

When Audio-Language Models Fail to Leverage Multimodal Context for Dysarthric Speech Recognition

Pehuén Moure^{1*}, Niclas Pokel^{1*}, Bilal Bounajma²,

Yingqiang Gao³, Roman Boehringer¹, Longbiao Cheng¹, Shih-Chii Liu¹

¹Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

²Department of Mathematics, ETH Zurich, Switzerland

³Department of Computational Linguistics, University of Zurich, Switzerland

{pehuen,npokel,roman,shih}@ini.ethz.ch, bbounajma@student.ethz.ch

longbiao@ini.uzh.ch, yingqiang.gao@cl.uzh.ch

Abstract

Automatic speech recognition (ASR) systems remain brittle on dysarthric and other atypical speech. Recent audio-language models raise the possibility of improving performance by conditioning on additional clinical context at inference time, but it is unclear whether these models can make use of such information. We introduce an evaluation protocol built on the Speech Accessibility Project (SAP) dataset that tests whether diagnosis labels, clinician-derived speech ratings, and progressively richer clinical descriptions improve transcription accuracy for dysarthric speech. Across matched comparisons on ten models, we find that current models do not meaningfully use this context: diagnosis-informed and clinically detailed prompts yield negligible improvements and often degrade word error rate, and counterfactual controls reveal that the models respond to the presence of prompt text rather than to its content. We complement the prompting analysis with context-dependent fine-tuning, showing that LoRA adaptation with a mixture of clinical prompt formats achieves a WER of 0.066, a 52% relative reduction over the frozen baseline, while preserving performance when context is unavailable. The residual benefit of context concentrates on the hardest utterances, and a difficulty-aware gate that applies clinical context only where it helps lowers WER by about 10% relative over the audio-only baseline on both architectures. These results clarify where current models fall short and provide a testbed for measuring progress toward more inclusive ASR.

1 Introduction

Modern ASR systems have achieved near-human performance on standard benchmarks (Radford et al., 2022; Zhang et al., 2022), but these gains are unevenly distributed. Speakers with dysarthria, a motor speech disorder arising from neurological conditions such as cerebral palsy (CP), amyotrophic lateral sclerosis (ALS), Parkinson’s disease, and stroke, experience dramatically higher word error rates (WER) (Tobin et al., 2024). This gap renders voice-based interfaces such as voice assistants, augmentative and alternative communication (AAC¹) devices, and clinical documentation tools effectively unusable for millions of people, excluding them from an increasingly voice-first digital infrastructure (Hoffman et al., 2014; Jaddoh et al., 2025).

A new opportunity has emerged with audio-native multimodal LLMs such as Gemini 2.0 (Gemini Team et al., 2025) and GPT-4o (OpenAI Team et al., 2024), which process raw audio directly alongside natural-language instructions. These models exhibit strong in-context learning on natural language processing (NLP) tasks (Dong et al., 2024); the open

*Equal contribution.

¹<https://www.asha.org/public/speech/disorders/aac/>

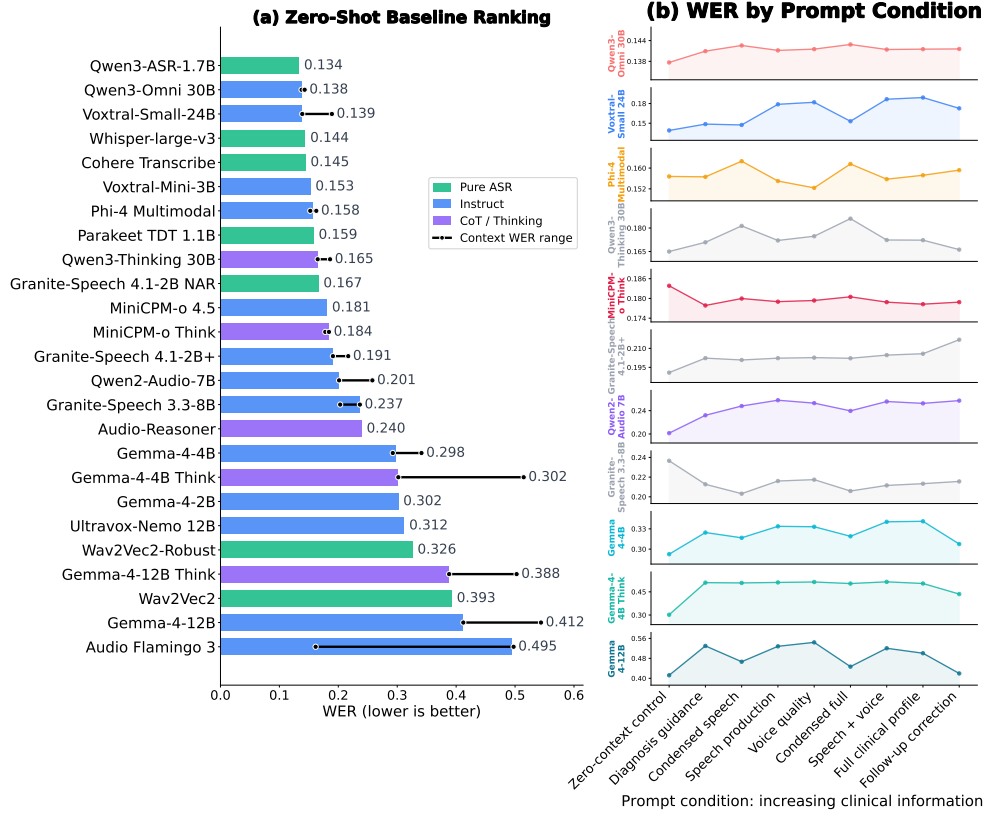


Figure 1: (a) ASR baseline ranking on dysarthric speech (zero-shot WER). Bars are colored by model type; black markers show the WER range across prompt conditions for promptable models. (b) Per-model WER across prompt conditions ($N = 11,218$ matched samples). Each facet shows one model on its own scale. Frozen models either remain flat or degrade as clinical context increases. AF3’s apparent gain with context reflects suppressed hallucination rather than genuine context use (Section 4.2).

question is whether this capability extends to audio understanding of atypical speech. Clinical metadata about a speaker’s condition (diagnosis, severity, articulatory characteristics) is routinely available in medical and therapeutic settings (Morris et al., 2024) and could, in principle, serve as a natural-language “adapter” at inference time without model retraining.

We make the following contributions. Working on the Speech Accessibility Project (SAP, Zheng et al. (2025)) dataset, first, we introduce a systematic evaluation protocol that isolates the effect of different clinical context types on dysarthric ASR, spanning pathology-informed prompting with progressively richer clinical profiles, follow-up correction, and counterfactual context controls (wrong clinical content, length-matched random text, and instruction-only filler). Second, we conduct the first (to the best of our knowledge) broad assessment of ten audio-language models under controlled clinical prompting conditions on dysarthric speech, within a zero-shot comparison of 24 recent ASR and audio-language systems, showing that frozen models consistently fail to exploit clinical context, with three distinct failure modes (robust, degrading, and format-dependent). Third, we show on two model families (Voxtral-Small-24B and Gemma-4-E2B) that context-dependent fine-tuning (LoRA adaptation (Hu et al., 2021) on a mixture of clinical prompt formats) removes the prompt-induced degradation and achieves the lowest WER of any system we evaluate while preserving performance when context is unavailable; counterfactual controls reveal that the fine-tuned models respond to the presence of clinical context rather than to its graded content. This indicates that the poor performance of multimodal models is due to the absence of similar data samples in their training data rather than some fundamental limitation. Fourth, we show that the residual benefit of context concentrates on the hardest utterances and formulate a difficulty-aware gate that applies clinical context only where it helps, lowering

WER by about 10% relative over the audio-only baseline on both architectures. We publicly release our evaluation code, prompt templates, and evaluation pipeline.

2 Related Work

ASR for dysarthric and atypical speech. Traditional approaches to dysarthric ASR include speaker-dependent GMM-HMM (Rabiner, 1989) and DNN-HMM (Bhat et al., 2016) systems, data augmentation strategies using tempo and pitch perturbation (Bhat & Strik, 2025), and TTS-based synthesis of dysarthric speech for training data expansion (Hermann & Magimai-Doss, 2023). More recent work has focused on fine-tuning self-supervised models such as wav2vec 2.0 (Hsu et al., 2021) or Whisper (Wagner et al., 2025) on dysarthric corpora, achieving substantial improvements over unadapted baselines. Recent work has also explored data-efficient Whisper-based personalization under limited-data conditions by constructing semantically structured training material from small amounts of user-specific speech (Pokel et al., 2025a). In parallel, parameter-efficient adaptation methods such as low-rank and Bayesian low-rank (Pokel et al., 2025b) fine-tuning have also been investigated for impaired speech recognition. A key limitation shared by all these approaches is the requirement for training or fine-tuning on data from the target speaker population, which is normally expensive to collect.

Closest to our setting are systems built on SAP data itself. The winner of the Interspeech 2025 SAP Challenge fully fine-tunes Parakeet-TDT and reaches 5.97%/8.11% WER on the challenge Test1/Test2 splits (Takahashi et al., 2025), and a concurrent proprietary system based on Gemini 2.5 Flash reaches about 5% WER through few-shot same-speaker in-context personalization (Agarwal et al., 2025). Both use full fine-tuning or speaker personalization and are measured on the challenge private test splits, not on our rated subset, so they serve as a complementary upper bound rather than a matched comparison. Anchored to a frozen Whisper baseline of comparable difficulty across splits (14–18% WER), our parameter-efficient context-dependent LoRA (Section 4.4) lands within one to two points of these systems at far lower adaptation cost (Appendix F). The personalization result independently supports same-speaker audio-based in-context learning, a direction we flag as future work.

Multimodal LLMs with audio understanding. The Gemini model family (Gemini Team et al., 2025) and GPT-4o (OpenAI Team et al., 2024) natively accept audio input alongside text, enabling audio understanding without a separate ASR front-end. Open-weight alternatives include Qwen2-Audio (Chu et al., 2024) and Qwen3-Omni (Xu et al., 2025). Evaluations of these models have focused predominantly on standard speech recognition benchmarks, speaker identification, and audio captioning. To our knowledge, no systematic evaluation exists for their performance on pathological or atypical speech under controlled prompting conditions.

In-context learning for speech and audio. In-context learning is well-studied in the text domain (Dong et al., 2024) and increasingly explored for vision tasks (Zhou et al., 2024). For speech and audio, in-context learning remains underexplored, with most work focusing on few-shot prompting for text-based NLP tasks rather than direct audio conditioning. A recent work (Alsayegh & Masood, 2025) evaluated commercial ASR systems and multimodal large language models on dysarthric speech using the TORGO corpus (Rudzicz et al., 2012), showing strong performance degradation with increasing severity and only limited benefits from simple prompting strategies. However, this line of work remains focused on zero-shot transcription and does not provide a systematic analysis of how different forms of clinical context affect recognition.

3 Evaluation Protocol

We present our evaluation protocol, consisting of a systematic clinical context conditioning assessment pipeline and a supervised fine-tuning adaptation study. The aim is to measure whether (and how) audio-language models can leverage increasing levels of clinical contextual information to improve dysarthric speech recognition (Figure 2).

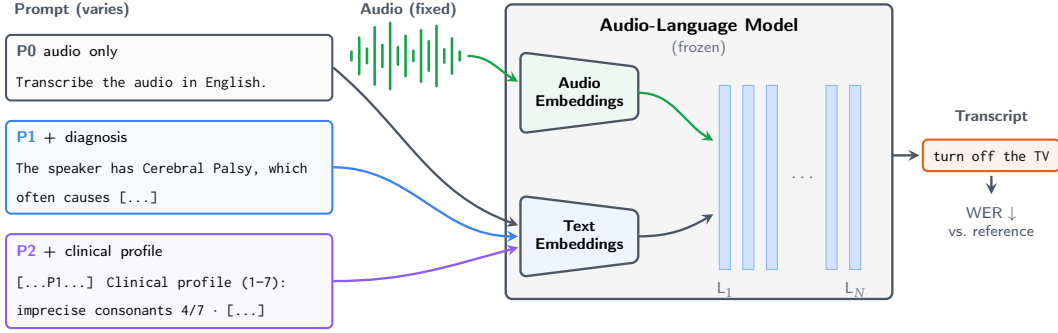


Figure 2: Overview of the evaluation protocol. The same frozen audio-language model receives identical audio while the text prompt varies from a bare transcription instruction (P0) to progressively richer clinical context (P1, P2; Section 3.1); outputs are scored by WER against reference transcripts. Counterfactual controls (Section 4.3) replace the clinical block with wrong, random, or instruction-only text of matched length.

We use standard ASR metrics (WER, CER) and SemScore (Phukon et al., 2025), detailed in the Appendix B, as well as the conventions from the SAP dataset (Zheng et al., 2025): we clip per-sample WER at 1.0, use dual-reference minimum scoring, and apply Whisper’s EnglishTextNormalizer (Radford et al., 2022).

3.1 Prompt Conditions

We define a hierarchy of prompt conditions with progressively richer clinical context, from a zero-context control (P0) through diagnosis guidance (P1) to full clinical profiles (P2); see Table 3 in Appendix A for the full prompts. The clinical ratings are drawn from two complementary profiles: the *speech production profile* captures articulatory characteristics such as imprecise consonants, slow rate, and hypernasality, and the *voice quality profile* captures phonation and vocal-control cues such as breathiness, tremor, and monopitch (full dimension lists in Appendix A).

Beyond this hierarchy, three counterfactual control conditions preserve the structure of the full clinical profile while manipulating its content: a *wrong-context* condition in which the diagnosis and per-sample ratings are replaced with those of a different speaker (cyclic rotation across etiologies), a *length-matched random text* condition in which the clinical block is replaced by unrelated text of matched length, and a *length-matched instruction-only* condition in which it is replaced by generic transcription instructions (templates in Appendix A).

3.2 Models Under Evaluation

We evaluate ten audio-language models under the full prompt-condition battery: Audio Flamingo 3 (Goel et al., 2025), MiniCPM-o 4.5 Think (Yu et al., 2025), Qwen2-Audio-7B (Chu et al., 2024), Qwen3-Omni-30B (Xu et al., 2025), Qwen3-Thinking-30B (Xu et al., 2025), Phi-4 Multimodal (Microsoft Team et al., 2025), Voxtral-Small-24B (Liu et al., 2025), Gemma-4-4B (Google DeepMind, 2026), Gemma-4-4B Think (Google DeepMind, 2026), and Gemma-4-12B (Google DeepMind, 2026) (a unified, encoder-free multimodal model). Five further audio-language models (Liu et al., 2025; Fixie AI, 2024; Google DeepMind, 2026; Yu et al., 2025; Xie et al., 2025) are included in the zero-shot ranking only (Section 4.1, Table 19), as compute constraints precluded the full prompt grid. As baselines, we report nine ASR systems (Radford et al., 2022; Shi et al., 2026; Rekesh et al., 2023; Cohere Labs, 2026; Saon et al., 2025; Hsu et al., 2021; Baevski et al., 2020), including Whisper large-v3, Parakeet-TDT-1.1B, and two legacy Wav2Vec2 models (complete list in Table 19). Supervised adaptation (Section 4.4) is conducted on Voxtral-Small-24B and replicated on a second model family, Gemma-4-E2B (listed as Gemma-4-2B in Table 19).

