Mechanistic Insights into Context Failures of Audio-Language Models for Impaired Speech

Pehuén Moure [*1,2] **Bilal Bounajma** [*3] **Niclas Pokel** [*1,2] **Yingqiang Gao** [4,2] **Roman Boehringer** [1]
Longbiao Cheng [1] **Gonçalo Guiomar** [2,1] **Shih-Chii Liu** [1]

Abstract

Frozen audio-language models can transcribe dysarthric speech *worse* when prompted with clinical context than with audio alone. We study this failure mode in Gemma-4-E2B on the Speech Accessibility Project (SAP) rated subset and trace it to the residual stream. To test whether the degradation is encoded in the model's internal state rather than in the audio representation itself, *Matched Residual Patching* (MRP) replaces the final-input-token residual at a middle decoder layer under the clinical prompt with the corresponding audio-only residual and closes 57–60% of the Word Error Rate (WER) gap bidirectionally. The repairable component lies mostly *outside* a sparse prompt-identity axis: ∼27 residual coordinates identify the prompt at 99.8% accuracy, but patching them recovers only 1–10% of the gap, while their 1,516-coordinate complement carries 40–59%. We then introduce *Self-Contrastive Residual Alignment* (SCRA), a one-pass inference-time edit that learns a low-rank linear corrector for the prompt-induced residual shift. At rank 128, SCRA matches the MRP ceiling, closes 60% of the aggregate gap, and slightly but significantly improves over the audio-only baseline on difficult samples.

1. Introduction

Voice interfaces are largely unusable for the millions of speakers with voice, speech, or language disorders (Hoffman et al., 2014; Jaddoh et al., 2025; Tobin et al., 2024). Native audio–text models (Chu et al., 2024; Gemini Team, Google, 2024; Google DeepMind, 2026; Liu et al., 2026; OpenAI, 2024) raise the prospect of *context-conditioned* recognition: a frozen model that adapts to a speaker at inference when given relevant language information (Wagner et al., 2025). Clinical context is the natural candidate when speech is impaired. Empirically, however, zero-shot evaluations on dysarthric corpora report uneven and degrading behavior under clinical prompting (Alsayegh & Masood, 2025; Moure et al., 2026), even when the context is accurate. The largest damage falls on utterances the audio-only baseline transcribes cleanly, so a useful repair must keep clinical context available while preventing it from disrupting audio-grounded recognition.

We study this failure mechanistically in Gemma-4-E2B (Google DeepMind, 2026) on the SAP rated dataset (Zheng et al., 2025), building on residual-stream analyses (Elhage et al., 2021), activation patching and causal tracing (Conmy et al., 2023; Geiger et al., 2021; Meng et al., 2022; Wang et al., 2023), and task-vector style localizations (Hendel et al., 2023; Todd et al., 2024). Our contributions are: (i) we localize a repairable residual-stream state with MRP, closing 57–60% of the WER gap bidirectionally with a single mid-stack patch; (ii) we show this state lies mostly outside the sparse coordinates that identify the prompt condition — a sparse/distributed split that explains why population-mean steering fails at the same depth; (iii) we introduce SCRA, a one-pass corrector that matches the MRP ceiling and slightly improves over audio-only on the difficult samples where clinical context would matter most.

2. Background

Speech recognition for atypical speakers. Articulation, voice quality, and intelligibility vary substantially by speaker and by underlying neurological condition (Tobin et al., 2024), and human listeners routinely rely on contextual information about the speaker to interpret ambiguous segments. Standard pipelines for non-normative automatic speech recognition remain brittle for atypical speakers and depend on speaker-dependent adaptation or personalized fine-tuning of self-supervised speech encoders (Wagner

[1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland [2]ETH AI Center, Zurich, Switzerland [3]ETH Zurich, Zurich, Switzerland [4]Department of Computational Linguistics, University of Zurich, Zurich, Switzerland. Correspondence to: Pehuén Moure <pehuen@ini.uzh.ch>, Niclas Pokel <npokel@ethz.ch>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

et al., 2025), both of which require target-population data that is expensive and difficult to collect at scale.

