ARTICLE | Article in Press

Simultaneous multi-task assessment of five neurological language disorders via cross-corpus multimodal learning

Byounghwa Lee · Hwa Jeon Song · Byung Ok Kang

Received: 1 July 2026 / Accepted: 13 October 2026
© The Author(s) 2026

Abstract

Automated analysis of spontaneous speech may offer a scalable, objective complement to clinical assessment of neurological language disorders, yet existing approaches are largely corpus- and disease-specific. We propose a unified model that simultaneously classifies five neurological language disorders—Aphasia, Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Traumatic Brain Injury (TBI), and Right Hemisphere Damage (RHD)—and a healthy Control class across five heterogeneous TalkBank corpora (1,638 samples). To our knowledge, this is the first model to combine multimodal input (text, speech, and tabular metadata), multi-corpus training, and multi-task output for this many disorders, fusing modalities through a Q-Former cross-attention module that feeds a shared representation to a classifier and three auxiliary clinical heads (WAB-AQ severity regression, disease-similarity retrieval, and longitudinal progression). Under five-seed evaluation the model attains test accuracy **90.98**% and Macro-F1 **86.19**%, outperforming all baselines, and the auxiliary heads estimate aphasia severity within the WAB-AQ's measurement error, retrieve same-disorder patients, and track recovery. Ablation identifies Q-Former fusion and class-frequency-weighted sampling as the principal performance drivers, and an inter-disorder correlation analysis explains the dominant MCI–Control confusion, recovers an etiology-based disorder grouping, and finds paralinguistic features among the top discriminators. Leave-One-Corpus-Out evaluation shows that zero-shot generalization to unseen disease classes remains fundamentally challenging.

Keywords Neurological language disorders · Multimodal learning · Spontaneous speech analysis · Q-Former · Cross-corpus generalization · Clinical NLP

Introduction

Neurological conditions that impair language—including post-stroke aphasia, Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Traumatic Brain Injury (TBI), and Right Hemisphere Damage (RHD)—collectively affect hundreds of millions of individuals worldwide[1]. These disorders manifest as characteristic disruptions in spontaneous speech and language production: reduced vocabulary diversity and increased disfluency

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

https://doi.org/10.1038/s41598-026-76495-9

Published online: 19 October 2026

scientific reports

in aphasia[2], semantic paraphasia and circumlocution in AD[3], impaired discourse coherence in TBI[4], and subtle pragmatic deficits in RHD[5]. These linguistic changes can precede or accompany structural neurological changes detectable by imaging, making speech and language biomarkers attractive for early screening.

Formal clinical assessment instruments such as the Western Aphasia Battery (WAB), Montreal Cognitive Assessment (MoCA), and picture description protocols (e.g., the Cookie Theft task) provide standardized measurements but require specialist administration, are time-consuming, and are inaccessible to many patients due to cost and availability. Natural Language Processing (NLP) and speech processing techniques offer a potential pathway toward scalable, objective, and remotely administrable screening tools, although any such tool requires prospective clinical validation before deployment. However, the field has developed largely in isolation: models targeting AD detection from DementiaBank/ADReSSo speech[6] do not transfer to aphasia classification from AphasiaBank[2], and vice versa. Existing work furthermore typically exploits a single modality (either lexicosyntactic features from text or acoustic features from audio), missing complementary information available across modalities.

We address these limitations with a unified multimodal model trained jointly across five TalkBank-derived corpora spanning five neurological disease categories alongside a healthy Control class. The model architecture draws on three complementary data streams: (i) high-dimensional contextual text representations from frozen RoBERTa-large[7], capturing semantic and syntactic patterns associated with each disorder; (ii) acoustic representations from frozen wav2vec2-large[8], encoding paralinguistic cues such as dysarthria, disfluency, and prosodic abnormalities; and (iii) tabular clinical metadata (age, sex, functional assessment availability), providing demographic context. These modality streams are integrated through a Q-Former cross-attention module[9], wherein eight learnable query tokens attend over all modalities, enabling the model to selectively weight evidence from each data stream. The resulting shared representation simultaneously feeds the classification head and three auxiliary task heads (aphasia severity regression, contrastive similarity retrieval, and longitudinal progression risk), enabling concurrent optimization of disease classification and clinically relevant auxiliary objectives.

We emphasize that the individual components of our architecture—frozen pretrained encoders, Q-Former cross-attention, class-frequency-weighted sampling, and auxiliary-task learning—are established techniques. The contribution of this work lies not in a new modeling primitive but in their integration and systematic evaluation across five heterogeneous clinical corpora and five disorders, a setting that, to our knowledge, no prior unified model has addressed.

This paper makes the following contributions:

1. To the best of our knowledge, the first unified model for the simultaneous assessment of five neurological language disorders, combining three modalities (*multimodal*), five heterogeneous corpora (*multi-corpus*), and a primary classifier with three auxiliary clinical heads (*multi-task*) in a single end-to-end architecture; the closest prior effort, NeuroScreen[10], is text-only and single-task.
2. A cross-corpus six-class classifier (five neurological language disorders plus healthy Control) trained on 1,638 samples from five heterogeneous corpora, achieving test accuracy 90.98% and Macro-F1 86.19%, together with three auxiliary heads sharing the same representation that estimate aphasia severity (WAB-AQ, within the instrument's own measurement error), retrieve same-disorder patients, and track longitudinal recovery.
3. A comprehensive evaluation framework—eight ablation variants, Leave-One-Corpus-Out cross-corpus generalization, corpus-confound analyses, and paired statistical significance tests—supporting reproducibility and methodological rigor.
4. A systematic ablation study identifying Q-Former cross-attention and class-frequency-weighted sampling as the principal drivers of performance (each significant by McNemar and bootstrap tests, see Results), and showing that contextual RoBERTa-large embeddings substantially outperform hand-crafted linguistic features.
5. An inter-disorder correlation analysis that explains the dominant MCI–Control confusion (near-identical raw-feature profiles, separated only by subtle spectral-variability cues), recovers a clinically coherent

etiology-based disorder grouping ({Control, MCI, AD} vs. {Aphasia, TBI, RHD}), and identifies paralinguistic features among the top discriminators—supporting the multimodal design.

The remainder of this paper is organized as follows. The Related Work section reviews prior work. The Method section describes the dataset, preprocessing pipeline, and model architecture. The Experimental Design section presents the experimental design. The Results section reports the full five-seed experimental results. The Discussion section interprets the findings and discusses limitations. The Conclusion section concludes with directions for future work.

Related work

Linguistic biomarkers in neurological disorders

Spontaneous speech is a rich source of non-invasive biomarkers for neurological conditions. Fraser et al.[3] discriminated Alzheimer's Disease from controls with 81.9% accuracy using lexicosyntactic features of Cookie Theft descriptions, and later work showed that BERT-based embeddings further improve AD detection[11]. The informative cues differ by disorder: discourse coherence for TBI[4], and subtle lexical-retrieval changes for MCI, whose speech is otherwise close to that of healthy controls[12]. This heterogeneity motivates a single unified model rather than a family of disorder-specific classifiers.

Most related to our work, NeuroScreen[10] is a text-only classifier of six conditions (left-hemisphere aphasia, dementia, MCI, TBI, RHD, and controls) built from 291 hand-crafted linguistic biomarkers over 1,394 TalkBank participants. We extend this direction by adding acoustic and tabular modalities and contextual RoBERTa-large embeddings fused via Q-Former cross-attention; a like-for-like numerical comparison is precluded by NeuroScreen's text-only features, random 80/20 split, and different corpus composition. More broadly, work that jointly models *multiple* co-occurring cognitive conditions remains scarce: Ehghaghi et al.[13] are among the few, separating depression from dementia by clustering acoustic and linguistic features, and note that such multi-disorder discrimination is rare despite its clinical importance. This gap motivates our unified five-disorder formulation.

Pre-trained speech and language models

Contextual language models—BERT[14], RoBERTa[7], and clinical variants[15,16]—provide rich token representations. A systematic search over frozen DistilBERT, BERT-base, RoBERTa-base, and RoBERTa-large identifies RoBERTa-large as the strongest text encoder here, its 1024-d embeddings aligning with wav2vec2-large for fusion. For acoustics, wav2vec2[8] learns self-supervised representations of phonetic and prosodic cues that complement text. We keep both encoders frozen and train only a lightweight head[17]: with 1,146 training samples against 355M RoBERTa-large parameters, full fine-tuning is unstable[18] and would overfit corpus-specific artifacts, whereas frozen features act as implicit regularization—a common choice in low-resource clinical NLP[16,19].

Multimodal fusion and cross-attention

The Q-Former, introduced in BLIP-2[9], is a lightweight "querying transformer" whose learnable query tokens extract task-relevant information from a frozen encoder via cross-attention. We adapt it to the clinical domain, letting the queries attend over the text, audio, and tabular streams. Among alternative fusion strategies—simple concatenation[20], attention pooling[21], and cross-modal transformers[22]—our ablation (A2) isolates the contribution of Q-Former cross-attention against direct fusion under matched settings.

