HINT-AD: HIERARCHICAL AND INTERPRETABLE TRANSFORMER FOR SPEECH-BASED ALZHEIMER'S DISEASE STAGING AND MARKER DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate staging within the mild cognitive impairment (MCI) continuum is challenging because progression from early MCI (eMCI) to late MCI (lMCI) manifests through subtle, interacting changes in speech, language, and cognition. Can eMCI and lMCI stages be distinguished from connected speech by explicitly modeling the hierarchy of cognitive-linguistic disruption while retaining interpretable evidence for each prediction? We introduce HINT-AD, a hierarchically interpretable Transformer that integrates phonological, syntactic, semantic, and cognitive-linguistic markers through marker-guided Transformer stages for eMCI–lMCI staging. Rather than treating heterogeneous markers as independent features, HINT-AD progressively models their interactions while retaining links from marker summaries to supporting acoustic representations and transcript spans.

Across 420 English and Chinese picture-description recordings from 218 participants, using MCI-stage labels transferred from the Alzheimer's Disease Neuroimaging Initiative (ADNI) reference cohort, with an additional 89 Spanish passage readings used for auxiliary training only, HINT-AD achieves 80% and 79% ROC-AUC and 69% and 74% average precision (AP) in English and Chinese, respectively. Compared with the strongest baselines, these results correspond to relative improvements of 8.1% and 5.3% in ROC-AUC and 7.8% and 10.4% in AP, respectively. Joint English–Chinese training improves ROC-AUC by 1.4% and 8.3% relative to language-specific training, while joint training with Spanish auxiliary data yields relative gains of 14.3% and 9.7% over the same baselines, respectively. Marker-group ablations further show that masking any linguistic level reduces performance in both languages, with the largest ROC-AUC reductions arising from semantics in English and cognition in Chinese. SHAP analysis identifies semantic markers as the strongest positive contributors in both languages, including role identification and lexical use, with additional language-specific contributions from scene integration and contrast use in English and thematic coherence and verb use in Chinese.

Together, these findings provide evidence that hierarchical modeling of connected-speech markers improves fine-grained staging within the MCI continuum while exposing interpretable cognitive-linguistic patterns associated with stage. HINT-AD reframes speech-based within-MCI staging from flat classification over heterogeneous biomarkers to evidence-linked hierarchical representation learning, offering a clinically motivated and interpretable framework for computational modeling of subtle neurocognitive change.

1 INTRODUCTION

Distinguishing early mild cognitive impairment (*e*MCI) from late MCI (*l*MCI) requires sensitivity to subtle changes within the MCI continuum. This distinction is relevant to assessment and cohort stratification in Alzheimer's disease (AD) research (Edmonds et al., 2019; Kasper et al., 2020). Connected speech provides a non-invasive source of evidence: a single narrative engages word retrieval, grammatical construction, acoustic fluency, and discourse organization (Eyigoz et al., 2020;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Clarke et al., 2021). The learning problem is to integrate these interacting signals while retaining interpretable evidence for each staging prediction.

These signals span linguistic levels and temporal scales. Pauses before content words can accompany lexical retrieval difficulty; sentence structure reflects grammatical organization; and omissions or inconsistent references affect the narrative as a whole (De Looze et al., 2021; Cho et al., 2022; Gao & He, 2024; Drummond et al., 2015). Pooled text–audio representations compress these changes into global features, while marker-score vectors summarize linguistic phenomena separately from their speech context. HINT-AD connects marker measurements to their supporting evidence and models interactions across levels.

Pretrained language and speech encoders support AD classification through single-modality and multimodal representations (Balagopalan et al., 2020; Qin et al., 2021; Cai et al., 2023; Ying et al., 2023). LLM-assisted methods provide linguistic descriptions (Mo et al., 2024; 2025), hierarchical speech models organize acoustic representations at multiple granularities (Chen et al., 2023; Dong et al., 2024), and cross-linguistic staging methods combine transcript generation with contrastive learning (Han et al., 2025b). We connect these directions by using linguistic markers to guide the integration of aligned acoustic and textual evidence through a shared hierarchy.

We introduce **HINT-AD**, a **H**ierarchical and **IN**terpretable **T**ransformer for within-MCI staging. Marker-specific queries retrieve acoustic representations or transcript spans, and successive phonological, syntactic, semantic, and cognitive-linguistic stages integrate the resulting summaries. Later stages retain earlier summaries, allowing higher-order representations to use lower-level acoustic-linguistic evidence. Final fusion combines these summaries with transcript and sentence-audio representations. The architecture preserves marker-to-evidence links and accommodates language-specific marker sets within the shared hierarchy.

We evaluate HINT-AD on 420 English and Chinese picture-description recordings from 218 participants, using ADNI-transferred stage labels and a participant-disjoint test set. An additional 89 Spanish passage readings provide auxiliary training data only. Our contributions are:

- **Evidence-linked hierarchical representation.** HINT-AD integrates four linguistic levels through marker-guided Transformer stages, modeling their interactions while retaining the acoustic and textual evidence supporting each summary.
- **Improved multilingual within-MCI staging.** HINT-AD achieves 80% and 79% ROC-AUC in English and Chinese, relative gains of 8.1% and 5.3% over the strongest baselines. Average precision (AP) improves by 7.8% and 10.4%, respectively. Joint and Spanish auxiliary training each improve both metrics.
- **Interpretable cognitive-linguistic patterns.** Marker-group ablations assess all four linguistic levels, and held-out SHAP analyses identify shared and language-specific marker contributions, with semantic markers providing the strongest positive attributions in both languages.

2 Related Work

Speech-based cognitive assessment. Reviews of speech-based AD assessment describe acoustic, lexical, syntactic, and discourse features, often elicited through picture description (De la Fuente Garcia et al., 2020; Petti et al., 2020; Ding et al., 2024). Earlier systems used handcrafted linguistic and acoustic features (Fraser et al., 2016; Ammar & Ayed, 2019); subsequent models learned text and audio representations (Balagopalan et al., 2020; Qin et al., 2021; Cai et al., 2023; Ying et al., 2023). HINT-AD studies the finer *e*MCI–*l*MCI distinction, organizing multimodal evidence around interpretable linguistic markers.

Linguistic markers and hierarchical representations. Studies of cognitive decline identify changes in speech timing, phonological fluency, lexical retrieval, and verb processing (De Looze et al., 2021; Olmos-Villaseñor et al., 2023; Pekkala et al., 2013; Williams et al., 2021). LLM-assisted marker extraction (Mo et al., 2024) and DECT (Mo et al., 2025) incorporate fine-grained linguistic knowledge into AD classification. SpeechFormer++ (Chen et al., 2023) and HAFFormer (Dong et al., 2024) organize acoustic computation hierarchically. HINT-AD combines marker-based and