In-context conditioning in audio-language models. Audio-native multimodal models (Chu et al., 2024; Gemini Team, Google, 2024; Google DeepMind, 2026; Liu et al., 2026; OpenAI, 2024) open a different route: a single frozen model that adapts to a speaker at inference using natural-language information about them. In-context learning is well studied for text but remains underexplored for speech and audio, where existing work focuses mostly on few-shot prompting of downstream NLP tasks rather than on direct conditioning of the audio-recognition pathway. Recent zero-shot evaluations on dysarthric corpora report limited or degrading benefits from clinical prompting (Alsayegh & Masood, 2025; Moure et al., 2026), even when the context is accurate and matched to the speaker. We take this degrading regime as our phenomenon and ask what changes inside the model can explain it.

Mechanistic interpretability for multimodal failure modes. Mechanistic interpretability has produced a broad toolkit for transformer analysis, including residual-stream analyses (Elhage et al., 2021), activation patching and causal tracing (Conmy et al., 2023; Geiger et al., 2021; Meng et al., 2022; Wang et al., 2023), and task-vector style localizations (Hendel et al., 2023; Todd et al., 2024). These tools have been developed largely in text-only language models, and their application to audio-language models is much less explored. We extend them here to a multimodal failure mode of a frozen audio-LM, with the explicit goal of identifying intervention sites that can be exploited at inference time without retraining.

3. Setup and Behavioral Effect

Model and dataset. Gemma-4-E2B (Google DeepMind, 2026) is a 35-layer audio-LM with hidden dimension 1,536 and a 4:1 local-to-global attention pattern (global layers $\{4, 9, 14, 19, 24, 29, 34\}$, local sliding window 512). We evaluate the 11,218 SAP rated-subset utterances from 437 speakers (Parkinson's, ALS, cerebral palsy, Down syndrome) for which matched outputs exist across every prompt condition. WER follows the SAP normalization protocol (Radford et al., 2022; Zheng et al., 2025) with per-utterance clipping at 1.0; *Difficult Samples* are utterances with $\mathrm{WER}_{\mathrm{P0}} \geq 0.1$ (58% of corpus), where only audio is presented to the model. All comparisons are paired at the utterance level.

Prompt grid. Prompts are aligned with a previous work on evaluating Audio-LM on the SAP dataset (Moure et al., 2026). P0 is audio-only with a generic prompt which asks for transcription; P1 adds the diagnosis; P2 adds the full clinical profile (etiology + per-sample 1–7 clinician ratings + listening guidance); P2a is a length-doubled rephrasing. Length-matched fillers replace the clinical block with non-clinical text (instruction-only repetition; random pseudo-text). A *wrong-content* control keeps the P2 structure but uses ratings from another speaker via etiology rotation.

Clinical context degrades the frozen model; the trigger is prompt structure. Aggregate WER rises from 0.327 (P0) to 0.395 (P2), a paired increase of $+0.068$ (SE$\approx$0.001 speaker-clustered) concentrated on items P0 transcribes cleanly (Figure 1B). Three controls localize the trigger. (i) *Random-text* filler reaches WER 0.375 and reproduces most of the easy-sample damage. (ii) *Instruction-only* filler of the same length lands at 0.322, indistinguishable from P0, ruling out length. (iii) *Wrong-content* P2 reaches 0.399, statistically indistinguishable from correct P2, indicating clinical-content accuracy does not contribute effectively to the gap. Per-speaker damage $\Delta_2{=}(\mathrm{P2}{-}\mathrm{P0})$ further correlates with $\Delta_{\mathrm{rand}}{=}(\mathrm{random}{-}\mathrm{P0})$ at $r{\approx}0.6$ (Figure 1C), consistent with a shared speaker-level susceptibility to non-instructional prompt content. Finally, under matched LoRA fine-tuning the same model reaches WER 0.196 at P2 and 0.200 at P0, both below frozen P0, indicating the failure is one of *integration* rather than of clinical content itself.

