads to reduced accuracy in Scenario 2. This suggests that few-shot prompts may introduce biases or hallucinations that adversely affect model behavior. Our observations are consistent with prior findings on text-only LLMs reported by Google (Jacovi et al., 2023).

5 Conclusion & Future Work

We present the first end-to-end benchmark for pediatric SLP, constructed within the HELM framework and encompassing four public corpora, five clinically grounded tasks, and a representative set of open- and closed-source LLMs. By standardizing evaluation across the diagnostic spectrum—from binary disorder screening to symptom-level tagging—this benchmark provides a rigorous and reproducible testbed for assessing the clinical viability of foundation models in SLP contexts.

Our empirical findings underscore the critical role of acoustic input for accurate clinical reasoning. Models with direct access to audio consistently outperform transcript-only pipelines on all tasks requiring fine-grained reasoning, with performance gaps ranging from several Macro-F1 points in binary classification to over 20 points in symptom-level tagging. However, audio grounding alone is insufficient: even the best-performing closed-source models fall short of clinical-grade reliability, revealing considerable room for improvement. Furthermore, although Whisper achieves significantly lower WER than most LLM-based ASR components, it underperforms on downstream clinical classification tasks, reinforcing that transcription fidelity alone is a poor proxy for diagnostic utility. Our fine-tuning experiments with the Qwen2.5 family demonstrate that performance can be substantially improved through knowledge transfer, particularly for the Qwen2.5-Omni 7B model. This highlights the effectiveness of task-specific adaptation and the potential for developing specialized SLP models that generalize well across tasks.

The fairness analysis reveals a consistent male-favored performance disparity in both screening and disorder classification tasks, indicating an urgent need for bias mitigation such as gender-balanced fine-tuning and targeted data augmentation. Cross-linguistic evaluations show that audio-grounded models maintain competitive diagnostic performance even when transcription accuracy deteriorates, as seen in the LeNormand dataset. This suggests that higher-order acoustic features may support language-agnostic reasoning capabilities. Our robustness experiments reveal that, despite modest absolute F1 scores, model performance remains stable under perturbation, indicating inherent resilience that could be enhanced through further optimization.

By integrating these evaluations into the HELM framework, our work transforms isolated model assessments into a transparent, extensible benchmark. It reveals both where foundation models already offer clinical utility and where substantial limitations persist—particularly in cross-lingual generalization, symptom-level precision, and reliability under real-world constraints.

Future work will extend this benchmark to continuous speech and conversational settings, expand coverage to low-resource languages and neurodiverse populations, and evaluate model explanations for clinical faithfulness. We also plan to investigate privacy-preserving fine-tuning paradigms, such as federated learning, to facilitate deployment in sensitive pediatric settings. Collectively, these directions aim to bridge the gap between promising laboratory advances and the development of clinically robust, ethically sound AI systems for SLP.

Limitations

Despite the promising results and the comprehensive scope of our benchmark, several limitations warrant discussion, particularly in the context of ethical and practical considerations for clinical deployment in SLP.

First, while our evaluation encompasses multiple clinically relevant tasks, the datasets employed—though diverse—remain limited in both scale and demographic representation. The majority of speech samples are drawn from English and French speakers, resulting in underrepresentation of other languages, dialects, and sociolinguistic backgrounds. This constraint may limit the generalizability of our findings to more linguistically and culturally diverse populations.

Second, our fairness analysis reveals systematic performance disparities across gender, with male speakers receiving consistently higher diagnostic accuracy. This pattern suggests the presence of gender-related biases, potentially inherited from pretraining corpora or upstream acoustic encoders. Such disparities pose ethical challenges, particularly when AI outputs inform clinical decisions, and underscore the importance of bias auditing and mitigation to ensure fair and just outcomes across patient groups.

