

Why Alzheimer’s Speech Screening Fails to Generalize: Bridging the Deployment Gap via Cross-Corpus Evidence Anchoring

Abstract. Speech-based screening is a promising, non-invasive approach for detecting Alzheimer’s disease and related cognitive risks. However, models trained on a single domain often generalize poorly to unseen languages, tasks, or recording protocols. This paper investigates this deployment gap using a leave-one-corpus-out evaluation across four distinct datasets. Among 70 interpretable speech and language features, 59 exhibit direction conflicts between healthy controls and risk groups across corpora, with pause, silence, and speech rate showing high protocol sensitivity. Furthermore, while an XLM-R text baseline model achieves strong average performance, its Area Under the ROC Curve (AUC) drops to 0.520 on the weakest held-out domain. To address this, we propose a fusion method that integrates XLM-R text baseline scores with evidence anchors selected during training. A balanced fusion achieves a 0.785 mean speaker AUC, while anchor-heavy fusion raises the worst-case speaker AUC to 0.615. This work highlights the need to audit feature transferability and report worst-case domain robustness in cognitive speech screening.

Keywords: Alzheimer’s disease detection · Cross-domain evaluation · Cross-corpus generalization · Explainable speech processing

1 Introduction

Speech offers a low-cost, non-invasive window into cognitive decline. Language production relies on memory, attention, lexical access, discourse planning, and executive control, functions that are frequently impaired in Alzheimer’s Disease (AD) and mild cognitive impairment [24, 10]. Unlike neuroimaging or extended clinical batteries, spontaneous speech can be collected remotely, tracked longitudinally, and analyzed with both interpretable markers and machine learning models.

Two paradigms dominate the field. The first designs interpretable features such as pauses, speech rate, silence ratio, lexical diversity, and discourse cohesion. ASR-derived measures have been used to detect mild cognitive impairment from spontaneous speech [24, 29, 15], temporal and acoustic parameters offer clinically meaningful insight into dementia speech [20], and narrative-speech linguistic features distinguish AD from healthy controls in picture-description tasks [10], with lexical and discourse markers further predicting probable or future AD status [22, 9]. The second paradigm uses deep representations such as BERT-style encoders [8], XLM-R text baseline [7], and self-supervised speech models

like wav2vec 2.0 [1], which capture complex patterns that escape single clinical markers [14].

While prior work confirms that speech carries a viable screening signal, it remains unclear whether this signal stays reliable once a screening system leaves a familiar domain. A deployed system meets new picture prompts, tasks, languages, microphones, segmentation policies, and ASR error profiles, and models tuned under random splits within one corpus risk fitting a mixture of disease evidence, task artifacts, and cohort bias. We show that both handcrafted markers and deep neural models fail to generalize reliably, though in different ways. Handcrafted features can be informative in one corpus yet heavily protocol-dependent in another: pause, silence, and speech rate are often treated as universal proxies for impairment, yet their direction can invert with task design, segmentation, or language, so single-marker rules are unsafe under domain shift. Deep representations pose a complementary risk, since strong average performance can mask severe failure on specific unseen corpora, which matters in practice because users experience their own deployment setting rather than a benchmark average. We therefore treat worst-case domain robustness as a primary evaluation criterion.

To address this, we propose a deployment-oriented framework for AD speech screening under external corpus shift. The core idea is to use heterogeneous existing datasets to simulate future, unseen deployment environments during model development. In a Leave-One-Corpus-Out (LOCO) setup, the model learns from multiple source domains and is evaluated on one fully held-out corpus, mimicking transfer across new languages, tasks, or recording protocols. Our framework has three stages: it audits the cross-corpus reliability of speech and language markers, selects a compact set of robust evidence anchors from source data, and fuses these anchor scores with the XLM-R text baseline score. This shifts the goal from maximizing performance on familiar benchmarks toward ensuring that both the evidence and the predictions remain reliable in unseen deployment scenarios. Our contributions are threefold. First, an audit of 70 interpretable markers across four corpora shows that 59 exhibit direction conflicts between healthy controls and risk groups, and these conflicts are structured, with lexical and POS markers stable while silence, speech rate, and pause duration are highly protocol-sensitive. Second, we show that deep representations do not fully mitigate this distribution shift, as the XLM-R text baseline achieves strong average performance but degrades sharply on the hardest held-out domain. Third, we introduce a transfer protocol that fuses audited evidence anchors with the XLM-R text baseline baseline, where a balanced configuration yields the best mean speaker AUC and an anchor-heavy configuration substantially raises performance on the weakest domain.