4. Localizing a Repairable Residual State

We use MRP as a causal diagnostic: for each utterance, replace an internal state under one prompt with the corresponding state from the same utterance under the alternative prompt. *Rescue* swaps P2 $\rightarrow$ P0, *induction* swaps P0 $\rightarrow$ P2. Concretely, a forward hook fired once during prefill replaces the residual-stream output of layer ℓ at the final prompt-token position before it enters layer $\ell{+}1$; subsequent autoregressive decoding proceeds from the modified KV cache. Following work that localizes high-level behavior to mid-stack residuals (Geva et al., 2021; Hendel et al., 2023; Meng et al., 2022; Todd et al., 2024), we target four candidate layers spanning both attention kinds: L14, L19, L24 (global) and L21 (local sliding-window).

A single final-input residual patch closes most of the gap. At L19, replacing the P2 state with the matched P0 state reduces WER from 0.395 to 0.354 (60% gap closure); the reverse patch raises WER from 0.327 to 0.366 (57%). The same internal state mediates a majority of the behavioral gap in both directions. The effective band is broad: L19, L21, and L24 give nearly identical rescue performance (within 0.001 WER), while the earlier global layer L14 is weaker (Figure 2D). The effect spans both attention kinds and holds within each etiology (52–66% per-etiology gap closure, no reversals). Position is sharp: shifting the patched position back even one token kills the effect (Figure 2B), and the

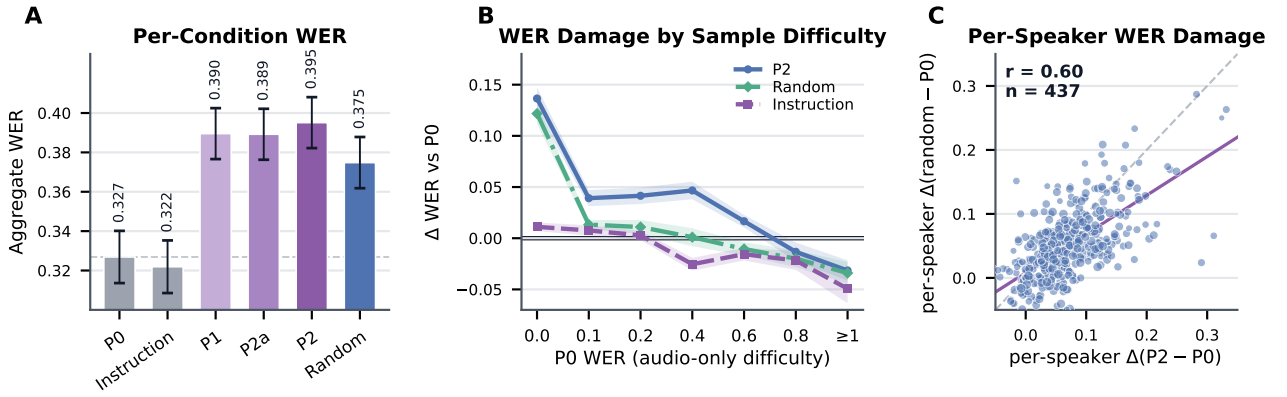

Figure 1. Behavioral effect of clinical context. *(A)* Per-condition aggregate WER with speaker-clustered ±SE caps. P2 and random-text filler raise WER above the P0 floor; instruction-only filler reproduces P0. *(B)* Per-utterance WER damage $\Delta(\text{cond} - \text{P0})$ profiled by P0-difficulty bucket, with ±SE bands. P2 and random-filler damage is concentrated on items P0 transcribes cleanly; instruction filler is at or below zero across the range. *(C)* Per-speaker damage under P2 vs. random filler, Pearson $r \approx 0.6$ across $n = 437$ speakers: vulnerability is shared across non-instructional prompt content, not diagnosis-specific.