3.3 Dataset

We conduct our clinical context conditioning evaluation on the Speech Accessibility Project (SAP, Zheng et al. (2025)) dataset, the largest available dysarthric speech corpus. SAP

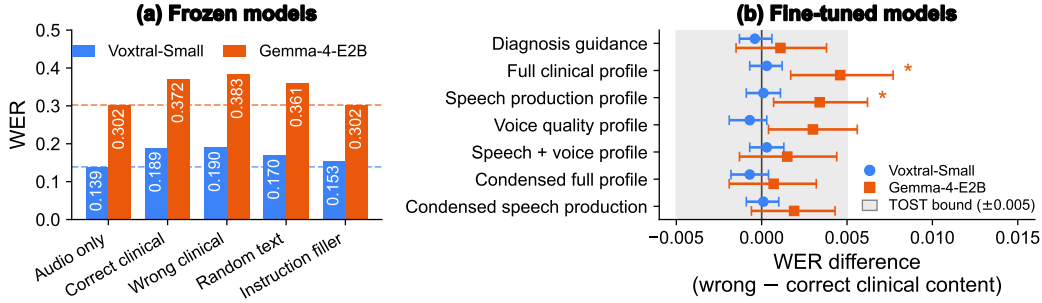


Figure 3: Counterfactual context controls on two architectures. (a) Frozen WER by prompt condition; dashed lines mark the audio-only baselines. (b) WER difference between wrong and correct clinical content for the fine-tuned context-dependent systems. Voxtral-Small is equivalent throughout, and Gemma-4-E2B shows a small correct-content gain at two conditions (*).

contains recordings from 959 speakers across five etiologies (Parkinson’s disease, ALS, cerebral palsy, Down syndrome, and stroke).

A subset of samples contains clinician-provided ratings across perceptual dimensions (e.g., imprecise consonants, breathy voice, monopitch), which form the basis for our clinical prompt conditions. The rated subset used for prompt-condition comparisons comprises $N = 11,218$ matched samples, drawn from four etiologies (Parkinson’s disease, ALS, cerebral palsy, and Down syndrome); stroke speakers are absent from this subset.

4 Results

4.1 Zero-Shot Baselines

Figure 1(a) and Table 4 rank all models by zero-shot WER on the SAP rated subset. Qwen3-ASR-1.7B achieves the lowest WER (0.134), followed by Qwen3-Omni-30B (0.138), Voxtral-Small-24B (0.139), Whisper large-v3 (0.144), and Cohere Transcribe (0.145). The conventional speech recognition system Wav2Vec2 (0.393) lags substantially behind (full results in Table 22).

4.2 Clinical Context Conditioning of Frozen Models

Across all ten models evaluated on the full set of prompt conditions, adding clinical context to the prompt either has a negligible effect or actively degrades WER. Figure 1(b) shows this trend (with the exception of Audio Flamingo 3, as discussed below). The full condition matrix and paired comparisons are in Tables 5 and 6 (Appendix C).²

The results reveal three failure modes. **Robust models** (Phi-4, Qwen3-Omni, MiniCPM-o Think, Qwen3-Thinking) show negligible WER change under clinical context (-0.006 to $+0.007$). **Degrading models** (Voxtral-Small, Qwen2-Audio, Gemma-4-4B, Gemma-4-4B Think, Gemma-4-12B) show significant WER increases under clinical prompting, with Gemma-4-4B Think exhibiting catastrophic failure (Δ WER $+0.202$, 42.7% of samples degraded) as the model’s chain-of-thought process compounds with prompt length and displaces the transcription itself (Section 5). **Format-dependent models** (Audio Flamingo 3) appear to improve markedly under prompting, but AF3’s unprompted WER of 0.495 reflects non-transcription output (audio descriptions, commentary) in the absence of a task instruction; the structured prompt merely restores baseline transcription behavior (≈ 0.162), and AF3 is nearly insensitive to the prompt’s clinical content (0.162–0.164).

Error decomposition (Table 10, Appendix C) gives a mechanistic account, with increased insertions for Voxtral-Small and Gemma-4-12B, whereas Gemma-4-4B Think degrades through truncation instead (category breakdown in Table 9; mechanisms in Section 5).

²All reported p -values use Benjamini–Hochberg false-discovery-rate (FDR) correction across the 50 model $\times$ condition comparisons.

4.3 Counterfactual Context Controls

The degradation documented above could stem from models reacting to the clinical *content* of the prompt or merely to the *presence* of additional prompt text. The control conditions (Section 3.1) separate these explanations on two architectures, Voxtral-Small and the smaller Gemma-4-E2B (Figure 3a). On both models, wrong clinical content is no better than correct content, length-matched random text degrades almost as much, and length-matched instruction-only filler lies closest to the audio-only baseline, ruling out prompt length as the cause. The frozen degradation therefore tracks the presence of non-instructional prompt text rather than its clinical accuracy, and the effect replicates across model families (Figure 3a; Table 11, Appendix E; fine-tuned controls in Section 4.4).

Severity and etiology stratification. Figure 10 (Appendix L) shows how baseline WER increases sharply with clinical severity across all models, with the steepest degradation for cerebral palsy and Down syndrome beyond severity level 2. The degradation pattern from prompting is broadly consistent across severity levels and etiologies. No subgroup shows reliable improvement under clinical prompting for frozen models. Full numerical breakdowns are in Appendix C Tables 7 and 8.

Iterative correction. Providing a prior-pass transcript alongside the full clinical profile does not systematically improve WER over single-pass prompting; Voxtral-Small and Qwen2-Audio remain above their zero-context baselines (Appendix L).

4.4 Supervised Adaptation

Given that frozen multimodal models uniformly fail to exploit clinical context, we investigate whether supervised fine-tuning can teach a model to use it. We fine-tune Voxtral-Small-24B with LoRA using 5-fold speaker-disjoint cross-validation across 437 unique speakers (87–88 held out per fold), yielding 11,218 utterances pooled out-of-fold. We train two systems. The **audio-only** system is fine-tuned with audio input only and no clinical prompt. The **context-dependent** system is fine-tuned on a mixture of audio-only, speech + voice profile, and condensed full profile samples, exposing the model to varied context formats during training (see Appendix D for full training details); all results are pooled out-of-fold predictions from the last checkpoint per fold.

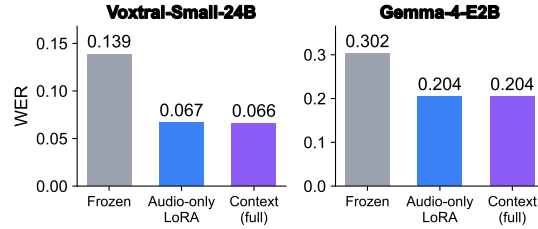


Figure 4: Fine-tuning comparison on both architectures (5-fold cross-validation, 437 speakers). LoRA halves WER on Voxtral-Small (0.139 $\rightarrow$ 0.067) and reduces Gemma-4-E2B from 0.302 to 0.204; the context-dependent systems match audio-only fine-tuning. Voxtral-Small values in Table 15.

The audio-only system reduces WER from 0.139 (frozen) to 0.067, a 52% relative reduction (Figure 4). The context-dependent system achieves the best overall WER of 0.066 when evaluated with the full clinical profile, a modest 0.8% relative gain over audio-only that does not reach significance at the speaker level ($p = 0.55$); the same weights produce WER 0.067 at zero-context ($p = 0.31$; Table 16). Notably, providing context incurs no penalty. The context-dependent model without context (0.067) performs comparably to audio-only (0.067, $p = 0.93$), indicating that context-dependent fine-tuning does not sacrifice audio-only quality.

Replication on a second model family. To test whether these findings are specific to Voxtral-Small, we replicate the fine-tuning protocol on Gemma-4-E2B, the weakest promptable model family in our frozen evaluation (training details in Appendix D.2). The pattern reproduces: audio-only LoRA reduces WER from 0.302 (frozen zero-context) to 0.204, the context-dependent system evaluated with the full clinical profile also reaches 0.204, and the

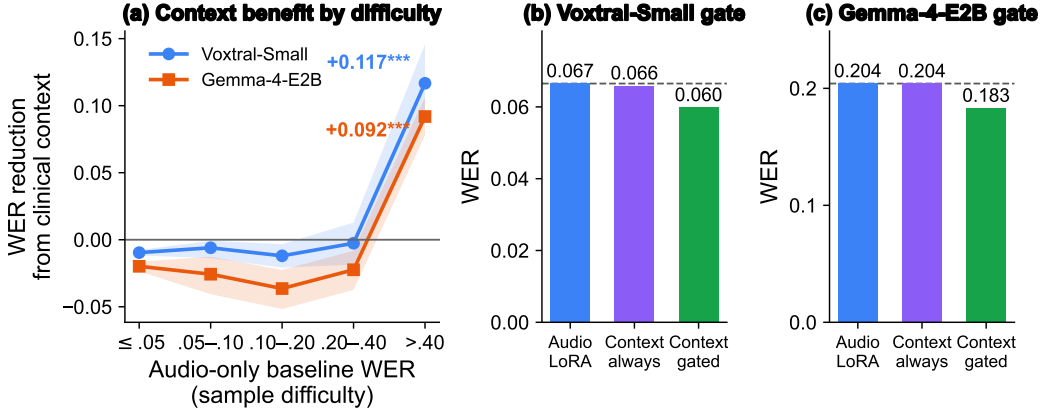


Figure 5: Clinical context yields the largest gains on the most difficult utterances (5-fold cross-validation, both architectures). (a) WER reduction from the full clinical profile as a function of audio-only sample difficulty; shaded bands are speaker-clustered 95% intervals. The benefit concentrates on the hardest utterances, while easy utterances incur a small penalty. (b, c) Applying context only to hard utterances (difficulty-aware gate, oracle difficulty signal) lowers WER relative to the audio-only LoRA on both architectures; the dashed line marks the audio-only baseline.

frozen-model degradation under clinical prompting (+0.070 at the full profile) disappears entirely.

Counterfactual controls on the fine-tuned systems. Does the context-dependent system read the clinical content, or only register its presence? We evaluate the same fine-tuned weights with correct versus wrong clinical profiles (wrong-context condition, Section 3.1) on the same utterances and test equivalence directly with two one-sided tests (TOST) at a bound of ± 0.005 WER (Appendix B.3). On Voxtral-Small, correct and wrong clinical content are statistically equivalent at every clinical prompt condition (TOST $p < 0.001$; Figure 3b), and this holds across three further training variants, including a curriculum-trained one (audio adaptation followed by context-only refinement). Gemma-4-E2B is informative in a different way: it shows a small but significant correct-over-wrong gain at the full clinical profile and speech production profile conditions (correct lower by about 0.004 WER, Wilcoxon $p \leq 0.005$) and equivalence elsewhere (Table 12, Appendix E).

Subgroup analysis. Table 26 (Appendix L) compares audio-only $\rightarrow$ context-dependent (full clinical profile) by etiology and severity, showing relative WER change.

Down syndrome speakers show a statistically significant 7.1% relative improvement ($p = 0.044$), and mild-severity speakers show a significant 2.7% relative improvement ($p = 0.033$). ALS speakers show a near-significant 4.6% improvement ($p = 0.066$). Cerebral palsy is the only etiology for which the context-dependent system slightly degrades (+0.006, not significant), consistent with the frozen-model findings.

Comparing these fine-tuning gains to the frozen-model results in Appendix C Table 8 reveals a striking reversal. Under frozen-model clinical prompting (P0 $\rightarrow$ P2), the average across the ten evaluated models is a Δ WER of +0.042 for Down syndrome and +0.029 for cerebral palsy speakers, with Down syndrome degrading on a majority of models and cerebral palsy on five of ten. After context-dependent fine-tuning, the same clinical information that degraded performance for Down syndrome speakers under a frozen model now yields the largest relative improvement (-7.1% , Table 26). The etiology ordering largely inverts: groups most harmed by frozen prompting are most helped by fine-tuning. This suggests that the clinical context contains useful signal for these populations that only becomes accessible once the model is trained to use it. Section 4.5 and the analysis in Section 5.1 trace these differences to baseline acoustic difficulty.

4.5 A Difficulty-Aware Context Gate

The aggregate parity between the audio-only and context-dependent systems hides a strong conditional effect. Holding the fine-tuned weights fixed, the benefit of the full clinical profile grows with sample difficulty, measured by the audio-only baseline WER (Spearman correlation, $p < 10^{-9}$; Figure 5a). On the hardest utterances, those the audio-only system transcribes at WER above 0.40, providing context lowers WER by 0.117 on Voxtral-Small and 0.092 on Gemma-4-E2B (Wilcoxon $p < 10^{-16}$, speaker-clustered intervals excluding zero). On easy utterances, context instead costs accuracy, slightly on Voxtral-Small and more on the weaker Gemma-4-E2B, so the two effects cancel in the aggregate.

The right policy is therefore to apply context only where it helps. Gating on sample difficulty, i.e. providing the clinical profile only for utterances the audio-only system finds hard, lowers overall WER from 0.0665 to 0.0599 on Voxtral-Small and from 0.2041 to 0.1833 on Gemma-4-E2B, a roughly 10% relative reduction over the audio-only baseline on each architecture (Figure 5b,c). We compute the gate with an oracle difficulty signal (the audio-only baseline WER), so these numbers bound what a deployable gate can achieve; the no-reference difficulty signals we tested (system disagreement, severity ratings, audio features) do not yet recover this headroom, leaving a practical gating mechanism as an open problem. The result nevertheless reframes the role of clinical context: as expected, it is usable, and most valuable, precisely on the severely degraded speech where accessibility needs are greatest.

5 Analysis and Discussion

5.1 Insight on the Failure of Frozen Models

The three failure modes identified in Section 4.2 suggest several distinct underlying mechanisms rather than a single explanation.

First, **weak conditioning on prompt content**. Robust models (Phi-4, Qwen3-Omni, MiniCPM-o Think, Qwen3-Thinking) may place much greater weight on the audio signal than on the semantic content of the text prompt, using the latter mainly to control output style or format. Audio Flamingo 3’s near-zero sensitivity to prompt *content* once output format is fixed is also consistent with this interpretation.

Second, **verbosity- and truncation-induced degradation**. For degrading models, longer clinical prompts shift the generation distribution away from faithful transcription. The error decomposition in Section 4.2 supports this finding: Voxtral-Small’s insertion rate increases by +0.212 (verbose reformulation), while Gemma-4-4B Think’s deletion rate increases by +0.174 (heavy truncation). The counterfactual controls (Section 4.3) refine the mechanism: the degradation is reproduced by length-matched random text but not by length-matched instruction-only filler, implicating non-instructional prompt content rather than clinical semantics or sheer prompt length.

Third, **distribution shift**. Clinical prompt formats are far from the training distribution of these models. Without exposure to the clinical-context-to-transcription mapping during training, the models cannot learn to condition on this information.

The supervised results support the third hypothesis directly. Once Voxtral-Small or Gemma-4-E2B is fine-tuned on clinical prompts, the prompt-induced degradation disappears, and the same weights yield a difficulty-conditioned benefit from context (Section 4.5). The counterfactual controls qualify what is learned: with the partial exception of the smaller Gemma model, the fine-tuned systems respond to the presence of clinical context rather than to its graded content. The limitation is not architectural; it is a matter of training exposure, though the exposure our recipe provides teaches models that context matters rather than what it says.