Beyond speech and text, a complementary line of work fuses language-model representations with neurophysiological and behavioral signals such as EEG and eye-tracking to decode reading comprehension and word-level

cognitive states[23–25]. These studies corroborate the premise that discriminative information about language processing is distributed across modalities, although—unlike our clinically oriented setting—they target healthy readers in controlled reading tasks.

Cross-corpus generalization in clinical NLP

Models trained on one corpus often fail to transfer to others because of differences in recording, task, transcription, and demographics[6]; those that lean on transcription artifacts (e.g., AphasiaBank CHAT disfluency markers) generalize especially poorly[11]. Leave-One-Corpus-Out (LOCO) evaluation quantifies this by training on all corpora but one and testing on the held-out corpus, as used in cross-corpus emotion recognition[26] and clinical speech[6]. We extend LOCO to the six-class setting, where excluding a single-disease corpus (e.g., Delaware/MCI) creates a zero-shot transfer scenario.

Method

Dataset

We compile data from five TalkBank-affiliated corpora, each associated with one or two neurological disease categories. Table 1 summarizes the corpus statistics. AphasiaBank[2] provides Aphasia and Control samples using the PWA (Person With Aphasia) discourse protocol. ADReSSo[6] provides Alzheimer's Disease and Control speech recordings. Coelho[27] provides TBI and Control narrative samples. RHD-English[28] provides Right Hemisphere Damage and Control discourse samples from RHDBank. Delaware[29] provides Mild Cognitive Impairment and Control samples from DementiaBank's Delaware corpus.

Elicitation tasks

The corpora rely on partially overlapping discourse protocols. AphasiaBank comprises personal narratives (stroke story and coping, important event), picture descriptions (Broken Window, Refused Umbrella, Cat Rescue), the Cinderella narrative, and a procedural task (peanut-butter-and-jelly sandwich)[2]; RHDBank shares the stroke story, Cat Rescue, Cinderella, and sandwich tasks and adds Cookie Theft, a first-encounter conversation, and question production[28]; Delaware includes Cookie Theft, Cat Rescue, a Norman Rockwell picture (*Going and Coming*), Cinderella, the sandwich task, and a personal narrative[29]; ADReSSo consists solely of Cookie Theft descriptions[6]. Every corpus except one thus shares its central tasks with at least two others, so task content provides no clean corpus signature. The exception is Coelho, whose tasks are shared by no other corpus (*The Bear and the Fly* retelling, *The Runaway* story generation, and conversation)[27].

The dataset is heavily imbalanced: Control comprises 49% of samples while RHD comprises only 2.3%. Per-class train/val/test counts are provided in Supplementary Table S1.

Table 1 Dataset statistics. *Text* and *Audio* denote the number of text transcripts and audio recordings per corpus

Corpus	Disease(s)	Text	Audio	Corpus	Disease(s)	Text	Audio
AphasiaBank	Aphasia, Control	814	789	RHD-English	RHD, Control	78	77
ADReSSo	AD, Control	237	237	Delaware	MCI, Control	405	407
Coelho	TBI, Control	104	104	Total		1,638	1,614

Data splits

Data are split by stratified random sampling within each class, with 70%/15%/15% allocated to train/validation/test sets respectively. Stratification operates at the sample level, so participants with multiple recordings (repeat sessions in AphasiaBank, repeat visits in Delaware) can contribute different recordings to different splits. The same fixed split is used across all experimental runs.

Preprocessing pipeline

CHAT-to-text conversion

TalkBank corpora are distributed in CHAT format [30]; AphasiaBank, Coelho, RHD-English, and Delaware corpora are normalized by a rule-based converter; ADReSSo is provided as audio recordings without accompanying transcripts, and plain-text transcripts are generated using OpenAI Whisper[31] prior to text feature extraction (see Supplementary Methods).

Linguistic and paralinguistic features are detailed in the Supplementary Methods (Supplementary Tables S2 and S3).

Pre-computed deep features

Two frozen pre-trained models are used to extract deep feature representations before training begins. Their parameters are not updated during training.

- RoBERTa-large (`roberta-large`, 355M parameters): Each transcript is tokenized (max length 512 tokens) and passed through the model; the `[CLS]` token embedding (1024-d) serves as the sentence representation. RoBERTa-large was selected as the strongest text encoder for this task, based on a systematic encoder search across four frozen candidates (DistilBERT, BERT-base, RoBERTa-base, RoBERTa-large).
- wav2vec2-large (`facebook/wav2vec2-large`, 317M parameters): Audio files are resampled to 16 kHz mono and segmented into 30-second chunks. Each segment is encoded independently, yielding segment-level 1024-d representations (B, S, 1024). In the final model, segment-level features (B, S, 1024) are retained for attention-weighted pooling (see the Architecture section). A corpus-scoped index prevents cross-corpus patient ID collisions.

Tabular features

Four clinical metadata features are constructed: (i) age normalized as $(age - 60)/15$, (ii) sex encoded as $\{0, 0.5, 1\}$ for {male, unknown, female}, (iii) a binary indicator for WAB Aphasia Quotient (WAB-AQ) availability, and (iv) normalized WAB-AQ score (available only for AphasiaBank participants; set to 0 when absent). Age and sex are individually recorded for AphasiaBank, Coelho, RHD-English, and Delaware (with cohort-level defaults for occasional missing entries), whereas all ADReSSo samples receive identical cohort-level defaults; the WAB-derived features are available only for AphasiaBank. The coverage pattern of these features thus partly encodes corpus membership; its quantitative consequences are examined in the Corpus-Confound Analysis (see also Limitations).

Model architecture

Figure 1 illustrates the overall architecture.

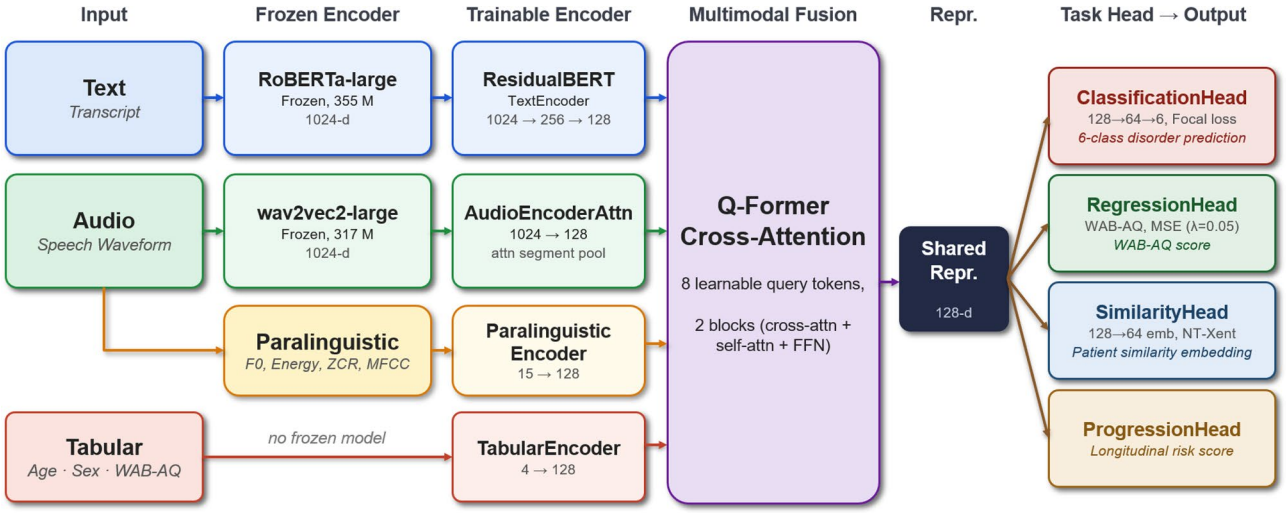

Fig. 1 Architecture of the proposed unified multimodal model. Four input streams (text, wav2vec2 acoustic, paralinguistic, and tabular) grouped into three semantic modalities are independently encoded to 128-d vectors and integrated by a Q-Former cross-attention module with $N_q = 8$ learnable query tokens. The resulting shared 128-d representation simultaneously feeds the task heads. Pre-trained encoders (RoBERTa-large, wav2vec2-large) are frozen; all other components are trainable (≈1.1M parameters total)

Modality encoders

Each modality stream is projected to a common $d = 128$ dimensional space by a lightweight encoder. The architecture uses four input streams—text (TEXT=0), tabular (TAB=1), wav2vec2 acoustic features (AUDIO=2), and hand-crafted paralinguistic features (PARA=3)—grouped into three semantic modalities (text, speech, and tabular). The two audio streams (AUDIO and PARA) represent complementary views of the same speech signal: deep neural phonetic embeddings from wav2vec2 and hand-crafted acoustic-prosodic features.