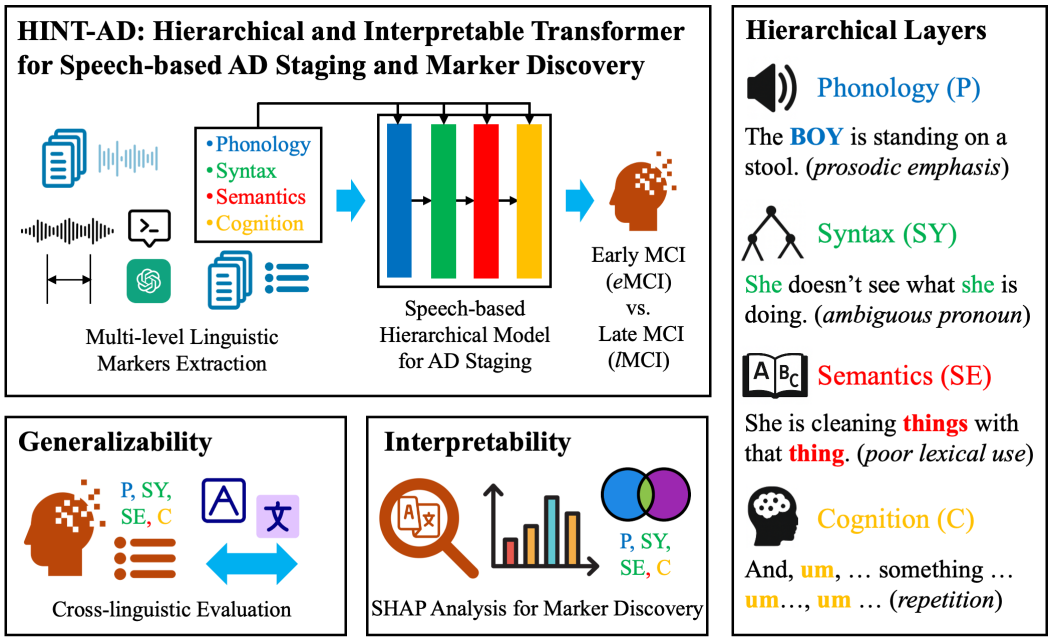


Figure 1: **Overview of HINT-AD.** Connected speech is represented through transcript and audio encoders together with four levels of interpretable linguistic markers. Marker-specific evidence is integrated progressively from phonology to syntax, semantics, and cognition before final multimodal prediction. The hierarchy is shared across languages, while higher-level marker sets can remain language-specific.

hierarchical modeling: marker scores form queries, aligned observations provide evidence, and successive linguistic stages integrate the resulting summaries before classification.

Cross-linguistic modeling. DementiaBank (Lanzi et al., 2023) and TAUKADIAL (Luz et al., 2024) support research across speech cohorts and languages. Prior work explores cross-language dementia recognition and transfer learning (Bertini et al., 2022; Pérez-Toro et al., 2022; Chen & Jin, 2024), as well as LLM-based MCI detection (Han et al., 2025a). Closest to our task, Han et al. (2025b) combine generated multilingual transcripts with a contrastive temporal–spatial model spanning several cognitive-status groups. HINT-AD uses recorded speech and its transcripts for within-MCI staging. We evaluate joint English–Chinese learning and Spanish auxiliary training separately, reporting performance for each evaluation language.

3 HINT-AD

3.1 Task and Representation

Given a connected-speech recording and its transcript, HINT-AD predicts *e*MCI or *l*MCI. Its inputs are transcript tokens X, sentence-level audio embeddings A, 12 acoustic features P, and structured marker measurements with supporting evidence M. Four stages integrate phonological, syntactic, semantic, and cognitive-linguistic evidence (Figure 1); we refer to these domains as phonology, syntax, semantics, and cognition. Each stage passes its marker summaries forward so that subsequent stages can model interactions with earlier evidence.

Multilingual BERT (mBERT) (Devlin et al., 2019) encodes the transcript. Qwen2-Audio (Chu et al., 2024) provides audio embeddings pooled at sentence boundaries identified through forced alignment. Acoustic features measure timing, disfluency, and prosodic variation. Higher-level markers use transcript-level preservation scores and supporting spans to describe syntax, semantics, and cognition. These measurements and evidence links guide the hierarchical representation.

Appendix A (Figure 2) details marker queries, evidence retrieval, summary accumulation, and final fusion.

3.2 Linguistic Markers and Supporting Evidence

The marker hierarchy spans acoustic realization, grammatical structure, semantic content, and narrative organization. We adapt the categories of Mo et al. (2024) to English and Chinese within-MCI staging. Marker selection used training transcripts; Appendix B describes this development step and lists the definitions.

Phonology. Twelve sentence-level features quantify speech rate, pauses, disfluency, emphasis, local repetition, and pitch, energy, and duration variability (Appendix Table 5). Forced alignment links words to audio intervals; part-of-speech and dependency annotations identify linguistic contexts, such as content words and clause boundaries. The features retain their acoustic measurement interpretation. Appendix C (Figure 3) illustrates word timings and pauses in short English and Chinese excerpts from both MCI stages.

Syntax, semantics, and cognition. English and Chinese each have five markers per higher-level domain. Syntax describes sentence construction and grammatical relations; semantics describes lexical choice, participant roles, and meaning; cognition describes discourse organization, including omission, repetition, and referential consistency. The shared domains accommodate language-specific phenomena, such as English article and preposition use and Chinese word order (Appendix Table 6).

For English and Chinese, an LLM assesses each domain in a separate whole-transcript call. Each marker receives a preservation score in $\{0, 0.25, 0.5, 0.75, 1\}$ and up to five supporting spans with line numbers and local scores. Higher values indicate better-preserved function; for omission, they indicate greater content coverage. Unassessable markers receive an explicit flag, a null score, and no evidence. Spanish passage reading permits LLM scoring of nonfunctional repetition only; other higher-level markers remain unassessable, while all 12 acoustic features are retained. HINT-AD learns staging predictions from these annotations together with the speech representations.

Annotations follow a fixed JSON format. We check responses for completeness and validity, regenerate nonconforming responses, and retain accepted annotations for reuse. Three extraction runs are aggregated by within-model consensus. DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) supplies the primary markers; GPT-5.6-Luna (OpenAI, 2026) supplies a second annotation set for the extractor-sensitivity analysis. Appendix D provides the prompt and response format (Figure 4).

3.3 Marker-Guided Hierarchical Computation

HINT-AD assigns two of the first eight mBERT layers to each linguistic stage and uses the final four layers for multimodal fusion. The stages integrate acoustic, grammatical, semantic, and discourse evidence in that order, progressively contextualizing the transcript with marker summaries.

Queries and evidence routing. At stage d, marker measurements form marker-specific queries Q_d. Evidence representations E_d are selected from the current transcript context and projected audio and acoustic features using marker-to-evidence associations. Cross-attention produces marker summaries:

$$Q_d = \mathrm{EmbedScore}_d(M_d), \qquad R_d = \mathrm{CrossAttn}_d(Q_d, E_d), \tag{1}$$

where M_d contains the measurements and evidence metadata for domain d. Phonological queries retrieve aligned acoustic information; higher-level queries retrieve supporting transcript spans and their context. Each summary thus connects a marker measurement to its evidence.

Accumulation across linguistic stages. Each stage retains the summaries from earlier stages. With S_0 empty, the accumulated representation after stage d is

$$S_d = [S_{d-1}; R_d], \qquad d \in \{1, 2, 3, 4\}, \tag{2}$$

where $[\,;\,]$ denotes concatenation. The stage's backbone layers contextualize the transcript with the accumulated summaries, allowing semantic and cognitive-linguistic processing to use earlier acoustic and syntactic evidence. Algorithm 1 in Appendix E gives the forward pass and training loop.

Final fusion and optimization. Let H_4 be the transcript representation after the four linguistic stages and U the projected sentence-audio and acoustic-feature tokens. Final fusion produces

$$Z = \mathrm{Fusion}(H_4, U, S_4), \qquad \ell = W Z_{\texttt{[CLS]}} + b, \tag{3}$$

where ℓ contains the two class logits. Classification uses the contextualized original transcript `[CLS]` token. AdamW minimizes class-weighted cross-entropy, learning how marker summaries and multimodal context jointly support staging.

3.4 Evidence Links and Marker Attributions

Evidence-routing attention links each marker summary to the acoustic representations or transcript spans used to form it. SHAP attributions estimate marker contributions to the prediction, and group-masking ablations measure performance changes when an entire linguistic level is withheld. These analyses connect supporting observations with marker-level and domain-level predictive contributions.