When is context beneficial? We regress per-sample ΔWER (audio-only $\rightarrow$ context-dependent at the full clinical profile) against baseline difficulty, mean severity, etiology, and individual clinical rating dimensions (Figure 6). Baseline audio-only WER is the strongest predictor (Spearman $\rho = -0.12$, $p < 0.001$; Pearson $r = -0.24$), etiology and mean severity

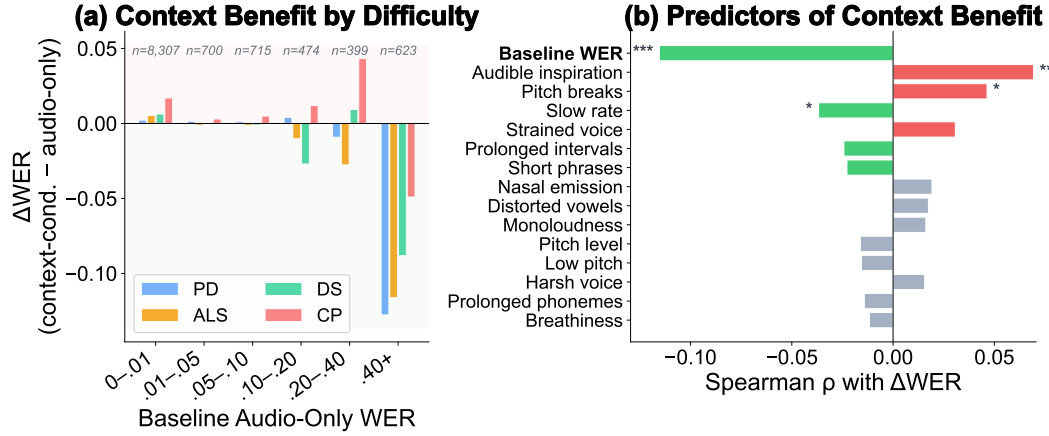


Figure 6: Predictors of per-sample context benefit (audio-only $\rightarrow$ context-dependent, 5-fold cross-validation, $N = 11,218$). (a) Mean Δ WER by baseline difficulty and etiology. (b) Spearman ρ between each predictor and Δ WER. Baseline WER dominates; individual clinical dimensions show near-zero correlation.

add little once difficulty is accounted for (additional $R^2 < 0.015$ each), and individual clinical rating dimensions show near-zero correlation with Δ WER (the strongest is audible inspiration at $\rho = +0.07$, and most fall below $|\rho| = 0.03$), and all clinical prompt conditions improve WER by similar amounts (Table 14). The subgroup findings of Table 26 are therefore better read as a consequence of baseline difficulty distributions across etiologies rather than etiology-specific use of the clinical profiles.

5.2 Hallucination, Verbosity and Refusal

Clinical prompting also changes *how* models fail. Defining a hallucination as an unclipped WER above 100%, zero-context rates range from 2.6% (Voxtral-Small) to 27.9% (Audio Flamingo 3, driven by a preamble artifact that prompting removes, 27.9% $\rightarrow$ 2.9%), while Voxtral-Small rises to 8.3% and Gemma-4-4B Think to 17.2% under prompting; the remaining models vary by ± 1 pp. Manual inspection identifies five recurring patterns: repetitive loops that echo the disfluencies described in the profile, verbose reformulation, language switching, refusal, and reasoning-chain leakage in CoT models. Only the first is directly attributable to clinical content, the model appearing to “anticipate” described disfluencies and reproduce them.

The unified Gemma-4-12B is the clearest instance of refusal. Clinical prompting raises its refusal rate from 2.4% at zero-context to roughly 24%, concentrated on short utterances such as digital-assistant commands (52% refused): rather than transcribing, the model treats the spoken target as an instruction addressed to it and declines (“I cannot fulfill this request”). The refusals come with verbose non-transcription output (P90 output-to-reference length ratio $\approx 8-10\times$), producing the largest insertion-rate increase of any model (+0.621, Table 10) and placing Gemma-4-12B among the degrading models despite competitive zero-context accuracy. Length ratios and fixed/induced decompositions are in Appendix L.

Word-level error decomposition. Aligning the two systems word by word shows that context-dependent fine-tuning repairs 17% of the consonant-word and 20% of the vowel-word errors the audio-only system missed; the repair is not specific to the matching clinical dimension and tracks acoustic difficulty rather than prompt content (Appendix L).

5.3 Limitations

Our evaluation is restricted to the Speech Accessibility Project (SAP) dataset, which, despite being the largest dysarthric speech corpus, presents several structural constraints: the clinical annotations are coarse, single-rater judgments on an ordinal 1–7 scale rather than standardized instruments (e.g., FDA-2), precluding inter-rater reliability analysis. The rated subset is roughly 3% of the dataset and artificially balanced toward more severe and diverse

speech; it overrepresents Parkinson’s disease (67%), provides limited Down syndrome (3%), and contains no stroke samples with complete profiles. The exclusive focus on US English-speaking adults prevents generalization to pediatric or cross-linguistic populations.

The population-level improvement from context is modest and not significant in aggregate, and the fine-tuned systems respond primarily to context presence (Section 5). The supervised experiments cover two architectures, and the speech encoder remains frozen during fine-tuning; retraining it may be important for atypical acoustics. The difficulty-aware gate is computed with an oracle difficulty signal, so it quantifies the benefit available from selective context use rather than a deployable system. Cerebral palsy is the only etiology for which the context-dependent system slightly degrades, possibly reflecting inter-speaker articulatory variability that clinical profiles do not capture, and speakers with severe impairment are underrepresented in the rated subset ($N = 21$ utterances).

The scarcity of dedicated audio-native reasoning models limits the generality of our chain-of-thought findings, since fundamental reasoning limits are hard to separate from model-specific artifacts such as reasoning-chain leakage. Finally, we evaluate isolated utterances with text-based context only, leaving audio-based in-context learning with same-speaker references for future work.

6 Conclusion

We presented a systematic evaluation of whether audio-language models can leverage clinical context for dysarthric speech recognition. Across ten models and nine prompt conditions, frozen models do not reliably use such context, and counterfactual controls show the degradation tracks the mere presence of non-instructional prompt text rather than its accuracy. The three failure modes we identify suggest a limitation of training exposure rather than architecture.

Context-dependent fine-tuning, evaluated on two model families, supports this interpretation. Voxtral-Small reaches a WER of 0.066 (a 52% relative reduction over the frozen baseline) while preserving performance when context is unavailable, and the pattern replicates on Gemma-4-E2B. Counterfactual controls show, however, that with the partial exception of the smaller Gemma model, correct and wrong clinical content are statistically equivalent for the fine-tuned systems: the models learn that context is present, not what it says. The benefit of context grows with acoustic difficulty and concentrates on the hardest utterances, and applying context only there, through a difficulty-aware gate, lowers WER by about 10% relative over the audio-only baseline on both architectures. Clinical context is thus most valuable exactly where accessibility needs are greatest. Future work should explore training objectives that encourage models to condition on individual clinical features, and no-reference difficulty estimators that make the gate deployable.

Any attempt to exploit clinical context must also weigh its ethical implications. Clinical labels may encourage a deficit-oriented framing of atypical speech, so such information should be optional, user-controlled support rather than a requirement for access; and because language models inherit social biases, identifying a speaker as disabled may itself affect transcription or downstream interpretation harmfully. A further risk is hallucinated correction: when a model reshapes speech toward a more fluent form, it may change the speaker’s intended meaning, which is especially problematic in AAC settings.

Two broader implications follow. Accessibility for dysarthric speech will not follow automatically from scaling general-purpose audio-language models (Hoffman et al., 2014; Jaddoh et al., 2025), even when contextual information is available, and this evaluation of open-weight models offers a reference point for future work, placing responsibility on developers of large-scale systems to evaluate accessibility directly.

Acknowledgments. We obtained official approval from the authors of the Speech Accessibility Project to evaluate audio-language models that do not violate the data redistribution regulations, and gratefully acknowledge the University of Illinois Urbana-Champaign, Beckman Institute for Advanced Science and Technology, for their permission and support in this study.

References

- Dhruuv Agarwal, Harry Zhang, Yang Yu, and Quan Wang. The universal personalizer: Few-shot dysarthric speech recognition via meta-learning, 2025. URL <https://arxiv.org/abs/2509.15516>.
- Ali Alsayegh and Tariq Masood. Zero-shot recognition of dysarthric speech using commercial automatic speech recognition and multimodal large language models, 2025. URL <https://arxiv.org/abs/2512.17474>.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.
- Chitralekha Bhat and Helmer Strik. Two-stage data augmentation for improved asr performance for dysarthric speech. *Computers in Biology and Medicine*, 189:109954, 2025. ISSN 0010-4825. doi: <https://doi.org/10.1016/j.combiomed.2025.109954>. URL <https://www.sciencedirect.com/science/article/pii/S0010482525003051>.
- Chitralekha Bhat, Bhargav Vachhani, and Sunil Kopparapu. Recognition of dysarthric speech using voice parameters for speaker adaptation and multi-taper spectral estimation. In *Proceedings of Interspeech 2016*, pp. 228–232, 2016. doi: 10.21437/Interspeech.2016-1085.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
- Cohere Labs. Cohere transcribe: State-of-the-art open-source speech recognition. <https://huggingface.co/CohereLabs/cohere-transcribe-03-2026>, March 2026. Accessed: 2026-03-28.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024. URL <https://arxiv.org/abs/2301.00234>.
- Fixie AI. Ultravox: An open, fast, and extensible multimodal model for audio understanding. <https://github.com/fixie-ai/ultravox>, 2024. Model card at https://huggingface.co/fixie-ai/ultravox-v0_5-llama-3_1-nemo-12b.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gura, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, Hyunjong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri,

Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihla, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara

Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezhadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josp Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styr, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudrit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjit Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Åhdel, Sujeewan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rządowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan

Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tume, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchenniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uribe, Yeongil Ko, Laura Knight, Amélie Hélie, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska,

- David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshov, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesch Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atlas, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikuś, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirsenschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>.
- Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025. URL <https://arxiv.org/abs/2507.08128>.
- Google DeepMind. Gemma 4: Byte for byte, the most capable open models. <https://blog.google/technology/developers/gemma-4/>, 2026. Model card and weights at <https://huggingface.co/google/gemma-4-E4B-it>.
- Eugen Hermann and Mathew Magimai-Doss. Few-shot dysarthric speech recognition with text-to-speech data augmentation. In *Proceedings of Interspeech 2023*, pp. 156–160, 2023. doi: 10.21437/Interspeech.2023-2481.
- Howard J. Hoffman, Chi-Ming Li, Katalin G. Losonczy, Mei-Shu Chiu, Jeffrey B. Lucas, and Kenneth O. St. Louis. Voice, speech, and language disorders in the u.s. population: The 2012 national health interview survey (nhis). In *Abstracts of the 47th Annual Meeting of the*

- Society for Epidemiologic Research*, pp. 156, Seattle, WA, jun 2014. Abstract No. 648, June 24–27, 2014.
- Wei-Ning Hsu, Anne Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- Aisha Jaddoh, Fernando Loizides, Khadijah Alreafai, and Omer Rana. Overcoming speech barriers: Non-verbal voice cue interaction technique for enhancing smart voice assistant accessibility for individuals with dysarthria. *ACM Transactions on Accessible Computing*, Volume 18, Issue 2, Article 9, 2025. URL <https://doi.org/10.1145/3726874>.
- Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmantlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélío Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Praveesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, Yihan Wan, and Yunhao Tang. Voxtral, 2025. URL <https://arxiv.org/abs/2507.13264>.
- Microsoft Team, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL <https://arxiv.org/abs/2503.01743>.
- Megan A. Morris, Cristina Sarmiento, and Kori Eberle. Documentation of disability status and accommodation needs in the electronic health record: A qualitative study of health care organizations’ current practices. *The Joint Commission Journal on Quality and Patient Safety*, Volume 50, Issue 1, Pages 16–23, 2024. URL <https://doi.org/10.1016/j.jcjq.2023.10.006>.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding, 2020. URL <https://arxiv.org/abs/1910.14599>.

OpenAI Team, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mađry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codis-poti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichen, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ika Lan, Ilya Kostikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCal-lum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljube, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter

- Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- Bornali Phukon, Xiuwen Zheng, and Mark Hasegawa-Johnson. Aligning asr evaluation with human and llm judgments: Intelligibility metrics using phonetic, semantic, and nli approaches, 2025. URL <https://arxiv.org/abs/2506.16528>.
- Niclas Pokel, Pehuén Moure, Roman Boehringer, and Yingqiang Gao. Adapting foundation speech recognition models to impaired speech: A semantic re-chaining approach for personalization of german speech, 2025a. URL <https://arxiv.org/abs/2506.21622>.
- Niclas Pokel, Pehuén Moure, Roman Boehringer, Shih-Chii Liu, and Yingqiang Gao. Variational low-rank adaptation for personalized impaired speech recognition, 2025b. URL <https://arxiv.org/abs/2509.20397>.
- Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.
- Dima Rekesh, Nithin Rao Koluguri, Samuel Krizan, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. Fast conformer with linearly scalable attention for efficient speech recognition, 2023. URL <https://arxiv.org/abs/2305.05084>.
- Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The TORGO Database of Acoustic and Articulatory Speech from Speakers with Dysarthria. *Language resources and evaluation*, 46(4):523–541, 2012.
- George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Moraes, Gakuto Kurata, Hagai Aronowitz, Ibrahim Ibrahim, Jeff Kuo, Kate Soule, Luis Lastras, Masayuki Suzuki, Ron Hoory, Samuel Thomas, Sashi Novitasari, Takashi Fukuda, Vishal Sunder, Xiaodong Cui, and Zvi Kons. Granite-speech: open-source speech-aware LLMs with strong English ASR capabilities, 2025. URL <https://arxiv.org/abs/2505.08699>. Model weights at <https://huggingface.co/ibm-granite>.
- Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.
- Kaito Takahashi, Keigo Hojo, Toshimitsu Sakai, Yukoh Wakabayashi, and Norihide Kitaoka. Fine-tuning Parakeet-TDT for dysarthric speech recognition in the Speech Accessibility Project Challenge. In *Proc. Interspeech 2025*, pp. 3304–3308, 2025. doi: 10.21437/Interspeech.2025-1484.

- J. Tobin, P. Nelson, B. MacDonald, R. Heywood, R. Cave, K. Seaver, A. Desjardins, P. P. Jiang, and J. R. Green. Automatic speech recognition of conversational speech in individuals with disordered speech. *Journal of Speech, Language, and Hearing Research*, 67(11):4176–4185, 2024. doi: 10.1044/2024_JSLHR-24-00045.
- Dominik Wagner, Ilja Baumann, Natalie Engert, Seanie Lee, Elmar Nöth, Korbinian Riedhammer, and Tobias Bocklet. Personalized fine-tuning with controllable synthetic speech from llm-generated transcripts for dysarthric speech recognition, 2025. URL <https://arxiv.org/abs/2505.12991>.
- Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models, 2025. URL <https://arxiv.org/abs/2503.02318>.
- Jin Xu et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL <https://arxiv.org/abs/2509.17765>.
- Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL <https://arxiv.org/abs/2509.18154>.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL <https://arxiv.org/abs/1904.09675>.
- Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou, Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang, Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran, Tara N. Sainath, Françoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu, Ruoming Pang, and Yonghui Wu. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1519–1532, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3182537. URL <http://dx.doi.org/10.1109/JSTSP.2022.3182537>.
- Xiwen Zheng, Bornali Phukon, Jonghwan Na, Ed Cutrell, Kyu Han, Mark Hasegawa-Johnson, Pan-Pan Jiang, Aadhrick Kuila, Colin Lea, Bob MacDonald, Gautam Mantena, Venkatesh Ravichandran, Leda Sari, Katrin Tomanek, Chang D. Yoo, and Chris Zwilling. The interspeech 2025 speech accessibility project challenge, 2025. URL <https://arxiv.org/abs/2507.22047>.
- Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models, 2024. URL <https://arxiv.org/abs/2402.11574>.