- ResidualBertTextEncoder: A two-path residual encoder from RoBERTa-large 1024-d to $d = 128$: (i) *main path*: $\text{Linear}_{1024\to256} \to \text{LN} \to \text{GELU} \to \text{Dropout}_{0.1} \to \text{Linear}_{256\to128} \to \text{LN} \to \text{GELU}$; (ii) *skip path*: $\text{Linear}_{1024\to128}$ (no bias); (iii) *output*: $\text{LayerNorm}(\text{main} + \text{skip})$. The direct linear skip from the 1024-d RoBERTa-large embedding to the 128-d space ensures stable gradient flow and prevents the deep main path from collapsing the representation.
- AudioEncoderAttn: Projects each segment representation $(B, S, 1024) \to (B, S, 128)$ via a linear layer, then learns scalar attention weights over the S segments using a single linear layer with softmax normalization, producing an attention-weighted pooled representation (B, 128). For samples without audio, a trainable *missing-audio token* $\mathrm{e}_\emptyset \in \mathbb{R}^{128}$ replaces the zero vector, allowing the model to distinguish "audio absent" from "audio present but silent." This token is optimized jointly with the rest of the network.
- ParalinguisticEncoder: Encodes a 15-dimensional paralinguistic feature vector—F0 (fundamental frequency) mean, F0 std, voiced ratio, RMS energy mean, RMS energy std, zero-crossing rate, spectral centroid, and MFCC 2–5 mean and std (4 coefficients × 2 statistics; see Supplementary Table S3 for full details)—to a 128-d representation: $\text{Linear}_{15\to32} \to \text{LN} \to \text{GELU} \to \text{Dropout}_{0.1} \to \text{Linear}_{32\to128} \to \text{LN} \to \text{GELU}$. This encoder enters the Q-Former as a fourth stream (PARA=3), complementing the wav2vec2-based AUDIO stream within the audio modality.
- TabularEncoder: $\text{Linear}_{4\to16} \to \text{LN} \to \text{GELU} \to \text{Dropout}_{0.1} \to \text{Linear}_{16\to32} \to \text{LN} \to \text{GELU} \to \text{Linear}_{32\to128}$

Q-Former cross-attention module

Following Li et al.[9], we maintain $N_q = 8$ learnable query tokens $\mathrm{Q} \in \mathbb{R}^{N_q \times d}$. The value $N_q = 8$ is selected based on ablation experiments (see the Ablation Study in Results). Each modality embedding is augmented with a trainable modality-token embedding e_m (TEXT=0, TAB=1, AUDIO=2, PARA=3) before the four streams are concatenated into a context matrix $\mathrm{C} \in \mathbb{R}^{4 \times d}$. The audio modality is represented by two complementary streams—neural embeddings from wav2vec2 (AUDIO=2) and hand-crafted acoustic-prosodic features (PARA=3)—both treated as sub-streams of the audio modality within the Q-Former. Two Q-Former blocks, each consisting of multi-head cross-attention ($N_q \times d$ queries, $4 \times d$ keys and values), followed by self-attention among queries and a position-wise FFN, update the query states:

$$\mathrm{Q}^{(\ell+1)} = \mathrm{QFormerBlock}^{(\ell)}(\mathrm{Q}^{(\ell)}, \mathrm{C}), \quad \ell = 0, 1. \tag{1}$$

The final query matrix is flattened and projected to $d = 128$:

$$\mathrm{h} = \mathrm{W}_p \cdot \mathrm{flatten}(\mathrm{Q}^{(2)}) \in \mathbb{R}^d, \tag{2}$$

where $\mathrm{W}_p \in \mathbb{R}^{d \times N_q d}$ is a learned projection. This shared representation h is the input to all four task heads.

Although the context matrix C contains only four vectors (one per modality stream), the $N_q = 8$ queries remain expressive: each query learns an independent attention pattern over the four modality tokens—weighting, for example, text more heavily for syntactic cues and paralinguistics more heavily for fluency—so the joint query bank can capture combinatorial modality relationships beyond a single attention-weighted sum. The subsequent self-attention layer further allows queries to share complementary information extracted from different modalities before the final pooling step.

The multi-head cross-attention uses $N_{\text{heads}} = 4$ heads.

Task heads

Table 2 summarizes the four task heads: the primary classification head and three auxiliary task heads (regression, similarity, and progression).

The classification head consists of two linear layers ($128 \to 64 \to 6$) with ReLU and Dropout(0.2). The regression head predicts WAB-AQ scores in the range 0–100 for AphasiaBank samples (WAB-AQ unavailable for other corpora). The similarity head projects h to a 64-dimensional embedding for NT-Xent contrastive learning, grouping same-class samples in the shared embedding space to support nearest-neighbor retrieval of similar patients. The progression head computes a risk score from the difference between two chronologically ordered patient embeddings.

Total learnable parameters

The total number of trainable parameters is approximately 1.1M (ResidualBertTextEncoder 430k + AudioEncoderAttn 132k + ParalinguisticEncoder 5k + TabularEncoder 5k + Q-Former 531k + task heads 29k). AudioEncoderAttn (∼132k) replaces the original AudioEncoder (∼230k) by eliminating the deep two-layer projection in favor of a linear segment projection plus a lightweight attention scorer; its input dimension matches

Table 2 Task heads, objectives, and training loss weights. ProgressionHead operates at inference only and carries no training loss

Head	Output	Loss	Weight (λ)
ClassificationHead	Logits ($B \times 6$)	focal loss ($\gamma = 2, \alpha$)	1.0
RegressionHead	WAB-AQ scalar	MSE	0.05
SimilarityHead	Embedding ($B \times 64$)	NT-Xent	0.1
ProgressionHead	Risk scalar $\in [0, 1]$	— (inference only)	—

wav2vec2-large (1024-d). The frozen RoBERTa-large and wav2vec2-large models together comprise ∼672M parameters that contribute no gradient computation.

Training

Multi-task loss

The training objective is:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}}, \tag{3}$$

with $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{reg}} = 0.05$, and $\lambda_{\text{sim}} = 0.1$. The auxiliary regression and similarity heads are trained jointly with classification, enabling the model to simultaneously predict WAB-AQ severity and learn a retrieval-ready embedding space.

Class imbalance

Clinical corpora are inherently imbalanced. The Control:RHD imbalance ratio in our dataset is approximately 21:1. We apply a two-tier strategy:

1. Focal loss[32] with class-specific $\alpha_c \propto n_c^{-1}$, clipped to [0, 5]: minority classes receive proportionally higher loss weight.
2. WeightedRandomSampler with sampling weight $w_c \propto n_c^{-0.5}$: the data loader oversamples rare classes within each mini-batch.

Optimizer and schedule

We use AdamW with weight decay 10^{-4}, linear warmup over 15 epochs to lr = 10^{-4}, and cosine decay over the remaining 135 epochs with gradient clipping at norm 1.0.

Hyperparameter search

A 72-configuration grid search (focal α power, sampler power, learning rate, label smoothing) on 50-epoch runs selects: lr = 10^{-4}, label smoothing = 0.10, focal power = 1.0, sampler power = 0.5; final runs use 150 epochs. The ResidualBertTextEncoder was selected via greedy ablation and RoBERTa-large via encoder search across four frozen encoders.

Experimental design

The full multi-seed pipeline uses five independent random seeds; all reported metrics are mean ± standard deviation across seeds.

Evaluation protocol

Multi-seed evaluation

Each configuration is trained independently five times with different random seeds. Performance metrics reported: test accuracy, macro-averaged F1, weighted F1, and per-class accuracy/F1/recall.

Prior-correction calibration

At inference, a class-prior correction term $\alpha \log(p_c C)$ is added to the logits of class c, where p_c is the training-set frequency of class c, $C = 6$ is the number of classes, and the scalar α is selected on the validation set (grid $[-0.5, 1.5]$ in steps of 0.1, maximizing the mean of accuracy and Macro-F1); the selected α is then applied unchanged to the test set. This calibration compensates for the mismatch between the sampler-balanced training distribution and the natural class distribution, and is applied identically to every neural-model configuration (proposed, ablations, and LOCO). Its effect on the final model is negligible (disabling it at evaluation changes test Macro-F1 by 0.1 pp, not significant); it is retained for consistency across configurations.

Baseline comparison

We compare against two groups of baselines, all evaluated on the same test set with identical train/val splits. The *trivial* baselines require no training and are computed analytically: Random (uniform class assignment), MostFrequent (always the majority Control class), and ClassPrior (sampling from the training class distribution). The *traditional ML* baselines receive 16-dimensional features (12 linguistic + 4 tabular) unless stated, with a BERT+tabular variant (772-d) also evaluated: an RBF-kernel SVM (SVM-RBF; $C = 10$, class-weighted), a RandomForest (300 estimators, class-balanced), and LogisticRegression (L2, $C = 1.0$, multinomial), each aggregated over five seeds.

Audio improvement variants

Separate from the component-level ablation below, we evaluate three audio improvements applied incrementally on top of a base mean-pooled wav2vec2-large configuration ($N_q = 8$). These are additive enhancements—not ablations of the proposed model—and their individual and combined results are summarized in Table 3.

Ablation study

Seven ablation configurations (A2–A8) each modify exactly one component of the proposed model; A1 serves as a lower-bound floor comparison using a minimal base configuration (Table 4). All configurations share the same multi-seed protocol, train/val/test split, and data pipeline. A3 (text+tabular only) isolates the contribution of the audio modality.