We apply grouped KernelSHAP (Lundberg & Lee, 2017) to the held-out test set using 256 coalitions per recording and fold. Each fold model's raw *e*MCI probability is attributed with transcript words and sentence audio fixed; averaging fold attributions explains the ensemble probability. Reference values are 0.5 for higher-level preservation scores and the language-specific training median for acoustic features. We first average recordings within participants, then summarize participants by language. Positive SHAP values raise the *e*MCI probability relative to the reference; signed means describe direction, and mean absolute values describe magnitude.

4 Experimental Setup

4.1 Cohorts, Staging Labels, and Data Split

The evaluation data comprise MCI picture-description recordings from three cohorts. Pitt (Becker et al., 1994) provides English descriptions, and TAUKADIAL (Luz et al., 2024) provides English and Chinese descriptions; both are accessed through DementiaBank (Lanzi et al., 2023).[1] PROCESS-2 (Pahar et al., 2026) contributes its English Cookie Theft Description task.[2] Access requires approval from the respective repositories. The Ivanova corpus (Ivanova et al., 2022) supplies Spanish passage readings for auxiliary training only.

Transferred staging labels. We transfer *e*MCI and *l*MCI labels from the Alzheimer's Disease Neuroimaging Initiative (ADNI) reference cohort, whose staging uses education-adjusted Logical Memory criteria (Edmonds et al., 2019). Separate 5-nearest-neighbor classifiers use standardized age, sex, and Mini-Mental State Examination (MMSE) variables for label transfer. English and Spanish recordings use the ADNI White reference subset; Chinese recordings use the Asian reference subset. HINT-AD predicts these ADNI-transferred stages from speech, providing a common staging reference across the speech cohorts.

The English and Chinese data contain 420 picture-description recordings from 218 participants: 137 *e*MCI and 283 *l*MCI recordings. English contributes 288 (104 *e*MCI, 184 *l*MCI), and Chinese contributes 132 (33 *e*MCI, 99 *l*MCI). Cohort counts are 27 for Pitt, 108 for PROCESS-2, and 285 for TAUKADIAL. A fixed participant-disjoint 80–20 split, stratified by stage, cohort, and language, assigns 336 recordings from 175 participants to training and validation and 84 recordings from 43 participants to testing. Separately, 89 Spanish passage readings from 89 participants (19 *e*MCI, 70 *l*MCI) enter auxiliary training only.

4.2 Speech Preprocessing

Qwen3-ASR-1.7B (Shi et al., 2026) transcribes all three languages. A rule-based procedure removes examiner interventions; we audit removed segments and retained speech for every prepared recording.

[1] DementiaBank: `https://talkbank.org/dementia/access.html`.

[2] PROCESS-2: `https://huggingface.co/datasets/CognoSpeak/PROCESS-2`.

The Montreal Forced Aligner (McAuliffe et al., 2017) supplies word-level audio–text boundaries for duration and pause measurements (Rousso et al., 2024).

Stanza (Qi et al., 2020) supplies English and Spanish part-of-speech and dependency annotations; HanLP (He & Choi, 2021) supplies the Chinese annotations. The 12 sentence-level acoustic features use aligned words, inserted pause intervals, linguistic annotations, and z-normalized word-level pitch, energy, and duration. Section 3.2 describes extraction of the eligible higher-level markers.

4.3 Comparisons and Training Protocol

Baselines and ablations. We evaluate six baselines: Qwen2-Audio, mBERT, concatenated pooled mBERT and Qwen2-Audio representations (flat fusion), a marker-only multilayer perceptron (MLP), SpeechFormer++, and HAFFormer. All audio-based models use Qwen2-Audio embeddings, and the hierarchical speech baselines are adapted to binary staging. The marker MLP uses the same four-domain markers without evidence routing. Ablations mask one marker group during both training and evaluation while retaining all backbone layers.

Joint and language-specific training. Joint models pool English and Chinese training data and use class-weighted cross-entropy. Auxiliary runs add all 89 Spanish passage readings to each fold's fitting subset, leaving English and Chinese validation and test membership unchanged. Language-specific HINT-AD models use only English or Chinese training data, without Spanish, and are tested on the corresponding held-out subset. All models in the primary performance and marker-ablation comparisons follow the full HINT-AD auxiliary-training policy.

Model selection and reporting. Five-fold stratified group cross-validation operates within the training partition, with participants disjoint across folds. Joint HINT-AD checkpoints maximize mean English–Chinese validation ROC-AUC, with mean AP breaking ties; language-specific runs use their corresponding metrics. The five selected models predict the same fixed held-out test set. We average their raw *e*MCI probabilities per recording before computing ROC-AUC and AP separately for each test language. AP complements ROC-AUC under class imbalance (Saito & Rehmsmeier, 2015).

Performance tables retain absolute scores on the 0–1 scale to two decimal places. Tables 1–3 also report relative percentage changes, $\Delta_{\mathrm{rel}} = 100(s - s_{\mathrm{ref}})/s_{\mathrm{ref}}$, with each reference specified in the caption. Relative changes are computed from the reported rounded scores and rounded to one decimal place.

HINT-AD uses a learning rate of 2×10^{-5}, batch size 8, and at most 10 epochs, with 10% linear warmup and early-stopping patience 3. Appendix F reports the remaining settings and computing environment.

5 Experimental Results

5.1 Held-Out Within-MCI Discrimination

HINT-AD achieves the highest ROC-AUC and AP among the evaluated models in both languages (Table 1). English ROC-AUC reaches 80%, an 8.1% relative improvement over the marker MLP; AP reaches 69%, a 7.8% improvement over flat fusion. Chinese ROC-AUC and AP reach 79% and 74%, relative improvements of 5.3% and 10.4% over flat fusion. Each comparison uses the strongest baseline for that language and metric.

Flat fusion improves on both single-modality encoders in each language. HINT-AD further improves ROC-AUC over flat fusion by 9.6% in English and 5.3% in Chinese and exceeds the marker MLP on both metrics. The following ablations quantify how each linguistic domain contributes to the complete marker-guided hierarchical system.

5.2 Effects of Joint and Auxiliary Training

Relative to language-specific training, joint English–Chinese training improves ROC-AUC by 1.4% in English and 8.3% in Chinese and AP by 18.2% and 3.4%, respectively (Table 2). Using the same

Table 1: Held-out performance with joint English–Chinese and Spanish auxiliary training. Absolute scores use the 0–1 scale. Underlined values identify the strongest baseline for each language–metric pair; the final row reports HINT-AD's relative gain over that value. DeepSeek-V4.1-Flash supplies the primary markers.

Model	**English**		**Chinese**	
	ROC-AUC	**AP**	**ROC-AUC**	**AP**
Qwen2-Audio	0.66	0.51	0.49	0.34
mBERT	0.63	0.56	0.72	0.50
mBERT + Qwen2-Audio	0.73	<u>0.64</u>	<u>0.75</u>	<u>0.67</u>
Marker MLP	<u>0.74</u>	0.52	0.66	0.49
SpeechFormer++	0.63	0.55	0.59	0.51
HAFFormer	0.68	0.54	0.56	0.48
HINT-AD	**0.80**	**0.69**	**0.79**	**0.74**
Relative gain vs. best baseline	**+8.1%**	**+7.8%**	**+5.3%**	**+10.4%**

language-specific references, joint training with Spanish auxiliary data yields relative gains of 14.3% and 9.7% in ROC-AUC and 25.5% and 27.6% in AP.

Table 2: HINT-AD performance by training regime. Absolute scores use the 0–1 scale; parentheses show relative gains over the corresponding language-specific model for both multilingual settings. Boldface highlights the best result for each language and metric.