2 Related Work

2.1 Speech and Language Markers for Cognitive Screening

DementiaBank and the wider TalkBank repository provide the field’s core transcribed clinical speech resources, mostly drawn from picture-description tasks

[18, 13, 26]. The Pitt Cookie corpus is a widely benchmarked English resource, while Mandarin corpora such as Chou offer important non-English validation [27, 6]. Reviews show that acoustic, prosodic, lexical, syntactic, and discourse markers all carry diagnostic value, but also report substantial variance across dataset characteristics, preprocessing, and evaluation protocols [11, 19]. Disfluency and interactional cues have further been used for spontaneous speech analysis, though these remain sensitive to transcription source and task format [21].

Interpretable markers are clinically attractive because features like pause ratios or pronoun frequencies can be visualized and clinically inspected [10]. **The critical risk is that clinical interpretability does not guarantee cross-corpus generalization.** Pause-based markers depend heavily on voice-activity-detection thresholds, microphone hardware, task, speaking time, and segmentation policy [20], and lexical markers are similarly sensitive to transcription pipeline, tokenization, and prompt design. We therefore separate two questions: whether a marker is statistically significant within a corpus, and whether its direction reliably transfers to an unseen domain.

2.2 Shared Tasks and Cross-Corpus Evaluation

Shared tasks such as ADReSS and ADReSSo standardized the field through balanced splits and unified baselines [16, 17]. This was extended to multilingual mild cognitive impairment detection by TAUkADIAL [3], while NCMMSC 2021 introduced a Chinese AD recognition evaluation with long and short speech tracks, serving as a rigorous test for Mandarin screening [28, 5]. Each corpus nonetheless remains tied to its own collection protocol, so robust evaluation should target held-out environments rather than in-distribution averages [12]. Since a deployed system inevitably meets unfamiliar tasks and cohorts, we adopt a rigorous LOCO framework rather than random splitting.

2.3 Deep Representations

Pre-trained text encoders such as BERT [8] and XLM-R [7] offer strong linguistic modeling, with XLM-R particularly advantageous in cross-lingual settings, and have also been compared directly against traditional features for AD detection [2]. On the acoustic side, wav2vec 2.0 [1] and wavLM [4] provide effective self-supervised speech representations, and multimodal ADReSSo systems show that acoustic, lexical, disfluency, and pause features yield complementary gains [25], echoing NCMMSC results where raw-waveform and ASR-driven deep architectures achieve competitive results on Chinese data [5, 23]. Despite this success, deep models face a severe deployment risk, since they easily overfit to corpus identity, transcription artifacts, channel conditions, or prompt structure [2]. Rather than relying on the deep model alone, we treat XLM-R as an audited text stream and counteract its risk by fusing it with robust, cross-corpus evidence anchors.

3 Data and Evaluation Protocol

3.1 Corpora

We use four heterogeneous corpora spanning Mandarin and English, picture-description tasks, and connected clinical speech:

- **NCMMSC**: A Mandarin dataset from the official Alzheimer’s Disease Recognition Evaluation challenge [28], using the *long audio* track of multi-institutional spontaneous speech elicited via clinical picture-description tasks.
- **Pitt Cookie**: From the DementiaBank Pitt Corpus, using spontaneous speech elicited via the “Cookie Theft” picture-description task [26].
- **TAUKADIAL Mandarin**: A subset of the multilingual mild cognitive impairment challenge, capturing connected speech from semi-structured clinical assessments [3].
- **Mandarin Chou**: A DementiaBank Mandarin corpus tracking language profiles of healthy controls and participants with amnesic mild cognitive impairment [27, 6].

Speaker IDs are prefixed by corpus before evaluation, and for speakers with multiple samples, sample risk scores are averaged before computing speaker AUC. We standardize the task into a binary setting, Healthy Control (HC) versus Cognitive Risk, collapsing AD, MCI, and other cognitive-impairment labels into a single Risk class. Table 1 summarizes the sample and speaker distributions.

Table 1. Corpora used for leave-one-corpus-out evaluation. HC and risk are sample-level counts.

Corpus	Language	Task	Samples	Speakers	HC	Risk
NCMMSC	Mandarin	Picture	246	116	90	156
Pitt Cookie	English	Picture	547	291	241	306
TAUKADIAL Mandarin	Mandarin	Connected	507	169	222	285
Mandarin Chou	Mandarin	Picture	261	87	120	141
Total			1561	663	673	888

3.2 Features and Models

The interpretable feature pool comprises 70 high-level speech and language features spanning temporal disfluency, pause dynamics, speech rate, acoustic intensity, lexical diversity, POS distributions, lexical concentration, repetition markers, information structure, and discourse cohesion, and a compact subset of 24 canonical features forms the traditional baseline marker set.