Third, although our robustness experiments indicate some degree of resilience to background noise, these evaluations are not exhaustive. Real-world clinical settings, especially pediatric and multilingual environments, often involve significant acoustic variability. Without extensive training on noisy or augmented data, the reliability of these models in such conditions remains uncertain, highlighting the need for further investigation.

Finally, ethical considerations surrounding privacy and consent are central to the deployment of AI systems in sensitive clinical domains. Our current setup does not yet incorporate privacy-preserving learning/evaluation framework. Addressing these concerns is essential to safeguard patient data and build trust among clinicians, patients, and caregivers.

References

Michael D. Abràmoff, Patrick T. Lavin, Mary Birch, Nicholas Shah, and John C. Folk. 2018. Pivotal trial of an autonomous AI–based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1:39.

American Speech-Language-Hearing Association. 2016. Speech sound disorders: Articulation and phonology. `https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/`. [Practice Portal].

Nicholas R. Benway, Jonathan L. Preston, Elaine R. Hitchcock, Adam Salekin, Harsh Sharma, and Tara McAllister. 2022. PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of R. In *Proceedings of Interspeech 2022*, pages 2408–2412, Incheon, Republic of Korea. International Speech Communication Association (ISCA).

Anish Bhattacharya and 1 others. 2024. Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14(1):15573.

Centers for Disease Control and Prevention. 2024. Why act early if you're concerned about development? `https://www.cdc.gov/ncbddd/actearly/whyActEarly.html`. Accessed May 10, 2025.

Claire Cordella, Manuel J. Marte, Hantian Liu, and Swathi Kiran. 2025. An introduction to machine learning for speech-language pathologists: Concepts, terminology, and emerging applications. *Perspectives of the ASHA Special Interest Groups*, 10(2):432–450.

Chinmoy Deka, Abhishek Shrivastava, Ajish K Abraham, Saurabh Nautiyal, and Praveen Chauhan. 2025. Ai-based automated speech therapy tools for persons with speech sound disorder: a systematic literature review. *Speech, Language and Hearing*, 28(1):2359274.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, and 1 others. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Elozino Egonmwan, Vittorio Castelli, and Md Arafat Sultan. 2019. Cross-task knowledge transfer for query-based text summarization. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 72–77, Hong Kong, China. Association for Computational Linguistics.

Paul E Engelhardt and 1 others. 2021. A review of automated speech and language features for assessment of cognitive and thought disorders. *Frontiers in Psychology*.

Ahmed Eshky, Mário S. Ribeiro, Jane Cleland, Korin Richmond, Zoe Roxburgh, James Scobbie, and Alan Wrench. 2018. Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *Proceedings of Interspeech*, pages 2342–2346, Hyderabad, India. International Speech Communication Association (ISCA).

Zhaoxi Fan and 1 others. 2024. A benchmark for automatic speech recognition on child speech. In *Proc. INTERSPEECH*.

S. C. Fanni, A. Marcucci, F. Volpi, S. Valentino, E. Neri, and C. Romei. 2023. Artificial intelligence–based software with ce mark for chest x–ray interpretation: Opportunities and challenges. *Diagnostics*, 13(12):2020.

M. E. Foster, A. L. Choo, and S. A. Smith. 2023. Speech-language disorder severity, academic success, and socioemotional functioning among multilingual and english children. *Frontiers in Psychology*, 14:1096145.

D. Gibbard, L. Coglan, and J. MacDonald. 2004. Cost-effectiveness analysis of a preschool speech and language therapy service. *International Journal of Language & Communication Disorders*, 39(1):1–11.

John Heilmann and 1 others. 2023. Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*.

E. R. Hitchcock, D. Harel, and T. M. Byun. 2015. Social, emotional, and academic impact of residual speech errors in school-age children. *Seminars in Speech and Language*, 36(4):283–294.

IBM Granite Team. 2025. Granite speech model documentation. `https://www.ibm.com/granite/docs/models/speech/`. Version last updated 14 May 2025; accessed 15 May 2025.

Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roee Aharoni, Bernd Bohnet, and Mor Geva. 2023. A comprehensive evaluation of tool-assisted generation strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13856–13878, Singapore. Association for Computational Linguistics.

Ondřej Klempíř and Radim Krupička. 2024. Analyzing wav2vec 1.0 embeddings for cross-database parkinson's disease detection and speech features extraction. *Sensors*, 24(17):5520. Shows wav2vec embeddings generalize across Italian and English PD datasets.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Jessica M. Lammert, Angela C. Roberts, Ken McRae, Laura J. Batterink, and Blake E. Butler. 2025. Early identification of language disorders using natural language processing and machine learning: Challenges and emerging approaches. *Journal of Speech, Language, and Hearing Research*, 68(2):705–718.

Marie-Thérèse Le Normand. 1997. Early morphological development in french children. In A. S. Olofsson and Sven Strömqvist, editors, *Cross-Linguistic Studies of Dyslexia and Early Language Development*, pages 59–79. Office for Official Publications of the European Communities, Luxembourg.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. In *Proc. INTERSPEECH*.

Umer Maqsood and 1 others. 2024. Large language models for dysfluency detection in stuttered speech. *arXiv preprint arXiv:2406.11025*.

Chirag Nagpal, Subhashini Venugopalan, Jimmy Tobin, Marilyn Ladewig, Katherine Heller, and Katrin Tomanek. 2025. Speech recognition with llms adapted to disordered speech using reinforcement learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Giacomo Sanguedolce and 1 others. 2024. Universal speech disorder recognition: Towards a foundation model for cross-pathology generalisation. `https://aphasia.talkbank.org/publications/2024/Sanguedolce24b.pdf`. Accessed May 2025.

G. Saon, A. Dekel, A. Brooks, T. Nagano, A. Daniels, A. Satt, A. Mittal, B. Kingsbury, and *et al.* 2025. Granite–speech: open-source speech-aware llms with strong english asr capabilities. *arXiv preprint*, arXiv:2505.08699.

Phyllis Schneider, Denyse Hayward, and Rebecca V. Dubé. 2006. Storytelling from pictures using the edmonton narrative norms instrument. *Journal of Speech-Language Pathology and Audiology*, 30:224–238.

Alyssa Schwartz and 1 others. 2024. Perceptions of artificial intelligence and chatgpt by speech-language pathologists and students. *American Journal of Speech-Language Pathology*.

Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. Personalizing asr for dysarthric and accented speech with limited data. In *Interspeech 2019*, pages 784–788.

Patti Simon and Sara Rosenbaum. 2016. *Speech and language disorders in children: Implications for the social security administration's supplemental security income program*. National Academies Press.

Hyewon Suh, Aayushi Dangol, Hedda Meadan, Carol A. Miller, and Julie A. Kientz. 2024. Opportunities and challenges for ai-based support for speech-language pathologists. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, CHIWORK '24, New York, NY, USA. Association for Computing Machinery.

Charalambos Themistocleous. 2024. Open brain AI. automatic language assessment. In *Proceedings of the Fifth Workshop on Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 45–53, Torino, Italia. ELRA and ICCL.

Dawn A. Tucker and Stella A. McKinnon. 2020. Addressing the shortage of speech–language pathologists in rural areas: Barriers and solutions. *Journal of Rural Health*, 36(4):620–628.

U.S. National Institute on Deafness and Other Communication Disorders. 2025. Quick statistics about voice, speech, language.

Qinyuan Ye. 2024. Cross-task generalization abilities of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 255–262, Mexico City, Mexico. Association for Computational Linguistics.