Training	**English**		**Chinese**	
	ROC-AUC	**AP**	**ROC-AUC**	**AP**
English only	0.70	0.55	—	—
Chinese only	—	—	0.72	0.58
Joint English–Chinese	0.71 (+1.4%)	0.65 (+18.2%)	0.78 (+8.3%)	0.60 (+3.4%)
Joint + Spanish auxiliary	**0.80 (+14.3%)**	**0.69 (+25.5%)**	**0.79 (+9.7%)**	**0.74 (+27.6%)**

5.3 Contributions of the Four Linguistic Levels

Masking any marker group lowers ROC-AUC and AP in both languages (Table 3). Relative to full HINT-AD, masking semantics produces the largest English ROC-AUC loss (7.5%), and masking phonology produces the largest English AP loss (7.2%). Masking cognition produces the largest Chinese losses in both ROC-AUC (12.7%) and AP (23.0%).

Table 3: Marker-group ablations with joint English–Chinese and Spanish auxiliary training. Absolute scores use the 0–1 scale; parentheses show relative changes from full HINT-AD. Negative values indicate losses; the largest loss per metric is bold. Each group is masked during training and evaluation, with all backbone layers retained.

Masked group	**English**		**Chinese**	
	ROC-AUC	**AP**	**ROC-AUC**	**AP**
None (full model)	**0.80**	**0.69**	**0.79**	**0.74**
Phonology	0.76 (−5.0%)	0.64 **(−7.2%)**	0.73 (−7.6%)	0.70 (−5.4%)
Syntax	0.75 (−6.3%)	0.66 (−4.3%)	0.71 (−10.1%)	0.67 (−9.5%)
Semantics	0.74 **(−7.5%)**	0.65 (−5.8%)	0.72 (−8.9%)	0.68 (−8.1%)
Cognition	0.78 (−2.5%)	0.65 (−5.8%)	0.69 **(−12.7%)**	0.57 **(−23.0%)**

The ablations show that each linguistic domain contributes predictive information to the complete system. Their differing magnitudes show that the shared hierarchy captures language-specific patterns of evidence use.

Table 4: Five largest positive and negative participant-balanced signed mean SHAP contributions per language on the held-out test set. Positive values shift raw *e*MCI probability upward; negative values shift it downward. Values are on the 0–1 probability scale.

Rank	Level	Marker	Mean SHAP
English: toward *e*MCI			
1	Semantics	Role identification	+0.00886
2	Semantics	Lexical use	+0.00706
3	Semantics	Scene integration	+0.00639
4	Semantics	Contrast use	+0.00388
5	Cognition	Omission	+0.00188
English: toward *l*MCI			
1	Phonology	Mean pause duration within clause	−0.00182
2	Phonology	Pitch variability	−0.00044
3	Phonology	Clause-final pause duration	−0.00042
4	Phonology	Lexical retrieval variability	−0.00037
5	Phonology	Repeated-word count	−0.00023
Chinese: toward *e*MCI			
1	Semantics	Thematic coherence	+0.00195
2	Semantics	Role identification	+0.00138
3	Semantics	Lexical use	+0.00121
4	Semantics	Verb use	+0.00117
5	Phonology	Clause-final pause duration	+0.00010
Chinese: toward *l*MCI			
1	Phonology	Disfluency count	−0.00015
2	Syntax	Conjunction use	−0.00014
3	Cognition	Repetition	−0.00013
4	Syntax	Sentence structure	−0.00011
5	Syntax	Pronoun use	−0.00010

5.4 Marker Contributions to Held-Out Predictions

Table 4 reports the five largest positive and negative participant-balanced signed mean SHAP values per language. Semantic markers provide the four strongest positive contributions in both languages. English role identification, lexical use, scene integration, and contrast use raise the *e*MCI probability on average relative to the specified marker references.

In English, mean within-clause pause duration has the strongest negative mean attribution (-0.00182 on the probability scale), shifting predictions toward *l*MCI relative to the acoustic reference. In Chinese, thematic coherence has the largest positive contribution, followed by role identification, lexical use, and verb use. Negative Chinese contributions span phonology, syntax, and cognition. These attributions describe the fitted model's prediction shifts relative to its marker references.

The speech markers highlighted by the attribution analysis were reviewed for clinical plausibility by a clinician with expertise in Alzheimer's disease.

5.5 Marker Repeatability and Extractor Sensitivity

For English and Chinese markers, DeepSeek-V4.1-Flash achieves 90.6% three-run agreement, 92.3–96.6% pairwise agreement, and Cohen's κ of 0.885–0.949. GPT-5.6-Luna achieves 81.8% three-run agreement, 87.4–88.3% pairwise agreement, and κ of 0.825–0.837. Spanish follows the separate sparse marker policy defined for passage reading and contributes through auxiliary training.

Repetition and omission scores from both extractors also show the expected associations with transcript-derived repetition and content coverage in English and Chinese (Appendix G, Table 7). Together with repeated-run agreement, these results provide complementary automated evidence for consistency of the repetition and omission annotations across both languages.

With GPT-derived markers, HINT-AD achieves 75% English and 72% Chinese ROC-AUC and 64% AP in both languages (Appendix H, Table 8). DeepSeek-derived markers yield 80% and 79% ROC-

AUC and 69% and 74% AP, respectively. Both annotation sources support within-MCI discrimination; DeepSeek combines higher repeatability with stronger held-out performance and supplies the primary results.

5.6 Interpretation of Speech Markers and Model Behavior

Semantic content and lexical retrieval. Role identification and lexical use contribute positively to the *e*MCI probability in both languages, linking the shared signal to how speakers identify scene participants and express relevant concepts. In longitudinal picture descriptions from autopsy-confirmed AD cases, Ahmed et al. (2013) found that semantic and lexical content tracked linguistic decline from MCI through later disease stages. The Chinese verb-use contribution has related support from Lai & Lin (2013): action fluency and action naming were associated with dementia severity in Chinese-speaking participants. These findings provide a linguistic basis for interpreting the semantic contributions in HINT-AD.

Narrative organization and reference. English scene integration and Chinese thematic coherence capture how details form a connected account. Drummond et al. (2015) found incomplete narratives and irrelevant propositions in amnestic MCI, with broader coherence and referential difficulties in AD. This evidence helps interpret the scene-level markers and the positive English omission attribution: higher omission scores denote better coverage of relevant content. Chinese conjunction use also has a discourse-level context. Lai & Lin (2012) observed reduced frequency, variety, and functional range of discourse markers in Chinese speakers with AD during picture description and personal narratives. These studies support examining both the content conveyed and the relations that organize it.

Speech timing and fluency. English within-clause pauses and Chinese disfluency count contribute toward *l*MCI relative to their acoustic references. De Looze et al. (2021) linked conversational timing to lexico-semantic and executive performance in MCI and AD, connecting temporal organization with the demands of language production. Clause-final pauses have different signed contributions across our two language models, motivating interpretation alongside linguistic context. HINT-AD retains these distinctions through separate pause measures and aligned acoustic and transcript evidence.

Implications for hierarchical modeling and marker discovery. The complementary contributions of meaning, discourse, and timing are consistent with the semantic, syntactic, acoustic, and information-content factors identified in DementiaBank speech by Fraser et al. (2016). In HINT-AD, semantic markers provide the strongest positive signed attributions in both languages, while masking cognition produces the largest Chinese group-ablation loss. SHAP describes prediction shifts in the fitted model; group ablations assess performance after retraining with an entire domain masked. Together with the multilingual training gains, these results support integrating shared linguistic domains while retaining language-specific markers. Auxiliary Spanish passage readings improve picture-description staging in both evaluation languages; because language, task, and sample count change together, this comparison does not isolate which factor drives the gain. The evidence links and attribution rankings prioritize interpretable candidate markers for further study.