For deep baselines we use frozen pre-trained encoders with a lightweight downstream classifier: 768-dimensional XLM-R text baseline embeddings extracted from transcripts, and 1024-dimensional representations from the intermediate (12th) and final (24th) layers of a pre-trained wav2vec 2.0 model. No

encoder is fine-tuned, so the comparison isolates domain robustness from representation drift.

All continuous features are preprocessed with parameters fitted strictly on the training partition, including winsorization at the 1st and 99th percentiles, median imputation, and z -score normalization. The downstream classifier is a weighted L_2 -regularized logistic regression ($\lambda = 1.0$), optimized via L-BFGS-B, with sample weights combining inverse class frequency and inverse per-speaker sample count. Decision thresholds are calibrated exclusively on training speakers, while all reported speaker AUC scores remain threshold-free.

3.3 Leave-One-Corpus-Out Protocol

Our core evaluation is a Leave-One-Corpus-Out (LOCO) cross-validation, where each fold holds out one corpus as an unseen external evaluation set and aggregates the remaining three as the training partition. All pipeline components, including feature selection, scaling, model fitting, and threshold calibration, are computed exclusively on the training corpora to prevent target-domain leakage. The primary metric is speaker-level speaker AUC, and we report both the macro-averaged mean speaker AUC across held-out folds and the worst-case domain speaker AUC.

To confirm that downstream classifiers capture genuine, transferable indicators rather than dataset artifacts, we implement four control experiments:

- **Label Shuffling:** training labels are randomly permuted before feature selection and model fitting, breaking any link between speech and clinical status.
- **Metadata and Length Controls:** models restricted to non-speech metadata (e.g., age, gender) or trivial recording parameters (transcript length, audio duration).
- **Corpus Bias Control:** a baseline using only categorical corpus-identity indicators as input.
- **Corpus Identity Probe:** a probing classifier trained to predict source corpus from the interpretable feature space, quantifying distributional distinctiveness among the four datasets.

4 Evidence Transferability Audit

We first audit whether interpretable speech and language markers maintain a consistent directional relationship between HC and Risk groups across corpora and task families.

4.1 Audit Definition

Rather than assessing standalone significance, this audit tests the domain transferability of feature directions. For marker m , let $d_m^{(c)}$ denote its within-corpus

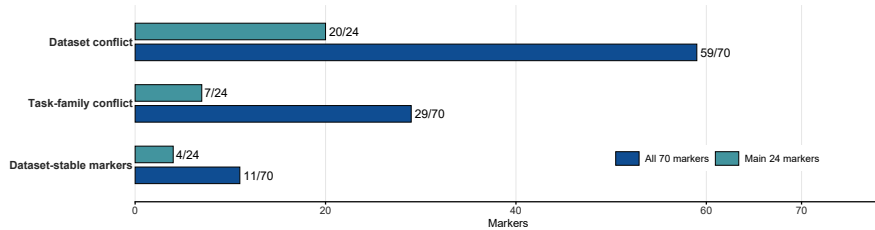


Fig. 1. Direction conflict prevalence across marker scopes.

standardized effect size (Cohen’s d) between the Risk and HC groups in corpus c :

$$d_m^{(c)} = \frac{\mu_{m,1}^{(c)} - \mu_{m,0}^{(c)}}{\sigma_m^{(c)}}, \quad (1)$$

where $\mu_{m,1}^{(c)}$ and $\mu_{m,0}^{(c)}$ are the Risk and HC feature means and $\sigma_m^{(c)}$ is the pooled within-corpus standard deviation. A positive value indicates elevated marker expression in the Risk group. We define a cross-dataset direction conflict as

$$\gamma_m^{\text{data}} = \mathbf{1} \left[\min_{c \in \mathcal{C}} d_m^{(c)} < 0 \wedge \max_{c \in \mathcal{C}} d_m^{(c)} > 0 \right]. \quad (2)$$

γ_m^{data} equals 1 when a marker shows opposing directional effects across corpora, and task-family conflicts are computed analogously over task categories. A marker is deemed non-transferable if its direction reverses across deployment settings, regardless of its localized magnitude $|d_m^{(c)}|$.

4.2 Marker Direction Conflicts

Traditional interpretable markers rarely generalize as rigid, single-feature diagnostic rules. Among all 70 markers, 59 exhibit clear direction conflicts between HC and Risk across the four datasets, and within the 24 canonical core markers, 20 suffer cross-dataset reversals. Task-family conflicts are less frequent but still substantial: 29 of 70 markers and 7 of the 24 core markers reverse direction between picture-description tasks and connected speech.