A Prompt Templates

All prompt conditions share a fixed preamble and task instruction. Context blocks between them are composed modularly depending on the condition.

Prompt Structure

Prompt Structure	
Preamble	(constant across all conditions)
Context Blocks	(composable, toggleable)
• condition — diagnosis + guidance • pathology — clinical speech ratings • follow_up — correction from prior pass	
Task Instruction	(constant across all conditions)

The **preamble** and **task instruction** are invariant.

Preamble: *You are transcribing speech from a person with a speech disorder. The audio may contain atypical pronunciation, rhythm, or voice quality. Use the provided clinical context to interpret ambiguous segments.*

Task instruction: *Transcribe the audio in English. Output ONLY the transcription, no explanations.*

P0, Zero-Shot Baseline

No clinical context is provided. The model receives only the audio and a plain transcription instruction.

P0: Zero-Shot

Transcribe this audio in English. Output only the transcription, no explanations.

P1, Condition Awareness

The model is informed of the speaker’s diagnosis and receives condition-specific guidance about expected speech characteristics. No per-sample clinical ratings are included.

P1: Condition Awareness (Cerebral Palsy example)

You are transcribing speech from a person with a speech disorder. The audio may contain atypical pronunciation, rhythm, or voice quality. Use the provided clinical context to interpret ambiguous segments.

The speaker has Cerebral Palsy. Cerebral Palsy often causes imprecise consonants, distorted vowels, and irregular speech rhythm. Words may sound slurred or have unusual stress patterns. Focus on the intended words rather than the surface-level distortions.

Transcribe the audio in English. Output ONLY the transcription, no explanations.

P2, Full Clinical Profile

In addition to the condition block from P1, the model receives per-sample clinical speech ratings across all annotation dimensions, sorted by severity.

P2: Full Clinical Profile (Cerebral Palsy example)

You are transcribing speech from a person with a speech disorder. The audio may contain atypical pronunciation, rhythm, or voice quality. Use the provided clinical context to interpret ambiguous segments.

The speaker has Cerebral Palsy.

Cerebral Palsy often causes imprecise consonants, distorted vowels, and irregular speech rhythm. Words may sound slurred or have unusual stress patterns. Focus on the intended words rather than the surface-level distortions.

Clinical speech profile (scale 1–7, 1=normal, 7=most severe):

- Naturalness: 5/7 (moderate-severe)
- Imprecise consonants: 4/7 (moderate)
- Monopitch: 4/7 (moderate)
- Distorted vowels: 3/7 (mild-moderate)
- Harsh voice: 3/7 (mild-moderate)
- Slow rate: 3/7 (mild-moderate)
- Intelligibility: 2/7 (mild)
- Low pitch: 1/7 (normal)

Transcribe the audio in English. Output ONLY the transcription, no explanations.

P3, Full Profile with Follow-Up Correction

Extends P2 with additional context from a prior transcription attempt, enabling iterative self-correction.

P3: Full Profile + Follow-Up (ALS example)

You are transcribing speech from a person with a speech disorder. The audio may contain atypical pronunciation, rhythm, or voice quality. Use the provided clinical context to interpret ambiguous segments.

The speaker has ALS.

ALS progressively weakens speech muscles, leading to slow, effortful speech with breathy or strained voice quality. Words may be prolonged or have nasal quality. Listen for the intended message through the motor speech difficulties.

Clinical speech profile (scale 1–7, 1=normal, 7=most severe):

- Imprecise consonants: 5/7 (moderate-severe)
- Breathiness: 4/7 (moderate)
- Slow rate: 4/7 (moderate)
- Hypernasality: 3/7 (mild-moderate)

Additional context from prior attempt: Previous transcription produced “I quite assure you I am too kind”. The correct phrasing is “I’m quite sure you are too kind-hearted”.

Transcribe the audio in English. Output ONLY the transcription, no explanations.

Ablation Sub-Conditions

The clinical ratings in P2 are organized into three tiers by ASR relevance. Sub-conditions isolate individual tiers to measure their contribution:

Condition-Specific Guidance

The condition block in P1–P3 includes a diagnosis-specific description tailored to each etiology in the SAP dataset:

Counterfactual Control Conditions

Three control conditions preserve the structure and length of the full clinical profile (P2) while manipulating its content (results in Section 4.3 and Section 4.4).

ID	Ratings Included	Dimensions
P2a	Tier 1 only	Imprecise consonants, distorted vowels, repeated/prolonged phonemes, speech rate, stress patterns, nasality, short phrases, inappropriate silences
P2b	Tier 2 only	Harsh/strained/breathy voice, voice tremor, pitch breaks, monopitch, monoloudness, low pitch, prolonged intervals
P2c	Tier 1 + 2	All actionable and voice quality dimensions
P2d	Condensed, all tiers	Compact key-value format: condition: Cerebral Palsy; speech_ratings: Imprecise consonants=4/7, Distorted vowels=3/7, ...
P2e	Condensed, Tier 1	Compact format, ASR-actionable dimensions only

Table 1: Ablation sub-conditions for the clinical ratings block. Tier 3 (meta-level judgments: naturalness, intelligibility, other) is included only in P2 and P2d.

Etiology	Guidance Text
Cerebral Palsy	Cerebral Palsy often causes imprecise consonants, distorted vowels, and irregular speech rhythm. Words may sound slurred or have unusual stress patterns. Focus on the intended words rather than the surface-level distortions.
ALS	ALS progressively weakens speech muscles, leading to slow, effortful speech with breathy or strained voice quality. Words may be prolonged or have nasal quality. Listen for the intended message through the motor speech difficulties.
Parkinson’s Disease	Parkinson’s Disease typically causes reduced loudness, monotone pitch, and sometimes rapid or mumbled speech. Words may run together or trail off. Pay close attention to softly spoken or rushed segments.
Down syndrome	Down syndrome can affect speech clarity through imprecise articulation and irregular speech rhythm. The speaker may have difficulty with certain consonant clusters. Focus on the overall message and common word patterns.
Stroke	Stroke can cause various speech difficulties including slurred speech, word-finding pauses, or sound substitutions. The speaker’s intended words may differ from how they sound on the surface.

Table 2: Etiology-specific guidance text included in the condition block (P1–P3).

- **Wrong clinical profile.** The diagnosis and per-sample ratings are replaced with those of a speaker with a different etiology, via cyclic rotation across etiologies (e.g., a cerebral palsy speaker receives an ALS speaker’s diagnosis, guidance, and ratings). Format, length, and rating-scale statistics are preserved; only the correctness of the content is destroyed.
- **Length-matched random text.** The condition and ratings blocks are replaced by fluent text with no relation to the task or the speaker, matched in length to the full clinical profile.
- **Length-matched instruction-only filler.** The condition and ratings blocks are replaced by generic transcription instructions (restating output-format requirements), matched in length and containing no non-instructional content.

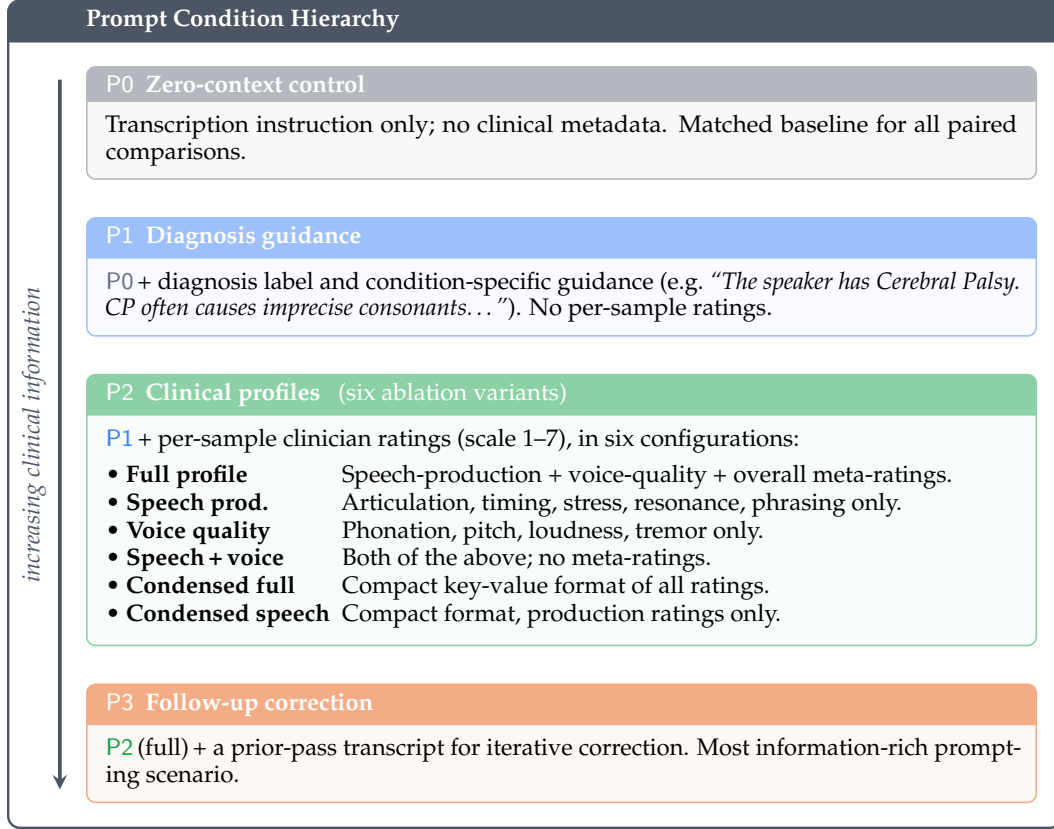
B Metrics

B.1 Evaluation Metrics

Word error rate (WER) and character error rate (CER) are the primary metrics. WER is defined as

$$\text{WER} = \frac{S + D + I}{N}$$

Table 3: Prompt condition hierarchy. Each level builds on the previous by adding clinical information. All conditions share the same $N = 11,218$ evaluation samples, enabling paired comparisons. Full example prompts can be found in Appendix A.



where S , D , I are substitutions, deletions, and insertions respectively, and N is the number of words in the reference. CER is computed analogously at the character level.

Text normalization. Following the SAP benchmark protocol (Zheng et al., 2025), both hypotheses and references are normalized using Whisper’s EnglishTextNormalizer (Radford et al., 2022). This includes number-to-word expansion, contraction normalization, British-to-American spelling conversion, filler word removal, and stripping of bracket-delimited prompt prefixes that appear in spontaneous speech references (e.g., [Tell us about your hobbies.]). Hypotheses are additionally truncated to 512 words after normalization.

Long-audio decoding. Models with a fixed 30 s audio input limit (Whisper and the Gemma-4 family) would otherwise silently truncate the 1,081 rated samples longer than 30 s. For these models, long samples are decoded in segments of at most 30 s, split at forced-alignment word gaps, with the same prompt applied to every segment and the segment transcripts concatenated.

Dual-reference minimum. The SAP dataset provides two reference transcriptions per utterance. The verbatim *transcript* r_t preserves disfluency markers (e.g., (um), (uh), incomplete words such as (w-)) and the *clean transcript* r_c has these markers removed. Since models are not expected to reproduce disfluency annotations, per-sample WER is computed against both references and the minimum is taken.

$$\text{WER}_{\min}^{(i)} = \min(\text{WER}(\hat{y}_i, r_t^{(i)}), \text{WER}(\hat{y}_i, r_c^{(i)}))$$

$\text{CER}_{\min}$ is defined analogously.

Clip-at-one. Per-sample WER and CER are clipped at 1.0 before averaging.

$$\overline{\text{WER}} = \frac{1}{N} \sum_{i=1}^N \min(\text{WER}_{\min}^{(i)}, 1.0)$$

This convention, adopted from the SAP challenge, prevents individual outliers from dominating the aggregate metric. Such outliers arise frequently in practice, as autoregressive models, particularly under sampling-based decoding or with long context prompts, can enter repetitive generation loops (e.g., producing “no no no no...” hundreds of times), yielding per-sample $\text{WER} \gg 1$. Since the severity of these hallucinations depends on the maximum generation length rather than on transcription quality, clipping ensures that a single degenerate sample does not overshadow systematic improvements across thousands of correctly transcribed utterances. We additionally report unclipped WER and hallucination rates (percentage of samples with $\text{WER}_{\text{raw}} > 1.0$) to characterize these failure modes separately.

B.2 SemScore

WER treats all word-level mismatches equally, penalizing semantically equivalent paraphrases (e.g., “what is” vs. “what’s”) as harshly as meaning-altering errors (e.g. “what is” vs. “watt is”). To capture semantic similarity, we additionally compute SemScore (Phukon et al., 2025), a composite metric that achieved $\rho = 0.89$ correlation with human intelligibility judgments on dysarthric speech in the SAP evaluation. SemScore combines, via learned weights, three sub-metrics.

$$\text{SemScore} = 0.40 \cdot \hat{s}_{\text{NLI}} + 0.28 \cdot \hat{s}_{\text{BERT}} + 0.32 \cdot \hat{s}_{\text{phon}}$$

where each $\hat{s}$ is min-max normalized to $[0, 1]$, consistently with the fixed thresholds from the SAP protocol.

- **NLI entailment** (s_{NLI}) measures bidirectional textual entailment probability from a RoBERTa-large model fine-tuned on SNLI, MNLI, FEVER, and ANLI (Nie et al., 2020). The score averages $P(\text{entailment} \mid r, \hat{y})$ and $P(\text{entailment} \mid \hat{y}, r)$, capturing whether prediction and reference mutually imply each other regardless of exact wording.
- **BERTScore** (s_{BERT}) measures token-level contextual embedding similarity (F1) using RoBERTa-large with baseline rescaling (Zhang et al., 2020).
- **Phonetic similarity** (s_{phon}) computes Jaro-Winkler similarity between Soundex encodings of prediction and reference, capturing phonetic resemblance independent of orthographic variation.

Following the dual-reference convention, SemScore is computed against both r_t and r_c , and the per-sample *maximum* is taken (higher SemScore is better, opposite to WER).

$$\text{SemScore}_{\max}^{(i)} = \max(\text{SemScore}(\hat{y}_i, r_t^{(i)}), \text{SemScore}(\hat{y}_i, r_c^{(i)}))$$

We report SemScore on a 0–100 scale. Unlike WER, SemScore is not subject to the hallucination inflation problem, as a repetitive output receives low semantic similarity scores regardless of length, making it a natural complement to the clipped WER metric.