6 Conclusion

These experiments show the value of connecting linguistic marker measurements to the speech evidence behind them for within-MCI staging. HINT-AD's four-level representation achieved higher held-out ROC-AUC and AP than the evaluated baselines in English and Chinese under ADNI-transferred stage labels, with further gains from joint-language learning and Spanish auxiliary training. Every marker level contributed in group ablations, although the largest losses arose from semantics in English and cognition in Chinese. SHAP attributions also revealed shared semantic signals, particularly role identification and lexical use, while the retained evidence links allow those signals to be examined in their speech context. The combination of shared gains and different marker effects suggests a useful direction for multilingual staging: preserve a common linguistic structure while allowing language-specific evidence to shape predictions. Longitudinal cohorts, expert assessment of the full marker set, and evaluations in additional languages and elicitation tasks will help establish the clinical and cross-task scope of these findings.

Ethics Statement

This study analyzes existing speech datasets under their access and data-use conditions; no new participants were recruited. Controlled-access recordings remain within their repository data-use conditions. HINT-AD is evaluated as a research framework for within-MCI staging using labels transferred from ADNI reference subsets as described in Section 4.1. Evaluation is reported separately for English and Chinese; Spanish passage readings are used only for auxiliary training.

Reproducibility Statement

Sections 3–4 specify the architecture, marker extraction, cohort composition, participant-disjoint split, and evaluation protocol. Appendices B and D provide marker definitions and the structured extraction prompt; Appendices E and F provide the training algorithm, hyperparameters, and computing environment. All models use the same held-out partition and group-aware validation design. Tables 1–3 specify the reference conditions used to calculate relative changes.

AI Use Statement

AI assistants supported literature search, code implementation, interpretation of results, drafting, and language revision. The authors are responsible for the content of the submitted manuscript.

References

Samrah Ahmed, Anne-Marie F. Haigh, Celeste A. de Jager, and Peter Garrard. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737, 2013. doi: 10.1093/brain/awt269.

Randa Ben Ammar and Yassine Ben Ayed. Evaluation of acoustic features for early diagnosis of Alzheimer disease. In *International Conference on Intelligent Systems Design and Applications*, pp. 172–181. Springer, 2019.

Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. In *Interspeech 2020*, pp. 2167–2171, 2020. doi: 10.21437/Interspeech.2020-2557.

James T Becker, François Boller, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594, 1994. doi: 10.1001/archneur.1994.00540180063015.

Flavio Bertini, Davide Allevi, Gianluca Lutero, Laura Calzà, and Danilo Montesi. A cross-language dementia classifier: a preliminary study. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pp. 438–443. IEEE, 2022. doi: 10.1109/MetroXRAINE54828.2022.9967558.

Hongmin Cai, Xiaoke Huang, Zhengliang Liu, Wenxiong Liao, Haixing Dai, Zihao Wu, Dajiang Zhu, Hui Ren, Quanzheng Li, Tianming Liu, and Xiang Li. Exploring multimodal approaches for Alzheimer's disease detection using patient speech transcript and audio data, 2023. URL https://arxiv.org/abs/2307.02514.

Guanlin Chen and Yun Jin. Cascaded transfer learning strategy for cross-domain Alzheimer's disease recognition through spontaneous speech. In *Proc. Interspeech*, pp. 907–911, 2024. doi: 10.21437/Interspeech.2024-627.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. SpeechFormer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:775–788, 2023. doi: 10.1109/TASLP.2023.3235194.

Sunghye Cho, Katheryn Alexandra Quilico Cousins, Sanjana Shellikeri, Sharon Ash, David John Irwin, Mark Yoffe Liberman, Murray Grossman, and Naomi Nevler. Lexical and acoustic speech features relating to Alzheimer disease pathology. *Neurology*, 99(4):e313–e322, 2022. doi: 10.1212/WNL.0000000000200581.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Natasha Clarke, Thomas R Barrick, and Peter Garrard. A comparison of connected speech tasks for detecting early Alzheimer's disease and mild cognitive impairment using natural language processing and machine learning. *Frontiers in Computer Science*, 3:634360, 2021. doi: 10.3389/fcomp.2021.634360.

Sofia De la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 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. doi: 10.3233/JAD-200888.

Céline De Looze, Amir Dehsarvi, Lisa Crosby, Aisling Vourdanou, Robert F Coen, Brian A Lawlor, and Richard B Reilly. Cognitive and structural correlates of conversational speech timing in mild cognitive impairment and mild-to-moderate Alzheimer's disease: Relevance for early detection approaches. *Frontiers in Aging Neuroscience*, 13:637404, 2021. doi: 10.3389/fnagi.2021.637404.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. doi: 10.18653/v1/N19-1423.

Kewen Ding, Madhu Chetty, Azadeh Noori Hoshyar, Tanusri Bhattacharya, and Britt Klein. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. *Artificial Intelligence Review*, 57(12):325, 2024. doi: 10.1007/s10462-024-10961-6.

Zhongren Dong, Zixing Zhang, Weixiang Xu, Jing Han, Jianjun Ou, and Björn W Schuller. HAF-Former: A hierarchical attention-free framework for Alzheimer's disease detection from spontaneous speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11246–11250. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446795.

Cláudia Drummond, Gabriel Coutinho, Rochele Paz Fonseca, Naima Assunção, Alina Teldeschi, Ricardo de Oliveira-Souza, Jorge Moll, Fernanda Tovar-Moll, and Paulo Mattos. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Frontiers in Aging Neuroscience*, 7:96, 2015. doi: 10.3389/fnagi.2015.00096.

Emily C Edmonds, Carrie R McDonald, Anisa Marshall, Kelsey R Thomas, Joel Eppig, Alexandra J Weigand, Lisa Delano-Wood, Douglas R Galasko, David P Salmon, Mark W Bondi, and Alzheimer's Disease Neuroimaging Initiative. Early versus late MCI: Improved MCI staging using a neuropsychological approach. *Alzheimer's & Dementia*, 15(5):699–708, 2019. doi: 10.1016/j.jalz.2018.12.009.

Elif Eyigoz, Sachin Mathur, Mar Santamaria, Guillermo Cecchi, and Melissa Naylor. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, 28:100583, 2020. doi: 10.1016/j.eclinm.2020.100583.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016. doi: 10.3233/JAD-150520.

Nan Gao and Qingshun He. A dependency distance approach to the syntactic complexity variation in the connected speech of Alzheimer's disease. *Humanities and Social Sciences Communications*, 11(1):995, 2024. doi: 10.1057/s41599-024-03509-0.

Yang Han, Jacqueline CK Lam, Victor OK Li, and Lawrence YL Cheung. A large language model based data generation framework to improve mild cognitive impairment detection sensitivity. *Data & Policy*, 7:e33, 2025a. doi: 10.1017/dap.2025.8.

Yang Han, Jacqueline CK Lam, Victor OK Li, and Lawrence YL Cheung. An LLM-based temporal-spatial data generation and fusion approach for early detection of late onset Alzheimer's disease (LOAD) stagings especially in Chinese and English-speaking populations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14977–14990, 2025b. doi: 10.18653/v1/2025.findings-emnlp.809.

Han He and Jinho D Choi. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5555–5577, 2021. doi: 10.18653/v1/2021.emnlp-main.451.

Olga Ivanova, Juan José G. Meilán, Francisco Martínez-Sánchez, Israel Martínez-Nicolás, Thide E. Llorente, and Nuria Carcavilla González. Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language. *Computer Speech & Language*, 73:101341, 2022. doi: 10.1016/j.csl.2021.101341.

Siegfried Kasper, Christian Bancher, Anne Eckert, Hans Förstl, Lutz Frölich, Jakub Hort, Amos D Korczyn, Reto W Kressig, Oleg Levin, and María Sagrario Manzano Palomo. Management of mild cognitive impairment (MCI): the need for national and international guidelines. *The World Journal of Biological Psychiatry*, 21(8):579–594, 2020. doi: 10.1080/15622975.2019.1696473.