Figure 1 quantifies these conflict distributions across marker scopes. This high conflict prevalence is a critical methodological vulnerability, since most clinical speech markers invert behavior under domain shift, which makes universal single-marker heuristics unreliable for general deployment. This directly challenges the common narrative that cognitive-decline speech is universally slower, more paused, less lexically rich, or more repetitive. While such patterns hold within specific cohorts or homogeneous tasks, our audit shows they should not be treated as invariant rules without protocol cross-verification.

Concrete features confirm this pattern. Lexical and structural markers such as pronoun ratio and function word ratio consistently elevate in Risk speech across all four corpora, while content-to-function word ratio and timestamp coverage ratio remain uniformly lower, a cross-corpus directional stability that qualifies

Because features diverge directionally across domains, directly pooling all 70 markers can cause a downstream classifier to overfit to training-distribution artifacts. To mitigate this, we pre-specify a compact, interpretable panel of 24 core indicators, avoiding any feature engineering driven by held-out performance. Features are retained only if they are computable across all four corpora under a uniform speaker-level aggregation protocol, which keeps the comparison independent of corpus-specific annotation granularity. The panel is organized into nine evidence families: word class, temporal fluency, discourse structure, information content, speech quantity, lexical diversity, alignment quality, disfluency, and repetition, which separates linguistic content from acoustic timing and transcript-alignment artifacts.

Figure 3 reports the domain transferability of each core indicator evaluated in isolation under LOCO, showing the macro-averaged mean speaker AUC with 95% confidence intervals on the left and the worst-case domain speaker AUC on the right. *Weak* denotes a minor or inconsistent single-indicator transfer signal, while *Shortcut* flags a marker that may exploit corpus-specific bias or fail catastrophically in the hardest fold. The top single indicator reaches only 0.658 mean speaker AUC, and many features yield worst-case speaker AUC below chance (0.50), reinforcing that the core 24 indicators should serve as auditable, complementary evidence channels for selection and fusion rather than standalone classifiers.

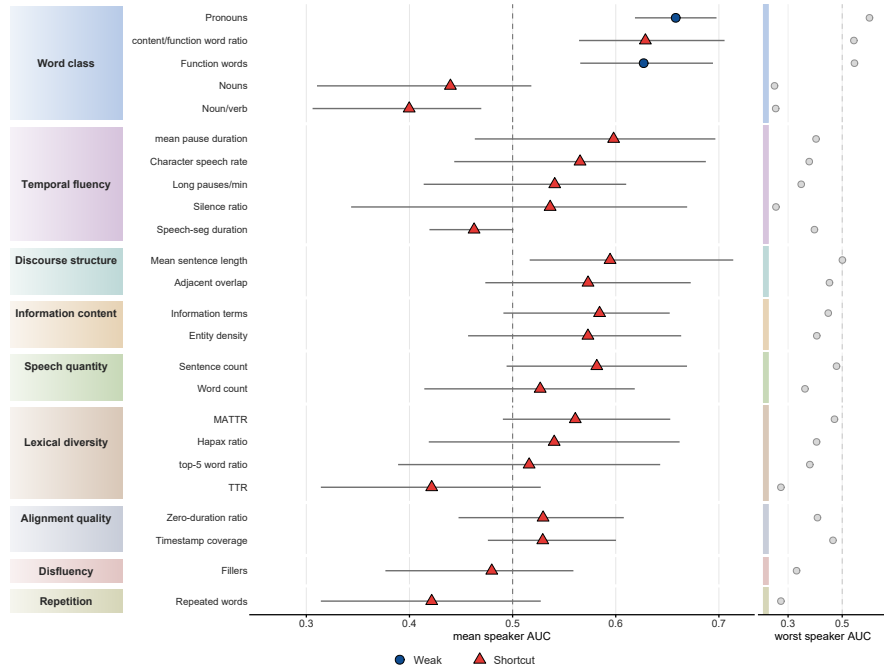


Fig. 3. Single-indicator external transfer of the 24 core indicators.

5 Deep Representations Under External Shift

The marker audit raises a further question: **whether deep representations avoid these transfer vulnerabilities?** We evaluate three frozen pre-trained representations under the identical LOCO protocol, the text-based XLM-R text baseline and the acoustic wav2vec 2.0 features from the intermediate (12th) and final (24th) layers, fitting only the downstream linear classifier on the designated source corpora in each fold.

Table 2. Deep-only baselines under leave one corpus out evaluation.