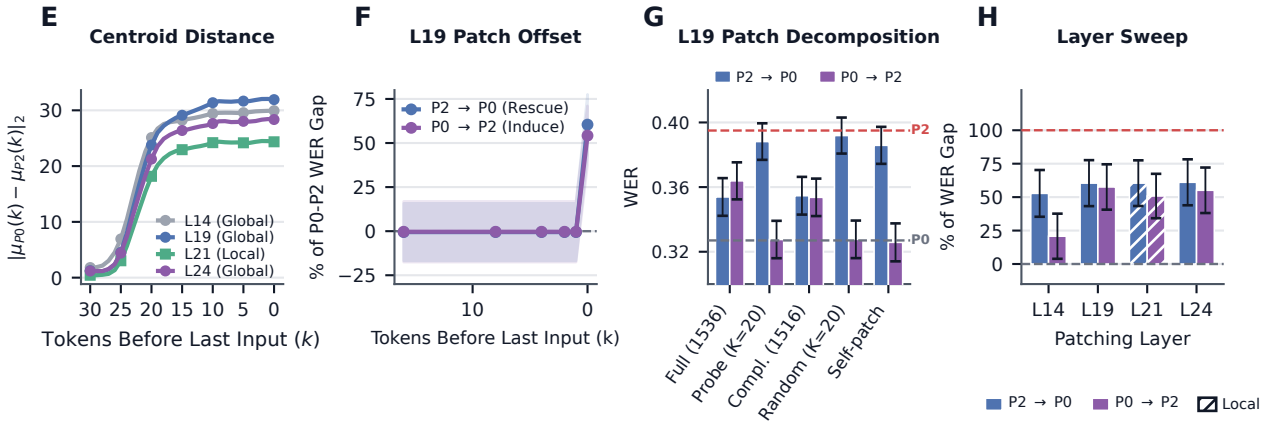

Figure 2. Last-input residual diagnostics. *(A)* Centroid distance $\|\mu_{\text{P0}}(k) - \mu_{\text{P2}}(k)\|_2$ over the last ∼30 input tokens accumulates and saturates at the patching target $k=0$ at every layer. *(B)* Position sweep at L19: patching is causally active only at $k=0$. *(C)* Feature-coordinate decomposition at L19: full residual swap recovers 60%/57% (rescue/induce); the top-$K=20$ probe coordinates that identify the prompt recover only 10%/1%; their 1,516-coordinate complement carries 59%/40%. *(D)* Layer sweep of single-position last-input patching: L19/L21/L24 close 53–62% of the gap; the early-mid global layer L14 lags. Hatching marks the local-attention layer (L21). Whiskers are ±SEM across speakers.

audio-end state alone does not move WER.

The repairable state is outside the sparse prompt-identity axis. At the final-input position, the prompt condition is almost perfectly identifiable from a sparse residual axis. An L_1-logistic probe trained under 5-fold speaker-disjoint cross validation (CV) distinguishes P0 from P2 at 99.8% accuracy using ∼27 active coordinates, and the population-mean directions across P1, P2, P2a are tightly aligned (pairwise cosines 0.91–0.99). One might therefore expect those coordinates to mediate the WER gap. However, using the probe coefficients as a coordinate mask (no refit), patching only the top $K=20$ probe coordinates recovers just 10% (rescue) and 1% (induction), while patching the complementary 1,516 coordinates recovers 59% and 40% (Figure 2C). The residual state therefore carries *two separable signals*: a sparse axis-aligned component that identifies which prompt was used, and a distributed component, orthogonal to it, that carries most of the behavioral effect. Consistent with this geometry, two natural low-dimensional interventions at the same depth fail: additive steering along the population-mean direction $\hat{d}_{\text{P2}}$ recovers $\leq 7\%$ at L19/L21/L24, and a mass-preserving rescaling (Wang et al., 2023) of audio-attention at the L21 dilution peak does not move WER. The mediator is per-sample and high-dimensional, not a fixed direction or an attention-magnitude effect. Methodologically, this dissociation matters: a near-perfect linear probe for the prompt condition does not localize the residual coordinates that mediate behavior, and population-mean direction analyses miss the per-utterance distributed mediator entirely. Probe targets and patching targets need not coincide.