Lin Zhang and 1 others. 2024. Speech recognition with llms adapted to disordered speech using reinforcement learning. *arXiv preprint arXiv:2501.00039*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts

A.1 Scenario 1: Binary Classification

Prompt `You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided, typically consisting of a speech prompt from a pathologist followed by a child's repetition. The prompt the child is trying to repeat is as follows: words. Based on your professional expertise: 1. Assess the child's speech in the recording for signs of typical development or potential speech-language disorder. 2. Conclude your analysis with one of the following labels only: typically developing or speech disorder. 3. Provide your response as a single letter without any additional explanation, commentary, or unnecessary text.`

A.2 Scenario 2: ASR-Based Classification

Prompt `You are a highly experienced Speech-Language Pathologist (SLP). An audio recording is provided to you, typically consisting of a speech prompt from a pathologist followed by a child's repetition. Based on your expertise transcribe the child's speech into text. Do not make any assumptions about the words the child is expected to say. Only transcribe based on the words that the child actually says. Only respond with the text transcription, no other text or commentary.`

A.3 Scenario 3: Transcription Accuracy

Prompt `You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided, typically consisting of a speech prompt from a pathologist followed by a child's repetition. Based on your expertise transcribe the child's speech into text. Try to understand what the child is expected to say. And only respond with the transcription of the child's speech. Not the pathologist's prompt or any other commentary. Only respond with the text transcription, no other text, commentary or punctuations.`

A.4 Scenario 4: Disorder Type Classification

Prompt `You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided, typically consisting of a speech prompt from a pathologist followed by a child's repetition. The prompt text the child is trying to repeat is as follows: words. Based on your professional` expertise: 1. Assess the child's speech in the recording for signs of typical development or potential speech-language disorder. 2. Conclude your analysis with one of the following labels only: A - 'typically developing' (child's speech patterns and development are within normal age-appropriate ranges), B - 'articulation' (difficulty producing specific speech sounds correctly, such as substituting, omitting, or distorting sounds), C - 'phonological' (difficulty understanding and using the sound system of language, affecting sounds of a particular type). 3. Provide your response as a single letter without any additional explanation, commentary, or unnecessary text

A.5 Scenario 5: Disorder Symptom Classification

Prompt You are a highly experienced Speech-Language Pathologist (SLP). An audio recording will be provided, typically consisting of a speech prompt from a pathologist followed by a child's repetition. The target phrase the child is attempting to repeat is: {words}. Based on your professional expertise, assess the child's speech in the recording and identify any abnormal features. These features can be one of the following: A — 'substitution' (the child replaces one word, syllable, or sound with another), B — 'omission' (the child omits a word, syllable, or sound), C — 'addition' (the child adds an extra word, syllable, or sound), D — 'typically developing' (the child's speech is appropriate for their age), or E — 'stuttering' (the child exhibits repetition, prolongation, or difficulty initiating speech). Provide your response as a single letter (A–E) only, without any additional explanation or commentary.

B Fine-tuning details

We perform supervised fine-tuning on three models, including Qwen2-Audio 7B, Qwen2.5-Omni 3B, and Qwen2.5-Omni 7B using LLaMA-Factory framework (Zheng et al., 2024). We set the same fine-tuning hyperparameters for those models, presented in Table 2 below. Regarding the prompts used for the three ablation settings for fine-tuning models, we present them as follows.

1. **ASR-only without asterisk**
   <audio>Transcribe this sound into text.

2. **ASR-only with asterisk**
   <audio>Transcribe this sound into text. If the speech is disordered, please mark the words with an asterisk.

Table 2: Finetuning hyperparameters

Hyperparameter	Value
LoRA rank	32
LoRA alpha	64
LoRA modules	all linear layers
Maximum token length	4096
Batch size	32
Epochs	3
Learning rate (LR)	0.0001
LR scheduling	cosine
Warm-up ratio	0.1

C Detail Results

In this section, we present our evaluation results of Scenrio 1 to Scenaario 5 in Table 3 to Table 4, respectively. All experiments are conducted once.