Representation	NCMMSC	Pitt	TAUKADIAL	Chou	Mean	Worst	Range
XLM-R text baseline	0.520	0.739		0.850	0.968	0.769	0.520 0.447
wav2vec 2.0 L12	0.462	0.502		0.835	0.955	0.688	0.462 0.493
wav2vec 2.0 L24	0.454	0.574		0.800	0.969	0.699	0.454 0.515

As Table 2 shows, pre-trained representations capture rich contextual information but do not inherently mitigate external domain shift. XLM-R text baseline is the strongest standalone deep baseline, reaching a mean speaker AUC of 0.769, yet its fold scores range from a near-chance 0.520 on NCMMSC to a near-perfect 0.968 on Mandarin Chou. This pattern is sharper for the acoustic baselines, since both wav2vec 2.0 layers are strong on Mandarin Chou and TAUADIAL Mandarin but collapse near chance on NCMMSC and remain weak on Pitt Cookie. Moving from the 12th to the 24th layer improves Pitt Cookie but degrades NCMMSC, showing that deeper layers alone do not resolve localized failure.

A high macro-averaged metric can therefore be driven by a few easily generalizable domains while masking catastrophic collapse elsewhere. The LOCO protocol acts as a multi-domain stress test by treating each corpus as an unseen deployment environment in turn. NCMMSC is the hardest fold here, but the broader issue is a systemic lack of worst-domain robustness under domain shift. Given its strong performance among the deep-only configurations, we select XLM-R text baseline as the primary deep network in our method, while the variances in Table 2 justify integrating an audited, interpretable evidence-selection framework.

6 Our Proposed Transfer Protocol

The audit above shows that domain-transferable evidence is highly sparse and corpus-dependent. Rather than treating the full feature repository as a static input matrix, we use the audited indicators to define a constrained anchor search space, dynamically extracting robust anchors from the active training partition within each LOCO fold and fusing the resulting anchor score with the XLM-R text baseline semantic representation.

6.1 Evidence Anchor Selection

Evidence Anchor Selection identifies a compact, reliable feature subset using exclusively the training partition of the active LOCO fold. To limit overfit-

Table 3. Traditional-indicator selection under leave one corpus out evaluation. All values are speaker AUC.

	P-value top four	Random top four	Evidence-ranked top four	All 24 core	All 70 indicators	Source-AUC anchors
k	4	4	4	24	70	4
Mean speaker AUC	0.506	0.542	0.555	0.521	0.538	0.629
Worst speaker AUC	0.363	0.195	0.520	0.418	0.395	0.588

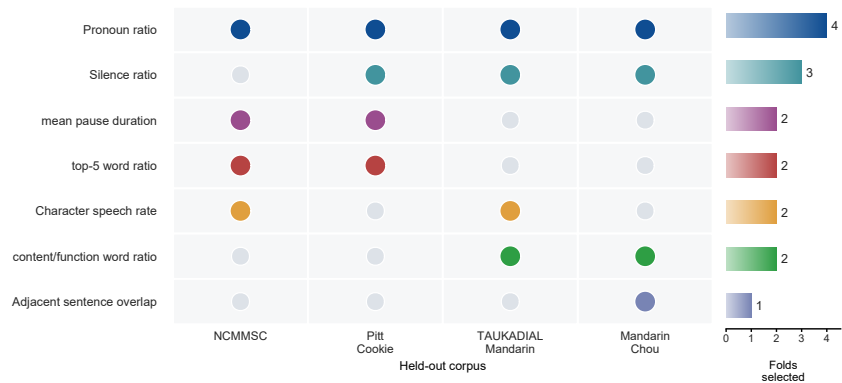


Fig. 4. Source-corpus anchor choices for each held out corpus.

ting, we restrict the candidate space to the 24 core indicators, rank them by macro-averaged training-corpus speaker AUC utility, and select the top $k = 4$ as domain anchors, then fit a weighted L_2 -regularized logistic regression solely on these anchors. The held-out corpus is never observed during anchor selection, regularization tuning, or threshold calibration.

Table 3 compares this against alternative selection strategies. The source-AUC anchor selector achieves a macro-averaged mean speaker AUC of 0.629 and worst-case speaker AUC of 0.588, which is substantially more stable than the unselected global configurations: all 24 indicators reach only 0.521 mean speaker AUC (0.418 worst), and all 70 markers reach 0.538 mean speaker AUC (0.395 worst). This confirms that blindly accumulating interpretable features introduces cross-domain noise, while targeted anchor selection safeguards generalization.