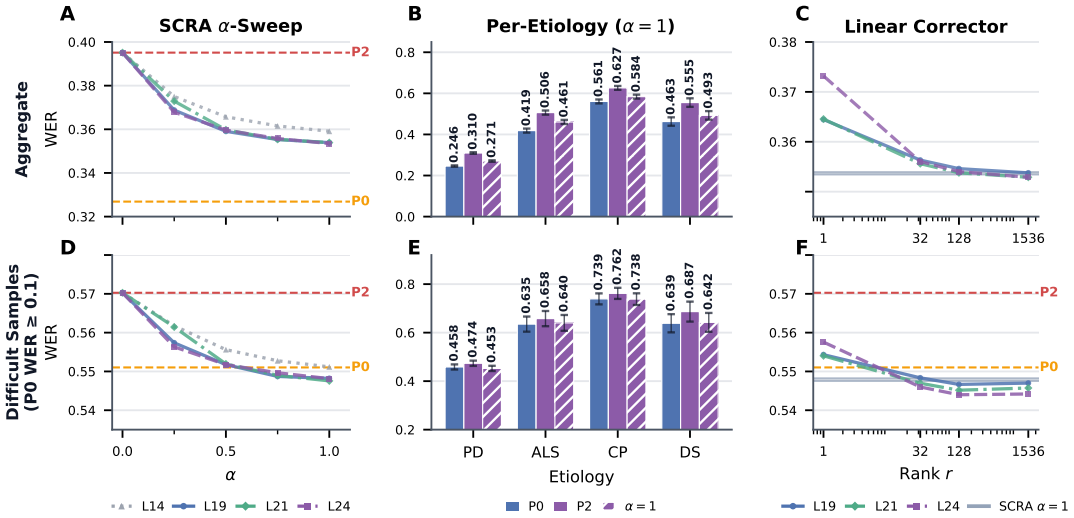

Figure 3. MRP and SCRA on the full corpus (top row) and Difficult Samples ($\mathrm{WER}_{\mathrm{P0}} \geq 0.1$, 58% of corpus, bottom). *(A,D)* Cached MRP α-sweep at four mid-stack layers; L19/L21/L24 within 0.001 WER at α=1. *(B,E)* Per-etiology rescue at α=1 with P0 and P2 brackets. *(C,F)* Single-pass corrector vs. MRP α=1: rank 128 matches the reference at all three trained layers.

5. Self-Contrastive Residual Alignment

MRP identifies $r^{(\ell)}_{\mathrm{P0}}$ as the target a repair should approximate, but it requires a cached paired P0 forward pass and is therefore a diagnostic, not a deployment method. We introduce *Self-Contrastive Residual Alignment* (SCRA), an inference-time edit that approximates this target without paired data.

Edit and variants. All variants apply, at the final-input token of layer ℓ,

$$r^{(\ell)}_{\mathrm{edit}}(\alpha) = (1-\alpha)\, r^{(\ell)}_{\mathrm{full}} + \alpha\, r^{(\ell)}_{\mathrm{target}}, \qquad \alpha \in [0,1], \quad (1)$$

through a forward hook fired once during prefill; decoding then proceeds autoregressively from the modified KV cache. The variants differ only in $r^{(\ell)}_{\mathrm{target}}$. MRP uses the cached audio-only residual $r^{(\ell)}_{\mathrm{P0}}$ (diagnostic). $\mathrm{SCRA}_{\mathrm{2p}}$ runs a second prefill on a stripped prompt p_{strip} (clinical block removed) and uses $r^{(\ell)}_{\mathrm{strip}}$ (two passes, no paired cache). The full SCRA method replaces the second pass with a learned estimate $r^{(\ell)}_{\mathrm{target}} = r^{(\ell)}_{\mathrm{full}} - W^{(\ell)} r^{(\ell)}_{\mathrm{full}}$, where $W^{(\ell)}$ is fit by ordinary least squares to predict $r^{(\ell)}_{\mathrm{full}} - r^{(\ell)}_{\mathrm{P0}}$ at the final-input position under 5-fold speaker-disjoint CV. Rank-r maps are obtained by SVD truncation, so a single decomposition per fold yields every $r \in \{1, 32, 128, 1536\}$.

Validating the corrector against the matched-patch ceiling. The cached MRP α-sweep is monotone at every layer; L19/L21/L24 are within 0.001 WER at α=1 (Figure 3A,D), confirming a mid-stack *band* effect rather than a single-layer effect. $\mathrm{SCRA}_{\mathrm{2p}}$ lands $\sim$0.010 WER above its cached counterpart at every layer (41–47% gap closure), so a paired P0 cache is not required to move WER substantially. The learned single-pass corrector matches the cached reference: at rank 128, all three trained layers are within 0.001 WER of MRP α=1 (Figure 3C). Per-etiology rescue at α=1 lies between each etiology's P0 and P2 brackets, with no reversal (Figure 3B,E).