Table 3: Model performance in Scenarios 1

Model	Macro F1↑	Micro F1↑	Exact Match↑
gemini-2.0-flash-lite	0.34	0.46	0.48
gemini-2.0-flash	0.33	0.46	0.46
gpt-4o-mini-audio	0.21	0.20	0.20
gpt-4o-audio	0.16	0.44	0.44
gpt-4o-mini-transcribe	0.34	0.56	0.56
gpt-4o-transcribe	0.37	0.47	0.47
whisper-gpt4o	0.37	0.47	0.47
qwen2.5-omni-7b	0.79	0.71	0.71
qwen2.5-omni-3b	0.59	0.42	0.42
qwen2-audio-7b-instruct	0.21	0.45	0.45
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.25	0.32	0.32
granite-speech-3.3-8b	0.00	0.00	0.00
granite-speech-3.3-2b	0.00	0.00	0.00
granite-speech-3.2-8b	0.00	0.00	0.00
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.93	0.95	0.95
qwen2.5-omni-3b (finetuned)	0.76	0.89	0.89
qwen2-audio-instruct (finetuned)	0.01	0.30	0.30
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.81	0.67	0.67
qwen2.5-omni-3b (finetuned)	0.20	0.25	0.25
qwen2-audio-instruct (finetuned)	0.01	0.36	0.36
Fewshot Prompting			
gpt-4o-mini-audio	0.19	0.23	0.23
gpt-4o-audio	0.21	0.71	0.71
gpt-4o-mini-transcribe	0.44	0.72	0.72
gpt-4o-transcribe	0.45	0.72	0.72
whisper-gpt4o	0.24	0.72	0.72

Table 4: Model performance in Scenarios 2

Model	Macro F1↑	Micro F1↑	Exact Match↑
gemini-2.0-flash-lite	0.01	0.01	0.01
gemini-2.0-flash	0.00	0.00	0.00
gpt-4o-mini-audio	0.03	0.07	0.07
gpt-4o-audio	0.01	0.02	0.02
gpt-4o-transcribe	0.03	0.13	0.13
whisper-gpt4o	0.00	0.00	0.00
qwen2.5-omni-7b	0.01	0.07	0.07
qwen2.5-omni-3b	0.00	0.01	0.01
qwen2-audio-7b-instruct	0.00	0.03	0.03
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.00	0.06	0.06
granite-speech-3.3-8b	0.00	0.03	0.03
granite-speech-3.3-2b	0.00	0.02	0.02
granite-speech-3.2-8b	0.00	0.04	0.04
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.06	0.44	0.44
qwen2.5-omni-3b (finetuned)	0.08	0.23	0.23
qwen2-audio-instruct (finetuned)	0.03	0.32	0.32
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.06	0.44	0.44
qwen2.5-omni-3b (finetuned)	0.08	0.23	0.23
qwen2-audio-instruct (finetuned)	0.16	0.44	0.44
Fewshot Prompting			
gpt-4o-mini-audio	0.03	0.09	0.09
gpt-4o-audio	0.04	0.12	0.12
gpt-4o-mini-transcribe	0.01	0.01	0.01
gpt-4o-transcribe	0.01	0.01	0.01
whisper-gpt4o	0.00	0.01	0.01