Anchor composition varies by fold. For the NCMMSC fold, the training corpora select mean pause duration, pronoun ratio, top-5 word ratio, and character speech rate. For the Pitt Cookie fold, they select silence ratio, pronoun ratio, mean pause duration, and top-5 word ratio. For TAU KADIAL Mandarin, they select silence ratio, pronoun ratio, character speech rate, and content-to-function word ratio. For Mandarin Chou, they select silence ratio, pronoun ratio, content-to-function word ratio, and adjacent sentence overlap. The pronoun ratio is the only indicator selected in all four folds.

Figure 4 illustrates this cross-fold variability. The pronoun ratio serves as a persistent cross-domain invariant, while the remaining anchor channels are selected dynamically based on the source partition mixture, confirming that

domain-transferable clinical evidence, though highly compact, has a composition that depends on the target deployment environment.

6.2 Anchor Fusion

The anchor-selection pipeline yields a compact, interpretable speaker-level score, while XLM-R text baseline captures dense, high-capacity multilingual semantic representations. We combine these under the same LOCO protocol. Let s_{anchor} and s_{xlmr} denote the speaker-level anchor and XLM-R text baseline scores, and compute

$$s_{\alpha} = \alpha s_{\text{anchor}} + (1 - \alpha) s_{\text{xlmr}}, \quad (3)$$

where α is fixed before evaluating the held-out corpus and applied to speaker-averaged risk scores. We report $\alpha = 0.50$ as the balanced point and $\alpha = 0.75$ as the anchor-heavy point. Figure 5 shows the mean-versus-weakest-corpus tradeoff across score normalizations: the balanced setting optimizes overall mean speaker AUC, and the anchor-heavy setting favors the hardest corpus.

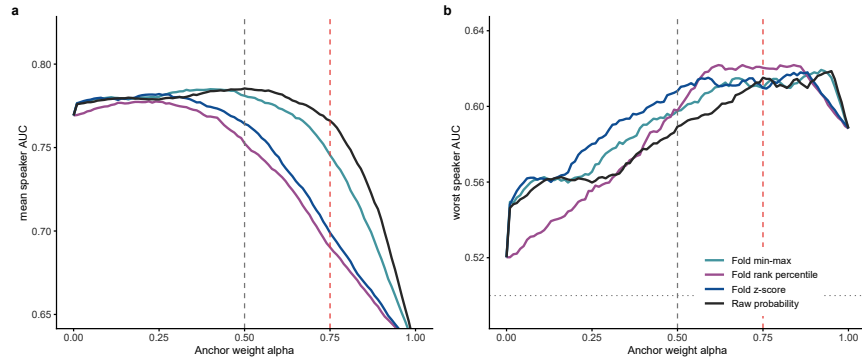


Fig. 5. Fusion weight sensitivity across score normalizations.

Figure 6 presents the primary cross-domain transfer evaluation. The anchor-only model yields lower macro-averaged performance than XLM-R text baseline alone but mitigates worst-case domain collapse, raising the lowest fold score from 0.520 to 0.588. Combining both modalities, our balanced fusion ($\alpha = 0.50$) achieves the top overall mean speaker AUC of 0.785, while the anchor-heavy regime ($\alpha = 0.75$) secures a worst-case domain speaker AUC of 0.615, a 0.095 absolute gain over the deep-only baseline in the hardest deployment environment. This validates the integration of robust clinical evidence anchors.

Figure 6 also shows the corpus-level speaker AUC matrix behind this tradeoff. XLM-R text baseline alone gives the highest average performance, while the anchors give a more resilient floor: on the near-saturated Mandarin Chou fold, adding anchors slightly hurts performance, but on NCMMSC the anchor model is markedly more robust and lets the anchor-heavy setting substantially recover performance. TAU KADIAL Mandarin benefits most from balanced fusion, while Pitt Cookie stays close to the deep baseline.

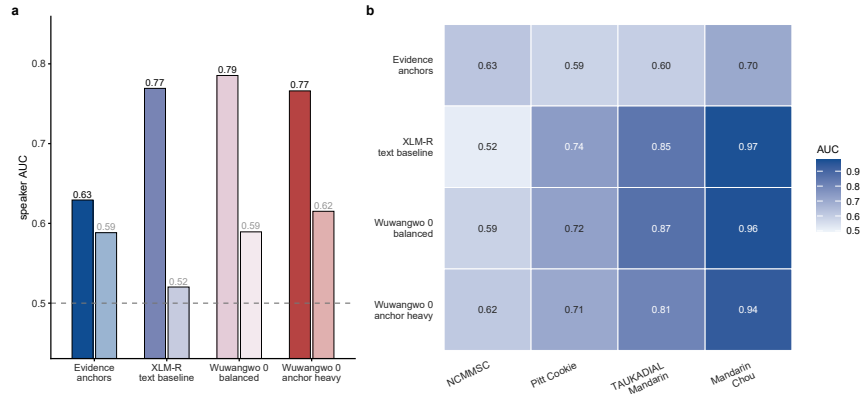


Fig. 6. Model comparison across held out corpora.