B.3 Statistical Testing

Paired frozen-model comparisons use Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction across the 50 model $\times$ condition comparisons (Section 4.2). For the correct-versus-wrong comparisons on the fine-tuned systems (Section 4.4) we test *equivalence* directly with two one-sided tests (TOST) on paired per-utterance WER differences, using an equivalence bound of ± 0.005 WER; $p < 0.05$ rejects the hypothesis that the true difference exceeds the bound. Speaker-clustered 95% confidence intervals are obtained by bootstrap resampling at the speaker level, which accounts for the within-speaker correlation of per-utterance

C Supporting Tables

Table 4: Zero-shot baseline WER on the SAP rated subset ($N = 11,218$). “Promptable” indicates whether the model accepts text instructions alongside audio. *Audio Flamingo 3’s high WER is due to non-transcription output without task instruction; the basic transcription instruction reduces it to 0.162 (see Section 4.2).

Model	Promptable	WER
Qwen3-ASR-1.7B	No	0.1338
Qwen3-Omni-30B	Yes	0.1377
Voxtral-Small-24B	Yes	0.1388
Whisper-large-v3	No	0.1443
Cohere Transcribe	No	0.1448
Voxtral-Mini-3B	Yes	0.1531
Phi-4 Multimodal	Yes	0.1577
Parakeet-TDT-1.1B	No	0.1585
Qwen3-Thinking-30B	Yes	0.1650
Granite-Speech 4.1-2B NAR	No	0.1667
MiniCPM-o 4.5	Yes	0.1812
MiniCPM-o Think	Yes	0.1841
Granite-Speech 4.1-2B+	Yes	0.1909
Qwen2-Audio-7B	Yes	0.2013
Granite-Speech 3.3-8B	Yes	0.2366
Audio-Reasoner	Yes	0.2400
Gemma-4-4B	Yes	0.2976
Gemma-4-4B Think	Yes	0.3018
Gemma-4-2B	Yes	0.3022
Ultravox-Nemo 12B	Yes	0.3123
Wav2Vec2-Robust	No	0.3264
Gemma-4-12B Think	Yes	0.3881
Wav2Vec2	No	0.3930
Gemma-4-12B	Yes	0.4122
Audio Flamingo 3*	Yes	0.4955

Table 5: WER across clinical prompt conditions ($N = 11,218$ matched samples per model). Columns are prompt conditions: P0 (zero-context control), P1 (diagnosis guidance), P2 (full clinical profile), P2a–P2e (profile ablations: speech-production, voice-quality, speech+voice, condensed-full, condensed-speech), and P3 (follow-up correction); see Table 3. Bold marks each model’s lowest (best) WER. No model improves meaningfully with clinical context. *Audio Flamingo 3’s zero-context (P0) WER of 0.162 reflects a large drop from its unprompted WER of 0.495; this is a format-correction effect (see Section 4.2).

Model	P0	P1	P2	P2a	P2b	P2c	P2d	P2e	P3
AF3*	0.1616	0.1615	0.1627	0.1624	0.1621	0.1626	0.1637	0.1635	0.1625
Gemma-4-4B	0.2925	0.3245	0.3412	0.3337	0.3331	0.3405	0.3190	0.3167	0.3075
Gemma-4-4B Think	0.3018	0.5095	0.5039	0.5108	0.5136	0.5147	0.5036	0.5076	0.4355
Gemma-4-12B	0.4122	0.5298	0.5005	0.5282	0.5440	0.5203	0.4468	0.4665	0.4195
Granite-3.3-8B	0.2366	0.2127	0.2133	0.2160	0.2174	0.2116	0.2059	0.2032	0.2156
Granite-4.1-2B+	0.1909	0.2023	0.2058	0.2022	0.2027	0.2047	0.2022	0.2008	0.2168
MiniCPM-o Think	0.1839	0.1779	0.1783	0.1790	0.1794	0.1789	0.1805	0.1800	0.1789
Phi-4	0.1568	0.1566	0.1572	0.1550	0.1524	0.1557	0.1616	0.1626	0.1592
Qwen2-Audio	0.2014	0.2320	0.2524	0.2578	0.2528	0.2556	0.2396	0.2479	0.2572
Qwen3-Omni	0.1377	0.1409	0.1415	0.1412	0.1415	0.1414	0.1429	0.1426	0.1416
Qwen3-Think	0.1650	0.1709	0.1722	0.1721	0.1747	0.1723	0.1860	0.1814	0.1662
Voxtral-S	0.1390	0.1486	0.1891	0.1788	0.1819	0.1866	0.1530	0.1474	0.1726

WER. The difficulty analysis (Section 4.5) bins utterances by the audio-only LoRA system’s WER and reports, within each bin, the paired difference between the audio-only system and the context-dependent system evaluated with the full clinical profile.

Table 6: Paired comparison from zero-context control to full clinical profile ($N = 11,218$). Positive delta indicates degradation. Bold indicates largest change. p -values use Benjamini-Hochberg false-discovery-rate (FDR) correction.

Model	Ctrl WER	Profile WER	Δ	Cohen d	Effect	p (FDR)	Degraded
Audio Flamingo 3	0.1616	0.1627	+0.0011	0.010	negligible	<0.0001	11.7%
MiniCPM-o Think	0.1839	0.1783	-0.0056	-0.040	negligible	0.3865	12.8%
Phi-4	0.1568	0.1572	+0.0004	0.003	negligible	<0.0001	13.1%
Qwen3-Omni-30B	0.1377	0.1415	+0.0038	0.035	negligible	<0.0001	12.3%
Voxtral-Small-24B	0.1390	0.1891	+0.0501	0.223	small	<0.0001	15.8%
Qwen2-Audio-7B	0.2014	0.2524	+0.0510	0.242	small	<0.0001	25.3%
Qwen3-Thinking-30B	0.1650	0.1722	+0.0072	0.049	negligible	<0.0001	10.7%
Gemma-4-4B	0.2925	0.3412	+0.0487	0.212	small	<0.0001	19.4%
Gemma-4-4B Think	0.3018	0.5039	+0.2021	0.575	medium	<0.0001	42.7%
Gemma-4-12B	0.4122	0.5005	+0.0883	0.233	small	<0.0001	27.4%
Gemma-4-12B Think	0.3881	0.5015	+0.1134	0.322	small	<0.0001	31.6%
Granite-Speech 3.3-8B	0.2366	0.2133	-0.0233	-0.097	negligible	<0.0001	12.3%
Granite-Speech 4.1-2B+	0.1909	0.2058	+0.0149	0.118	negligible	<0.0001	17.7%

Table 7: WER delta (zero-context to full clinical profile) by severity bin. Positive values indicate degradation. Bold indicates largest change. Italics indicates largest improvement.

Model	Mild (1-2)	Moderate (3-4)	Severe (5-7)
AF3	+0.002	-0.000	+0.035
Phi-4	-0.007	+0.014	+0.029
MiniCPM-o Think	-0.003	-0.011	+0.004
Qwen2-Audio	+0.059	+0.037	+0.014
Qwen3-Omni	+0.003	+0.005	+0.035
Qwen3-Think	+0.006	+0.009	+0.047
Gemma-4-4B	+0.068	+0.012	+0.014
Gemma-4-4B Think	+0.224	+0.162	+0.093
Voxtral-S	+0.066	+0.020	+0.029
Gemma-4-12B	+0.107	+0.054	+0.081

D Fine-Tuning Details

D.1 Context-dependent Fine-Tuning Protocol

We fine-tune Voxtral-Small-24B using LoRA (rank 16, $\alpha = 32$, dropout 0.05) applied to the query and value projection matrices of the language model layers. The audio encoder is frozen throughout.

Training uses 5-fold speaker-disjoint cross-validation. In each fold, approximately 80% of speakers are used for training and the remaining 20% (87-88 speakers per fold, 437 unique speakers total) are held out for evaluation. Folds are constructed so that no speaker appears in both training and evaluation within the same fold, preventing speaker-level data leakage.

The **audio-only** system trains exclusively on audio-transcription pairs with no clinical prompt. The **context-dependent** system trains on a mixture of three prompt formats per utterance, including audio-only (no clinical context), speech + voice profile (P2c), and condensed full profile (P2d). During each training epoch, the prompt format is sampled uniformly at random for each utterance, so the model sees the same audio with different context configurations across epochs. This exposure diversity is intended to teach the model to use context when present without degrading performance when it is absent.

Both systems are trained for 3 epochs per fold with a learning rate of 2×10^{-4} (cosine schedule, 100 warmup steps), effective batch size of 8 (gradient accumulation over 4 steps), and bf16 mixed precision. We report the last checkpoint per fold. All evaluation results in the main text are pooled out-of-fold predictions across all five folds ($N = 11,218$ utterances).

D.2 Gemma-4-E2B Replication

We replicate the protocol on Gemma-4-E2B (google/gemma-4-E2B-it; listed as Gemma-4-2B in Table 19) with the same LoRA configuration (rank 16, $\alpha = 32$, dropout 0.05) applied

Table 8: WER delta (zero-context to full clinical profile) by etiology. Positive values indicate degradation. Bold indicates largest change. Italics indicates largest improvement.

Model	ALS	CP	Down S.	Parkinson's
AF3	+0.004	-0.003	-0.009	+0.002
Phi-4	+0.004	-0.007	-0.002	+0.002
MiniCPM-o Think	<i>-0.013</i>	-0.018	<i>-0.032</i>	-0.000
Qwen2-Audio	+0.029	+0.036	+0.036	+0.060
Qwen3-Omni	+0.001	<i>-0.019</i>	-0.009	+0.010
Qwen3-Think	+0.001	-0.010	+0.031	+0.011
Gemma-4-4B	+0.068	+0.047	+0.040	+0.046
Gemma-4-4B Think	+0.175	+0.147	+0.196	+0.221
Voxtral-S	+0.057	+0.046	+0.077	+0.048
Gemma-4-12B	+0.057	+0.072	+0.091	+0.099

Table 9: WER delta (zero-context to full clinical profile) by utterance category. Positive values indicate degradation. Bold indicates largest change. Italics indicates largest improvement.

Model	Asst. Cmd	Non-spont.	Novel Sent.	Spont.
AF3	-0.002	-0.005	+0.008	-0.002
Phi-4	-0.018	<i>-0.097</i>	+0.011	+0.015
MiniCPM-o Think	<i>-0.019</i>	-0.011	+0.007	<i>-0.004</i>
Qwen2-Audio	+0.039	+0.011	+0.089	+0.028
Qwen3-Omni	-0.007	-0.054	+0.020	+0.002
Qwen3-Think	+0.001	-0.032	+0.011	+0.012
Gemma-4-4B	+0.137	+0.114	+0.000	-0.003
Gemma-4-4B Think	+0.256	+0.119	+0.135	+0.215
Voxtral-S	+0.109	+0.124	+0.008	+0.024
Gemma-4-12B	+0.168	+0.005	+0.053	+0.042

to the attention projections of the language-model layers ($\approx 3.1\text{M}$ trainable parameters, 0.06% of the model), keeping the audio encoder frozen. Training uses the same 5-fold speaker-disjoint cross-validation splits and the same prompt-format mixture as the Voxtral-Small context-dependent system, with bf16 precision. Frozen Gemma-4-E2B reaches WER 0.302 at zero-context and degrades to 0.372 under the full clinical profile; after fine-tuning, both the audio-only and the context-dependent system reach WER 0.204, removing the degradation entirely (Section 4.4). The frozen evaluations use chunked decoding for long audio (Appendix B.1); the LoRA evaluations predate this fix, and since chunking only lowers WER on the affected long samples, the reported fine-tuning improvements are, if anything, understated.

D.3 Hardware

Inference and fine-tuning for models up to 12B parameters were conducted on a single NVIDIA GeForce RTX 5090 (32 GB VRAM) with an Intel Core i7-6900K. Larger models (Qwen3-Omni-30B, Qwen3-Thinking-30B) required a Project DIGITS developer kit with an NVIDIA GB10 (128 GB unified memory).

E Counterfactual Control Results

Table 11 lists the frozen counterfactual control results (Section 4.3) and Table 12 the correct-versus-wrong comparison for the fine-tuned context-dependent systems (Section 4.4).

F Positioning Against Specialized SAP Systems

Figure 7 places our fine-tuned systems next to the strongest published SAP system. The comparison is approximate: the SAP Challenge winner (Takahashi et al., 2025) is evaluated on the challenge private Test1/Test2 splits under full fine-tuning of Parakeet-TDT, whereas our systems are evaluated on the rated subset under parameter-efficient LoRA. The Gemini 2.5 Flash personalization system (Agarwal et al., 2025) reaches about 5% WER with

Table 10: Change in error rates from zero-context control to full clinical profile (Δ rates normalised by total reference words). Bold indicates largest change. Italics indicates largest improvement.

Model	Δ Sub	Δ Ins	Δ Del	Δ Hit
AF3	-0.006	+0.001	+0.007	-0.000
Phi-4	+0.025	+0.043	-0.007	-0.018
MiniCPM-o Think	+0.012	+0.269	-0.005	-0.007
Qwen2-Audio	+0.023	+0.005	+0.027	-0.050
Qwen3-Omni	-0.008	+0.030	-0.002	<i>+0.009</i>
Qwen3-Think	+0.006	<i>-0.031</i>	+0.015	-0.021
Gemma-4-4B	+0.008	+0.063	+0.000	-0.008
Gemma-4-4B Think	<i>-0.010</i>	-0.025	+0.174	-0.165
Voxtral-S	+0.011	+0.212	-0.001	-0.010
Gemma-4-12B	+0.031	+0.621	<i>-0.012</i>	-0.019

Table 11: Frozen counterfactual controls (WER, $N = 11,218$ rated samples). Wrong clinical content is no better than correct content, length-matched random text degrades almost as much, and length-matched instruction-only filler sits on the audio-only baseline.

Prompt condition	Voxtral-Small	Gemma-4-E2B
Audio only (zero-context)	0.139	0.302
Correct clinical profile	0.189	0.372
Wrong clinical profile	0.190	0.383
Random text (length-matched)	0.170	0.361
Instruction filler (length-matched)	0.153	0.302

few-shot same-speaker enrollment but is omitted from the figure because its evaluation split is not directly comparable. A frozen Whisper anchor of comparable difficulty across splits (14–18% WER) indicates that the evaluation sets are of similar hardness, so the comparison is approximate positioning rather than a matched benchmark.

G Full Paired Comparisons

Table 13 presents paired comparisons for the five primary prompt conditions across all models from the clinical context conditioning evaluation.

H Context-dependent Evaluation Grid

Table 14 presents the full evaluation grid for the Voxtral-Small context-dependent system (5-fold CV), comparing against the audio-only system. Green triangles mark conditions where the context-dependent system beats the audio-only baseline.

I Models

Prompt placement and the unified Gemma-4-12B. Gemma-4-12B is the only *unified* (encoder-free) model in our set: audio is projected directly into the language model without a separate audio encoder. Like the other promptable models it degrades under clinical prompting supplied in the user turn (Table 5), but the degradation is unusually large. In exploratory runs we observed only minor differences between placing the clinical prompt in the user turn versus a system turn for most models, whereas for the unified 12B a system-turn placement substantially reduced the degradation. We did not pursue this further; the effect may relate to how the unified architecture routes system versus user tokens, which we leave to future mechanistic analysis. Throughout the paper we report the user-turn condition, as it most closely reflects how clinical context would be supplied at deployment.

J Model Size and Performance

Figure 8 plots model parameter count against zero-shot WER on the SAP rated subset. A log-linear OLS fit (excluding Audio Flamingo 3, whose high WER reflects output formatting

Table 12: Correct vs. wrong clinical content for the fine-tuned context-dependent systems: WER difference (wrong – correct, positive = correct better) with speaker-clustered 95% CIs. Voxtral-Small is TOST-equivalent (± 0.005 bound, $p < 0.001$) at every condition. **Bold**: significant correct-over-wrong gain (Wilcoxon $p \leq 0.005$).