Yi-Hsiu Lai and Yu-Te Lin. Discourse markers produced by Chinese-speaking seniors with and without Alzheimer's disease. *Journal of Pragmatics*, 44(14):1982–2003, 2012. doi: 10.1016/j.pragma.2012.09.002.

Yi-Hsiu Lai and Yu-Te Lin. Factors in action-object semantic disorder for Chinese-speaking persons with or without Alzheimer's disease. *Journal of Neurolinguistics*, 26(2):298–311, 2013. doi: 10.1016/j.jneuroling.2012.10.003.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438, 2023. doi: 10.1044/2022_AJSLP-22-00281.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Saturnino Luz, Sofia De La Fuente Garcia, Fasih Haider, Davida Fromm, Brian MacWhinney, Alyssa Lanzi, Ya-Ning Chang, Chia-Ju Chou, and Yi-Chien Liu. Connected speech-based cognitive assessment in Chinese and English. In *Interspeech 2024*, pp. 947–951, 2024. doi: 10.21437/Interspeech.2024-1807.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, volume 2017, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

Tingyu Mo, Jacqueline CK Lam, Victor OK Li, and Lawrence YL Cheung. Leveraging large language models for identifying interpretable linguistic markers and enhancing Alzheimer's disease diagnostics. *medRxiv*, 2024. doi: 10.1101/2024.08.22.24312463. URL `https://www.medrxiv.org/content/10.1101/2024.08.22.24312463v1`. Preprint, version 1.

Tingyu Mo, Jacqueline CK Lam, Victor OK Li, and Lawrence YL Cheung. DECT: Harnessing LLM-assisted fine-grained linguistic knowledge and label-switched and label-preserved data generation for diagnosis of Alzheimer's disease. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24885–24892, 2025. doi: 10.1609/aaai.v39i23.34671.

Rocio Olmos-Villaseñor, Consuelo Sepulveda-Silva, Teresa Julio-Ramos, Eduardo Fuentes-Lopez, David Toloza-Ramirez, Rodrigo A Santibañez, David A Copland, and Carolina Mendez-Orellana. Phonological and semantic fluency in Alzheimer's disease: A systematic review and meta-analysis. *Journal of Alzheimer's Disease*, 95(1):1–12, 2023. doi: 10.3233/JAD-221272.

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

OpenAI. GPT-5.6 Luna model, 2026. URL `https://developers.openai.com/api/docs/models/gpt-5.6-luna`.

Madhurananda Pahar, Caitlin H. Illingworth, Bahman Mirheidari, Hend Elghazaly, Fritz Peters, Sophie Young, Wing-Zin Leung, Labhpreet Kaur, Daniel Blackburn, and Heidi Christensen. PROCESS-2: A benchmark speech corpus for early cognitive impairment detection. *arXiv preprint arXiv:2605.14888*, 2026. URL https://arxiv.org/abs/2605.14888.

Seija Pekkala, Debra Wiener, Jayandra J Himali, Alexa S Beiser, Loraine K Obler, Yulin Liu, Ann McKee, Sanford Auerbach, Sudha Seshadri, Philip A Wolf, and Rhoda Au. Lexical retrieval in discourse: An early indicator of Alzheimer's dementia. *Clinical linguistics & phonetics*, 27(12): 905–921, 2013. doi: 10.3109/02699206.2013.815278.

Paula Andrea Pérez-Toro, Philipp Klumpp, Abner Hernandez, Tomás Arias-Vergara, Patricia Lillo, Andrea Slachevsky, Adolfo Martín García, Maria Schuster, Andreas K. Maier, Elmar Nöth, and Juan Rafael Orozco-Arroyave. Alzheimer's detection from English to Spanish using acoustic and linguistic embeddings. In *Interspeech 2022*, pp. 2483–2487, 2022. doi: 10.21437/Interspeech.2022-10883.

Ulla Petti, Simon Baker, and Anna Korhonen. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797, 2020.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108, 2020. doi: 10.18653/v1/2020.acl-demos.14.

Ying Qin, Wei Liu, Zhiyuan Peng, Si-Ioi Ng, Jingyu Li, Haibo Hu, and Tan Lee. Exploiting pre-trained ASR models for Alzheimer's disease recognition through spontaneous speech. In *Proceedings of NCMMSC 2021*, 2021. URL https://arxiv.org/abs/2110.01493.

Rotem Rousso, Eyal Cohen, Joseph Keshet, and Eleanor Chodroff. Tradition or innovation: A comparison of modern ASR methods for forced alignment. *arXiv preprint arXiv:2406.19363*, 2024.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432, 2015.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report, 2026. URL https://arxiv.org/abs/2601.21337.

Eric Williams, Megan McAuliffe, and Catherine Theys. Language changes in Alzheimer's disease: A systematic review of verb processing. *Brain and Language*, 223:105041, 2021. doi: 10.1016/j.bandl.2021.105041.

Yangwei Ying, Tao Yang, and Hong Zhou. Multimodal fusion for Alzheimer's disease recognition. *Applied Intelligence*, 53(12):16029–16040, 2023. doi: 10.1007/s10489-022-04255-z.

APPENDIX: SUPPLEMENTARY MATERIAL

A TECHNICAL OVERVIEW AND WORKING MECHANISM

Figure 2 complements Figure 1 with the full architecture and an expanded marker-guided stage. It shows how marker-specific queries retrieve evidence, how summaries accumulate across linguistic levels, and how final fusion combines these summaries with transcript and audio representations.