Cross-corpus generalization (leave-one-corpus-out)

LOCO experiments evaluate cross-domain generalization by training on all corpora except one and evaluating on the held-out corpus. Because each disease class occurs in exactly one corpus, this protocol serves two complementary purposes: it measures transfer under corpus shift for the class whose labels are preserved in training

Table 3 Audio improvement variant configurations. E0 is the base; E1, E2, E3 are individual additions; E1+E2+E3 is the final proposed model

	Name	Description
E0	Base audio	Fixed mean-pooling over wav2vec2-large segments; no audio improvements.
E1	Missing token	Trainable missing-audio token replaces zero-gating for samples without audio (≈1.5%).
E2	Paralinguistic	15-dim paralinguistic features (F0, energy, ZCR, MFCC) added as a fourth Q-Former input stream (see Supplementary Table S3).
E3	Attn pooling	Attention-weighted pooling over 30-second wav2vec2 segments replaces fixed mean-pooling.
E1+E2+E3	**Proposed**	All three improvements combined; the final proposed model.

Table 4 Ablation study configurations. A1 is a lower-bound floor comparison; A2–A8 each modify exactly one component of the proposed model

	Name	Description
A1	Floor (base config)	DirectFusion with the minimal base configuration: DistilBERT-base text encoder, mean-pooled audio without enhancements (E0). Serves as a lower-bound floor for the proposed model.
A2	No Q-Former (full)	Replace Q-Former with DirectFusion under the full proposed configuration (RoBERTa-large, hyperparameter-optimized); isolates the pure fusion-mechanism contribution.
A3	Text+Tabular Only	Disable wav2vec2 and paralinguistic encoders; use text and tabular features only.
A4	No focal loss	Replace focal loss with standard CrossEntropyLoss.
A5	No Sampler	Replace WeightedRandomSampler with uniform random sampling.
A6	LING Encoder	Replace frozen RoBERTa-large features with the 12-d linguistic feature vector via a lightweight encoder.
A7	$N_q = 16$	Q-Former with 16 query tokens instead of 8.
A8	$N_q = 4$	Q-Former with 4 query tokens instead of 8.

Table 5 Leave-One-Corpus-Out configurations. In all four settings the excluded disease class is strictly zero-shot (present only in the excluded corpus), while the Control class remains in the training set; zero-shot transfer applies to the disease class only ("partial")

Excluded	Disease	Evaluation focus
ADReSSo	AD / Control	AD vs. Control; picture description, ASR transcripts
Coelho	TBI / Control	TBI vs. Control; story retelling, young adults
RHD-English	RHD / Control	RHD vs. Control; right-hemisphere damage
Delaware	MCI / Control	MCI vs. Control; shared discourse tasks, older adults

(Control, present in all corpora), and it provides an honest report of the failure mode for disease classes entirely absent from training, rather than a test of disease-class generalization per se. The four exclusion configurations are described in Table 5.

For each LOCO run, the remaining samples are split into training and validation sets; the excluded corpus forms the entire test set. LOCO experiments are run with five independent random seeds.

Statistical significance testing

To assess whether performance differences are statistically significant, we apply McNemar's test[33] and bootstrap permutation testing[34] to per-sample correct/incorrect arrays. Significance is declared at $p < 0.05$. Methodological details and full pairwise results are provided in the Supplementary Methods and Supplementary Table S5.

Error analysis

We conduct a structured error analysis on the proposed model's test predictions using four analyses:

1. Misclassification inventory: All incorrect predictions are tabulated with patient ID, corpus, true/predicted label, prediction confidence, and 12 linguistic feature values.
2. Corpus-specific error patterns: Per-corpus error rates quantify domain-specific difficulty and identify corpus-level factors (recording protocol, training-set size, task format) driving misclassifications.
3. MCI–Control confusion analysis: The MCI and Control populations are divided into four groups (MCI-correct, MCI-as-Control, Control-correct, Control-as-MCI) and their mean linguistic features are compared. This identifies which acoustic-linguistic features most discriminate the confusion pairs.
4. Confusion matrix: The full 6×6 confusion matrix reveals systematic inter-class confusions (e.g., MCI predicted as Control, or AD predicted as Control).

Table 6 Main classification results. Deterministic methods (SVM, LogisticRegression, Trivial) report no standard deviation. RandomForest reports mean ± std across five seeds. All methods use the identical 70/15/15 train/val/test split

Group	Model	Macro-F1 (%)	Accuracy (%)
Trivial	Random	11.3	15.0
	MostFrequent (Control)	11.1	49.6
	ClassPrior	16.1	31.3
Traditional ML	SVM-RBF (ling+tab)	69.8	71.5
	SVM-RBF (BERT+tab)	66.6	78.1
	RandomForest (ling+tab)	55.4 ± 1.7	79.0 ± 0.8
	LogisticRegression (ling+tab)	59.6	61.4
Proposed	Multimodal Q-Former (RoBERTa-large, $N_q = 8$)	86.19 ± 0.99	90.98 ± 0.18

Table 7 Audio improvement variant results (mean ± std, five seeds, 150 epochs). E0 is the base mean-pooled configuration; E1+E2+E3 is the final proposed model

	Configuration	Macro-F1 (%)	Accuracy (%)
E0	Base (mean-pooled wav2vec2)	81.94 ± 2.99	90.57 ± 0.73
E1	Missing token	82.24 ± 4.82	90.08 ± 0.84
E2	Paralinguistic	84.12 ± 1.68	89.59 ± 1.30
E3	Attn pooling	84.16 ± 1.80	90.73 ± 0.60
E1+E2+E3	Proposed (RoBERTa-large, $N_q = 8$)	86.19 ± 0.99	90.98 ± 0.18

Results

Main classification results

The proposed multimodal Q-Former model (RoBERTa-large text encoder, wav2vec2-large audio encoder, $N_q = 8$, with audio enhancements E1+E2+E3) achieves Macro-F1 $86.19\% \pm 0.99\%$ and accuracy $90.98\% \pm 0.18\%$ on the six-class test set. Against two groups of baselines—trivial heuristics and traditional ML—the proposed model achieves the highest Macro-F1 and accuracy (Table 6).

The proposed model's Macro-F1 (86.19%) and accuracy (90.98%) substantially exceed all baselines.

Audio improvement variants

Table 7 reports the five-seed results for audio improvement variants, evaluated independently and in combination. **E1**: a learned missing-audio token replaces zero-gating for samples without audio (approximately 1.5% of samples, mainly AphasiaBank). **E2**: 15-dimensional paralinguistic features (F0 mean/std/voiced ratio, RMS energy mean/std, zero-crossing rate, spectral centroid, MFCC 2–5 mean/std; see Supplementary Table S3) enter the Q-Former as a fourth input stream. **E3**: attention-weighted segment pooling replaces fixed mean-pooling over 30-second wav2vec2 segments.

See the Discussion section for interpretation of the audio modality contribution.

Ablation study

Ablation results are reported in Table 8. A2 holds all settings identical to the proposed model except the fusion mechanism, isolating the pure Q-Former advantage. A3 disables all audio input, using text and tabular only. A7 and A8 vary $N_q \in \{4, 8, 16\}$ under the full configuration.

A1 establishes a lower-bound floor using a minimal base configuration (DistilBERT-base, DirectFusion, mean-pooled audio): Macro-F1 $67.01\% \pm 5.24\%$ (-19.18 pp vs proposed), where the gap reflects the cumulative effect of all simultaneously downgraded components rather than any single factor. A2 isolates the pure Q-Former contribution under matched settings: -5.17 pp ($81.03\% \pm 2.53\%$), significant by both tests (McNemar $p = 0.013$; bootstrap $p = 0.003$), confirming that Q-Former cross-attention provides a consistent advantage over DirectFusion when the fusion mechanism is the only variable.

Table 8 Ablation study results (mean ± std, five seeds, 150 epochs). ΔMF1 and ΔAcc: difference from the proposed model (86.19% / 90.98%) in percentage points. ⋆†: significant by both McNemar ($p < 0.05$) and bootstrap ($p < 0.05$) vs proposed. Full pairwise tests in Supplementary Table S5

	Configuration	Macro-F1 (%)	ΔMF1	Accuracy (%)	ΔAcc
A1	w/o Q-Former (base)⋆†	67.01 ± 5.24	−19.18	81.79 ± 3.58	−9.19
A2	w/o Q-Former (full)⋆†	81.03 ± 2.53	−5.17	89.11 ± 0.78	−1.87
A3	Text+Tabular only	84.08 ± 2.47	−2.11	90.16 ± 0.45	−0.82
A4	w/o focal loss	83.84 ± 4.81	−2.35	89.84 ± 3.49	−1.14
A5	w/o WeightedRandomSampler⋆†	79.44 ± 8.56	−6.76	87.32 ± 3.39	−3.66
A6	LING encoder (12-d)⋆†	73.0 ± 5.6	−13.19	81.63 ± 2.95	−9.35
A7	Q-Former ($N_q = 16$)	84.34 ± 1.13	−1.85	90.49 ± 0.36	−0.49
A8	Q-Former ($N_q = 4$)	84.59 ± 2.30	−1.60	90.49 ± 1.10	−0.49

Table 9 Leave-One-Corpus-Out results (mean ± std across five seeds). In all four settings the excluded disease class is strictly zero-shot (present only in the excluded corpus); the Control class is present in the training set in each case (partial zero-shot)