Eval condition	Voxtral-Small	Gemma-4-E2B
Diagnosis guidance	−0.0004 [−0.0013, +0.0006]	+0.0011 [−0.0015, +0.0038]
Full clinical profile	+0.0003 [−0.0007, +0.0012]	+0.0046 [+0.0017, +0.0077]
Speech production profile	+0.0001 [−0.0009, +0.0011]	+0.0034 [+0.0007, +0.0062]
Voice quality profile	−0.0007 [−0.0019, +0.0003]	+0.0030 [+0.0004, +0.0056]
Speech + voice profile	+0.0003 [−0.0007, +0.0013]	+0.0015 [−0.0013, +0.0044]
Condensed full profile	−0.0007 [−0.0018, +0.0004]	+0.0007 [−0.0019, +0.0032]
Condensed speech production	+0.0001 [−0.0009, +0.0010]	+0.0019 [−0.0006, +0.0043]

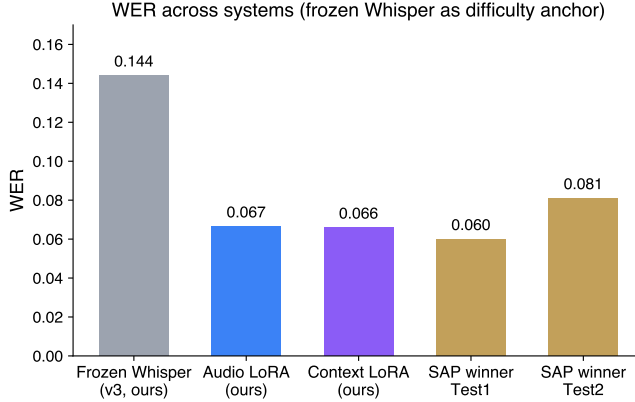


Figure 7: WER across systems with frozen Whisper as a difficulty anchor. The systems use different SAP splits (our rated subset vs. challenge Test1/Test2) and normalization conventions; this is approximate positioning rather than a matched comparison. Challenge results from [Takahashi et al. \(2025\)](#).

rather than transcription quality; see Section 4.2) yields $R^2 = 0.07$, indicating that model size alone is a weak predictor of dysarthric speech performance. Several smaller dedicated ASR models (Qwen3-ASR at 1.7B, Cohere Transcribe at 2.0B) match or outperform much larger instruct models (Qwen2-Audio at 7B, Ultravox-Nemo at 12B), suggesting that architecture and training data composition matter more than scale for this task. Notably, the three Gemma-4 variants (4B, 4B Think, and the unified 12B) cluster well above the trend line while Voxtral-Small-24B and Qwen3-Thinking-30B lie below it at larger parameter counts, further reinforcing that family and training recipe dominate over raw size.

The weak pooled fit is best understood as a confound with model class rather than a genuine size effect. Splitting by class, the correlation between log parameters and WER is strongly negative for the pure-ASR models ($r = -0.95$, $R^2 = 0.91$) but essentially absent for the instruct models ($r = -0.14$, $R^2 = 0.02$), with the chain-of-thought group too small to interpret ($N = 4$). The ASR trend, moreover, is an artifact of the two legacy Wav2Vec2 systems (0.3B, WER > 0.32) anchoring one end against modern 1–2B ASR models at the other; it does not reflect a smooth size–accuracy relationship. The dedicated ASR models are small yet strong precisely because they are purpose-built and forgo the large language-model backbone that the instruct and CoT systems carry, whereas among those LM-backbone models WER is uncorrelated with scale (Voxtral-Small-24B at 0.139 versus Gemma-4-12B at 0.412). We therefore treat model size as globally uninformative for dysarthric ASR once model class is accounted for.

K Metric Breakdown and Agreement

This section provides detailed metric breakdowns. Table 20 shows the correlation between ASR performance and clinical severity. Table 21 stratifies performance by severity level. Table 22 gives the full zero-shot results with WER, CER, and SemScore by etiology. Ta-

Table 13: Paired comparisons vs. zero-context control for the five primary prompt conditions. p -values use Benjamini–Hochberg false-discovery-rate (FDR) correction across all 50 comparisons. $N = 11,218$. Bold indicates the largest change ($|\Delta|$, $|d|$, and degradation rate) within each condition block.

Condition	Model	Δ	d	Effect	p (FDR)	Degraded
Diagnosis guidance	AF3	−0.0001	−0.001	negl.	<0.0001	10.9%
	Phi-4	−0.0002	−0.001	negl.	<0.0001	13.0%
	MiniCPM-o Think	−0.0060	−0.043	negl.	0.2112	12.3%
	Qwen2-Audio	+0.0305	+0.167	negl.	<0.0001	22.7%
	Qwen3-Omni	+0.0033	+0.029	negl.	<0.0001	12.1%
	Qwen3-Think	+0.0059	+0.039	negl.	<0.0001	10.5%
	Gemma-4-4B	+0.0321	+0.165	negl.	<0.0001	17.1%
	Gemma-4-4B Think	+0.2076	+0.574	medium	<0.0001	42.2%
	Voxtral-S	+0.0096	+0.075	negl.	<0.0001	11.1%
	Gemma-4-12B	+0.1177	+0.304	small	<0.0001	29.0%
Speech prod. profile	AF3	+0.0008	+0.007	negl.	<0.0001	11.4%
	Phi-4	−0.0018	−0.011	negl.	<0.0001	13.4%
	MiniCPM-o Think	−0.0048	−0.035	negl.	0.9011	13.2%
	Qwen2-Audio	+0.0564	+0.258	small	<0.0001	26.6%
	Qwen3-Omni	+0.0035	+0.032	negl.	<0.0001	12.0%
	Qwen3-Think	+0.0071	+0.047	negl.	<0.0001	10.9%
	Gemma-4-4B	+0.0413	+0.192	negl.	<0.0001	18.7%
	Gemma-4-4B Think	+0.2090	+0.584	medium	<0.0001	43.0%
	Voxtral-S	+0.0398	+0.193	negl.	<0.0001	14.8%
	Gemma-4-12B	+0.1160	+0.299	small	<0.0001	29.8%
Voice quality profile	AF3	+0.0004	+0.004	negl.	<0.0001	11.5%
	Phi-4	−0.0044	−0.028	negl.	0.0014	12.5%
	MiniCPM-o Think	−0.0045	−0.032	negl.	0.7944	13.2%
	Qwen2-Audio	+0.0514	+0.251	small	<0.0001	26.2%
	Qwen3-Omni	+0.0038	+0.035	negl.	<0.0001	12.1%
	Qwen3-Think	+0.0098	+0.064	negl.	<0.0001	10.9%
	Gemma-4-4B	+0.0407	+0.191	negl.	<0.0001	18.3%
	Gemma-4-4B Think	+0.2118	+0.586	medium	<0.0001	43.6%
	Voxtral-S	+0.0429	+0.203	small	<0.0001	15.2%
	Gemma-4-12B	+0.1319	+0.335	small	<0.0001	30.7%
Condensed full profile	AF3	+0.0021	+0.019	negl.	<0.0001	11.2%
	Phi-4	+0.0048	+0.026	negl.	<0.0001	11.8%
	MiniCPM-o Think	−0.0034	−0.024	negl.	0.3184	13.0%
	Qwen2-Audio	+0.0381	+0.195	negl.	<0.0001	23.6%
	Qwen3-Omni	+0.0052	+0.047	negl.	<0.0001	12.9%
	Qwen3-Think	+0.0210	+0.118	negl.	<0.0001	13.6%
	Gemma-4-4B	+0.0265	+0.145	negl.	<0.0001	16.9%
	Gemma-4-4B Think	+0.2018	+0.557	medium	<0.0001	42.0%
	Voxtral-S	+0.0140	+0.101	negl.	<0.0001	11.8%
	Gemma-4-12B	+0.0347	+0.098	negl.	<0.0001	21.7%
Follow-up correction	AF3	+0.0009	+0.011	negl.	<0.0001	7.8%
	Phi-4	+0.0024	+0.017	negl.	<0.0001	9.2%
	MiniCPM-o Think	−0.0050	−0.035	negl.	0.7873	13.1%
	Qwen2-Audio	+0.0557	+0.300	small	<0.0001	23.5%
	Qwen3-Omni	+0.0039	+0.045	negl.	<0.0001	9.1%
	Qwen3-Think	+0.0012	+0.013	negl.	0.3957	2.3%
	Gemma-4-4B	+0.0151	+0.168	negl.	<0.0001	8.0%
	Gemma-4-4B Think	+0.1337	+0.454	small	<0.0001	26.7%
	Voxtral-S	+0.0336	+0.185	negl.	<0.0001	11.0%
	Gemma-4-12B	+0.0073	+0.021	negl.	0.6453	11.8%

bles 23, 24, and 25 break down the effect of context prompting on each metric by model and etiology, showing that WER and SemScore generally agree on the direction of prompting effects.

Table 14: Context-dependent evaluation grid (Voxtral-Small, 5-fold CV, 11,218 pooled utterances). Audio-only WER is 0.0665 for reference. Bold indicates the best (lowest WER / largest improvement) value in each numeric column.

Eval condition	Ctx-cond. WER	Δ vs zero-context	Δ vs audio-only	Rel.% vs audio-only
Zero-context control	0.0673	—	+0.0008	−1.2%
Diagnosis guidance	0.0663	−0.0010	−0.0002	+0.2%
Full clinical profile	0.0659	−0.0014	−0.0005	+0.8%
Speech production profile	0.0664	−0.0009	−0.0000	+0.0%
Voice quality profile	0.0662	−0.0010	−0.0002	+0.3%
Speech + voice profile	0.0662	−0.0010	−0.0002	+0.3%
Condensed full profile	0.0663	−0.0010	−0.0002	+0.3%
Condensed speech prod.	0.0661	−0.0012	−0.0003	+0.5%

Table 15: Supervised fine-tuning results (5-fold CV, 437 speakers, 11,218 pooled utterances).

System	Eval condition	WER	CER
Voxtral-Small (frozen)	Unprompted	0.1388	0.0980
audio-only LoRA	audio-only	0.0665	0.0506
Context-cond. LoRA	Zero-context control	0.0673	0.0511
Context-cond. LoRA	Full clinical profile	0.0659	0.0497

L Additional Analyses

This appendix collects analyses summarised in the main text: the iterative-correction condition, the deployment profile of context-dependent fine-tuning, the word-level error decomposition, and supporting figures and subgroup tables.

L.1 Hallucination and Verbosity: Full Analysis

A persistent concern with context prompting is whether clinical information simultaneously increases catastrophic failures. We analyze hallucination rates, verbosity, and failure modes across all frozen models under zero-context (P0), diagnosis-only (P1), and full clinical profile (P2) conditions (Figure 12, Appendix L).

We define a hallucination as any sample where unclipped WER exceeds 100%, indicating the prediction contains more erroneous words than the reference. Under zero-context decoding, hallucination rates range from 2.6% (Voxtral-Small) to 27.9% (Audio Flamingo 3), with the latter driven by a systematic preamble artifact ("The spoken content of the audio is '...'") that inflates word counts. Context prompting has divergent effects. For Audio Flamingo 3, diagnosis guidance reduces the hallucination rate from 27.9% to 2.9%, nearly eliminating the preamble mode. Voxtral-Small, by contrast, increases from 2.6% to 8.3% under P2 due to verbose reformulations triggered by the clinical profile. Gemma-4-4B Think shows the most extreme case, with the hallucination rate increasing from 3.8% at zero-context to 17.2% under diagnosis guidance and 14.7% under the full clinical profile. Qwen3-Thinking, by contrast, remains stable across conditions (3.3–3.4%), and the remaining models also show ± 1 pp variation.

The 90th percentile of output-to-reference length ratios (Figure 12b) confirms that most models maintain a P90 near 1.0 regardless of condition. Audio Flamingo 3 under P0 is the exception at $3.3\times$, normalizing to 1.0 once prompting constrains the output format, and Gemma-4-4B Think inflates from 1.0 at zero-context to $2.3\text{--}2.6\times$ under prompting, consistent with its catastrophic-degradation behaviour.

Decomposing the rate change into samples *fixed* (hallucinating under P0 but not P1) versus *induced* (not hallucinating under P0 but hallucinating under P1), Audio Flamingo 3 fixes approximately 25% of all samples (Figure 12c). Gemma-4-4B Think shows the opposite pattern: diagnosis guidance induces hallucinations in 14.6% of samples while fixing almost none. Among the remaining models, induced and fixed rates are small and roughly balanced (1–3% each), indicating that context prompting does not systematically create new failure modes in well-behaved instruct models.

Table 16: Headline paired comparisons for supervised systems (5-fold CV, 437 speakers). 95% CI at the speaker level.

Comparison	Base WER	New WER	Δ	Rel.%	p_{spkr}	95% CI (spkr)
Frozen $\rightarrow$ audio-only LoRA	0.1388	0.0665	-0.0723	+52.1	<.0001	—
audio-only $\rightarrow$ Context-cond. (full clin. prof.)	0.0665	0.0659	-0.0005	+0.8	0.5517	[-.0015, +.0023]
Context-cond. zero-context $\rightarrow$ full clin. prof. (same weights)	0.0673	0.0659	-0.0014	+2.0	0.3149	[-.0027, +.0007]

Table 17: audio-only $\rightarrow$ context-dependent (full clinical profile) by severity bin.

Severity	N	audio-only	Ctx-cond.	Δ	p_{spkr}
Mild (1–2)	7,304	0.031	0.030	-0.001	0.033
Moderate (3–4)	3,893	0.131	0.131	+0.000	0.900
Severe (5–7)	21	0.616	0.605	-0.011	0.750

Manual inspection reveals five recurring failure patterns. (1) *Repetitive loops*, where the model echoes disfluency patterns described in the clinical profile. (2) *Verbose reformulation*, where the model paraphrases rather than transcribes. (3) *Language switching*, producing fragments in German or Chinese when confronted with severely distorted speech. (4) *Refusal* (“I’m sorry, I cannot...”). (5) *Reasoning chain leakage*, unique to CoT models, where internal deliberation text appears in the output. Of these, only (1) is directly attributable to the clinical context. The model appears to “anticipate” described disfluency patterns and inadvertently reproduce them.

L.2 Iterative Correction

The follow-up correction condition (Table 3 in Appendix A) tests whether models can iteratively improve a transcription when given both the full clinical profile and a prior-pass transcript. This two-pass setup represents the prompting scenario where the model receives the audio, complete clinical context, and an initial transcription attempt to refine.

Across models where this condition was evaluated, follow-up correction does not systematically improve WER over single-pass prompting. Comparing follow-up correction to the full clinical profile in Table 5, WER is effectively unchanged for AF3 (-0.0002), Phi-4 (+0.0020), and Qwen3-Omni (+0.0001); Voxtral-Small (-0.0165) recovers some ground from the prompt-induced verbose output seen under the full profile; Qwen2-Audio shows no such recovery and instead worsens further (+0.0048 vs. full profile), with both models still underperforming the zero-context baseline by +0.0336 and +0.0558 WER respectively. The prior transcript does not systematically improve recognition, consistent with the broader finding that frozen models do not meaningfully condition on text-based context when processing audio, and suggests that iterative probing alone is unlikely to overcome this limitation.