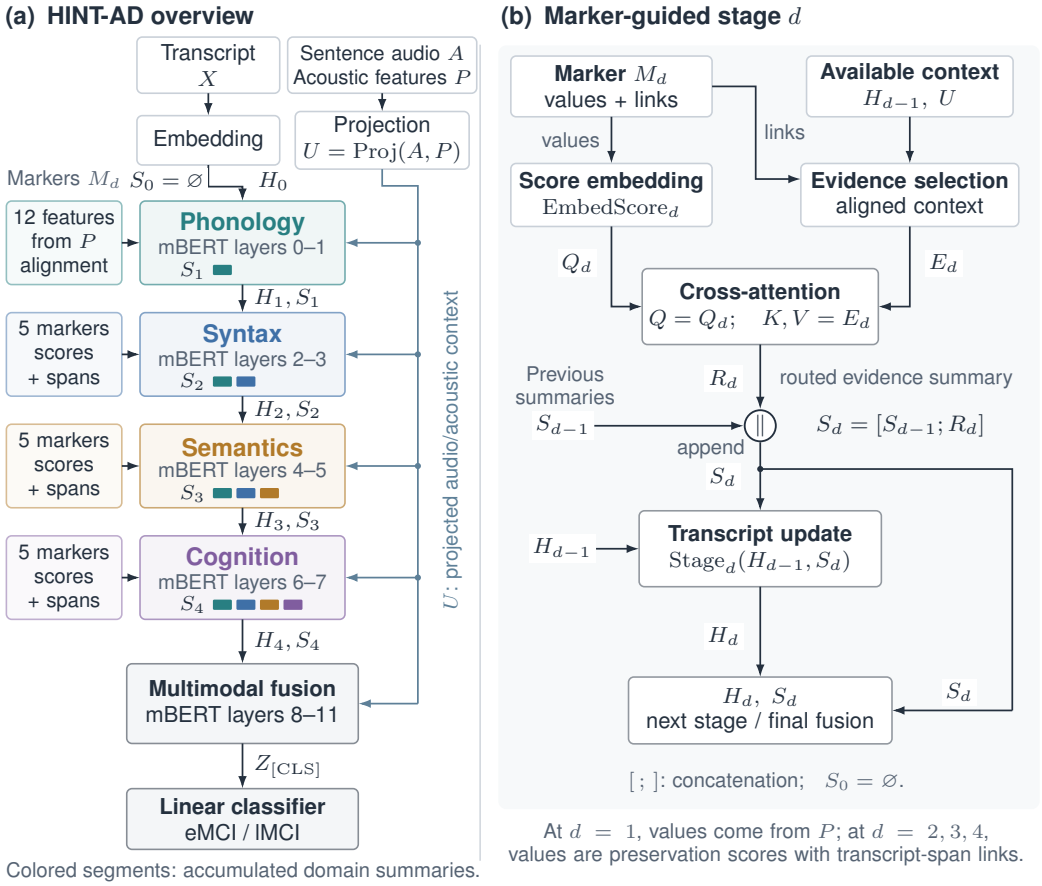


Figure 2: **HINT-AD architecture and marker-guided evidence routing.** (a) Transcript tokens X are embedded as H_0, and sentence-level audio embeddings A and acoustic features P are projected into U. Four linguistic stages pass updated transcript context H_d and accumulated marker summaries S_d forward. Final fusion combines H_4, U, and S_4 before classification from the transcript [CLS] representation. (b) Marker values form queries Q_d; marker links select evidence E_d from the available context. Cross-attention yields R_d, which is concatenated with S_{d-1} before the transcript update, following Algorithm 1. Phonology uses measured acoustic features; higher levels use language-specific preservation scores and transcript spans. The hierarchy is shared across languages. Colored segments denote domain-summary groups, not individual tokens. Layer indices are zero-based.

B SPEECH MARKER DEFINITIONS AND DEVELOPMENT

We adapted the linguistic-marker categories of Mo et al. (2024), originally developed for AD-versus-normal-control classification, into descriptive marker names. For example, "narrative inconsistencies" became "narrative sequencing." Using these categories as a reference, GPT-4o (OpenAI, 2024) summarized syntactic, semantic, and cognitive markers from English and Chinese *e*MCI and *l*MCI training transcripts. The resulting marker sets were then scored using the extraction models specified in Appendix D.

Table 5: Phonological features shared by English and Chinese. Values describe acoustic measurements, not language-preservation scores.

Name	Description
Speech rate	Words per second, excluding pause segments
Mean pause duration within clause	Average pause duration before function words (e.g., pronouns, prepositions)
Clause-final pause duration	Average pause duration following clause boundaries (e.g., punctuation, conjunctions)
Disfluency count	Total count of filler words (e.g., "uh", "um")
Mean pause before content words	Average pause duration preceding content words (nouns, verbs, adjectives, adverbs)
Emphasis on content words	Mean pitch and energy of content words minus the global mean
Lexical retrieval variability	Standard deviation of pause durations before content words
Pause ratio	Proportion of total speaking time spent in pauses
Pitch variability	Standard deviation of pitch across words
Energy variability	Standard deviation of energy across words
Duration variability	Standard deviation of word durations
Repeated-word count	Count of locally repeated words within a sentence

Table 6: Syntactic, semantic, and cognitive marker sets for English and Chinese. Each assessable marker receives a preservation score from 0 to 1.

Domain	English	Chinese
Syntax	Sentence structure; tense use; conjunction use; pronoun use; article and preposition use	Sentence structure; tense or time reference; conjunction use; pronoun use; word order
Semantics	Lexical use; role identification; synonym use; contrast use; scene integration	Lexical use; role identification; synonym use; verb use; thematic coherence
Cognition	Repetition; omission; narrative sequencing; referential use; story structure	Repetition; omission; narrative sequencing; referential use; confusion

The marker-development prompt was: *Syntax/Grammar: e.g., structure, tense, conjunction, object vs. subject, pronoun. Semantics: e.g., meaning, synonyms, antonyms, roles distinctions (e.g., mother vs. sister). Cognition: e.g., memory loss, omissions, confusions (she vs. he), repetitions. Develop speech marker categories to guide syntax/semantics/cognition tokenization layers (with reference to the table for AD vs. NC: [AD vs. NC table]. Categorize speech markers observed in [language] transcripts of picture description tasks (e.g., Cookie Theft) into distinct categories to differentiate early mild cognitive impairment (eMCI) from late MCI (lMCI). Output a table for each level (Syntax/Grammar, Semantics, Cognition).* We also prompted GPT-4o to suggest acoustic features corresponding to the identified categories. These suggestions informed the phonological marker set; the feature values were computed from aligned recordings.

Tables 5 and 6 define the acoustic features and higher-level marker sets. English and Chinese share 12 phonological features and each use five markers per higher-level domain. Higher-level scores measure preservation of function; acoustic values retain their measurement interpretation.

C Illustrative Speech Examples

Figure 3 shows short excerpts of the aligned speech inputs described in Section 4.2.

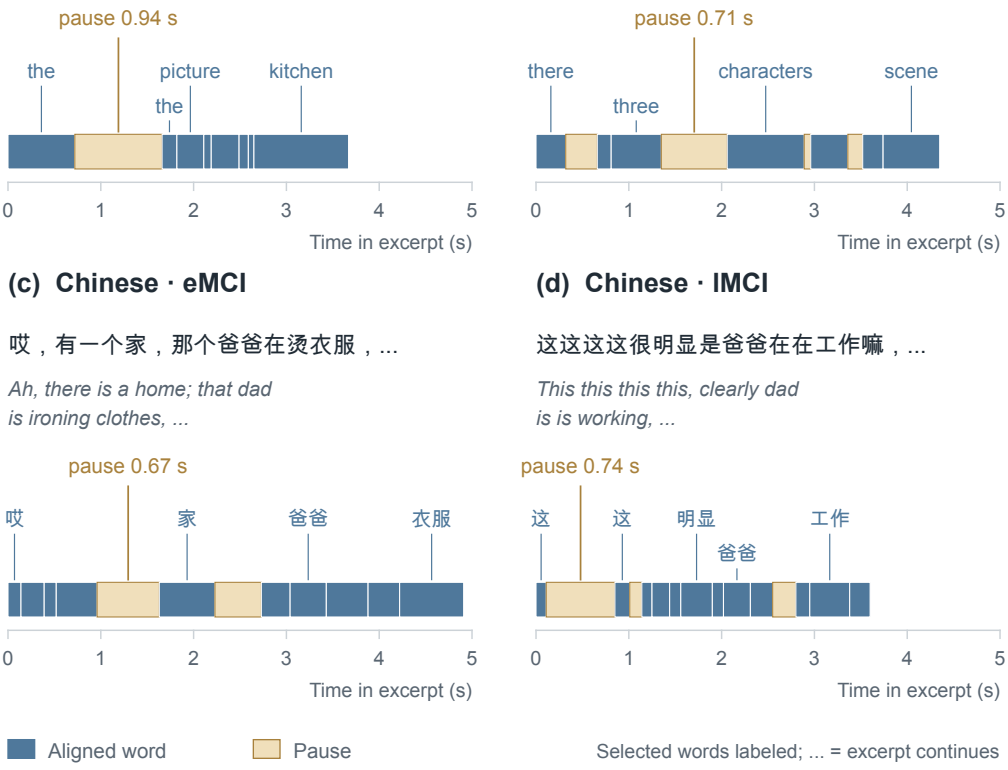


Figure 3: **Illustrative speech excerpts and forced alignment.** Short excerpts from four different participants cover English and Chinese *e*MCI and *l*MCI. Blue blocks show aligned word intervals; shaded blocks show pauses. Selected words and one pause duration per panel are labeled for readability. All panels use the same time scale, with time reset to the first displayed word; ellipses indicate that the original utterance continues. Chinese glosses are provided only as reading aids. These paired text and timing observations illustrate the inputs used to construct HINT-AD's sentence-level representations and acoustic features. Participant identifiers and recording metadata are omitted. The excerpts illustrate input alignment and are not intended to establish characteristic differences between stages.