Excluded	Disease	Macro-F1 (%)	Accuracy (%)	Control-F1 (%)
ADReSSo	AD / Control	11.58 ± 0.70	41.43 ± 5.34	66.9 ± 1.7
Coelho	TBI / Control	9.31 ± 1.89	20.38 ± 8.00	46.5 ± 8.5
RHD-English	RHD / Control	15.75 ± 1.73	46.15 ± 2.03	65.5 ± 1.6
Delaware	MCI / Control	21.09 ± 4.54	62.32 ± 0.54	77.1 ± 0.2

The two class-imbalance countermeasures (A4, A5) behave asymmetrically under our severely skewed label distribution (Control:RHD ≈ 21:1). Removing the WeightedRandomSampler (A5) inflicts the single largest component-level drop (−6.76 pp; 79.44% ± 8.56%), significant by both tests (McNemar $p < 0.001$; bootstrap $p = 0.001$) and accompanied by the highest variance of any configuration; without explicit oversampling, rare disorders such as RHD (2.3% of samples) are overwhelmed by the majority Control class, so class-frequency-weighted sampling is essential for stable minority-class recall on this imbalanced dataset. Removing focal loss (A4) produces a smaller, non-significant drop (−2.35 pp; 83.84% ± 4.81%; McNemar $p = 0.235$; bootstrap $p = 0.073$): loss-level down-weighting of easy majority examples still helps numerically but is partly redundant once the sampler already rebalances each mini-batch. Together, A4 and A5 indicate that on a 21:1-imbalanced clinical corpus the dominant driver of imbalance robustness is explicit minority-class oversampling rather than reweighting the loss alone.

A6 (LING encoder) is significant by both tests (McNemar $p = 0.012$; bootstrap $p = 0.001$), validating frozen contextual embeddings over hand-crafted features.

The remaining ablations do not reach significance and confirm the robustness of the proposed configuration. Disabling the audio modality (A3, text+tabular only) costs −2.11 pp (84.08% ± 2.47%; bootstrap $p = 0.174$), a small non-significant margin examined in detail in the Discussion. Varying the query count leaves performance essentially unchanged (A7, $N_q = 16$: −1.85 pp, 84.34% ± 1.13%; A8, $N_q = 4$: −1.60 pp, 84.59% ± 2.30%; bootstrap $p \geq 0.20$), indicating that $N_q = 8$ sits on a broad plateau and the model is insensitive to the exact number of query tokens within this range; we adopt $N_q = 8$ as the proposed configuration.

Leave-one-corpus-out generalization

Cross-corpus generalization is assessed by leaving out each corpus in turn and evaluating on the held-out set. Performance across four exclusion settings is shown in Table 9.

All LOCO configurations produce Macro-F1 of 9–21%. For reference, a uniform random 6-class predictor on these 2-class test sets yields approximately 8.3% Macro-F1 (the four absent classes contribute F1 = 0, and only the two present classes contribute non-zero F1); all point estimates exceed this chance level. In all four LOCO settings, the excluded disease class is strictly zero-shot; the model has no training examples for it, making

disease-class generalization structurally unachievable without zero-shot transfer capabilities. Delaware/MCI achieves the highest Macro-F1 (21.09%), consistent with its discourse tasks being shared with the retained corpora (see the Dataset section). Coelho/TBI achieves the lowest Macro-F1 (9.31%), reflecting the complete absence of TBI-specific training data together with elicitation tasks that no retained corpus shares. ADReSSo/AD (11.58%) and RHD-English (15.75%) also show limited transfer, consistent with ADReSSo's ASR-generated transcripts and RHD's subtle pragmatic deficits making zero-shot generalization difficult. The Control-F1 column isolates the transfer of the one class whose label overlap is preserved; it is interpreted, together with a complementary sub-corpus hold-out that preserves *disease*-label overlap, in the Corpus-Confound Analysis below. AphasiaBank is excluded from LOCO: as the source of 50% of training samples, its removal would render cross-corpus generalization results uninterpretable.

Corpus-confound analysis

Because each disorder is drawn from a single corpus, high within-corpus accuracy could reflect corpus fingerprints rather than disorder signal. We assess this from three angles: corpus predictability from the learned representation, whether classification *decisions* follow corpus identity where the two dissociate, and the contribution of the tabular modality.

Corpus predictability (representation level)

A multinomial logistic-regression probe on the 128-dimensional shared representation (fit on train+validation, evaluated on test; five seeds) identifies a sample's source corpus with $97.8 \pm 0.3\%$ accuracy, and $97.4 \pm 1.3\%$ ($93.0 \pm 5.0\%$ Macro-F1; 44.3% majority chance) on Control-only samples, where all disorder signal is absent. The same probe on the 27 interpretable features computed from the same recordings (12 linguistic + 15 paralinguistic) already reaches 96.3–96.7%: corpus identity is nearly fully decodable from shallow features alone, and the learned representation neither amplifies nor removes it. Whether this predictability contaminates predictions is a separate, decision-level question.

Decision-level dissociation (Control recall by corpus)

Control, present in all five corpora, provides a built-in dissociation test: if decisions followed corpus fingerprints, minority-corpus Controls would be assigned to their corpus's disorder. Instead, test-split Control recall is $90.4 \pm 3.4\%$ (AphasiaBank, n=54), $95.5 \pm 0.0\%$ (Delaware, n=44), $78.7 \pm 6.5\%$ (ADReSSo, n=15), and 100% (RHD-English, n=2), and held-out Control recall under LOCO reaches 85.4–99.0% (corresponding Control-F1 in Table 9). The exception on both counts is Coelho ($57.1 \pm 9.0\%$, n=7, errors all assigned to TBI; 42.4% under LOCO): with its disjoint tasks and distinct recording conditions, TBI/Coelho is the setting where corpus and disorder signal are least separable.

Tabular-modality contribution

Permuting the four tabular features at inference (100 permutations) lowers Macro-F1 by 17.6 ± 1.6 pp, and retraining without the TabularEncoder (five seeds, identical protocol) yields $73.8 \pm 4.2\%$ versus $86.2 \pm 1.0\%$ (-12.4 pp; McNemar $p < 0.02$ in every seed), with the drop confined to the three smallest classes (recall: MCI $90.0 \rightarrow 55.5\%$, RHD $76.7 \rightarrow 53.3\%$, TBI $87.5 \rightarrow 72.5\%$). Fixing only the two WAB-derived features—the pair most directly encoding AphasiaBank membership—to their population defaults costs 3.2 ± 1.8 pp, with the drop confined to the Aphasia and Control classes, and AD is unaffected despite ADReSSo's constant default tabular vector. The tabular contribution is thus substantial, dominated by age and sex rather than by the WAB fingerprint, and partly corpus-correlated (see Limitations).

Table 10 AphasiaBank sub-corpus hold-out (mean ± std across five seeds). Each row holds out one entire collection site; Aphasia and Control both remain in the training distribution (preserved label overlap), unlike the LOCO protocol

Held-out site	*n* (Aph/Ctl)	Aphasia recall	Aphasia F1	Control recall	Control F1
Richardson	16 / 63	78.8 ± 3.1	84.6 ± 2.6	86.0 ± 12.6	89.7 ± 7.3
UNH	30 / 20	100.0 ± 0.0	94.0 ± 0.6	81.0 ± 2.0	89.5 ± 1.2
UMD	15 / 26	88.0 ± 2.7	83.6 ± 2.8	36.2 ± 27.6	45.1 ± 29.2

Table 11 Per-class precision, recall, and F1 (%) for the proposed model ($N_q = 8$; mean ± std across five seeds). TBI ($n = 8$) and RHD ($n = 6$) test sets are small; class-level estimates carry substantial uncertainty

Class	Prec (%)	Recall (%)	F1 ± std (%)	Support
Control	92.7	89.0	90.8 ± 0.3	122
Aphasia	93.2	98.0	95.5 ± 0.3	70
AD	81.2	84.4	82.6 ± 1.1	18
TBI	70.3	87.5	77.9 ± 2.7	8
RHD	84.3	76.7	77.4 ± 5.7	6
MCI	96.1	90.0	93.0 ± 1.5	22
Macro avg	86.3	87.6	86.2	246

Sub-corpus hold-out with preserved label overlap

LOCO cannot preserve disease-label overlap, but AphasiaBank aggregates over twenty independently collected sub-corpora, so we additionally hold out each of the three collection sites with at least ten patients per class and retrain (five seeds each; Table 10). Held-out Aphasia recall reaches 78.8–100.0%: disorder detection transfers to unseen collection sites once the label is represented in training. Control recall transfers at 81.0–86.0% for two of the three sites but drops sharply for UMD (36.2 ± 27.6%); notably, the UMD errors flow to RHD rather than to Aphasia, indicating substantial site-level variability rather than a shortcut toward the held-out site's own disorder.