L.3 Context-dependent Fine-Tuning in Practice

The context-dependent fine-tuning results (Sections 4.4 and 4.4) show that clinical context becomes useful once the model has been trained to expect it. The subgroup pattern (largest gains for Down syndrome and mild severity, no gain for moderate, slight degradation for cerebral palsy) is consistent with a model that benefits from context where the audio-only system struggles. The difficulty analysis (Section 4.5) makes this precise: the benefit grows with baseline difficulty and peaks on the hardest utterances, while clearly transcribable speech gains nothing and can pay a small penalty.

From a deployment standpoint, the approach has a favourable risk profile. When clinical context is unavailable, the context-dependent model performs comparably to the audio-only baseline ($p = 0.93$). When context is available, it helps. This asymmetry, with no penalty for missing context and modest benefit when present, makes context-dependent fine-tuning a strategy for clinical settings where metadata availability varies across patients and sessions. The difficulty-aware gate sharpens this further: rather than applying context indiscriminately, a deployed system should invoke it selectively on utterances where the audio-only pass is likely failing (Section 4.5).

Table 18: audio-only $\rightarrow$ context-dependent (full clinical profile) by etiology. p values from Wilcoxon signed-rank test at the speaker level.

Etiology	N	Spkrs	audio-only	Ctx-cond.	Δ	p_{spkr}
Down syndrome	358	21	0.152	0.141	-0.011	0.044
ALS	1,594	70	0.059	0.057	-0.003	0.066
Parkinson’s disease	7,504	249	0.037	0.036	-0.001	0.231
Cerebral palsy	1,762	97	0.180	0.185	+0.006	0.121

Table 19: Models evaluated for dysarthric speech transcription on the SAP dataset. Voxtral-Small-24B and Gemma-4-2B (Gemma-4-E2B) are additionally fine-tuned with LoRA (Section 4.4).

Model	Type	Params	HuggingFace Model ID
<i>Pure ASR</i>			
Qwen3-ASR-1.7B	Pure ASR	1.7B	Qwen/Qwen3-ASR-1.7B
Granite-Speech 4.1-2B NAR	Pure ASR	2B	ibm-granite/granite-speech-4.1-2b-nar
Cohere Transcribe	Pure ASR	2.0B	CohereLabs/cohere-transcribe-03-2026
Whisper Large v3	Pure ASR	1.5B	openai/whisper-large-v3
NVIDIA Parakeet TDT 1.1B	Pure ASR	1.1B	nvidia/parakeet-tdt-1.1b
wav2vec2-large-robust	Pure ASR	0.3B	facebook/wav2vec2-large-robust-ft-libri-960h
wav2vec2-large-960h	Pure ASR	0.3B	facebook/wav2vec2-large-960h
<i>Instruct (audio + text input)</i>			
Voxtral-Small-24B	Instruct	24B	mistralai/Voxtral-Small-24B-2507
Voxtral-Mini-3B	Instruct	3B	mistralai/Voxtral-Mini-3B-2507
Phi-4-multimodal	Instruct	5.6B	microsoft/Phi-4-multimodal-instruct
Qwen2-Audio-7B	Instruct	7B	Qwen/Qwen2-Audio-7B-Instruct
MiniCPM-o 4.5	Instruct	9B	openbmb/MiniCPM-o-4_5
Qwen3-Omni-30B-A3B	Instruct	30B [†]	Qwen/Qwen3-Omni-30B-A3B-Instruct
Ultravox-v0.4.1-Nemo	Instruct	12B	fixie-ai/ultravox-v0_4_1-mistral-nemo
NVIDIA Audio Flamingo 3	Instruct	8B	nvidia/audio-flamingo-3-hf
Granite-Speech 3.3-8B	Instruct	8B	ibm-granite/granite-speech-3.3-8b
Granite-Speech 4.1-2B+	Instruct	2B	ibm-granite/granite-speech-4.1-2b-plus
Gemma-4-2B	Instruct	2B	google/gemma-4-E2B-it
Gemma-4-4B	Instruct	4B	google/gemma-4-E4B-it
Gemma-4-12B [‡]	Instruct	12B	google/gemma-4-12B-it
<i>Chain-of-Thought (CoT)</i>			
Qwen3-Omni-30B-A3B-Thinking	CoT	30B [†]	Qwen/Qwen3-Omni-30B-A3B-Thinking
MiniCPM-o 4.5 Think	CoT	9B	openbmb/MiniCPM-o-4_5
Gemma-4-4B Think	CoT	4B	google/gemma-4-E4B-it
Gemma-4-12B Think [‡]	CoT	12B	google/gemma-4-12B-it
Audio-Reasoner	CoT	7B	zhifeixie/Audio-Reasoner

[†] Mixture-of-Experts architecture with 3B active parameters per forward pass. [‡] Gemma-4-12B uses a unified (encoder-free) multimodal architecture: audio is projected directly into the language model without a separate audio encoder, unlike the E2B/E4B variants.

L.4 Word-Level Error Decomposition

To characterize what the context-dependent system repairs, we align the audio-only and context-dependent predictions word by word against the reference, type each error by CMU-dictionary phonological class (consonant, vowel, word-boundary, other), and measure how often the context-dependent system recovers a reference word the audio-only system missed (speaker-level statistics, $N = 11,218$ utterances).

Context-dependent fine-tuning does repair pronunciation errors: the context-dependent system recovers 17% of the audio-only system’s consonant-word errors and 20% of its vowel-word errors. The net effect, however, is architecture-dependent (Figure 13a). On Voxtral-Small the recoveries are offset by new errors, and the net change per 1,000 reference words is null for all four error types (every interval crossing zero). On Gemma-4-E2B, context significantly reduces consonant errors (about 10 per 1,000 words, interval excluding zero), consistent with the weaker model using context to repair consonant-heavy speech.

The repair is not specific to the matching clinical dimension (Figure 13b). If the model extracted graded information, samples rated high on *imprecise consonants* should show a

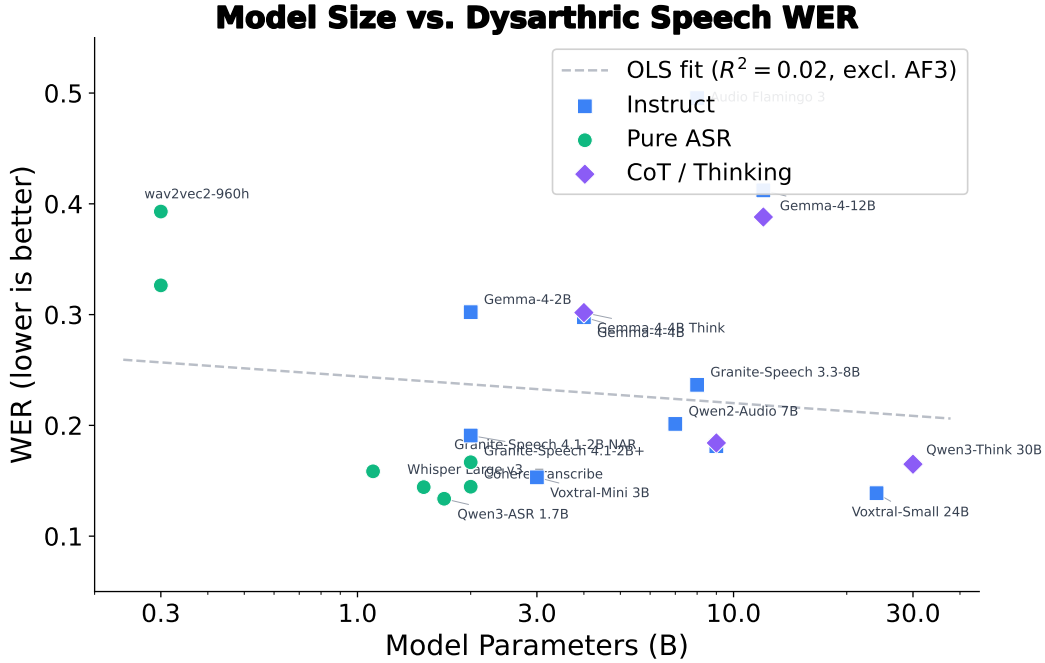


Figure 8: Model size (parameters) vs. zero-shot WER on dysarthric speech. Dashed line shows OLS fit on log-scale parameters (excluding AF3). Model size is a weak predictor of performance ($R^2 = 0.07$); architecture and training data matter more.

Table 20: Correlation between ASR model performance and human-assessed speech severity on rated samples ($n=11,218$). Spearman ρ is reported between per-sample mean severity rating (scale 1–7) and each metric. All correlations are significant at $p < 10^{-100}$. Intelligibility-specific correlation uses only samples with an explicit intelligibility dimension rating ($n=3,738$).

Model	Mean Severity ρ			Intelligibility
	WER	CER	SemScore	WER ρ
Qwen3-ASR-1.7B	.420	.426	−.394	.494
Whisper Large v3	.401	.412	−.392	.461
Cohere Transcribe	.380	.392	−.366	.448
Parakeet TDT 1.1B	.366	.379	−.340	.428
wav2vec2-robust	.368	.412	−.406	.386
wav2vec2-960h	.351	.403	−.419	.371

larger consonant fix-rate gap than samples rated low, and analogously for *distorted vowels*. The dimension-aligned gaps are negligible on Voxtral-Small (−0.001 for imprecise consonants, +0.010 for distorted vowels, +0.002 for the prosodic control) and show no clean pattern on Gemma-4-E2B either (−0.010 and −0.024, with the slow-rate control at −0.048). A counterfactual cross-check settles the mechanism: when the clinical profile is replaced with a different speaker’s, the high-versus-low gap by the sample’s *true* rating persists rather than disappearing. The recovery therefore tracks the acoustic difficulty of the sample, not the content of the prompt. Our reading is that the fine-tuned model uses clinical context as a coarse signal that triggers lexical recovery on degraded audio, concentrated where the audio-only model is failing, rather than as graded per-dimension information, consistent with the difficulty-conditioned benefit of Section 4.5 and with the hallucinated-correction risk discussed in Section 6.

As a further probe we also ran the chain-of-thought (*Think*) variant of the unified 12B on P0–P2 (Tables 4, 6). Enabling thinking *lowers* the zero-context WER (0.388 vs. 0.412) and suppresses the refusals almost entirely (from $\approx 24\%$ to under 2%), but it does not rescue the prompted condition: the profile-induced degradation is if anything slightly larger (Δ WER

Table 21: ASR performance stratified by mean clinical severity level (rounded). Severity 1 = normal range, 5 = most severe in the rated subset. All models are evaluated on the same 11,218 rated samples. The steepest degradation occurs between severity 2 and 3 across all models and metrics.

Model	WER _{norm_min} (↓)					CER _{norm_min} (↓)					SemScore (↑)				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Qwen3-ASR	.044	.113	.357	.477	.736	.028	.076	.261	.358	.657	87.9	75.0	49.2	40.6	40.7
Cohere Trans.	.063	.125	.353	.444	.770	.039	.085	.265	.352	.722	86.1	73.9	49.7	41.6	42.0
Whisper v3	.051	.126	.362	.483	.805	.033	.087	.270	.365	.690	86.7	73.1	48.0	39.1	40.5
Parakeet	.088	.133	.374	.448	.766	.056	.092	.293	.364	.739	81.5	70.9	45.7	39.6	27.6
w2v2-robust	.225	.293	.627	.706	.957	.104	.161	.400	.477	.758	74.8	61.5	32.4	24.9	17.2
w2v2-960h	.291	.356	.712	.775	.946	.146	.203	.469	.525	.776	70.3	56.7	26.9	20.8	20.9
<i>n samples</i>	2,519	7,111	1,436	142	10										

Table 22: Zero-shot ASR performance on the SAP rated subset ($N = 11,218$). WER and CER follow the SAP evaluation protocol (normalized, dual-reference minimum, clipped at 1.0). SemScore is the SAP composite semantic similarity metric (0–100, higher is better). All models are evaluated truly unprompted (audio only), except that Audio Flamingo 3’s raw unprompted output is reported (the model produces non-transcription output without a prompt; the basic transcription instruction reduces it to 0.162). Best result per column in **bold**. [†]Mixture-of-Experts with 3B active parameters.

Model	Type	Params	Overall			WER by Etiology				SemScore by Etiology			
			WER↓	CER↓	Sem↑	PD	ALS	DS	CP	PD	ALS	DS	CP
Pure ASR Models													
Qwen3-ASR-1.7B	Pure ASR	1.7B	.134	.093	74.1	.073	.176	.227	.333	79.8	72.5	59.7	54.4
Granite-Speech 4.1-2B NAR	Pure ASR	2B	.167	.120	69.4	.086	.254	.286	.405	76.3	64.2	53.0	47.8
Cohere Transcribe	Pure ASR	2.0B	.145	.102	73.1	.076	.191	.289	.365	79.3	71.4	54.0	52.3
Whisper Large v3	Pure ASR	1.5B	.144	.102	72.5	.083	.182	.260	.348	78.3	71.5	55.5	52.3
Parakeet TDT 1.1B	Pure ASR	1.1B	.159	.114	69.6	.088	.216	.258	.385	76.0	67.0	54.4	48.0
wav2vec2-robust	Pure ASR	0.3B	.326	.183	60.3	.234	.401	.526	.612	67.7	56.6	40.3	35.9
wav2vec2-960h	Pure ASR	0.3B	.393	.229	55.4	.294	.486	.598	.688	63.5	50.6	34.2	29.7
Instruct Audio LLMs													
Voxtral-Small 24B	Instruct	24B	.139	.098	73.3	.075	.183	.257	.347	79.2	71.8	56.7	53.0
Voxtral-Mini 3B	Instruct	3B	.153	.106	72.1	.087	.199	.282	.366	78.1	70.3	54.2	51.4
Phi-4 Multimodal	Instruct	5.6B	.158	.119	72.6	.093	.211	.253	.367	78.4	70.0	58.4	52.8
Qwen3-Omni 30B	Instruct	30B†	.138	.103	73.4	.068	.194	.239	.365	79.7	70.7	58.5	51.8
Qwen2-Audio 7B	Instruct	7B	.201	.153	68.7	.129	.256	.340	.433	75.3	65.7	50.7	46.8
MiniCPM-o 4.5	Instruct	9B	.181	.132	70.0	.105	.256	.297	.416	76.6	65.9	54.7	48.4
Granite-Speech 4.1-2B+	Instruct	2B	.191	.137	69.8	.103	.291	.327	.447	77.3	63.9	52.3	46.8
Granite-Speech 3.3-8B	Instruct	8B	.237	.183	66.9	.169	.304	.302	.449	73.4	62.2	53.8	46.3
Ultravox-Nemo 12B	Instruct	12B	.312	.244	54.1	.242	.360	.478	.533	60.1	50.6	36.6	35.3
Gemma-4-4B	Instruct	4B	.298	.210	59.7	.211	.391	.444	.553	66.7	54.2	43.2	38.2
Gemma-4-2B	Instruct	2B	.302	.213	59.5	.216	.397	.479	.550	66.4	54.0	41.4	38.7
Gemma-4-12B	Instruct	12B	.412	.339	51.8	.324	.487	.595	.683	58.9	47.0	33.1	29.3
Audio Flamingo 3	Instruct	8B	.495	.465	47.7	.428	.557	.701	.686	52.4	43.9	35.8	33.8
Chain-of-Thought													
MiniCPM-o 4.5 Think	CoT	9B	.184	.134	69.8	.106	.264	.301	.420	76.4	65.6	54.5	48.6
Qwen3-Thinking 30B	CoT	30B†	.165	.122	72.8	.086	.240	.259	.414	79.5	68.9	58.8	50.4
Gemma-4-4B Think	CoT	4B	.302	.215	60.3	.220	.391	.446	.542	67.2	55.0	43.2	39.2
Audio-Reasoner	CoT	7B	.240	.194	67.6	.161	.302	.385	.491	74.2	64.1	50.9	45.9
Gemma-4-12B Think	CoT	12B	.388	.325	55.6	.313	.466	.503	.612	61.9	50.5	41.0	36.4

+0.113 vs. +0.088), converging to the same ≈ 0.50 WER. This within-family contrast is itself informative, for the smaller Gemma-4-4B, thinking instead *amplifies* the degradation (Table 6), and points to future mechanistic analysis of when a reasoning trace aids versus displaces the transcription.