D Structured Linguistic-Marker Extraction

For English and Chinese, each extraction call assesses one higher-level domain using the full transcript. This preserves narrative context while applying domain-specific marker definitions. The Spanish passage-reading prompt assesses only nonfunctional repetition; other higher-level markers remain null. DeepSeek-V4.1-Flash supplies the primary scores, and GPT-5.6-Luna is used for extractor-sensitivity analysis.

Transcript Presentation. Each transcript is presented as a numbered list of sentences. Numbering supports direct links between a marker score and its supporting transcript spans.

Marker Definitions. For each language, the marker definitions specify the phenomenon to assess and the interpretation of preserved performance. The LLM returns one transcript-level score per marker from $\{0, 0.25, 0.5, 0.75, 1\}$, where larger values indicate better-preserved language function.

It may attach up to five supporting spans, each with a line number, quoted text, and local score. A marker that cannot be assessed is returned with an explicit flag and no score or evidence.

```
System:
You are a careful linguistic annotator of connected speech. Return a JSON response in the requested format.
User:
Score the requested {domain} markers using the full transcript. Higher scores indicate better-preserved
language function.
Transcript:
{numbered_lines}
Return only valid JSON. Do not include markdown, comments, explanations, or extra keys.
Required response format:
- Return exactly one object for every marker in {marker_set}.
- Each object contains: marker, assessable, score, evidence.
- score must be one of 0, 0.25, 0.5, 0.75, or 1 when assessable.
- evidence contains at most five items with line_number, text, and local_score.
- If assessable is false, score is null and evidence is empty.
- Evidence text must be copied from the specified transcript line.
Return the full JSON object now.
```

Figure 4: Condensed prompt template for score-based marker extraction. The numbered transcript, linguistic domain, and language-specific marker set are supplied for each analysis.

Prompt and Response Format. The model returns only the structured JSON response specified in Figure 4. We check completeness and validity, regenerate nonconforming responses, and retain accepted annotations for reuse. Three runs are combined by within-model consensus.

E Training Algorithm

Algorithm 1 summarizes supervised training. Mini-batches contain transcript tokens X, sentence-level audio embeddings A, acoustic features P, marker measurements and evidence M, and stage labels y. Four linguistic stages accumulate marker summaries before final fusion with transcript and audio representations. Checkpoint selection follows Section 4.3.

Algorithm 1 HINT-AD Model Training

Require: Transcripts X; sentence audio A; acoustic features P; marker scores and evidence M; labels y
Require: Model weights θ; learning rate α; number of epochs N; batch size K
Ensure: Fitted model parameters θ

Initialize model parameters θ
for epoch $= 1$ to N **do**
  for each mini-batch (X, A, P, M, y) **do**
    // Forward pass
    Encode transcript tokens: $H \leftarrow \mathrm{Embed}(X)$
    Project multimodal context: $U \leftarrow \mathrm{Proj}(A, P)$; initialize summaries $S \leftarrow \emptyset$
    for $d \in \{\mathrm{phon}, \mathrm{syn}, \mathrm{sem}, \mathrm{cog}\}$ **do**
      Create marker queries $Q_d \leftarrow \mathrm{EmbedScore}(M_d)$
      Select marker-aligned evidence $E_d \leftarrow \mathrm{Evidence}(H, U, M_d)$
      Route evidence $R_d \leftarrow \mathrm{CrossAttn}(Q_d, E_d)$ and append R_d to S
      Update transcript context $H \leftarrow \mathrm{Stage}_d(H, S)$
    end for
    Fuse modalities: $Z \leftarrow \mathrm{Fusion}(H, U, S)$
    Compute predictions: $\tilde{y} \leftarrow \mathrm{Classifier}(Z_{\texttt{[CLS]}})$
    // Loss and backpropagation
    Compute loss: $\mathcal{L} = \mathrm{WCE}(\tilde{y}, y; w)$, where WCE denotes class-weighted cross-entropy with class weights w.
    Update parameters θ using AdamW with learning rate α
  end for
end for
return fitted parameters θ; select a checkpoint using validation ROC-AUC

F Training Details and Computing Infrastructure

A fixed random seed determines the participant-disjoint train–test split. All models use the same held-out partition and group-aware validation design.

HINT-AD uses AdamW with a learning rate of 2×10^{-5}, batch size 8, and at most 10 epochs. Training uses class-weighted cross-entropy, a linear schedule with 10% warmup, zero weight decay, and early stopping with patience 3. Joint checkpoints maximize mean English–Chinese validation ROC-AUC with mean AP as a tie-breaker; language-specific checkpoints use the corresponding metrics. SpeechFormer++ and HAFFormer use learning rates of 1×10^{-4} and 2×10^{-3}, respectively, following their original settings (Chen et al., 2023; Dong et al., 2024).

Training used an NVIDIA A100 40 GB GPU on a Linux system through Google Colab. The reported software environment comprises Python 3.12, PyTorch 2.12.0, and Transformers 5.13.0.

DeepSeek-V4.1-Flash was the primary model for LLM-based marker extraction. The GPT-5.6-Luna sensitivity analysis used the same prompts, response checks, retention of accepted annotations, and three-run consensus procedure.

HINT-AD used uncased multilingual BERT,[3] while the flat text baselines used its cased variant.[4] All audio embeddings were obtained from Qwen2-Audio (Chu et al., 2024).[5] English, Chinese, and Spanish transcription used Qwen3-ASR-1.7B (Shi et al., 2026).[6]

G Deterministic Marker Consistency

We compare consensus repetition and omission scores with deterministic transcript measures across the 420 English and Chinese picture-description recordings. Repetition is the fraction of lexical tokens repeated within the preceding three tokens of a sentence. Content coverage is the fraction of task-specific lexical content units found in a transcript for supported picture-description tasks. Because higher marker scores indicate better-preserved function, repetition scores should correlate negatively with repeated-token frequency and omission scores positively with content coverage. Table 7 reports Spearman correlations by language and extractor.

Table 7: Spearman correlations between marker scores and transcript-derived repetition or content coverage.

Language	Marker	DeepSeek	GPT
English	Repetition	-0.43	-0.36
English	Omission	$+0.57$	$+0.54$
Chinese	Repetition	-0.54	-0.47
Chinese	Omission	$+0.68$	$+0.65$

Both extractors show the expected directions in both languages. The omission association remains positive after accounting for task and transcript length (DeepSeek: 0.37 in English and 0.62 in Chinese; GPT: 0.34 and 0.61). These associations support the consistency of repetition and omission scores with independently computed transcript measures across extractors and languages.

[3] https://huggingface.co/google-bert/bert-base-multilingual-uncased

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased

[5] https://huggingface.co/Qwen/Qwen2-Audio-7B

[6] https://huggingface.co/Qwen/Qwen3-ASR-1.7B-hf

H SENSITIVITY TO THE MARKER-EXTRACTION MODEL

Table 8 compares the two marker sources under the same joint English–Chinese and Spanish auxiliary training protocol. DeepSeek-derived markers yield 80% English and 79% Chinese ROC-AUC, compared with 75% and 72% for GPT-derived markers, and higher AP in both languages. Together with the repeatability analysis in Section 5.5, this experiment quantifies sensitivity to the marker-extraction model.

Table 8: Sensitivity to marker-extraction model choice (0–1 scale). Both marker sources are evaluated with joint English–Chinese and Spanish auxiliary training.

Marker model	**English**		**Chinese**	
	ROC-AUC	**AP**	**ROC-AUC**	**AP**
DeepSeek-V4.1-Flash	0.80	0.69	0.79	0.74
GPT-5.6-Luna	0.75	0.64	0.72	0.64