Deployment performance. On the full corpus, rank-128 SCRA closes 60% of the aggregate P0–P2 gap. On Difficult Samples ($\mathrm{WER}_{\mathrm{P0}} \geq 0.1$, 58% of corpus), SCRA reaches WER 0.547 vs. 0.551 under P0 and 0.570 under P2, a small but statistically significant improvement over the audio-only baseline (paired 95% CI $[-0.0076, -0.0012]$ for $\Delta(\mathrm{SCRA} - \mathrm{P0})$). The one-pass corrector therefore preserves the full clinical-context prompt while matching or slightly improving on audio-only performance on exactly the items where additional context would be most valuable.

6. Limitations and Ethics

Our analysis is limited to a single frozen architecture, and the SAP rated subset overrepresents Parkinson's disease, so per-etiology uniformity is most reliable for the larger groups. Supervised LoRA adaptation removes the degradation entirely, suggesting the limitation is one of training exposure rather than architecture and consistent with cross-model evidence (Moure et al., 2026). Clinical labels should be treated as optional support rather than a requirement for access; hallucinated correction toward fluent speech is especially problematic in augmentative communication. Accessibility for atypical speech is unlikely to follow automatically from scaling (Hoffman et al., 2014; Jaddoh et al., 2025); targeted training or inference-time intervention is required.

7. Acknowledgment

For this work, we obtained official approval from the authors of the Speech Accessibility Project to evaluate both closed-weight and open-weight audio-language models (in particular Gemma 4) that do not violate the data redistribution regulations. We gratefully acknowledge the University of Illinois Urbana-Champaign, Beckman Institute for Advanced Science and Technology, for their permission and support in this study.

References

Alsayegh, A. and Masood, T. Zero-shot recognition of dysarthric speech using commercial automatic speech recognition and multimodal large language models, 2025. URL https://arxiv.org/abs/2512.17474.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.14997.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Geiger, A., Lu, H., Icard, T. F., and Potts, C. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Gemini Team, Google. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5484–5495, 2021.

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.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Hoffman, H. J., Li, C.-M., Losonczy, K. G., Chiu, M.-S., Lucas, J. B., and St. Louis, K. O. 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, 2014.

Jaddoh, A., Loizides, F., Alreafai, K., and Rana, O. Overcoming speech barriers: Non-verbal voice cue interaction technique for enhancing smart voice assistant accessibility for individuals with dysarthria. ACM Transactions on Accessible Computing, Vol. 18, Iss. 2, Art. 9, 2025.

Liu, A. H., Ehrenberg, A., Lo, A., Sun, C.-Y., Lample, G., Delignon, J.-M., Chandu, K. R., von Platen, P., Muddireddy, P. R., Arora, R., Gandhi, S., Subramanian, S., and Ghosh, S. Voxtral realtime, 2026. URL https://arxiv.org/abs/2602.11298.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

Moure, P., Pokel, N., Bounajma, B., Gao, Y., Boehringer, R., Cheng, L., and Liu, S.-C. When audio-language models fail to leverage multimodal context for dysarthric speech recognition, 2026. URL https://arxiv.org/abs/2605.02782.

OpenAI. Gpt-4o system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Tobin, J., Nelson, P., MacDonald, B., Heywood, R., Cave, R., Seaver, K., Desjardins, A., Jiang, P. P., and Green, J. R. 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.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Wagner, D., Baumann, I., Engert, N., Lee, S., Nöth, E., Riedhammer, K., and Bocklet, T. Personalized fine-tuning with controllable synthetic speech from llm-generated transcripts for dysarthric speech recognition, 2025. URL https://arxiv.org/abs/2505.12991.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Zheng, X., Phukon, B., Na, J., Cutrell, E., Han, K., Hasegawa-Johnson, M., Jiang, P.-P., Kuila, A., Lea, C., MacDonald, B., Mantena, G., Ravichandran, V., Sari, L., Tomanek, K., Yoo, C. D., and Zwilling, C. The interspeech 2025 speech accessibility project challenge, 2025. URL https://arxiv.org/abs/2507.22047.