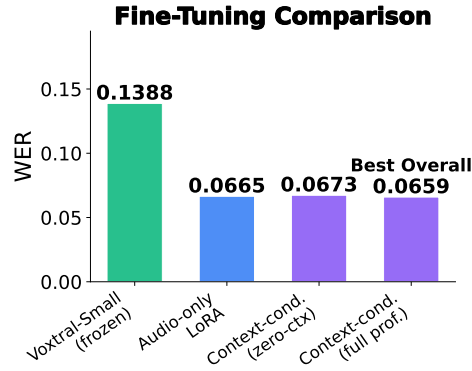


Figure 9: Fine-tuning comparison (5-fold cross-validation, 437 speakers). The context-dependent system achieves the best overall WER. Numerical values are in Table 15.

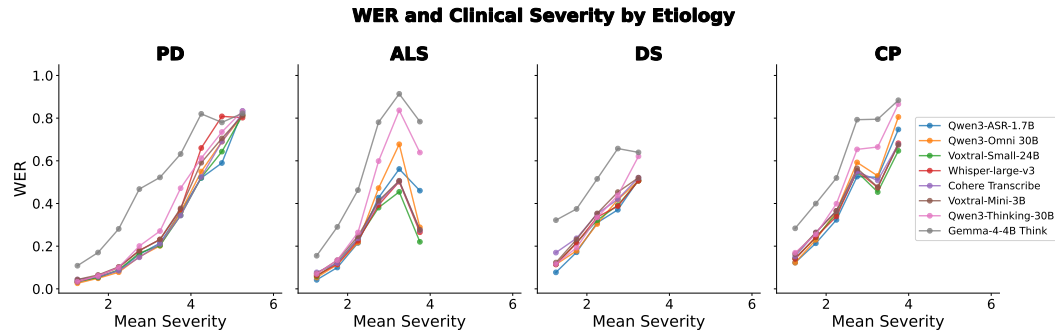


Figure 10: WER vs. clinical severity by etiology. Performance degrades beyond severity level 2 across models, with the steepest decline for cerebral palsy and Down syndrome.

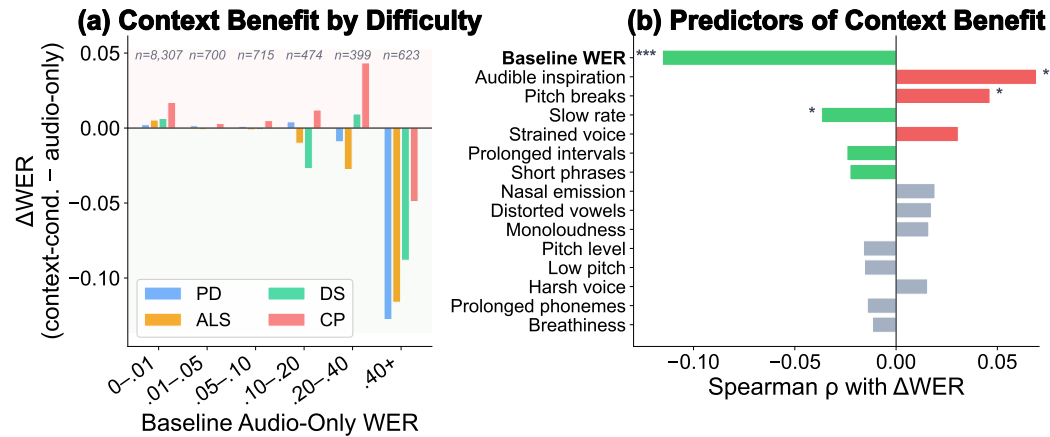


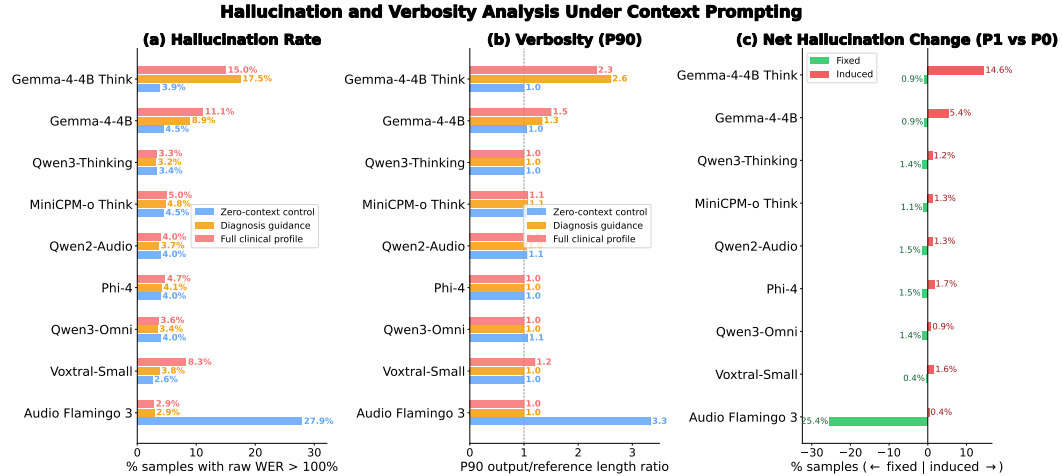
Figure 11: Predictors of per-sample context benefit (audio-only $\rightarrow$ context-dependent, 5-fold cross-validation, $N = 11,218$). (a) Mean ΔWER by baseline difficulty and etiology. (b) Spearman ρ between each predictor and ΔWER . Baseline WER dominates; individual clinical dimensions show near-zero correlation.

Table 23: Effect of context prompting on $WER_{\text{norm_min}}$ ($\downarrow$). Values show absolute P0 and $\Delta = P_x - P_0$. Green indicates improvement, red degradation. For WER, negative Δ is an improvement.

Model	P0	P1 (condition only)					P2 (full profile)				
		ΔAll	ΔPD	ΔALS	ΔDS	ΔCP	ΔAll	ΔPD	ΔALS	ΔDS	ΔCP
Audio Flamingo 3	.498	-.336***	-.333***	-.346***	-.431***	-.322***	-.335***	-.331***	-.345***	-.435***	-.320***
MiniCPM-o Think	.184	-.006	-.000***	-.015**	-.033***	-.016***	-.006	-.000***	-.013*	-.032***	-.018***
Qwen3-Omni 30B	.138	+.003***	+.010***	+.003***	-.018**	-.020***	+.004***	+.010***	+.001***	-.009	-.019***
Phi-4 Multimodal	.157	-.000***	+.005***	-.003	-.031**	-.015	+.000***	+.001***	+.004***	-.002	-.007
Voxtral-Small 24B	.139	+.010***	+.010***	+.014*	+.011	+.002	+.050***	+.048***	+.057***	+.077***	+.046***
Qwen2-Audio 7B	.201	+.031***	+.033***	+.028***	+.035*	+.023***	+.051***	+.060***	+.029***	+.036***	+.036***
Qwen3-Thinking 30B	.165	+.006***	+.013***	-.006	+.005	-.012	+.007***	+.012***	+.001	+.031*	-.010
Gemma-4-4B	.292	+.032***	+.029***	+.043***	+.038***	+.034***	+.049***	+.045***	+.068***	+.040***	+.047***
Gemma-4-4B Think	.302	+.208***	+.225***	+.174***	+.178***	+.172***	+.202***	+.221***	+.175***	+.196***	+.147***
Gemma-4-12B	.412	+.118***	+.125***	+.107***	+.123***	+.094***	+.088***	+.099***	+.057***	+.090***	+.072***

Table 24: Effect of context prompting on $CER_{\text{norm_min}}$ ($\downarrow$). Conventions as in Table 23.

Model	P0	P1 (condition only)					P2 (full profile)				
		ΔAll	ΔPD	ΔALS	ΔDS	ΔCP	ΔAll	ΔPD	ΔALS	ΔDS	ΔCP
Audio Flamingo 3	.467	-.349***	-.337***	-.365***	-.476***	-.363***	-.348***	-.336***	-.363***	-.478***	-.358***
MiniCPM-o Think	.134	-.004***	+.000	-.010**	-.021***	-.013***	-.004***	+.000	-.009***	-.017**	-.015***
Qwen3-Omni 30B	.103	-.002***	+.005***	-.004	-.014*	-.023***	-.001***	+.005***	-.006	-.008	-.022***
Phi-4 Multimodal	.118	-.001***	+.004***	-.002	-.028***	-.014	+.001***	+.000***	+.007***	+.001	-.002
Voxtral-Small 24B	.098	+.011***	+.009***	+.020***	+.021*	+.011*	+.054***	+.047***	+.068***	+.093***	+.062***
Qwen2-Audio 7B	.153	+.016***	+.020***	+.010***	+.008	+.002	+.037***	+.046***	+.017***	+.018	+.022***
Qwen3-Thinking 30B	.122	+.004***	+.010***	-.002	+.004	-.013	+.006***	+.010***	+.003	+.029*	-.012
Gemma-4-4B	.205	+.028***	+.026***	+.036***	+.034**	+.031***	+.050***	+.045***	+.076***	+.046**	+.052***
Gemma-4-4B Think	.215	+.226***	+.234***	+.203***	+.221***	+.216***	+.223***	+.233***	+.204***	+.250***	+.190***
Gemma-4-12B	.339	+.145***	+.148***	+.135***	+.168***	+.138***	+.116***	+.122***	+.085***	+.137***	+.116***

Figure 12: Hallucination and verbosity under context prompting ($N = 11,218$). (a) Hallucination rate (raw WER > 100%). (b) P90 output-to-reference length ratio; dashed line marks parity. (c) Samples fixed vs. induced by diagnosis guidance.Table 25: Effect of context prompting on SemScore ($\uparrow$, scale 0–100). Conventions as in Table 23. For SemScore, positive Δ is an improvement.

Model	P0	P1 (condition only)					P2 (full profile)				
		ΔAll	ΔPD	ΔALS	ΔDS	ΔCP	ΔAll	ΔPD	ΔALS	ΔDS	ΔCP
Audio Flamingo 3	51.1	+18.3***	+19.2***	+20.5***	+15.0***	+13.3***	+18.3***	+19.1***	+20.4***	+15.6***	+13.1***
MiniCPM-o Think	69.8	+0.6	+0.2***	+1.2	+2.7***	+1.5***	+0.9***	+0.5	+1.5*	+2.8***	+1.6***
Qwen3-Omni 30B	73.4	+0.6***	+0.1***	+0.9***	+2.1***	+2.3***	+0.6***	+0.1***	+1.0***	+1.3***	+2.1***
Phi-4 Multimodal	72.6	-.05***	-1.0***	+0.3*	+2.4	+0.4	-.05***	-.06***	-.01*	+0.2	-.05
Voxtral-Small 24B	73.1	-.02	-.02*	-0.4	-0.3	+0.1***	-3.6***	-3.3***	-4.6***	-5.6***	-3.5***
Qwen2-Audio 7B	68.7	-3.9***	-4.5***	-3.1***	-2.6***	-2.3***	-5.3***	-6.5***	-2.9***	-2.6***	-2.9***
Qwen3-Thinking 30B	72.8	-0.4	-0.9***	+0.2	+0.0	+1.0***	-0.5***	-0.9***	+0.1	-1.7**	+0.5
Gemma-4-4B	60.9	-2.8***	-2.7***	-3.3***	-3.1***	-3.0***	-4.6***	-4.3***	-6.5***	-3.2**	-4.4***
Gemma-4-4B Think	60.3	-15.6***	-16.4***	-15.1***	-12.5***	-12.8***	-13.5***	-14.9***	-12.5***	-11.9***	-9.2***
Gemma-4-12B	51.8	-10.4***	-11.0***	-10.2***	-9.0***	-8.2***	-8.3***	-9.0***	-6.5***	-7.1***	-7.2***

Table 26: Subgroup analysis: audio-only vs. context-dependent. Rel.% is the aggregate relative WER change, $(\overline{\text{WER}}_{\text{ctx}} - \overline{\text{WER}}_{\text{a0}}) / \overline{\text{WER}}_{\text{a0}}$ (negative = improvement). p from Wilcoxon signed-rank at speaker level.

By Etiology				By Severity			
	N	Rel.%	p		N	Rel.%	p
Down syndrome	358	−7.1	.044	Mild (1–2)	7,304	−2.7	.033
ALS	1,594	−4.6	.066	Moderate (3–4)	3,893	+0.0	.900
Parkinson’s	7,504	−2.7	.231	Severe (5–7)	21	−1.8	.750
Cerebral palsy	1,762	+3.1	.121				

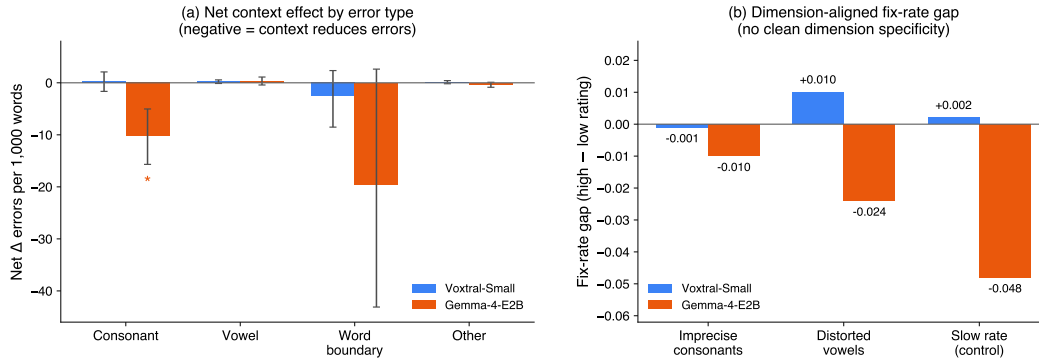


Figure 13: Word-level error decomposition of the context-dependent vs. audio-only LoRA systems (5-fold cross-validation, $N = 11,218$). (a) Net change in errors per 1,000 reference words from adding the full clinical profile (negative = context reduces that error type), with bootstrap intervals. On Voxtral-Small all four types are null; on Gemma-4-E2B consonant errors drop significantly (*, interval excluding zero). (b) Fix-rate gap between samples rated high vs. low on each clinical dimension. The dimension-aligned gaps are small and the prosodic control (slow rate) is comparable, so the residual recovery is not directed by the prompt’s content.