Statistical significance

Full pairwise statistical test results (proposed model vs each ablation/baseline) are reported in Supplementary Table S5. Four of eight ablation comparisons reach significance. A1 (floor comparison) is significant by both tests (McNemar $\chi^2 = 63.0$, $p < 0.001$; bootstrap $p = 0.001$), though its 19.18 pp gap reflects multiple simultaneously downgraded components, not Q-Former alone. A2 (no Q-Former, matched config) isolates the pure fusion effect and is significant by both tests (McNemar $p = 0.013$; bootstrap $p = 0.003$). A5 (no WeightedRandomSampler) is significant by both tests (McNemar $p < 0.001$; bootstrap $p = 0.001$), confirming that class-frequency-weighted sampling is essential. A6 (LING encoder) is significant by both tests (McNemar $p = 0.012$; bootstrap $p = 0.001$), validating contextual RoBERTa embeddings over hand-crafted features. SVM-RBF is significant by both tests (McNemar $p = 0.007$; bootstrap $p = 0.001$), confirming the proposed model statistically outperforms the strongest ML baseline. A3 ($p = 0.326$/0.174) and A4 ($p = 0.235$/0.073) are not significant.

Per-class performance and error analysis

Per-class precision, recall, and F1 for the proposed model are reported in Table 11 (mean across five seeds).

The 6×6 confusion matrix for a representative evaluation run is provided in Supplementary Table S4. The dominant error pattern involves Control–MCI confusion. Aphasia achieves the highest recall (98.0%) and strong precision (93.2%), reflecting its distinctive lexicosyntactic signature in AphasiaBank transcripts. AD recall (84.4%) is reasonable given the limited test set size (18 samples), demonstrating effective classification despite the small ADReSSo test partition. MCI precision (96.1%) and F1 (93.0% ± 1.5%) are among the highest, while recall (90.0%) reflects residual MCI–Control overlap. TBI F1 (77.9% ± 2.7%) and RHD F1 (77.4% ± 5.7%) remain variable due to very small test sets ($n = 8$ and $n = 6$ respectively): a single misclassification shifts class-level recall by 12–17 pp, so

these figures should be treated as indicative estimates rather than reliable performance metrics; definitive assessment for TBI and RHD requires substantially larger evaluation cohorts.

Corpus-specific error patterns

For the four corpora where a single disease class maps one-to-one, class recall in Table 11 directly equals corpus disease-recognition accuracy. RHD-English shows the highest disease-class error rate (RHD recall 76.7%, error rate 23.3%), attributable to the subtlety of right-hemisphere pragmatic deficits and the very small test partition ($n = 6$). ADReSSo (AD error rate 15.6%) is the only corpus distributed without transcripts, so its text inputs are Whisper-generated ASR transcripts that introduce recognition noise absent from the human-transcribed corpora. Coelho (TBI error rate 12.5%) is constrained by a small training partition ($n_{\text{train}} = 38$); the within-seed variability (F1 $77.9\% \pm 2.7\%$) reflects sensitivity to individual training examples rather than a systematic representational failure. Delaware (MCI error rate 10.0%) errors cluster specifically in MCI–Control confusion: both groups engage the same discourse tasks, and the linguistic boundary between early-stage cognitive decline and healthy aging is inherently narrow. AphasiaBank differs structurally from the four corpora above: it comprises over twenty sub-corpora spanning diverse aphasia subtypes and task formats, and its Control participants are drawn from the same corpus mix. The aggregate Aphasia recall (98.0%) is therefore an average across heterogeneous sub-corpora, and performance on any individual sub-corpus may deviate from this figure depending on aphasia severity and elicitation task.

Inter-disorder correlation structure

To characterize how the five disorders relate to one another in speech and language—complementing the per-class error analysis above and echoing transdiagnostic studies[10,13]—we analyze inter-disorder structure from two viewpoints: the raw input features and the model's learned representation.

Raw feature profiles (Figure 2). We z-score the 27 interpretable features (12 linguistic + 15 paralinguistic), form a mean profile per disorder, and correlate profiles pairwise. Two findings stand out. First, Control and MCI have the strongest positive profile correlation ($r = +0.63$; $+0.90$ on linguistic-only and $+0.82$ on paralinguistic-only features), directly explaining the dominant Control–MCI confusion: at the population-mean level MCI mirrors Control across both modalities, and its only salient signature is elevated spectral variability (MFCC 2–5 standard deviations, $z \approx +1.1$). A per-feature contribution analysis attributes their residual separability mainly to paralinguistic dynamics (F0 and voicing) rather than to lexical measures. Second, hierarchical clustering recovers a clinically coherent split by etiology—a cognitive/diffuse group {Control, MCI, AD} versus a focal acquired-injury group {Aphasia, TBI, RHD} (Aphasia–Control $r = -0.88$; TBI–RHD $r = +0.50$). These per-disorder signatures match established clinical descriptions—reduced lexical diversity with elevated fillers in aphasia, sharply reduced verbal output in AD, the highest pause density in TBI, a Control-like but verbose and referentially imprecise profile in RHD, and an MCI deviation confined largely to elevated spectral variability rather than lexical measures[2–5,12]. In this dendrogram (average linkage on $1 - r$), the dashed line is a coloring cut at $0.7\times$ the maximum cophenetic distance: links merging below it are colored to mark a cluster, and links above are drawn in grey.

Learned representation (Figure 3). We extract the shared 128-d Q-Former representation for every test sample and compute, per disorder, the centroid and pairwise centroid cosine similarity (averaged over five seeds). In the embedding space the model separates the six classes well: all off-diagonal similarities are negative except Control–RHD ($+0.37$) and Control–TBI ($+0.08$). Aphasia is the most distinct (most negative similarity to Control, -0.32), consistent with its high recall. Because this space is optimized for discrimination rather than for preserving raw-feature similarity, the most confusable pair—Control and MCI, near-identical in the raw features—is driven *further* apart (-0.22) than the more easily separated Control–AD pair (-0.11): the model allocates the largest margin where separation is hardest, by design rather than in error. Centroid cosine measures the separation of class *means*, not the absence of sample-level overlap, so this margin coexists with the residual Control–MCI confusion

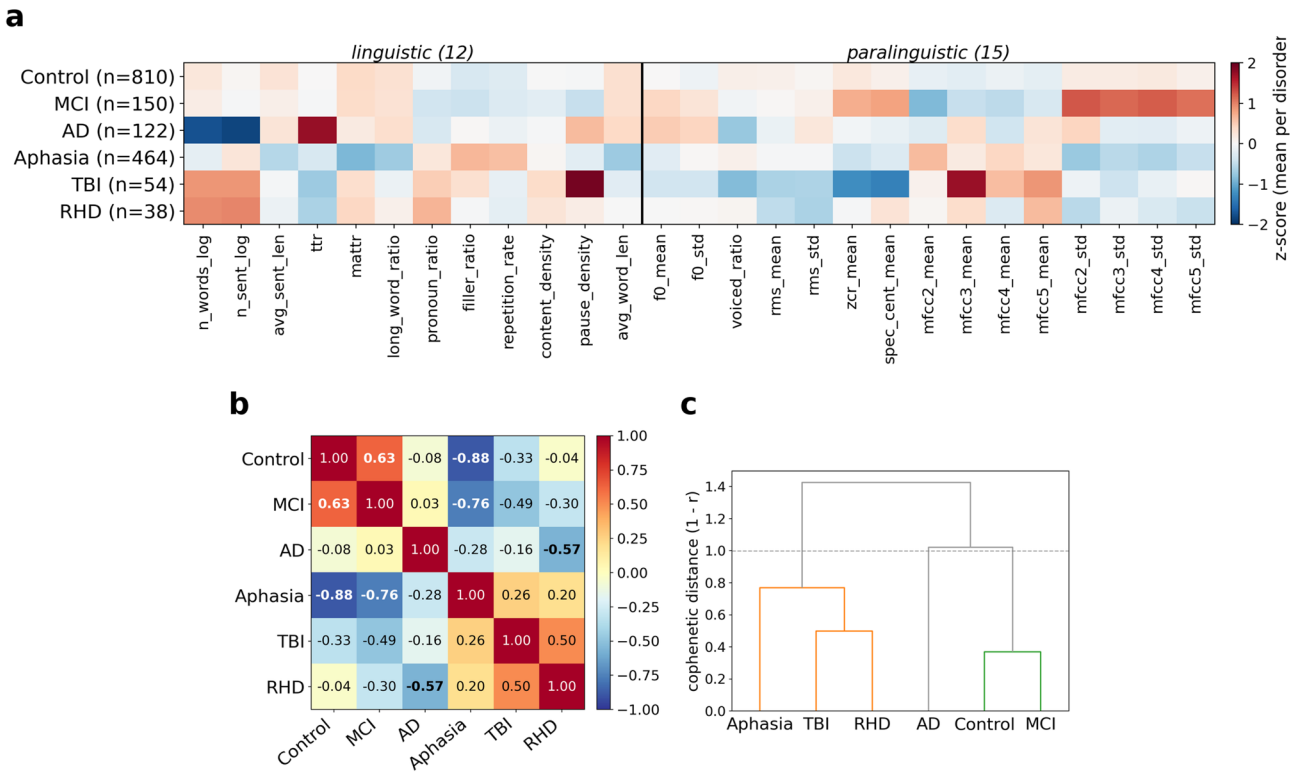

Fig. 2 Disorder profiles in the raw feature space (all 1,638 samples; 12 linguistic + 15 paralinguistic features): (**a**) per-disorder mean profile (z-scored; red/blue = above/below the grand mean, vertical rule separates the two feature blocks), (**b**) pairwise profile correlation (Pearson over 27 features), and (**c**) hierarchical clustering

noted above. In the embedding dendrogram (Figure 3d, average linkage on $1 - \cos$; the cut and coloring follow the same convention as Figure 2c), only {Control, RHD} merges below the cut; every other disorder remains separate—that is, the learned space is disorder-discriminative apart from a residual RHD–Control (and weaker TBI–Control) proximity attributable to the very small RHD/TBI training support rather than to feature inseparability.