Table 4. Negative controls and corpus separability checks under leave one corpus out evaluation. Values are mean speaker AUC with 95 percent intervals.

Control	Mean speaker-level AUC	95% CI
Label shuffle	0.490	0.465 to 0.514
Metadata only	0.469	0.416 to 0.500
Corpus only	0.500	0.500 to 0.500
Length only	0.580	0.458 to 0.694
Corpus identity probe	0.899	0.864 to 0.934

We treat the weakest fold as an out-of-distribution stress test rather than a privileged target domain, and here NCMMSC is the bottleneck. A 2,000-resample speaker-level bootstrap gives baseline XLM-R text baseline an speaker AUC of 0.520 (95% CI [0.407, 0.627]), versus 0.615 (95% CI [0.509, 0.715]) for anchor-heavy fusion. The paired bootstrap CI for this 0.095 improvement is $[-0.004, 0.195]$, confirming that heavier weight on the audited anchor stream raises the performance floor in severely shifted settings.

The two streams are also empirically distinct: the mean absolute Spearman correlation between anchor and XLM-R text baseline scores across folds is low ($\rho = 0.235$, orthogonality index 0.765), and the audit phase pre-specifies which evidence channels are active in each fold before fusion with the deep stream.

Table 4 confirms these gains are not artifacts. Models trained on shuffled labels, metadata alone, or corpus-identity markers default to near-random chance, while transcript length carries negligible signal. The corpus identity probe, however, reaches high accuracy, confirming that the four datasets remain highly separable in marker space, which is why LOCO evaluation is required.

7 Discussion and Conclusion

This study shows that real-world AD speech screening is fundamentally an external domain-transfer problem, where overlapping task, linguistic, and environ-

mental shifts cause traditional acoustic-lexical markers to reverse direction. Deep representations like XLM-R text baseline deliver high average performance but remain vulnerable to severe localized collapse on unseen domains. Our protocol addresses this gap through systematic evidence auditing and dynamic anchor selection, giving a controllable tradeoff between mean performance ($AUC = 0.785$ under balanced fusion) and weak-domain robustness ($AUC = 0.615$ under an anchor-heavy regime).

Our evaluation is limited to four English and Mandarin corpora, a simple fusion architecture, and fixed deep representations. Even so, the results support a practical conclusion: moving speech-based cognitive screening into deployable clinical workflows requires evidence auditing and worst-case domain testing, not just stronger deep classifiers.

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. In: *Advances in Neural Information Processing Systems* 33. pp. 12449–12460 (2020)
2. Balagopalan, A., Eyre, B., Rudzicz, F., Novikova, J.: To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer’s disease detection. In: *Proc. Interspeech 2020*. pp. 2167–2171 (2020). <https://doi.org/10.21437/Interspeech.2020-2557>
3. Barrera-Altuna, B., Lee, D., Zarnaz, Z., Han, J., Kim, S.: The Interspeech 2024 TAUADIAL challenge: Multilingual mild cognitive impairment detection with multimodal approach. In: *Proceedings of Interspeech 2024*. pp. 967–971 (2024). <https://doi.org/10.21437/Interspeech.2024-1352>
4. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* **16**(6), 1505–1518 (2022)
5. Chen, X.C., Zhang, W.Q., Ma, Y.: Raw waveform-based end-to-end Alzheimer’s disease detection method. *Acta Electronica Sinica* **51**(12), 3582–3590 (2023). <https://doi.org/10.12263/DZXB.20220162>
6. Chou, C.J., Chang, C.T., Chang, Y.N., Lee, C.Y., Chuang, Y.F., Chiu, Y.L., Liang, W.L., Fan, Y.M., Liu, Y.C.: Screening for early Alzheimer’s disease: Enhancing diagnosis with linguistic features and biomarkers. *Front. Aging Neurosci.* **16**, 1451326 (2024). <https://doi.org/10.3389/fnagi.2024.1451326>
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pp. 8440–8451 (2020). <https://doi.org/10.18653/v1/2020.acl-main.747>
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 4171–4186 (2019). <https://doi.org/10.18653/v1/N19-1423>

9. Eyigöz, E., Mathur, S., Santamaria, M., Cecchi, G., Naylor, M.: Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine* **28**, 100583 (2020). <https://doi.org/10.1016/j.eclinm.2020.100583>
10. Fraser, K.C., Meltzer, J.A., Rudzicz, F.: Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimer's Dis.* **49**(2), 407–422 (2016). <https://doi.org/10.3233/JAD-150520>
11. de la Fuente Garcia, S., Ritchie, C.W., Luz, S.: Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease* **78**(4), 1547–1574 (2020). <https://doi.org/10.3233/JAD-200888>
12. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balasubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S.M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., Liang, P.: WILDS: A benchmark of in-the-wild distribution shifts. In: *Proc. 38th Int. Conf. Mach. Learn. Proc. Mach. Learn. Res.*, vol. 139, pp. 5637–5664 (2021), <https://proceedings.mlr.press/v139/koh21a.html>
13. Lanzi, A.M., Saylor, A.K., Fromm, D., Liu, H., MacWhinney, B., Cohen, M.L.: DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology* **32**(2), 426–438 (2023). https://doi.org/10.1044/2022_AJSLP-22-00281
14. Li, Y., Zhang, Z., Han, J., Bell, P., Lai, C.: Semi-supervised cognitive state classification from speech with multi-view pseudo-labeling. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2025)
15. Liu, Y.L., Li, Y., Feng, R., Chen, J., Wang, Y., Chen, Y.A., Ding, N., Yuan, J., Ling, Z.H.: Beyond manual transcripts: Exploring the potential of automatic speech recognition errors in improving alzheimer's disease detection. *Journal of Biomedical Informatics* p. 104968 (2025)
16. Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B.: Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In: *Proceedings of Interspeech 2020*. pp. 2172–2176 (2020). <https://doi.org/10.21437/Interspeech.2020-2571>
17. Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B.: Detecting cognitive decline using speech only: The ADReSSo challenge. In: *Proceedings of Interspeech 2021*. pp. 3780–3784 (2021). <https://doi.org/10.21437/Interspeech.2021-1220>
18. MacWhinney, B., Fromm, D., Forbes, M., Holland, A.: AphasiaBank: Methods for studying discourse. *Aphasiology* **25**(11), 1286–1307 (2011). <https://doi.org/10.1080/02687038.2011.589893>
19. Martínez-Nicolás, I., Llorente, T.E., Martínez-Sánchez, F., Meilán, J.J.G.: Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: A systematic review article. *Front. Psychol.* **12**, 620251 (2021). <https://doi.org/10.3389/fpsyg.2021.620251>
20. Meilán, J.J.G., Martínez-Sánchez, F., Carro, J., López, D.E., Millian-Morell, L., Arana, J.M.: Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dement. Geriatr. Cogn. Disord.* **37**(5–6), 327–334 (2014). <https://doi.org/10.1159/000356726>
21. Nasreen, S., Rohanian, M., Hough, J., Purver, M.: Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Front. Comput. Sci.* **3**, 640669 (2021). <https://doi.org/10.3389/fcomp.2021.640669>

22. Orimaye, S.O., Wong, J.S.M., Golden, K.J., Wong, C.P., Soyiri, I.N.: Predicting probable Alzheimer’s disease using linguistic deficits and biomarkers. *BMC Bioinformatics* **18**(1), 34 (2017). <https://doi.org/10.1186/s12859-016-1456-0>
23. Qin, Y., Liu, W., Peng, Z., Ng, S.I., Li, J., Hu, H., Lee, T.: Exploiting pre-trained ASR models for Alzheimer’s disease recognition through spontaneous speech. *arXiv:2110.01493* (2021), posted 4 October 2021
24. Roark, B., Mitchell, M., Hosom, J.P., Hollingshead, K., Kaye, J.: Spoken language derived measures for detecting mild cognitive impairment. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2081–2090 (2011). <https://doi.org/10.1109/TASL.2011.2112351>
25. Rohanian, M., Hough, J., Purver, M.: Alzheimer’s dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. In: *Proc. Interspeech 2021*. pp. 3820–3824 (2021). <https://doi.org/10.21437/Interspeech.2021-1633>
26. TalkBank: DementiaBank English Pitt Corpus. <https://talkbank.org/dementia/access/English/Pitt.html> (2026), DOI: 10.21415/CQCW-1F92. Last accessed 3 July 2026
27. TalkBank: DementiaBank Mandarin Chou Corpus. <https://talkbank.org/dementia/access/Mandarin/Chou.html> (2026), DOI: 10.21415/X3Q2-GG37. Last accessed 3 July 2026
28. THU Speech and Audio Technology Lab: AD2021: Alzheimer’s Disease Recognition Evaluation 2021. <https://github.com/THUatlab/AD2021> (2021), last accessed 3 July 2026
29. Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., Pákási, M., Kálmán, J.: A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* **15**(2), 130–138 (2018). <https://doi.org/10.2174/1567205014666171121114930>