Table 5: Model performance in Scenarios 3

Model	WER↓	MER↓	WIP↑
gemini-2.0-flash-lite	0.83	0.68	0.21
gemini-2.0-flash	0.94	0.71	0.24
gpt-4o-mini-audio	1.40	0.68	0.26
gpt-4o-audio	2.25	0.70	0.26
gpt-4o-mini-transcribe	1.54	0.75	0.19
gpt-4o-transcribe	1.31	0.74	0.23
whisper-gpt4o	2.84	0.75	0.18
qwen2.5-omni-7b	2.17	0.74	0.22
qwen2.5-omni-3b	4.98	0.75	0.22
qwen2-audio-7b-instruct	4.98	0.75	0.22
qwen-audio-chat	12.3	0.90	0.08
phi-multimodal	6.36	0.76	0.20
granite-speech-3.3-8b	13.50	0.93	0.05
granite-speech-3.3-2b	4.13	0.89	0.07
granite-speech-3.2-8b	6.64	0.66	0.14
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	1.40	0.52	0.41
qwen2.5-omni-3b (finetuned)	0.97	0.53	0.39
qwen2-audio-instruct (finetuned)	0.58	0.43	0.50
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	1.76	0.46	0.47
qwen2.5-omni-3b (finetuned)	0.95	0.49	0.43
qwen2-audio-instruct (finetuned)	0.52	0.38	0.56
Fewshot Prompting			
gpt-4o-mini-audio	1.58	0.65	0.28
gpt-4o-audio	1.73	0.62	0.30
gpt-4o-mini-transcribe	1.08	0.80	0.12
gpt-4o-transcribe	1.01	0.79	0.14
whisper-gpt4o	1.89	0.77	0.16

Table 6: Model performance in Scenarios 4

Model	Macro F1↑	Micro F1↑	Exact Match↑
gemini-2.0-flash-lite	0.17	0.19	0.19
gemini-2.0-flash	0.20	0.46	0.46
gpt-4o-mini-audio	0.12	0.15	0.15
gpt-4o-audio	0.14	0.36	0.36
gpt-4o-mini-transcribe	0.28	0.42	0.46
gpt-4o-transcribe	0.32	0.41	0.41
whisper-gpt4o	0.33	0.43	0.41
qwen2.5-omni-7b	0.79	0.71	0.71
qwen2.5-omni-3b	0.56	0.44	0.44
qwen2-audio-7b-instruct	0.20	0.33	0.33
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.18	0.37	0.37
granite-speech-3.3-8b	0.00	0.00	0.00
granite-speech-3.3-2b	0.00	0.00	0.00
granite-speech-3.2-8b	0.00	0.01	0.01
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.93	0.95	0.95
qwen2.5-omni-3b (finetuned)	0.76	0.89	0.89
qwen2-audio-instruct (finetuned)	0.02	0.21	0.21
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.28	0.40	0.40
qwen2.5-omni-3b (finetuned)	0.24	0.36	0.36
qwen2-audio-instruct (finetuned)	0.05	0.27	0.27

Table 7: Model performance in Scenarios 5

Model	Macro F1↑	Micro F1↑	Exact Match↑
gemini-2.0-flash-lite	0.19	0.43	0.43
gemini-2.0-flash	0.09	0.22	0.22
gpt-4o-mini-audio	0.10	0.39	0.39
gpt-4o-audio	0.20	0.49	0.49
gpt-4o-mini-transcribe	0.15	0.26	0.26
gpt-4o-transcribe	0.13	0.28	0.28
whisper-gpt4o	0.18	0.36	0.36
qwen2.5-omni-7b	0.79	0.71	0.71
qwen2.5-omni-3b	0.56	0.44	0.44
qwen2-audio-7b-instruct	0.08	0.10	0.10
qwen-audio-chat	0.00	0.00	0.00
phi-multimodal	0.09	0.13	0.13
granite-speech-3.3-8b	0.00	0.00	0.00
granite-speech-3.3-2b	0.00	0.00	0.00
granite-speech-3.2-8b	0.06	0.20	0.20
Finetuned Models with Asterisk			
qwen2.5-omni-7b (finetuned)	0.93	0.95	0.95
qwen2.5-omni-3b (finetuned)	0.76	0.89	0.89
qwen2-audio-instruct (finetuned)	0.00	0.07	0.07
Finetuned Models without Asterisk			
qwen2.5-omni-7b (finetuned)	0.16	0.34	0.34
qwen2.5-omni-3b (finetuned)	0.08	0.16	0.16
qwen2-audio-instruct (finetuned)	0.01	0.08	0.08