Multimodal evidence. The discriminative cues are distributed across both modalities. A one-way ANOVA over the 27 features ranks a paralinguistic feature highest (mfcc2_std, $F = 163.5$), and the top ten comprise six linguistic and four paralinguistic features: even though the audio modality is not significant in aggregate (A3), specific acoustic features carry decisive per-disorder information—and MCI, the hardest class, is set apart almost entirely by such cues (its spectral-variability signature). This cross-modal distribution of evidence supports the multimodal design.

Auxiliary heads

WAB-AQ severity estimation

The model optimizes a WAB-AQ regression head jointly with the classification objective. Evaluated on AphasiaBank test samples with available severity labels ($n = 70$), the regression head achieves Mean Absolute Error (MAE) = 2.92 and Root Mean Square Error (RMSE) = 3.42 on the 0–100 WAB-AQ scale. The WAB-AQ has an established test-retest reliability of approximately ±3–5 points [35]; the model's MAE therefore falls within the instrument's own measurement uncertainty, indicating that, within the AphasiaBank collection conditions, aphasia severity can be estimated from speech alone without administering the full clinical assessment. Whether this

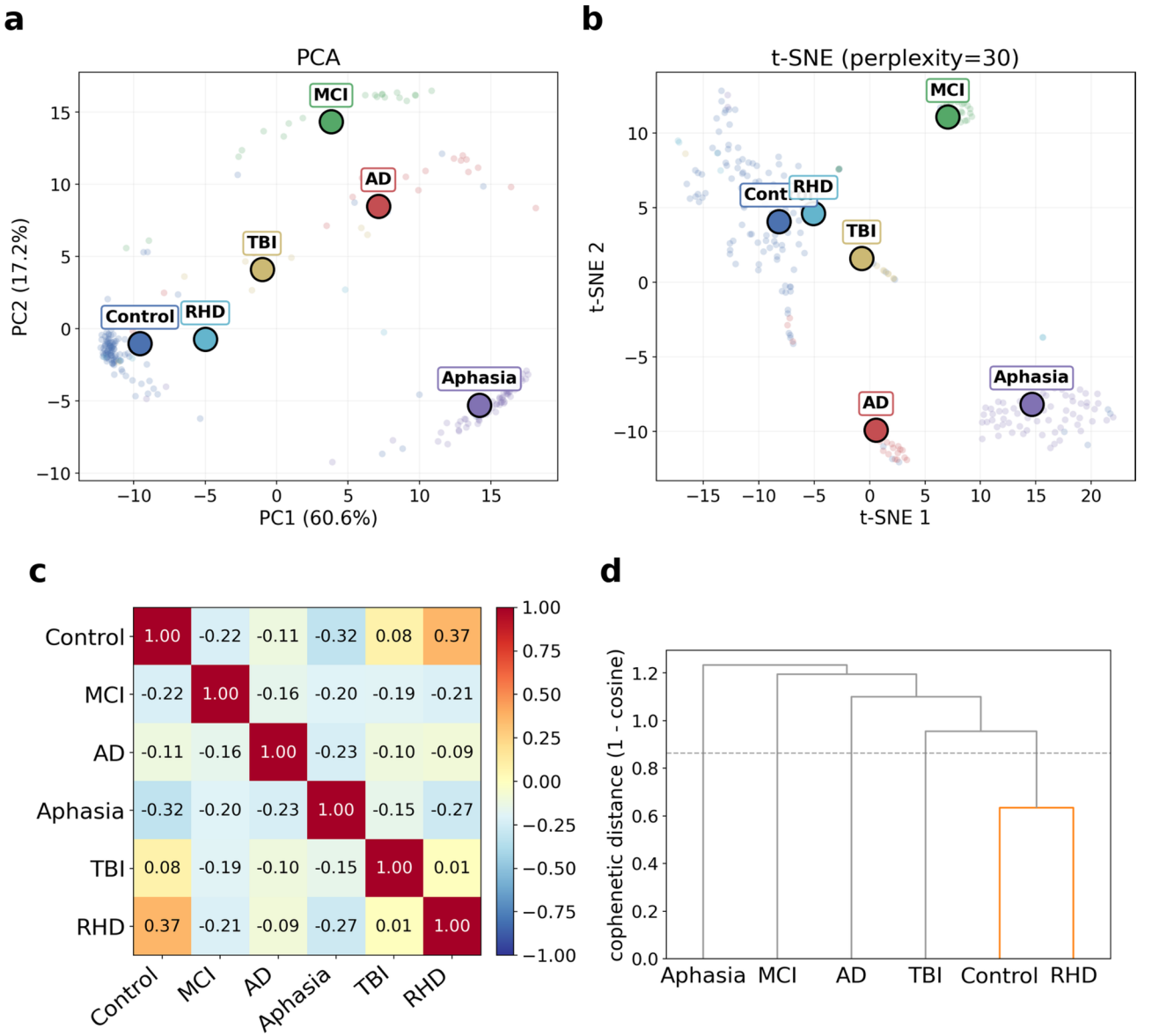

Fig. 3 Inter-disorder structure in the learned 128-d representation (proposed model, test set): (**a**) PCA and (**b**) t-SNE projections (faint points: samples; outlined markers: per-disorder centroids), (**c**) centroid cosine similarity (mean over five seeds), and (**d**) hierarchical clustering. Negative similarities indicate well-separated disorders; the dendrogram cut and coloring are described in the text

capability extends to independently collected recordings and other clinical settings remains to be established prospectively.

Disease similarity retrieval

The shared Q-Former embedding space supports nearest-neighbor retrieval as a secondary inference mode. Precision@5 on the full test set achieves 74.6%: given a query patient, approximately four of the five nearest neighbors in the embedding space belong to the same disease class. This precision is achieved by a single representation trained jointly across all five disorders simultaneously, demonstrating that the embedding encodes disease-specific structure despite serving multi-task classification and severity regression concurrently. This structured

embedding space suggests nearest-neighbor retrieval as a potential direction for clinical decision support—for example, surfacing previously diagnosed patients with similar linguistic profiles as references for a new case—although prospective validation would be required before any clinical use.

Longitudinal trajectory tracking

The TBI samples used in the main classification experiment come from the Coelho corpus[27], a cross-sectional dataset that does not include longitudinal follow-up. For longitudinal analysis, we apply the trained encoder—without retraining—to the separate Togher TBI cohort[4,36] ($n = 58$ patients, 5 timepoints: 3–24 months post-injury; 237 transcripts), which has no class labels. We test whether Glasgow Coma Scale (GCS) score at admission—recorded in the Togher corpus metadata and not used in model training—predicts the model's Control-class probability at the latest available follow-up (12 or 24 months). Spearman correlation yields $\rho = +0.397$ ($p = 0.0047$, $n = 49$): less severe injury (higher GCS) is associated with greater latent-space shift toward the Control cluster over time, demonstrating that the learned representation encodes clinically meaningful recovery gradients without longitudinal supervision. The analysis uses a single external cohort without model retraining, and cross-sectional confounds within each timepoint cannot be excluded; prospective replication on additional cohorts remains an open direction.

Discussion

Role of Q-former multimodal fusion

A2 isolates the pure Q-Former contribution by replacing Q-Former with DirectFusion (modality concatenation followed by linear projection) while holding all other settings identical to the proposed model (RoBERTa-large, hyperparameter-optimized, E1+E2+E3). Under this matched single-factor comparison, Q-Former achieves +5.17 pp Macro-F1 over DirectFusion (McNemar $p = 0.013$; bootstrap $p = 0.003$), confirming that learned cross-modal attention provides a consistent advantage over simple concatenation-projection. The gain reflects Q-Former's ability to selectively weight modalities via dynamic cross-attention, whereas DirectFusion collapses all information into a fixed projection that cannot adapt to per-sample evidence availability.

Audio modality contribution

The text+tabular-only ablation (A3) achieves Macro-F1 $84.08\% \pm 2.47\%$, below the proposed model by 2.11 pp (bootstrap $p = 0.174$, not significant). The audio improvements (E3 attention pooling, E2 paralinguistic features, E1 learned missing token) together yield Macro-F1 $86.19\% \pm 0.99\%$ and accuracy $90.98\% \pm 0.18\%$ (+4.25 pp over base mean-pooled configuration). The 1024-dimensional wav2vec2-large representations align with RoBERTa-large, enabling meaningful cross-modal attention: E1+E2 alone (85.31%; not shown) already exceeds the text+tabular baseline (84.08%, A3). E3 (attention segment pooling) further improves F1: softmax attention over wav2vec2 segments differentially weights regions by diagnostic relevance, rather than treating all segments equally as mean pooling does.

Limitations

1. Small TBI and RHD test sets. The TBI and RHD test sets contain only 8 and 6 samples respectively, so a change of one sample shifts class-level recall by 12–17 pp; per-class metrics for these two disorders are therefore statistically unreliable, and expanding their data collection is a priority for future work.

2. LOCO evaluation scope. Each LOCO test set contains only two classes (one disease plus Control), so the six-class Macro-F1 primarily reflects Control recognition rather than zero-shot disease detection; disease-class F1 is structurally near zero without in-distribution training examples.
3. Tabular modality coverage. WAB-derived metadata exist only for AphasiaBank, and all ADReSSo samples carry identical demographic defaults, while age and sex are individually recorded for the remaining corpora. Removing the tabular modality costs 12.4 pp Macro-F1 (concentrated in MCI, RHD, and TBI), versus 3.2 pp for the WAB features alone: the tabular stream contributes substantial signal dominated by age and sex, part of which—given systematic demographic differences between corpora—may act as corpus-correlated rather than patient-specific information.
4. Generalization scope and corpus–disorder coupling. The reported results demonstrate discrimination within the corpora and data-collection conditions represented during training; robust transfer to independently collected cohorts, alternative recording conditions, or new clinical populations has not been established, and the model should be regarded as a research prototype rather than a validated screening tool. Moreover, because each disorder is drawn from a single corpus, the one-to-one correspondence between disorders and corpora is a structural limitation of the present data composition; fully disentangling disorder-specific from corpus-specific signal will require multi-site corpora containing the same disorder. The Corpus-Confound Analysis partially constrains this concern: corpus identity is nearly fully decodable from shallow features and the learned representation alike, yet for the Control class—the one case where corpus and label dissociate—decisions follow corpus identity only for Coelho, the sole corpus whose elicitation tasks are shared by no other; its TBI results warrant particular caution.

Conclusion

We presented a unified multimodal, multi-corpus, multi-task Q-Former model for the simultaneous six-class assessment of neurological language disorders across five heterogeneous TalkBank corpora, achieving test accuracy 90.98% and Macro-F1 86.19%. Ablation analysis attributes the main gains to Q-Former cross-modal attention and class-frequency-weighted sampling, with contextual RoBERTa-large embeddings clearly outperforming hand-crafted linguistic features and the three audio enhancements adding a further, modest improvement; a complementary correlation analysis shows that the learned representation separates disorders that overlap in the raw features—most notably MCI and Control—and that paralinguistic cues carry disorder-specific information beyond text.

Cross-corpus zero-shot generalization to unseen disease classes remains the principal open challenge, and—although corpus-confound analyses find no evidence that decisions reduce to corpus identity where this is testable, Coelho excepted—the disorder–corpus coupling itself can only be resolved by multi-site, same-disorder corpora. Future work includes domain-adapted fine-tuning of the text and audio encoders, parameter-efficient adaptation (e.g., LoRA) for few-shot transfer to new corpora, and prospective validation on independently collected clinical cohorts.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-76495-9.

Acknowledgements This research was supported by the National Research Council of Science & Technology (NST) grant funded by the Korean government (MSIT) (No. CAP21055-300) and by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (No. 26CB1200).

Author contributions B.L. conceived the study, designed and implemented the model, conducted the experiments, analyzed the results, and wrote the manuscript. H.J.S. and B.O.K. contributed to the study design and to the analysis and interpretation of the results. All authors reviewed and approved the final manuscript.

Funding This research was supported by the National Research Council of Science & Technology (NST) grant funded by the Korean government (MSIT) (No. CAP21055-300) and by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (No. 26CB1200).

Data availability The data used in this study are drawn from the TalkBank repository (https://talkbank.org); individual corpora require registration and agreement to the respective usage guidelines. The ADReSSo corpus is available via the Interspeech 2021 ADReSSo challenge repository. Derived features are available from the corresponding author upon reasonable request, subject to the TalkBank data-use agreement.

Code availability The complete source code for data preprocessing, model training, and evaluation is publicly available on GitHub (https://github.com/bh0903lee/cross-corpus-neurolang) and permanently archived on Zenodo (https://doi.org/10.5281/zenodo.21930364).

Declarations

Competing interests The authors declare no competing interests.

Ethical approval This work used only existing, publicly available de-identified data from TalkBank. Therefore, no approval or waiver from an institutional review board (IRB) was required.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Steinmetz, J. D. et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. *Lancet Neurol.* **23**, 344–381 (2024).
2. MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. AphasiaBank: Methods for studying discourse. *Aphasiology* **25**, 1286–1307 (2011).
3. Fraser, K. C., Meltzer, J. A. & Rudzicz, F. Linguistic features identify Alzheimer's disease in narrative speech. *J.Alzheimer's Dis.* **49**, 407–422 (2016).
4. Togher, L. et al. Incog recommendations for management of cognition following TBI, part IV: Cognitive communication. *J. Head Trauma Rehab.* **29**, 353–368 (2014).
5. Blake, M. L. Perspectives on treatment for communication deficits associated with right hemisphere brain damage. *Am. J. Speech-Language Pathol.* **16**, 331–342 (2007).
6. 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*, 3780–3784 (2021).
7. Liu, Y. *et al.* RoBERTa: A robustly optimized BERT pretraining approach. Preprint at https://arXiv.org/abs/1907.11692 (2019).
8. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst. (NeurIPS)*. **33**, 12449–12460 (2020).
9. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *Proceedings of the International Conference on Machine Learning (ICML)*, 19730–19742 (2023).
10. Themistocleous, C. & Stark, B. C. Language biomarker screening using AI: a transdiagnostic approach to the brain. *Sci. Rep.* **16**, 4207 (2026).
11. Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F. & Novikova, J. Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. *Front. Aging Neurosci.* **13**, 635945 (2021).

12. König, A. et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **1**, 112–124 (2015).
13. Ehghaghi, M., Rudzicz, F. & Novikova, J. Data-driven approach to differentiating between depression and dementia from noisy speech and language data. In: *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, 24–37 (2022).
14. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*, 4171–4186 (2019).
15. Alsentzer, E. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop* 72–78 (2019).
16. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
17. Radford, A. Learning transferable visual models from natural language supervision. In: *Proceedings of the International Conference on Machine Learning (ICML)* 8748–8763 (2021).
18. Mosbach, M., Andriushchenko, M. & Klakow, D. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In: *International Conference on Learning Representations (ICLR)* (2021).
19. Gururangan, S. Don't stop pretraining: Adapt language models to domains and tasks. In: *Annual Meeting of the Association for Computational Linguistics (ACL)* (2020).
20. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Tran. Pattern Analysis Mach. Intell.* **41**, 423–443 (2019).
21. Tsai, Y..-H. H. Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* 6558–6569 (2019).
22. Lu, J., Batra, D., Parikh, D. & Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32 (2019).
23. Zhang, Y. et al. Integrating large language model, EEG, and eye-tracking for word-level neural state classification in reading comprehension. *IEEE Trans. Neural Syst. Rehab. Eng.* **32**, 3465–3475. https://doi.org/10.1109/TNSRE.2024.3435460 (2024).
24. Shubi, O., Meiri, Y., Hadar, C. A. & Berzak, Y. Fine-grained prediction of reading comprehension from eye movements. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3372–3391, https://doi.org/10.18653/v1/2024.emnlp-main.198 (2024).
25. Quach, B. M., Gurrin, C. & Healy, G. DERCo: A dataset for human behaviour in reading comprehension using EEG. *Sci. Data* **11**, 1104. https://doi.org/10.1038/s41597-024-03915-8 (2024).
26. Schuller, B. et al. Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Transactions on Affective Computing* **1**, 119–131 (2010).
27. Coelho, C. A. Story narratives of adults with closed head injury and non-brain-injured adults: Influence of socioeconomic status, elicitation task, and executive functioning. *J. Speech, Language, Hearing Res.* **45**, 1232–1248 (2002).
28. Minga, J., Johnson, M., Blake, M. L., Fromm, D. & MacWhinney, B. Making sense of right hemisphere discourse using RHDBank. *Topics Language Disord.* **41**, 99–122 (2021).
29. Lanzi, A. M. et al. DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *Am. J. Speech-Language Pathol.* **32**, 426–438 (2023).
30. MacWhinney, B. & Snow, C. The Child Language Data Exchange System: An update. *J. Child Language* **17**, 457–472 (1990).
31. Radford, A. Robust speech recognition via large-scale weak supervision. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2023).
32. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2980–2988 (2017).
33. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).
34. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall/CRC, New York, 1994).
35. Kertesz, A. *Western Aphasia Battery* (Grune & Stratton, New York, 1982).
36. Elbourn, E., Kenny, B., Power, E. & Togher, L. Psychosocial outcomes of severe traumatic brain injury in relation to discourse recovery: A longitudinal study up to 1 year post-injury. *Am. J. Speech-Language Pathol.* **28**, 1463–1478 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Byounghwa Lee[1] · Hwa Jeon Song[1] · Byung Ok Kang[1]

✉ Byounghwa Lee
byounghwa.lee@etri.re.kr

[1] Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea