frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 06 November 2025
DOI 10.3389/frai.2025.1669896

Check for updates

OPEN ACCESS

EDITED BY
Nicola Zeni,
University of Bergamo, Italy

REVIEWED BY
Fabio Tamburini,
University of Bologna, Italy
Madhurananda Pahar,
The University of Sheffield, United Kingdom

*CORRESPONDENCE
Maryam Zolnoori
✉ mz2825@cumc.columbia.edu;
✉ m.zolnoori@gmail.com

RECEIVED 20 July 2025
ACCEPTED 13 October 2025
PUBLISHED 06 November 2025

CITATION
Zolnour A, Azadmaleki H, Haghbin Y, Taherinezhad F, Nezhad MJM, Rashidi S, Khani M, Taleban A, Sani SM, Dadkhah M, Noble JM, Bakken S, Yaghoobzadeh Y, Vahabie A-H, Rouhizadeh M and Zolnoori M (2025) LLMCARE: early detection of cognitive impairment via transformer models enhanced by LLM-generated synthetic data.
Front. Artif. Intell. 8:1669896.
doi: 10.3389/frai.2025.1669896

COPYRIGHT
© 2025 Zolnour, Azadmaleki, Haghbin, Taherinezhad, Nezhad, Rashidi, Khani, Taleban, Sani, Dadkhah, Noble, Bakken, Yaghoobzadeh, Vahabie, Rouhizadeh and Zolnoori. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

LLMCARE: early detection of cognitive impairment via transformer models enhanced by LLM-generated synthetic data

Ali Zolnour[1], Hossein Azadmaleki[1], Yasaman Haghbin[1], Fatemeh Taherinezhad[1], Mohamad Javad Momeni Nezhad[1], Sina Rashidi[1], Masoud Khani[2], AmirSajjad Taleban[2], Samin Mahdizadeh Sani[3], Maryam Dadkhah[1], James M. Noble[1,4], Suzanne Bakken[5,6,7], Yadollah Yaghoobzadeh[3], Abdol-Hossein Vahabie[3], Masoud Rouhizadeh[8] and Maryam Zolnoori[1,5,7]*

[1]Columbia University Irving Medical Center, New York, NY, United States, [2]University of Wisconsin-Milwaukee, Milwaukee, WI, United States, [3]School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran, [4]Department of Neurology, Taub Institute for Research on Alzheimer's Disease and The Aging Brain, GH Sergievsky Center, Columbia University, New York, NY, United States, [5]School of Nursing, Columbia University, New York, NY, United States, [6]Department of Biomedical Informatics, Columbia University, New York, NY, United States, [7]Data Science Institute, Columbia University, New York, NY, United States, [8]College of Pharmacy, University of Florida, Gainesville, FL, United States

Background: Alzheimer's disease and related dementias (ADRD) affect nearly five million older adults in the United States, yet more than half remain undiagnosed. Speech-based natural language processing (NLP) provides a scalable approach to identify early cognitive decline by detecting subtle linguistic markers that may precede clinical diagnosis.

Objective: This study aims to develop and evaluate a speech-based screening pipeline that integrates transformer-based embeddings with handcrafted linguistic features, incorporates synthetic augmentation using large language models (LLMs), and benchmarks unimodal and multimodal LLM classifiers. External validation was performed to assess generalizability to an MCI-only cohort.

Methods: Transcripts were obtained from the ADReSSo 2021 benchmark dataset (n = 237; derived from the Pitt Corpus, DementiaBank) and the DementiaBank Delaware corpus (n = 205; clinically diagnosed mild cognitive impairment [MCI] vs. controls). Audio was automatically transcribed using Amazon Web Services Transcribe (general model). Ten transformer models were evaluated under three fine-tuning strategies. A late-fusion model combined embeddings from the best-performing transformer with 110 linguistically derived features. Five LLMs (LLaMA-8B/70B, MedAlpaca-7B, Ministral-8B, GPT-4o) were fine-tuned to generate label-conditioned synthetic speech for data augmentation. Three multimodal LLMs (GPT-4o, Qwen-Omni, Phi-4) were tested in zero-shot and fine-tuned settings.

Results: On the ADReSSo dataset, the fusion model achieved an F1-score of 83.32 (AUC = 89.48), outperforming both transformer-only and linguistic-only baselines. Augmentation with MedAlpaca-7B synthetic speech improved performance to F1 = 85.65 at 2 × scale, whereas higher augmentation volumes reduced gains. Fine-tuning improved unimodal LLM classifiers (e.g., MedAlpaca-7B, F1 = 47.73 → 78.69), while multimodal models demonstrated lower

performance (Phi-4 = 71.59; GPT-4o omni = 67.57). On the Delaware corpus, the pipeline generalized to an MCI-only cohort, with the fusion model plus 1 × MedAlpaca-7B augmentation achieving F1 = 72.82 (AUC = 69.57).
Conclusion: Integrating transformer embeddings with handcrafted linguistic features enhances ADRD detection from speech. Distributionally aligned LLM-generated narratives provide effective but bounded augmentation, while current multimodal models remain limited. Crucially, validation on the Delaware corpus demonstrates that the proposed pipeline generalizes to early-stage impairment, supporting its potential as a scalable approach for clinically relevant early screening. All codes for LLMCARE are publicly available at: GitHub.

KEYWORDS

Alzheimer's disease, mild cognitive impairment (MCI), large language models, data augmentation, transformers, natural language processing

1 Introduction

Alzheimer's disease and related dementias (ADRD) pose a major public health challenge, affecting approximately five million individuals—11% of older adults—in the United States (Alzheimer's Association, 2013; Zolnoori et al., 2023). Despite national efforts, over half of patients remain undiagnosed and untreated (Boise et al., 2004; Tóth et al., 2018; National Institute on Aging, n.d.). With an expected 13.2 million cases by 2050 (Nichols et al., 2017), the National Institute on Aging has prioritized the development of effective screening tools (National Institute on Aging, 2021; National Institute on Aging, n.d.). Meeting this need requires an interdisciplinary approach spanning neuroscience, data science, and speech-language pathology.

One promising direction involves leveraging natural language processing (NLP) to analyze spontaneous speech (Zolnoori et al., 2024a), which can reveal subtle cognitive changes often missed by traditional screening instruments. Early linguistic impairments—such as word-finding difficulties (Meilán et al., 2020; Aramaki et al., 2016), syntactic disorganization (Sung et al., 2020), and reduced fluency (Meilán et al., 2020)—may be detectable through tasks like picture descriptions. Although speech-based screening has shown potential, progress is limited by scarce labeled clinical speech data and poor model generalizability across populations and clinical settings (Rashidi et al., 2025; Zolnoori et al., 2024b).

Transformer-based NLP models—particularly BERT (Devlin et al., 2018) and its variants—capture linguistic context well and have achieved strong results in classifying cognitive impairment in corpora such as DementiaBank (Lanzi et al., 2023). However, variations in fine-tuning protocols, validation sets, and downstream classifiers lead to inconsistent findings on how well these models encode linguistic markers of cognitive decline (see Table 4 in Appendix A as an example of this variation) (Zolnoori et al., 2023). Progress is further constrained by the small size of available speech datasets, which limits both model training and rigorous model validation.

Recent work suggests that large language models (LLMs), such as GPT-4 (OpenAI, 2023), can generate synthetic clinical data resembling real-world datasets. Compared to generative adversarial networks (Goodfellow et al., 2014), LLMs are more accessible and require less technical expertise. Yet, their effectiveness for downstream tasks varies. For instance, synthetic mental health interviews significantly improved ML-based depression detection (Kang et al., 2025), while only marginal gains were observed for named entity recognition in social determinants of health (e.g., Macro-F1 improvement <1%) (Guevara et al., 2024). In autism detection, synthetic data increased recall by 13% but reduced precision by 16% (Woolsey et al., 2024). These mixed results highlight that LLM-generated data must preserve linguistic complexity, align with real data distributions, and support generalization.

Beyond text, emerging multimodal LLMs extend these capabilities by jointly modeling language and audio inputs, enabling them to capture both what is said and how it is said—such as prosody (Peng et al., 2024). These acoustic-linguistic features may be critical in early cognitive impairment detection. However, their application in dementia research remains limited.

This study addresses these gaps through a multi-component design evaluated on two datasets: the ADReSSo 2021 benchmark, which includes participants across a range of cognitive impairment severity (from mild cognitive impairment [MCI] to severe dementia) versus cognitively healthy [controls], and the Delaware corpus, which is restricted to clinically diagnosed MCI versus controls.

1.1 Component 1: developing the screening algorithm

We systematically evaluated BERT-based and newer transformers (e.g., BGE) on the picture-description task to identify the optimal model for encoding linguistic cues. We then combined embeddings from the top-performing model with handcrafted features (e.g., lexical richness) to develop a screening algorithm, hypothesizing that integration would enhance detection accuracy.

1.2 Component 2: leveraging LLMs to generate synthetic speech

We evaluated state-of-the-art LLMs, including open-weight (LLaMA, MedAlpaca, Ministral) and commercial (GPT-4), to assess their ability to learn linguistic markers of cognitive impairment and generate synthetic speech faithful to patient language. We then tested whether augmenting training data with synthetic speech improved screening performance.

1.3 Component 3: evaluation of LLMs as classifiers

We assessed the diagnostic capabilities of LLMs in zero-shot and fine-tuned settings to establish baseline and advanced benchmarks, examining whether model size and training improve classification compared to pre-trained transformers.

1.4 Component 4: evaluating multimodal LLMs for integrated speech and text analysis

We explored whether multimodal LLMs that jointly process linguistic and acoustic inputs improve detection of cognitive impairment compared to text-only models.

1.5 Component 5: validation on the Delaware corpus

Finally, we validated the pipeline on the Delaware dataset, restricted to clinically diagnosed MCI versus controls, to assess generalizability beyond the mixed-severity cohort in ADReSSo 2021 and to test the performance of Components 1–4 in early-stage cognitive impairment screening.

This study makes several contributions to speech-based ADRD detection. We systematically evaluate ten transformer architectures on the ADReSSo 2021 benchmark and show that combining transformer embeddings with 110 linguistic features in a late-fusion design improves generalization. We introduce a controlled augmentation framework using distributionally aligned LLM-generated narratives, which enhances performance without compromising validity. We also benchmark unimodal and multimodal LLMs, demonstrating that well-tuned linguistic transformers remain competitive with large-scale LLMs. Finally, we validate the pipeline on the Delaware dataset, the first evaluation of this approach on clinically diagnosed MCI versus controls, providing evidence of generalizability to early-stage impairment.

2 Method

2.1 Pipeline overview

Figure 1 illustrates the methodology for developing the screening algorithm using the fusion of pre-trained transformer model and handcrafted lexical features, process of synthetic text generation using state-of-the-art LLMs, and measuring the performance of both unimodal and multimodal LLMs as classifiers for ADRD detection.

2.2 Dataset and cohorts

We used the ADReSSo 2021 benchmark dataset, derived from DementiaBank's Pitt Corpus and introduced as a standardized benchmark for dementia detection. The dataset includes 237 participants performing the Cookie-Theft picture description and is divided into an official training set (166 participants) and an official test set (71 participants), balanced for age and gender. For model development, we further split the training portion into 116 training and 50 validation participants using stratified sampling, with all hyperparameter tuning performed on the validation set only. The final results are reported on the official ADReSSo 2021 test set, ensuring comparability with prior work.

Participant characteristics are summarized in Table 1. All individuals underwent comprehensive neuropsychological evaluations, including verbal tasks and the Mini-Mental State Examination (MMSE). Diagnoses were assigned by clinical specialists (neurologists and neuropsychologists) following full clinical assessments. All participants were older than 53 years, and females comprised more than 60% of each group. MMSE scores in the case group ranged from 7–28 (training), 3–27 (validation), and 5–27 (test), reflecting mild to severe cognitive impairment, while scores in the control group ranged from 24–30, consistent with normal cognition. On average, control participants produced more words and had shorter recordings than cases, consistent with expected language production differences in dementia.

2.2.1 Transcription and preprocessing

To avoid reliance on manual transcripts, we re-transcribed the audio data with Amazon Transcribe (general model). We minimized normalization (automatic grammar rewriting disabled), enabled disfluency/hesitation tokens and partial-word emission, and retained word-level timestamps. All study transcripts were produced with this configuration without human edits.

In prior benchmarking on the same audio against Amazon Transcribe (medical), Whisper-Large, and a fine-tuned wav2vec2, the Amazon general model achieved a competitive English Word Error Rate (WER) $\approx 13\%$ and most faithfully preserved verbatim phenomena—fillers ("uh/um"), lexical repetitions, and fragmented/partial words—compared with the alternatives. These cues are central to our feature extraction and fusion classifier.

2.3 Component 1: developing the screening algorithm using pre-trained transformer models and domain-related linguistic features

2.3.1 Linguistic transformers baselines

We systematically evaluated ten transformer models commonly cited in healthcare NLP literature to assess their ability to detect subtle linguistic cues indicative of cognitive decline. Leveraging attention mechanisms, transformers can identify disfluencies such as repetitions, syntactic errors, and filler words—key linguistic cues of cognitive impairment.

Our evaluation included five general-purpose models—BERT, DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), BGE (Xiao et al., 2023), and Longformer (Beltagy et al., 2020)—pretrained on corpora like Wikipedia, BookCorpus, and online health forums. We also tested five domain-specific models—BioBERT (Lee et al., 2020), BioClinicalBERT (Alsentzer et al., 2019), ClinicalBigBird (Li et al., 2022), and BlueBERT (Peng et al., 2019)—trained on biomedical and clinical texts. We hypothesized that domain-specific models may be less sensitive to disfluent,

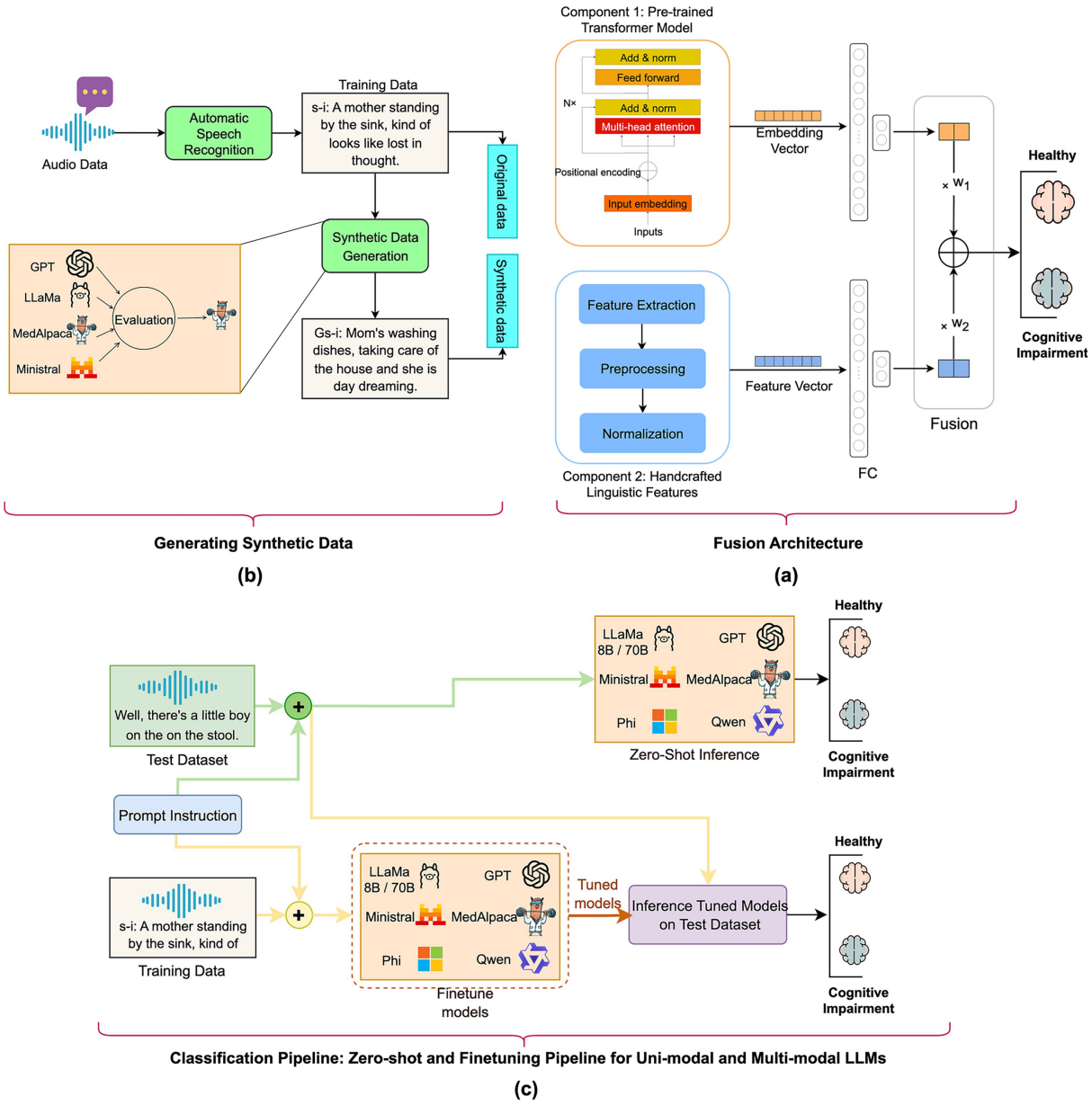

FIGURE 1
Overview of the study's methodology: **(a)** fusion model: automatic speech-recognition transcripts are transformed into deep transformer embeddings and merged with handcrafted linguistic features in a fully connected fusion layer; **(b)** generative synthetic data: GPT-4, LLaMA-8B/70B, MedAlpaca-7B, and Ministral-8B create additional transcripts that mirror the original distribution for augmenting the training dataset; **(c)** classification pipeline—zero-shot and fine-tuning for unimodal and multimodal LLMs: Both unimodal and multimodal LLMs were evaluated in both zero-shot and fine-tuning regimes to assign "cognitively healthy" or "cognitively impaired" labels on a held-out test set.

conversational speech, limiting their ability to capture nuanced impairments.

To evaluate fine-tuning strategies, we tested three configurations: (i) no fine-tuning (frozen transformer as feature extractor), (ii) full fine-tuning (updating parameters of all layers), and (iii) last-layer fine-tuning (updating only the final transformer layer). The frozen model served as a baseline. Although full fine-tuning can boost performance, it risks overfitting on small datasets. Last-layer tuning retains ~90% of the performance gain while preserving generalizable features and reducing computational cost. Fine-tuning intermediate layers (e.g., layers 6–7 in BERT) was not pursued due to minimal added benefit and increased complexity.

Each transformer was paired with a two-layer multilayer perceptron (MLP) classifier. Embeddings were fed into the MLP with 256 hidden units and a 0.4 dropout rate. For both full fine-tuning and last-layer fine-tuning approaches, models were trained using AdamW (batch size = 8, learning rate = 2×10^{-5}, weight decay = 2×10^{-3}) for 50 epochs. We selected the best-performing epoch based on the highest F1-score on the validation dataset. To reduce variance from

TABLE 1 Baseline demographic, cognitive, and speech characteristics of participants across training, validation, and test cohorts.

Attribute	Train		Validation		Test	
	Case	Control	Case	Control	Case	Control
Participants (*N*)	60	56	27	23	35	36
Gender (F/M)	39/21	37/19	19/8	15/8	21/14	23/13
Age (Mean ± Std)	69.33 ± 7.14	66.27 ± 6.81	70.59 ± 6.01	65.48 ± 4.72	68.51 ± 7.12	66.11 ± 6.53
Age range	53–79	54–80	60–80	56–74	56–79	56–78
MMSE (Mean ± Std)	17.80 ± 5.04	29.04 ± 1.13	16.63 ± 5.94	28.87 ± 1.22	18.86 ± 5.8	28.91 ± 1.25
MMSE range	7–28	26–30	3–27	26–30	5–27	24–30
Recording length (s), (Mean ± Std)	87.20 ± 48.35	68.98 ± 25.85	88.52 ± 43.27	68.25 ± 25.43	79.42 ± 36.79	66.35 ± 28.17
Recording length (s), range	35.26–268.49	22.79–168.61	39.91–219.5	26.16–121.47	28.39–150.15	22.35–135.68
Word count (Mean ± Std)	82 ± 43	114 ± 78	101 ± 55	111 ± 43	92 ± 57	111 ± 53
Word count range	22–189	21–523	31–284	54–197	27–256	45–243

Recording length is reported in seconds (s).

random initialization, each experiment was repeated five times with different seeds; Next, we reported the average F1-score on the held-out test set using the best validation epoch.

2.3.2 Handcrafted linguistic features

Transformer embeddings can detect linguistic patterns but often lack transparency. To improve interpretability, we extracted 110 handcrafted lexical features across four dimensions (Zolnoori et al., 2023): (1) lexical richness was measured with established diversity metrics to gauge reliance on high-frequency vocabulary (Paganelli et al., 2003; Fraser et al., 2016; Meteyard et al., 2014); (2) syntactic complexity was assessed through part-of-speech tagging to reflect grammatical structure (Calzà et al., 2021; Khodabakhsh et al., 2015); (3) semantic coherence and fluency were quantified by measuring word repetition and filler words (Nicholas et al., 1985; Tomoeda et al., 1996); (4) psycholinguistic cues were extracted using LIWC 2015, which groups commonly used words into 11 top-level categories (e.g., affective, social, cognition) relevant to cognitive decline (see Appendix B for details of these lexical features) (O'Dea et al., 2021; Burkhardt et al., 2021; Asgari et al., 2017; Collins et al., 2009; Glauser et al., 2020).

We built a lexical feature-based model using 110 lexical features as input and two-layer MLP with 64 hidden neurons, trained with AdamW (learning rate = 8×10^{-3}, weight decay = 1×10^{-3}) for up to 50 epochs. We selected the best-performing epoch based on the highest F1-score on the validation dataset. This allowed us to evaluate the standalone utility of handcrafted features.

2.3.3 Late fusion classifier

To combine the strengths of transformer representations and domain-informed features, we developed a fusion classifier (Figure 1). Embeddings from the top-performing transformer were passed through a two-layer MLP with 256 hidden units; linguistic features entered a separate two-layer MLP with 128 units. We applied a late fusion strategy by combining the two outputs using a learnable weighted sum. The fusion model was trained using AdamW (learning rate = 2×10^{-5}, weight decay = 2×10^{-3}) for 50 epochs. Each experiment was repeated five times with different random seeds. We report the mean and 95% confidence intervals of F1-score and AUC-ROC on the validation and test sets.

2.4 Component 2: LLM-based synthetic text for augmentation

To generate synthetic descriptions reflecting speech of cognitively impaired or cognitively healthy, we adopted a label-conditioned language modeling framework, where each token is generated based on the prior context and the target label. This approach allows LLMs to learn and reproduce label-specific linguistic features—such as repetition or disfluency—that are critical for data augmentation in classification tasks.

$$C = \{(S_i, y_i) | i = 1,2,\ldots,M\},$$

where each sequence $S_i = (w_1^i, w_2^i, \ldots, w_{T_i}^i)$ is paired with a cognitive status label $y_i \in \{Case, Control\}$, our goal is to model the conditional probability

$$P_\theta(S_i | y_i) = \prod_{t=1}^{T_i} P_\theta\left(w_t^i | w_1^i, \ldots, w_{t-1}^i, y_i\right)$$

It is important to note that synthetic transcriptions were generated exclusively for training augmentation; all validation and test evaluations were conducted on real, held-out participants.

2.4.1 LLM models and tuning

We evaluated five LLMs spanning model sizes and training data: LLaMA 3.1 8B Instruct (Grattafiori et al., 2024): Balanced in size and quality, suitable for generating coherent narratives with class-specific variation (e.g., reduced vocabulary); MedAlpaca 7B (Han et al., 2025): A clinically fine-tuned model included to test whether exposure to biomedical language improves generation of patient-like language and terminology alignment; Ministral 8B Instruct (Mistral, 2024): Offers strong sentence-level coherence and low-latency inference, suitable for

generating fluent but compact narratives typical of non-cognitively impaired individuals; LLaMA 3.3 70B Instruct (Grattafiori et al., 2024): Used to evaluate whether increased model capacity improves simulation of complex or disorganized language patterns in cognitively impaired speech; GPT-4o (Hurst et al., 2024) (text-only mode): Used as a benchmark for fluency and coherence, capable of mimicking subtle disfluencies when prompted.

We fine-tuned open-weight LLMs (LLaMA 3.1 8B, MedAlpaca 7B, Ministral 8B, and LLaMA 3.3 70B) using the Quantized Low-Rank Adapter (QLoRA) framework, which inserts lightweight adapters into frozen models to enable memory-efficient training. We tested LoRA ranks of 64 and 128, with scaling factors set to $\alpha = 2 \times$ rank, and applied dropout rates of 0 or 0.1 within adapter layers to mitigate overfitting. For LLaMA 3.1 8B, MedAlpaca 7B, and Ministral 8B, adapters were inserted into all linear layers. For LLaMA 3.3 70B, model weights were quantized to 4-bit precision before fine-tuning, and adapters were placed only in the query, key, and value (QKV) projection layers to reduce memory usage. Other fine-tuning parameters for open-weight LLMs included the PagedAdamW optimizer with mixed-precision (float16) training, a cosine learning rate scheduler, and learning rates of 2e-4 or 1e-4. For GPT-4o, only batch size (16 or 20) and learning rate multipliers (2.5 or 3) were tuned. All LLMs were trained for 10 epochs.

2.4.2 Prompt design and inference

For fine-tuning, we used prompts that incorporated label-specific linguistic cues—for example, "advanced sentence structures" for cognitively healthy (control) participants and "repetition and filler words" for cognitively impaired (case) participants. This improved the model's ability to generate class-consistent outputs. However, relying on a single fixed prompt reduced transcription diversity and limited generalizability. To address this, we created 10 prompt variations that differed in the assigned role (e.g., "language and cognition specialist" vs. "speech pathologist") while keeping task instructions consistent. With 116 training samples, each prompt was applied to about 11–12 samples, ensuring controlled variation without introducing excessive noise (see Figure 2).

For inference, we initially tested prompts that also included label-specific cues, but these produced repetitive and unnatural outputs. To encourage more spontaneous and generalizable speech, we instead used neutral prompts, allowing models to rely on the linguistic patterns learned during fine-tuning (see Figure 2; also for more details about prompt engineering see Appendix C). Inference hyperparameters were tuned to balance coherence and diversity of generated text. We tested values for top-p (0, 0.9, 0.95, −1), top-k (40, 50), and temperature (0.5, 0.7, 0.9, 1.0, 2.0). Optimal settings were: top-p = 0.95, top-k = 50, and temperature = 1 for LLaMA-8B, LLaMA-70B, and MedAlpaca-7B; top-p disabled for Ministral-8B; and temperature = 1 for GPT-4o, consistent with OpenAI's single-parameter guidance.

2.4.3 Synthetic data evaluation and scaling

Evaluation metrics for measuring the quality of the synthetic generated data included:

1. F1-score on validation dataset: For each LLM and fine-tuning epoch, we generated a synthetic dataset ($N = 116$) using the inference prompt. We then retrained the fusion-based screening algorithm on the combined original training data and synthetic data and measured its performance on the validation set. The highest F1-score determined the optimal configuration for each LLM. The optimal configuration for each LLM is presented in Table 5 in Appendix B.
2. BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020): BLEU measured syntactic similarity by computing n-gram overlap ($n = 1$–4) between generated and reference transcriptions in the validation dataset, while BERTScore assessed semantic similarity using contextualized embeddings. These metrics provided additional insight into the extent to which the generated transcriptions preserved structural and semantic properties of original patient speech.
3. t-SNE (van der Maaten and Hinton, 2008) visualization: We applied t-SNE to sentence-level embeddings from synthetic data, original training set, the validation set, and the held-out test set to visualize overlap and distribution similarity in embedding space.

Using the best configuration for each LLM, we generated synthetic data at 1x to 5x the size of the original training set and measured the fusion-based screening model's performance. This assessed whether larger volumes of synthetic text enhanced generalizability while preserving diagnostic cues.

2.5 Component 3: LLMs as classifiers (text-only)

We evaluated whether LLMs—LLaMA (variants), MedAlpaca, Ministral, and GPT-4—can classify transcripts as "Cognitively healthy" or "Cognitively impaired" both without task-specific training (zero-shot) and with fine-tuning.

2.5.1 Zero-shot prompting

To identify an effective prompt, we tested several prompting formulations and selected one that: (a) assigns the model the role of a cognitive-and-language expert; (b) specifies that the input is a transcript of spontaneous speech; (c) instructs a binary decision ("cognitively healthy" vs. "cognitively impaired"); and (d) omits explicit linguistic cues, encouraging the model to rely on its internal reasoning and general language knowledge (see Appendix B for the exact prompt). For inference, open-weight models used temperature = 0 (deterministic outputs), and GPT-4 used temperature = 0.7 per platform guidance.

2.5.2 Fine-tuning

We also fine-tuned each LLM to classify transcripts as Healthy or ADRD. Unlike Component 2, where generation and inference prompts differed, here we used the same prompt during fine-tuning and inference to promote stability and consistency. The hyperparameter search mirrored Component 2. Each model was trained for 10 epochs, and the best checkpoint was chosen by the highest validation F1-score. Final performance was reported on the held-out test set using F1-score, precision, and recall.

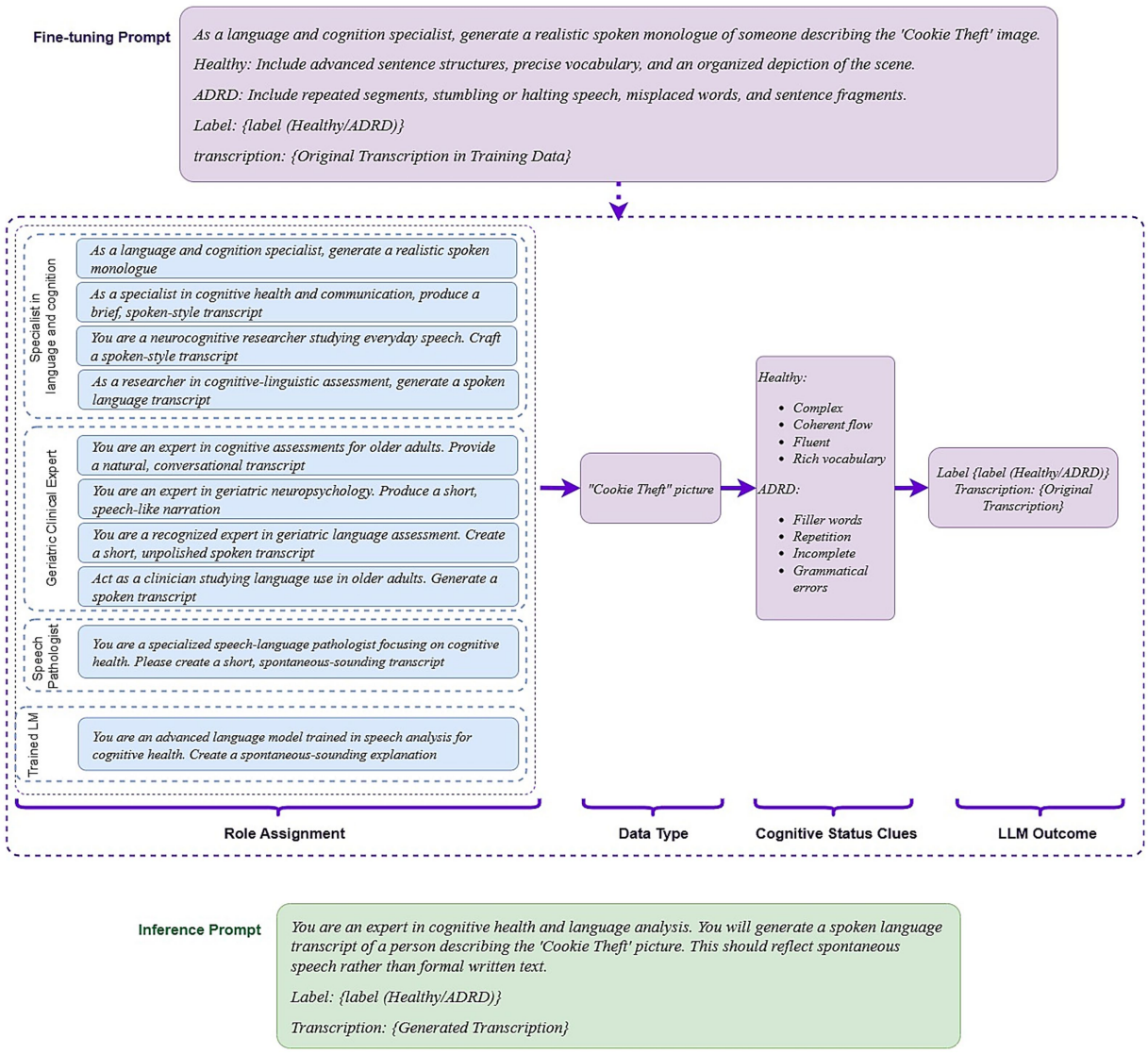

FIGURE 2
Prompt-engineering workflow for synthetic transcript generation and classification. Fine-tuning prompt: A role-specific instruction directs the LLM to describe the Cookie-Theft picture in spoken language, embedding class-defining cues—advanced syntax and fluent flow for cognitively healthy (healthy) speech, repetition and grammatical slips for cognitively impaired (ADRD) speech; Role templates: Ten expert personas (e.g., language-and-cognition specialist, geriatric clinician, speech-language pathologist) provide prompt diversity while the task wording remains constant; data & cues: the prompt explicitly references the cookie-Theft image and the cognitive cues the model should express; output: each generated transcript is saved with its label (Healthy or ADRD); inference prompt: a neutral expert persona requests a transcript label without repeating class hints, encouraging the model to rely on patterns learned during fine-tuning.

2.6 Component 4: evaluating multimodal LLMs as classifiers

We evaluated three state-of-the-art audio–text multimodal models:

1. Qwen 2.5-Omni (Xu et al., 2025) (7B–8.4B parameters): an open-weight "Thinker-Talker" architecture that natively processes text, audio, image, and video, supporting real-time speech responses and full fine-tuning via Hugging Face checkpoints.
2. Phi-4-Multimodal (Abouelenin et al., 2025) (5.6B parameters): Microsoft's successor to the Phi-series, unifying speech, vision, and language encoders into a single network, offering 128 K-token context. We used its open-weight version for domain-specific fine-tuning.
3. GPT-4o ("omni"): OpenAI's flagship closed-weight model with sub-300 ms speech latency, capable of processing any mix of text, audio, image, and video. We tested it only in zero-shot mode due to unavailable API for fine-tuning.

We evaluated Qwen and Phi in both zero-shot and fine-tuning settings, whereas GPT-4o was assessed in zero-shot only. This design allowed comparison of joint acoustic-linguistic modeling against text-only baselines and evaluation of the benefits of multimodal fine-tuning.

2.7 Component 5: external generalizability evaluation—DementiaBank Delaware corpus

We evaluated performance of the pipeline on the DementiaBank Delaware corpus, which includes three picture-description tasks (Cookie Theft, Cat Rescue, Rockwell), a Cinderella story recall, and a procedural discourse task from 205 English-speaking participants (99 MCI, 106 controls). Labels were binary (clinically diagnosed MCI vs. control).

We applied a participant-level split (~60% train: $n = 124$; ~20% validation: $n = 40$; ~20% test: $n = 41$), ensuring recordings from each individual appeared in only one partition. Initial experiments showed that Cat Rescue and Rockwell tasks provided limited discriminatory signals (based on F1 scores on the validation set). We therefore focused on Cookie Theft, Cinderella recall, and procedural discourse, which yielded stronger MCI-detection performance.

2.7.1 Screening algorithm with transformers and linguistic features

(1) Pre-trained Transformer Baselines: We fine-tuned four top-performing transformers on the ADReSSo dataset—BERT, DistilBERT, Longformer, and BioBERT. Each transformer fed embeddings into a two-layer MLP classifier (256 hidden units; dropout = 0.4). Last layer of models was trained with AdamW (batch size = 8, learning rate = 2×10^{-5}, weight decay = 2×10^{-3}) for 50 epochs. The best epoch was selected by the highest F1-score on the validation set.

(2) Fusion of Transformer Embeddings and Handcrafted Linguistic Features: Embeddings from the top-performing transformer were passed through a two-layer MLP (256 hidden units). Handcrafted linguistic features were processed by a separate two-layer MLP (64 hidden units). We applied late fusion via a learnable weighted sum of the two outputs. The fusion model was trained with AdamW (learning rate = 2×10^{-5}, weight decay = 2×10^{-3}) for 50 epochs.

2.7.2 LLM-based synthetic augmentation

(1) Leveraging LLMs to generate synthetic data: we fine-tuned MedAlpaca-7B—the best model identified in ADReSSo—for data augmentation, using the same prompting strategy as for the ADReSSo dataset. MedAlpaca-7B was fine-tuned separately for each task, and synthetic samples were generated per task.

(2) Assessing augmentation effects: we fine-tuned DistilBERT + linguistic features with 1 × and 2 × augmentation to evaluate the impact of LLM-generated synthetic data on MCI detection.

2.7.3 Text-based LLMs (unimodal) and multimodal (audio + text) LLMs classifier

2.7.3.1 Text-based LLMs

Since LLaMA 3.1 8B, LLaMA 3.3 70B, and GPT-4o emerged as the top three text-only LLMs on the ADReSSo 2021 dataset—achieving the best classification performance across both zero-shot and fine-tuning strategies—we selected these models for external evaluation on the Delaware dataset.

2.7.3.2 Multimodal LLMs

For multimodal classification, we opted for GPT-4o (omni) and Phi-4, again applying both zero-shot and fine-tuning strategies, as these models demonstrated strong performance and robust handling of multimodal inputs.

3 Result

3.1 Component 1: developing the screening algorithm using pre-trained transformer models and domain-related linguistic features

3.1.1 Performance of transformer models

Figure 3A reports results for six general-purpose transformers and Figure 3B for four clinical/biomedical transformers, each evaluated under three strategies: no fine-tuning, last-layer fine-tuning, and full fine-tuning. General-purpose models—particularly BERT and DistilBERT—improved substantially with last-layer fine-tuning, with BERT achieving the highest test F1 = 82.76 ± 4.51 on the ADReSSo Benchmark. By contrast, full fine-tuning often degraded performance (notably for RoBERTa and XLNet), consistent with overfitting on a limited dataset. Clinical/biomedical models showed smaller gains across all strategies, and although ClinicalBigBird and BlueBERT benefited modestly from full fine-tuning, their F1 scores remained below those of general-domain models. Overall, these findings suggest that pretraining on general-domain text better captures conversational disfluencies relevant to cognitive impairment than pretraining on structured clinical text.

3.1.2 Performance of handcrafted linguistic features

The classifier using 110 handcrafted linguistic features, capturing lexical richness/diversity, syntactic complexity, discourse fluency, and psycholinguistic categories, performed well on the validation set (F1 = 81.29) but did not generalize to the held-out test set (F1 = 66.83), indicating limited robustness to unseen speakers (Table 2).

3.1.3 Performance of late fusion classifier

Combining fine-tuned BERT embeddings with the same feature set in a late-fusion architecture reduced the generalization gap, achieving F1 = 83.32 and AUC = 89.48 on the test set (validation: F1 = 78.17, AUC = 82.05). Relative to the linguistic-only model, fusion improved test F1 by 24.7% and AUC by 19.6% (Table 2). Compared with BERT alone, fusion yielded a 0.7% increase in F1 with a − 0.6% change in AUC (Table 2).

3.2 Component 2: LLM-generated synthetic text for augmentation

3.2.1 Per-model augmentation

Figure 4A shows validation F1 after retraining the fusion model with synthetic transcripts from each LLM on the ADReSSo Benchmark. MedAlpaca-7B achieved the highest score (81.02), surpassing the baseline fusion model (F1 = 78.17, Table 2). LLaMA-8B and GPT-4 followed, while Ministral-8B provided modest gains and LLaMA-70B

performed lowest (77.21), indicating that increased model capacity did not yield more informative synthetic data. These findings suggest that clinically tuned models such as MedAlpaca-7B generate synthetic speech that more effectively reinforces class-specific linguistic patterns.

Figures 4B,C compare synthetic to real transcripts using semantic similarity (BERTScore) and lexical overlap (BLEU-1–4) for this benchmark. LLaMA-8B achieved the highest BERTScore (0.61), reflecting strong semantic alignment. GPT-4 and LLaMA-70B followed (≈0.58). Although LLaMA-70B attained the highest BLEU-1 (0.87) and BLEU-2 (0.64), this did not translate into improved classification (Figure 4A). MedAlpaca-7B maintained a high BERTScore (0.59) and consistently outperformed GPT-4 on BLEU, indicating that capturing class-specific structure and semantics is more important than surface lexical overlap.

Figure 4D visualizes embeddings via t-SNE. Synthetic samples from MedAlpaca-7B and LLaMA-8B are interspersed with train, validation, and test data, consistent with their strong F1 and alignment metrics on this dataset. In contrast, GPT-4 and LLaMA-70B form distinct clusters, mirroring their lower semantic similarity, while Ministral-8B shows diffuse but less specific overlap. These patterns align with the performance differences observed in.

3.2.2 Scaling augmentation

Figure 5 evaluates the effect of augmentation scale for MedAlpaca-7B (Figure 5A) and GPT-4 (Figure 5B) on the ADReSSo benchmark. For GPT-4, a modest 1 × augmentation yielded a slight improvement in F1 (83.32 ± 2.78 → 84.14 ± 1.92), but performance declined at 2 × (80.76 ± 5.16) and fluctuated between 80–82 thereafter, consistent with t-SNE evidence of drift away from the real-speech manifold. In contrast, MedAlpaca-7B improved at 1 × (85.35 ± 1.96), peaked at 2 × (85.65 ± 1.64), and then declined with larger volumes (3 × = 81.87 ± 3.03; 4 × = 80.45 ± 4.36). These findings suggest that augmentation is beneficial only while synthetic data remains distributionally aligned, with effective limits of approximately 2 × for MedAlpaca-7B and 1 × for GPT-4 (for detailed results of 1 × to 5 × augmentation for both LLMs, see Appendix D).

3.2.3 Operating characteristics before vs. after augmentation

Figure 6 compares the performance of the fusion model before (green) and after two-fold augmentation using MedAlpaca-7B synthetic speech (pink) on ADReSSo benchmark. The ROC curves (Panel a) remain nearly identical, with AUC

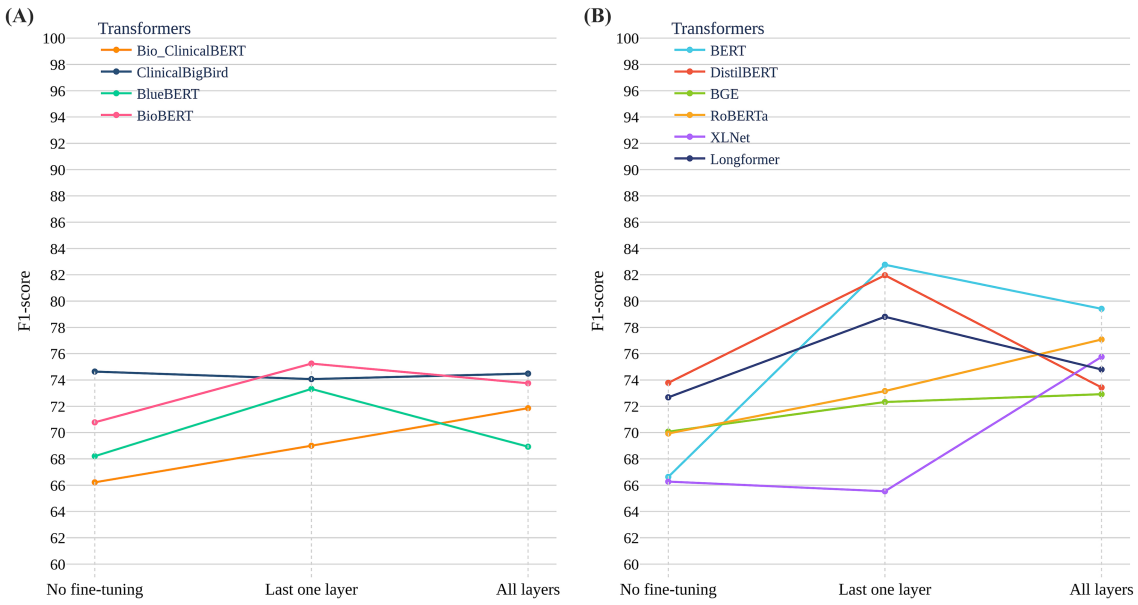

FIGURE 3
Performance of general-purpose and clinical-domain transformer models across fine-tuning strategies on the ADReSSo Benchmark. **(A)** F1-scores of six general-purpose models across three strategies: no fine-tuning, last-layer fine-tuning, and full fine-tuning. BERT and DistilBERT showed the largest gains with last-layer fine-tuning. **(B)** F1-scores of four clinical-domain models pretrained on biomedical and clinical corpora (e.g., PubMed, MIMIC-III). Performance gains were modest across all strategies, with overall F1-scores lower than those of general-purpose models.

TABLE 2 **Performance comparison of BERT, linguistic feature-based, and fusion models on validation and test sets of the ADReSSo benchmark.**

Models	Validation F1, mean ± 95%CI	Validation AUC, mean ± 95%CI	Test F1, mean ± 95%CI	Test AUC, mean ± 95%CI
BERT	78.97 ± 1.84	80.55 ± 1.26	82.76 ± 4.51	90.03 ± 1.29
Linguistic features	81.29 ± 1.16	78.10 ± 1.41	66.83 ± 4.32	74.83 ± 1.48
Fusion model (BERT + linguistic features)	78.17 ± 1.29	82.05 ± 2.52	83.32 ± 2.78	89.48 ± 4.40

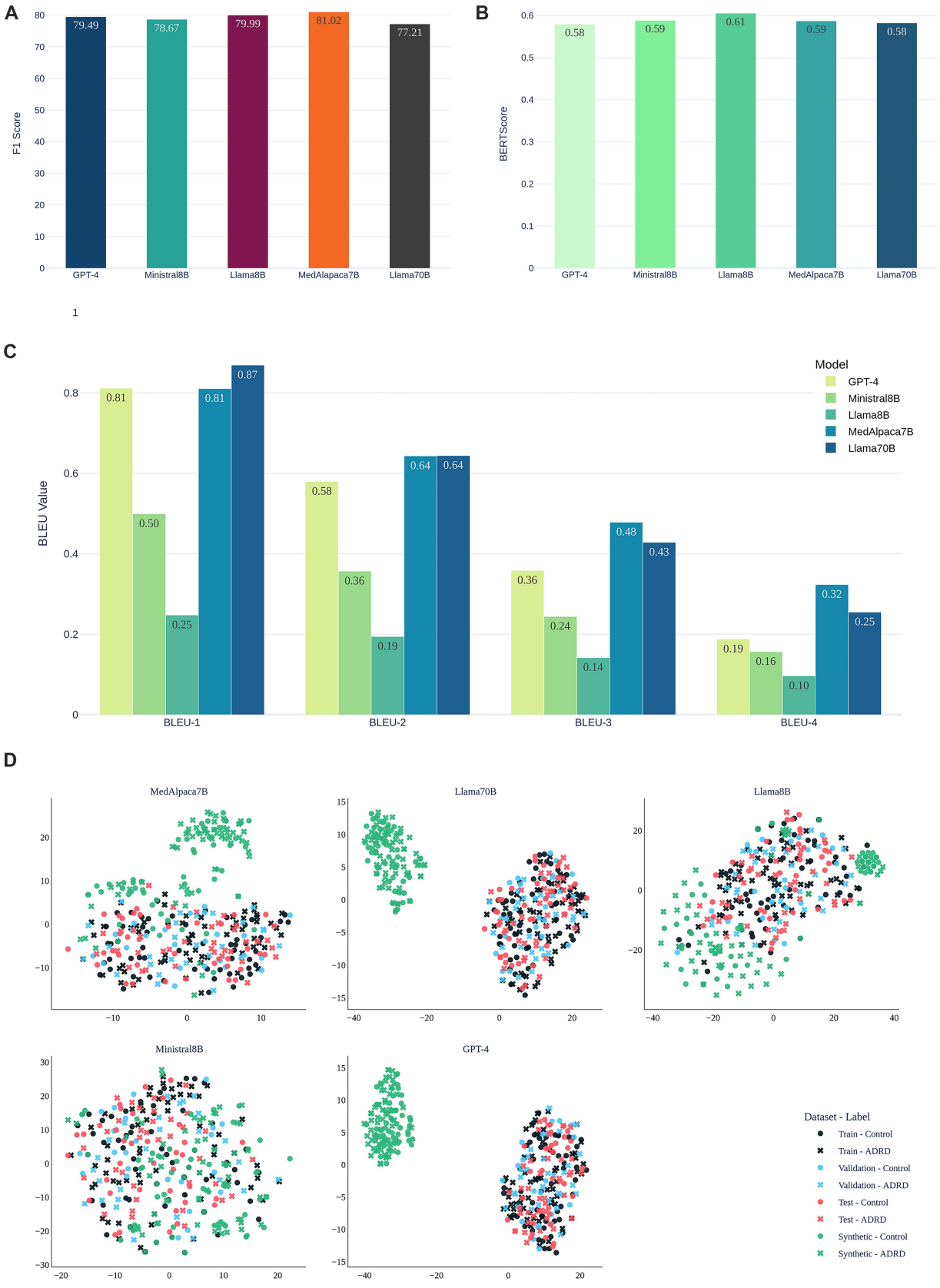

FIGURE 4
Evaluation of synthetic speech generated by LLMs for data augmentation on the ADReSSo benchmark. **(A)** Validation F1-scores of the fusion-based screening model after augmenting the training set with LLM-generated transcripts; **(B)** Semantic similarity of synthetic and human transcripts measured by BERTScore; **(C)** Lexical similarity evaluated using BLEU-1 to BLEU-4 scores; **(D)** visualization of the embedding space of original and synthetic narratives using t-SNE.

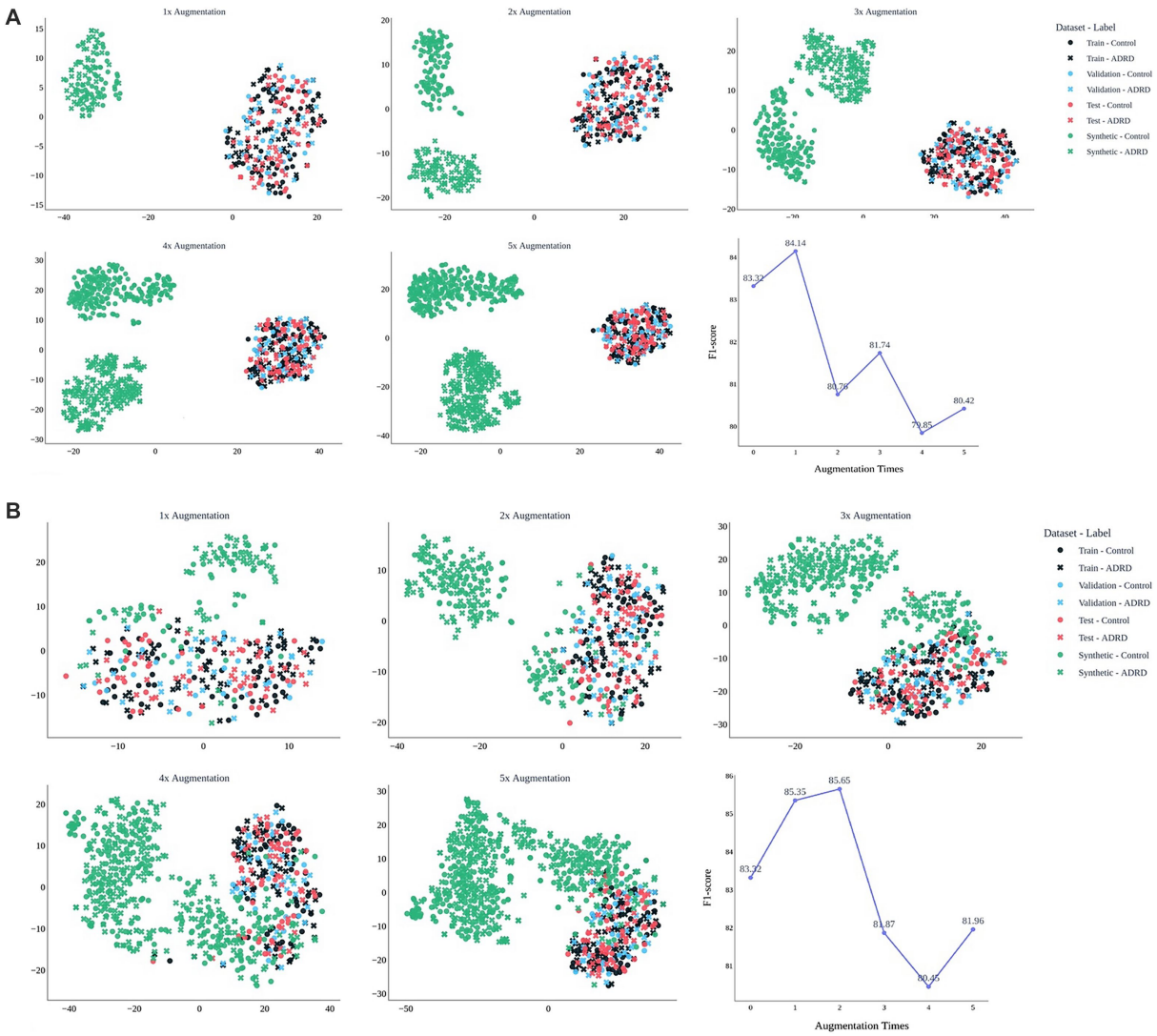

FIGURE 5
Effect of synthetic data volume on embedding structure and model performance. t-SNE plots show how synthetic narratives from MedAlpaca-7B **(A)** and GPT-4 **(B)** integrate with real data across 1 × to 5 × augmentation. MedAlpaca-7B remains aligned up to 2×, supporting peak F1 (85.7), while GPT-4 drifts after 1×, reducing effectiveness. Line plots show corresponding F1-scores of the fusion -based screening model on the held-out test dataset. The results are based on the ADReSSo benchmark.

improving marginally from 89.48 ± 4.40 to 89.56 ± 2.32, indicating stable overall discrimination. The precision–recall curve (Panel b) shows improved precision at lower recall levels, and the cumulative gains curve (Panel c) demonstrates enhanced early retrieval of positive cases, particularly between the 40–70% sample range. Panel d indicates that the positive-predictive-value profiles of the pre- and post-augmentation models overlap across most probability percentiles, with only a slight dip for the augmented model at lower thresholds, while Panel e shows closely matching sensitivity curves, together implying that the added synthetic data left PPV largely intact and fully preserved sensitivity. The prediction density plots further support these findings: before augmentation (Panel f), class distributions overlapped considerably around the decision boundary, whereas after augmentation (Panel g), class 0 (cognitively healthy [control]) and class 1 (cognitively impaired [case]) predictions became more concentrated and more separable, with reduced uncertainty near the 0.5 threshold.

3.3 Components 3 and 4: text-based LLMs (unimodal) and multimodal (audio + text) LLMs classifier

As shown in Table 3, which reports F1-scores with 95% confidence intervals, in the zero-shot setting, the best performance among text-only LLMs came from GPT-4o (F1-score = 73.05), followed closely by LLaMA 3.3 70B (72.93). For multimodal models, GPT-4o (omni) achieved 67.57, and Qwen 2.5 Omni reached 67.31.

With fine-tuning, all text-based LLMs improved. MedAlpaca-7B increased from 47.73 to 78.69 (improvement of 64.9%), and LLaMA 3.1 8B from 68.54 to 81.08 (+18.3%). More modest gains were observed

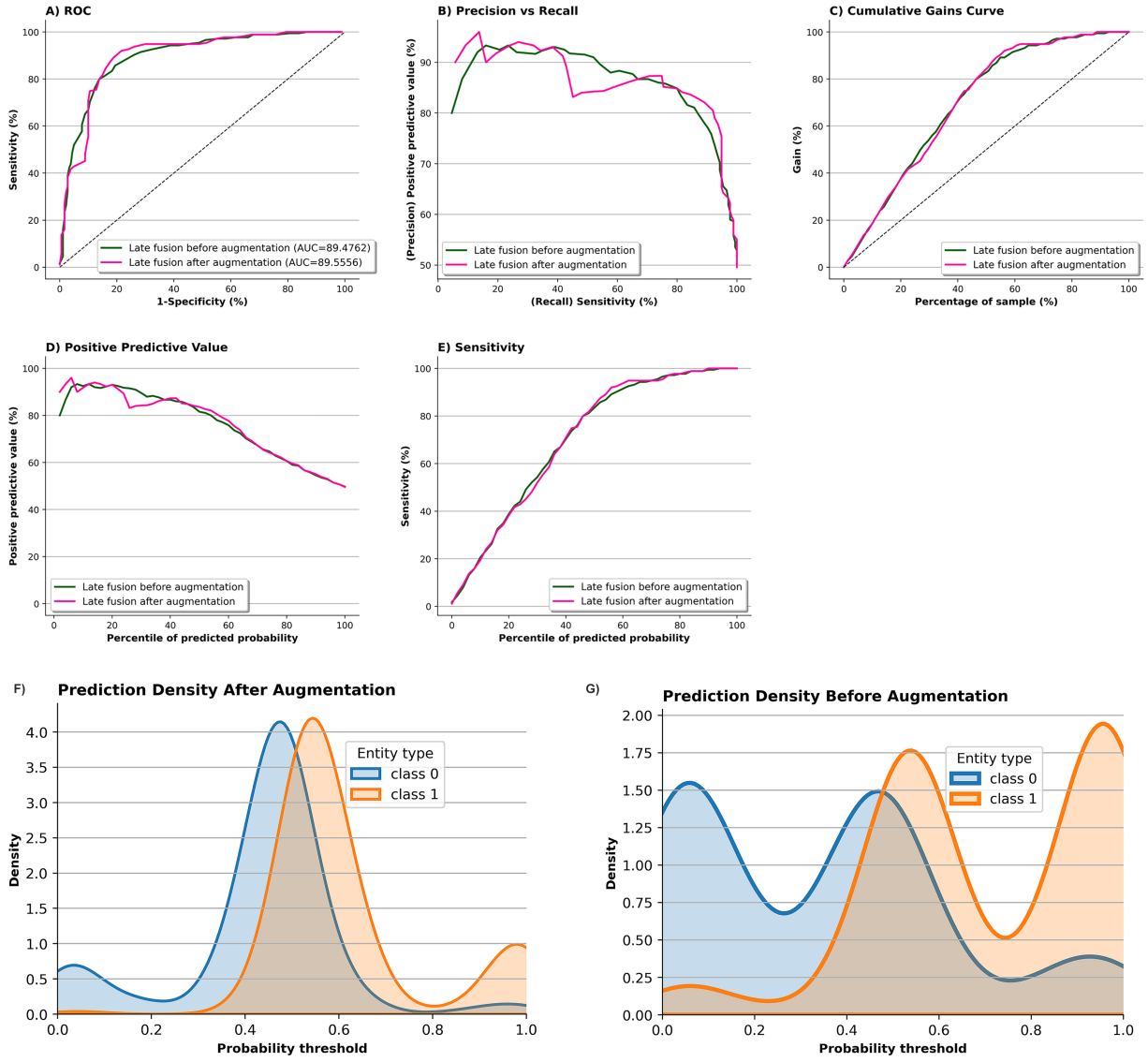

FIGURE 6
Impact of MedAlpaca-7B synthetic data on screening model performance and prediction confidence on ADReSSo benchmark. Panels a–e compare the fusion-based screening model before (green) and after 2 × augmentation with MedAlpaca-7B synthetic speech (pink). ROC curves **(A)** show stable discrimination (AUC: 89.48 → 89.56). The precision–recall curve **(B)** shows improved precision at lower recall. Cumulative gains **(C)** indicate better early retrieval of positives (notably between 40–70%). PPV **(D)** and sensitivity **(E)** profiles remain nearly identical. Density plots **(F,G)** show that post-augmentation predictions are more concentrated and better separated across classes, with reduced uncertainty around the 0.5 threshold.

for LLaMA 3.3 70B (+10.2%) and GPT-4o (+3.1%). Among multimodal models, Phi-4 showed one of the strongest improvements overall (+76.3%), whereas Qwen 2.5 Omni declined substantially (−26.8%).

Overall, Table 3 demonstrates that fine-tuning provides the greatest benefit for smaller or domain-specific text models, while multimodal models remain inconsistent, with outcomes ranging from major improvements (Phi-4) to sharp declines (Qwen 2.5 Omni).

3.4 Component 5: external generalizability evaluation—DementiaBank Delaware corpus

3.4.1 Fusion of transformers and handcrafted linguistic features

Among the four transformers fine-tuned on the Delaware dataset, DistilBERT achieved the best performance (F1 = 66.02 ± 1.23). Integrating DistilBERT embeddings with handcrafted features in a late-fusion classifier further improved results, yielding F1 = 68.05 ± 3.16 and AUC = 62.90 ± 10.99.

3.4.2 LLM-based synthetic augmentation

Incorporating MedAlpaca-7B synthetic samples with the original data improved performance with 1 × augmentation (F1 = 72.82 ± 6.72; AUC = 69.57 ± 10.02). However, performance declined with 2 × augmentation (F1 = 70.43 ± 2.27). These findings suggest that augmentation is beneficial only when synthetic data remains aligned with the real speech distribution, with effectiveness extending up to approximately 1 × for MedAlpaca-7B on the Delaware corpus.

3.4.3 Text-based and multimodal LLM classifiers

As shown in Table 4, among the text-only LLMs, GPT-4o achieved the highest improvement (+11.1%), while LLaMA 3.3–70B and

TABLE 3 Zero-shot versus fine-tuned F1 performance of unimodal (text-only) and multimodal LLM classifiers on official test of ADReSSo benchmark.

Model	Zero-shot F1 (%), mean ± 95%CI	Fine-tuned F1(%), mean ± 95%CI	Improvement (%)
Unimodal (text)			
LLaMA 3.1 8B Instruct	68.54 ± 2.1	81.08 ± 0.94	+18.3
MedAlpaca 7B	47.73 ± 1.03	78.69 ± 2.31	+64.9
Ministral 8B (2410)	66.58 ± 8.55	73.7 ± 3.08	+10.7
LLaMA 3.3 70B Instruct	72.93 ± 1.2	80.35 ± 1.92	+10.2
GPT-4o (2024-08-06)	73.05 ± 2.04	75.28 ± 0.95	+3.1
Multimodal			
GPT-4o (omni)	67.57 ± 1.08	-	-
Qwen 2.5-Omni	67.31 ± 0	49.3 ± 0	−26.8
Phi-4	40.6 ± 3.77	71.59 ± 0.83	+76.3

TABLE 4 Zero-shot versus fine-tuned F1 performance of unimodal (text-only) and multimodal LLM classifiers on held-out test set of Delaware.

Model	Zero-Shot F1 (%), Mean ± 95%CI	Fine-Tuned F1(%), Mean ± 95%CI	Improvement (%)
Unimodal (text)			
LLaMA 3.1 8B instruct	66.44 ± 1.54	68.23 ± 0.43	+2.7
LLaMA 3.3 70B instruct	61.7 ± 2.21	64.52 ± 0.65	+4.6
GPT-4o (2024-08-06)	60.47 ± 2.82	67.18 ± 1.4	+11.1
Multimodal			
GPT-4o (omni)	60.23 ± 1.96	-	-
Phi-4	55.76 ± 3.37	65.57 ± 0	+17.6

LLaMA 3.1–8B showed smaller gains. For multimodal models, Phi-4 benefited most from fine-tuning (+17.6%). In contrast, GPT-4o (omni) was only evaluated in the zero-shot setting, as fine-tuning multimodal models was not possible through the API.

4 Discussion

This study systematically evaluated pretrained transformer models and handcrafted linguistic features to develop a fusion model for early detection of cognitive impairment using spontaneous speech from the ADReSSo 2021 benchmark dataset. Prior studies have explored transformers and linguistic features separately, but our work is among the first to comprehensively assess ten transformer architectures across fine-tuning strategies and integrate their embeddings with 110 domain-informed linguistic features in a late-fusion design. Among the models tested, BERT with last-layer fine-tuning achieved the highest test performance (F1 = 82.76). Linguistic features alone showed strong internal validity (validation F1 = 81.29) but limited generalization (test F1 = 66.83), whereas the fusion of transformer embeddings and linguistic features improved robustness and achieved the strongest overall generalization (test F1 = 83.32).

Notably, BERT, one of the earliest transformer architectures, outperformed more recent and larger models. This result likely reflects both pretraining domain and dataset scale. BERT was trained on broad-domain English text that closely matches the conversational style of DementiaBank, while newer domain-specific models (e.g., BioBERT, ClinicalBERT) were optimized on biomedical documentation, which differs from spontaneous speech and is less effective at capturing disfluencies and syntactic irregularities. In addition, BERT's smaller architecture was better suited to the limited dataset size, whereas larger LLMs such as LLaMA-70B or GPT-4o typically require far larger task-specific corpora to realize their advantages and risk overfitting when fine-tuned on small samples. These findings suggest that for speech-based dementia detection, alignment between pretraining data, model capacity, and dataset scale may be more important than architectural novelty or size.

Building on this foundation, we further evaluated five state-of-the-art LLMs—LLaMA-3.3 8B, MedAlpaca, LLaMA-70B, Ministral, and GPT-4—and three leading multimodal models (text + speech) in both zero-shot and fine-tuned configurations. Zero-shot prompting allowed models to leverage latent knowledge of linguistic cues, whereas fine-tuning enabled adaptation to task-specific patterns. Substantial improvements followed fine-tuning, with the largest gains observed for MedAlpaca-7B (F1: 47.73 → 78.69), followed by LLaMA-8B and Ministral-8B, indicating that smaller open-weight models respond particularly well to targeted training. By contrast, multimodal LLMs underperformed, with Phi achieving the highest F1 = 71.59, suggesting that current audio–text architectures are not yet optimized for detecting cognitive-linguistic markers in spontaneous speech.

Training augmentation with LLM-generated transcripts was effective only when synthetic speech remained semantically and structurally aligned with real data. MedAlpaca-7B produced synthetic samples embedded within the real-data manifold (t-SNE overlap; BERTScore = 0.59), raising the fusion model's validation F1 from 78.17 to 81.02. The highest test F1 = 85.65 was achieved when synthetic data equaled twice the original training size; performance

declined with further augmentation ($3\times - 5\times$). LLaMA-8B and GPT-4 yielded smaller but stable improvements. In contrast, LLaMA-70B, despite strong lexical overlap (top BLEU scores), formed a separate embedding cluster and reduced performance to 77.2. These results confirm that lexical similarity alone is insufficient; effective augmentation must preserve cognitively salient features—such as repetition, disfluency, and syntactic errors—that carry diagnostic value. Thus, augmentation should be limited to volumes that maintain distributional alignment, accompanied by embedding-space validation to avoid degrading signal quality.

External evaluation on the Delaware corpus, restricted to clinically diagnosed MCI vs. controls, provides strong evidence for early-stage screening. The late-fusion pipeline outperformed transformer-only and feature-only baselines, and limited augmentation with MedAlpaca-7B ($1\times$) further improved performance, whereas larger augmentation ($\geq 2\times$) introduced distributional drift and reduced gains. These findings, combined with stable ROC characteristics, indicate that augmentation is effective only within bounded scales and should be accompanied by embedding-space monitoring. Importantly, successful transfer to an MCI-only cohort demonstrates the pipeline's generalizability beyond the mixed-severity ADReSSo benchmark, directly addressing concerns about disease severity and reinforcing its relevance for early detection. Future work should expand MCI samples across sites and incorporate standardized prompts with pre-specified calibration to improve transportability.

Recent regulatory advances underscore the growing relevance of multimodal screening. In May 2025, the FDA approved Fujirebio's Lumipulse G pTau217/β-amyloid 1–42 (U.S. Food and Drug Administration, 2025) blood test for Alzheimer's disease, offering a minimally invasive biomarker assay. While biologically informative, such tests do not reflect how cognitive decline manifests in everyday communication. Language changes—reduced fluency, disorganized sentences—often appear early and may signal real-world functional decline that biological tests cannot detect. Transformer models and LLMs offer a scalable solution by analyzing short voice recordings to identify subtle communication deficits. Their ability to detect linguistic cues offers a critical complement to biomarker testing (Zhang et al., 2025; Hosseini et al., 2025). Combining biological data with speech-based analysis may yield a fuller clinical picture, supporting earlier and more informed decisions on referrals, imaging, and intervention.

The potential of speech processing algorithms for cognitive screening in healthcare is significant, emphasizing the need for comprehensive research on its integration into clinical workflow (Azadmaleki et al., 2025). This calls for interdisciplinary studies to understand clinical facilitators and barriers, including compatibility with existing workflows, clinician attitudes, and operational challenges. It is essential to evaluate the technical, logistical, and financial viability of deploying speech-processing tools in clinical settings, considering their fit with current practices and their ability to improve cognitive health assessments (Zhang et al., 2025). Overcoming these challenges by leveraging government support is essential for harnessing AI's potential to advance patient care and outcomes for patients with cognitive impairment (Hosseini et al., 2025).

Some ASR systems normalize transcripts and suppress diagnostic cues (Zolnoori et al., 2024b; Taherinezhad et al., 2025). In our prior benchmarking on the Pitt audio, Amazon Transcribe (general) maintained a competitive English WER (~13%) and preserved verbatim cues (fillers, repetitions, fragmented/partial words) more reliably than other transcription systems (Amazon Whisper, wav2vec2). For real-world automatic screening pipeline, we recommend ASR settings that retain disfluencies/partial words and avoid grammar-rewriting post-processors, with periodic audits of cue rates and threshold recalibration as ASR systems evolve.

Research on digital linguistic biomarkers spans a continuum from handcrafted linguistic features to transformer-based text models, acoustic and multimodal approaches, and, more recently, LLM-centric methods (Ding et al., 2024; Martínez-Nicolás et al., 2021). On the DementiaBank "Cookie Theft" task, early transcript-only studies relied on feature sets indexing lexical richness/diversity, syntactic complexity, discourse coherence, and filler/repetition rates, establishing that language organization itself provides clinically informative signal (Lanzi et al., 2023; Guo et al., 2021). Building on this foundation, BERT and its derivatives applied to transcripts—typically with no tuning or last-layer fine-tuning—achieved performance in the range of ≈75–84 (F1/accuracy), capturing disfluencies and local syntactic irregularities more effectively than purely hand-engineered sets (Balagopalan et al., 2020; Qiao et al., 2021). Other transcript-based approaches explored within the same benchmark include stacking ensembles of linguistic complexity, disfluency features, and pretrained transformers (Qiao et al., 2021) (F1 ≈ 82), as well as functionals of deep textual embeddings enriched with silence-segment preprocessing and fusion model (Syed et al., 2021) (F1 ≈ 84.45). In parallel, acoustic and multimodal methods have combined prosodic or MFCC-based features with representations from speech transformers such as wav2vec 2.0 or Whisper, often using late or attention-based fusion with text embeddings. These systems commonly report performance in the ≈82–87 range, depending on the linguistic model, acoustic feature extraction, and fusion strategy (Lin and Washington, 2024; Pan et al., 2021; Shao et al., 2025). More recently, LLMs have been explored in zero-shot, few-shot, or fine-tuned configurations for tasks such as dementia detection or disfluency scoring, although their clinical utility remains under investigation (Bang et al., 2024). Within this landscape, our BERT (last-layer fine-tuned) achieved F1 = 82.76, matching or exceeding state-of-the-art LLM fine-tuning baselines; late fusion with handcrafted linguistic features increased performance to F1 = 83.32. With distribution-aligned synthetic augmentation, the fusion model further improved to F1 = 85.65, surpassing all LLM and audio-LLM fine-tuning baselines in our evaluation.

4.1 Limitations

This study has several limitations. First, we focused primarily on textual representations of speech and did not include acoustic transformer models, such as Whisper or wav2vec 2.0, which may capture additional prosodic or phonatory cues relevant to cognitive impairment. Second, while we evaluated several leading unimodal and multimodal LLMs, the multimodal models were limited to those with open or partially accessible APIs, and we could not fully fine-tune closed models like Google Gemini. Third, our evaluation was restricted to English-language transcripts from a structured task, which may limit generalizability to more diverse or naturalistic speech settings. Finally, although we used standard metrics and t-SNE for embedding analysis, future work should incorporate more rigorous interpretability methods to examine which linguistic and acoustic features drive model predictions.

5 Conclusion

This study demonstrates the potential of combining transformer-based embeddings with handcrafted linguistic features to improve the early detection of cognitive impairment from spontaneous speech. By systematically evaluating a range of pretrained transformer models and LLMs—including both unimodal and multimodal architectures—we identified configurations that maximize generalization and classification performance. Our results show that clinically tuned LLMs like MedAlpaca-7B not only adapt well to task-specific fine-tuning but also generate synthetic speech that meaningfully augments training data when distributionally aligned with real speech. These findings support the use of speech-based AI tools as a scalable and interpretable complement to biomarker-driven approaches for dementia screening and highlight the need for continued development of linguistically sensitive, clinically integrated NLP models.

Future work will build on these findings by addressing current limitations. We plan to integrate acoustic transformer models (e.g., wav2vec 2.0, Whisper) to capture prosodic and phonatory cues, enhance multimodal development through open-weight models with distillation from closed systems, and extend evaluation to multilingual and naturalistic conversations through multi-site validation. We will also strengthen the augmentation pipeline by combining LLM-guided narrative generation with prosody-controllable text-to-speech (TTS) to produce distributionally aligned synthetic speech. In addition, we aim to improve interpretability with feature attribution, psycholinguistic mapping, and fairness audits, providing insights into model decisions. Together, these efforts will help overcome current constraints and move this line of work toward practical, clinically useful screening tools.

Data availability statement

The data are available from two sources: (i) the ADReSSo 2021 benchmark dataset, derived from the Pitt Corpus in DementiaBank, comprising 237 participants labeled as cognitively impaired or cognitively healthy, and (ii) the Delaware corpus, also derived from DementiaBank, comprising 205 English-speaking participants labeled as mild cognitive impairment (MCI) or cognitively healthy.

Author contributions

AZ: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. HA: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. YH: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. FT: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. MN: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. SR: Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. MK: Methodology, Writing – review & editing. AT: Methodology, Writing – review & editing. SS: Writing – review & editing. MD: Writing – review & editing. JN: Writing – review & editing. SB: Writing – review & editing. YY: Writing – review & editing. A-HV: Writing – review & editing. MR: Writing – review & editing. MZ: Data curation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the "Development of a Screening Algorithm for Timely Identification of Patients with Mild Cognitive Impairment and Early Dementia (Project no. R00AG076808) in Home Healthcare" from National Institute on Aging, and the Columbia Center for Interdisciplinary Research on Alzheimer's Disease Disparities (CIRAD) (Project no. P30AG059303).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1669896/full#supplementary-material

References

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. Available online at: https://arxiv.org/abs/2503.01743 (2025).

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., et al. Publicly Available Clinical BERT Embeddings. Available online at: https://arxiv.org/abs/1904.03323 (2019).

Alzheimer's Association (2013). Alzheimer's disease facts and figures. *Alzheimers Dement.* 9, 208–245. doi: 10.1016/j.jalz.2013.02.003

Aramaki, E., Shikata, S., Miyabe, M., and Kinoshita, A. (2016). Vocabulary size in speech may be an early indicator of cognitive impairment. *PLoS One* 11:e0155195. doi: 10.1371/journal.pone.0155195

Asgari, M., Kaye, J., and Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer Dementia* 3, 219–228. doi: 10.1016/j.trci.2017.01.006

Azadmaleki, H., Haghbin, Y., Rashidi, S., Momeni Nezhad, M. J., Naserian, M., Esmaeili, E., et al. (2025). S: harnessing multimodal innovation to transform cognitive impairment detection-insights from the National Institute on Aging Alzheimer's speech challenge. *Stud. Health Technol. Inform.* 329, 1856–1857. doi: 10.3233/SHTI251249

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To BERT or not to BERT: Comparing speech and language-based approaches for Alzheimer's disease detection. 2167–2171. doi: 10.48550/arXiv.2008.01551

Bang, J., Han, S., and Kang, B. (2024). Alzheimer's disease recognition from spontaneous speech using large language models. *ETRI J.* 46, 96–105. doi: 10.4218/etrij.2023-0356

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer arXiv. Ithaca, NY, USA. (2020).

Boise, L., Neal, M. B., and Kaye, J. (2004). Dementia assessment in primary care: results from a study in three managed care systems. *J. Gerontol. A Biol. Sci. Med. Sci.* 59, M621–M626. doi: 10.1093/gerona/59.6.M621

Burkhardt, H. A., Alexopoulos, G. S., Pullmann, M. D., Hull, T. D., Areán, P. A., and Cohen, T. (2021). Behavioral activation and depression symptomatology: longitudinal assessment of linguistic indicators in text-based therapy sessions. *J. Med. Internet Res.* 23:e28244. doi: 10.2196/28244

Calzà, L., Gagliardi, G., Favretti, R. R., and Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Comput. Speech Lang.* 65:101113. doi: 10.1016/j.csl.2020.101113

Collins, S. E., Chawla, N., Hsu, S. H., Grow, J., Otto, J. M., and Marlatt, G. A. (2009). Language-based measures of mindfulness: initial validity and clinical utility. *Psychol. Addict. Behav.* 23, 743–749. doi: 10.1037/a0017579

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference 1, 4171–4186.

Ding, K., Chetty, M., Noori Hoshyar, A., Bhattacharya, T., and Klein, B. (2024). Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. *Artif. Intell. Rev.* 57:325. doi: 10.1007/s10462-024-10961-6

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimer's Dis* 49, 407–422. doi: 10.3233/JAD-150520

Glauser, T., Santel, D., DelBello, M., Faist, R., Toon, T., Clark, P., et al. (2020). Identifying epilepsy psychiatric comorbidities with machine learning. *Acta Neurol. Scand.* 141, 388–396. doi: 10.1111/ane.13216

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial networks. *Sci Robot* 3, 2672–2680.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. The Llama 3 herd of models. arXiv. Ithaca, NY, USA. (2024).

Guevara, M., Chen, S., Thomas, S., Chaunzwa, T. L., Franco, I., Kann, B. H., et al. (2024). Large language models to identify social determinants of health in electronic health records. *NPJ Digit. Med.* 7, 1–14. doi: 10.1038/s41746-023-00970-0

Guo, Y., Li, C., Roan, C., Pakhomov, S., and Cohen, T. (2021). Crossing the "cookie theft" corpus chasm: applying what BERT learns from outside data to the ADReSS challenge dementia detection task. *Front Comput Sci* 3:642517. doi: 10.3389/fcomp.2021.642517

Han, T., Adams, L. C., Papaioannou, J. M., Grundmann, P., Oberhauser, T., Löser, A., et al. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. Available online at: https://arxiv.org/abs/2304.08247 (2025).

Hosseini, S. M. B., Nezhad, M. J. M., Hosseini, M., and Zolnoori, M. (2025). Optimizing entity recognition in psychiatric treatment data with large language models. *Stud. Health Technol. Inform.* 329, 784–788. doi: 10.3233/SHTI250947

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., et al. GPT-4o System Card. Available online at: https://arxiv.org/abs/2410.21276 (2024).

Ilias, L., and Askounis, D. (2022). Explainable identification of dementia from transcripts using transformer networks. *IEEE J. Biomed. Health Inform.* 26, 4153–4164. doi: 10.1109/JBHI.2022.3172479

Kang, A., Chen, J. Y., Lee-Youngzie, Z., and Fu, S. Synthetic data generation with LLM for improved depression prediction. arXiv. Ithaca, NY, USA. (2025).

Khodabakhsh, A., Yesil, F., Guner, E., and Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *Eurasip J. Audio Speech Music Process.* 9:Article number 9. doi: 10.1007/978-1-4939-1985-7_11

Koo, J., Lee, J. H., Pyo, J., Jo, Y., and Lee, K. (2020). Exploiting multi-modal features from pre-trained networks for Alzheimer's dementia recognition. 2217–2221. doi: 10.21437/Interspeech.2020-3153

Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., and Cohen, M. L. (2023). DementiaBank: theoretical rationale, protocol, and illustrative analyses. *Am. J. Speech Lang. Pathol.* 32, 426–438. doi: 10.1044/2022_AJSLP-22-00281

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi: 10.1093/bioinformatics/btz682

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., and Luo, Y. (2022). Clinical-Longformer and clinical-BigBird: Transformers for long clinical sequences. *arXiv*. doi: 10.48550/arXiv.2201.11838

Lin, K., and Washington, P. Y. (2024). Multimodal deep learning for dementia classification using text and audio. *Sci. Rep.* 14:13887. doi: 10.1038/s41598-024-64438-1

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. A robustly optimized BERT Pretraining approach. arXiv. Ithaca, NY, USA. (2019).

Martínez-Nicolás, I., Llorente, T. E., Martínez-Sánchez, F., and Meilán, J. J. G. (2021). Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: a systematic review article. *Front. Psychol.* 12:620251. doi: 10.3389/fpsyg.2021.620251

Meilán, J. J. G., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T. E., and Carro, J. (2020). Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behav. Neurol.* 2020, 1–10. doi: 10.1155/2020/4683573

Meteyard, L., Quain, E., and Patterson, K. (2014). Ever decreasing circles: speech production in semantic dementia. *Cortex* 55, 17–29. doi: 10.1016/j.cortex.2013.02.013

Mistral, AI Team. Un Ministal, des Ministraux: Introducing the world's best edge models. Available online at: https://mistral.ai/news/ministraux (2024).

National Institute on Aging. The National Institute on Aging: Strategic Directions for Research, 2020–2025. Available online at: https://www.nia.nih.gov/about/aging-strategic-directions-research/goal-health-interventions#c2 (2021).

National Institute on Aging. Assessing Cognitive Impairment in Older Patients. Available online at: https://www.nia.nih.gov/health/assessing-cognitive-impairment-older-patients.

Nicholas, M., Obler, L. K., Albert, M. L., and Helm-Estabrooks, N. (1985). Empty speech in Alzheimer's disease and fluent aphasia. *J. Speech Lang. Hear. Res.* 28, 405–410. doi: 10.1044/jshr.2803.405

Nichols, L. O., Martindale-Adams, J., Zhu, C. W., Kaplan, E. K., Zuber, J. K., and Waters, T. M. (2017). Impact of the REACH II and REACH VA dementia caregiver interventions on healthcare costs. *J. Am. Geriatr. Soc.* 65, 931–936. doi: 10.1111/jgs.14716

O'Dea, B., Boonstra, T. W., Larsen, M. E., Nguyen, T., Venkatesh, S., and Christensen, H. (2021). The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: a longitudinal study. *PLoS One* 16:e0251787. doi: 10.1371/journal.pone.0251787

OpenAI GPT-4 Technical Report. San Francisco, CA, USA: OpenAI. (2023).

Paganelli, F., Vigliocco, G., Vinson, D., Siri, S., and Cappa, S. (2003). An investigation of semantic errors in unimpaired and Alzheimer's speakers of Italian. *Cortex* 39, 419–439. doi: 10.1016/S0010-9452(08)70257-0

Pan, Y., Mirheidari, B., Harris, J. M., Thompson, J. C., Jones, M., Snowden, J. S., et al. Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech. in Interspeech 3810–3814. Baixas, France: International Speech Communication Association (ISCA). (2021).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. in Proceedings of the 40th annual meeting of the Association for Computational Linguistics (eds. Isabelle, P., Charniak, E. & Lin, D.) 311–318 (Association for Computational Linguistics, Philadelphia, Pennsylvania, USA) (2002).

Pappagari, R., Cho, J., Moro-Velazquez, L., and Dehak, N. (2020). Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity: Interspeech, Baixas, France: International Speech Communication Association (ISCA), 2177–2181.

Peng, Y., Yan, S., and Lu, Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. in Proceedings of the 2019 workshop on biomedical natural language processing (BioNLP 2019) 58–65 (2019).

Peng, J., Wang, Y., Li, B., Guo, Y., Wang, H., Fang, Y., et al. A survey on speech large language models. arXiv. Ithaca, NY, USA. (2024).

Qiao, Y., Yin, X., Wiechmann, D., and Kerz, E. (2021). Alzheimer's disease detection from spontaneous speech through combining linguistic complexity and (dis)fluency features with pretrained language models. 3805–3809. doi: 10.48550/arXiv.2106.08689

Rashidi, S., Azadmaleki, H., Zolnour, A., Nezhad, M. J. M., and Zolnoori, M. (2025). "SpeechCura: a novel speech augmentation framework to tackle data scarcity in healthcare" in Stud health Technol inform, vol. 329, 1858–1859. doi: 10.3233/SHTI251250

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv. Ithaca, NY, USA. (2019).

Shao, H., Pan, Y., Wang, Y., and Zhang, Y. (2025). Modality fusion using auxiliary tasks for dementia detection. *Comput. Speech Lang.* 95:101814.

Sung, J. E., Choi, S., Eom, B., Yoo, J. K., and Jeong, J. H. (2020). Syntactic complexity as a linguistic marker to differentiate mild cognitive impairment from Normal aging. *J. Speech Lang. Hear. Res.* 63, 1416–1429. doi: 10.1044/2020_JSLHR-19-00335

Syed, Z. S., Syed, M. S. S., Lech, M., and Pirogova, E. Tackling the ADRESSO challenge 2021: the MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 6, 4231–4235 (2021).

TaghiBeyglou, B., and Rudzicz, F. (2024). Context is not key: detecting Alzheimer's disease with both classical and transformer-based neural language models. *Nat. Lang. Proc. J.* 6:100046. doi: 10.1016/j.nlp.2023.100046

Taherinezhad, F., Nezhad, M. J. M., Karimi, S., Rashidi, S., Zolnour, A., Dadkhah, M., et al. (2025). Speech-based cognitive screening: A systematic evaluation of LLM adaptation strategies. *arXiv*. doi: 10.48550/arXiv.2509.03525

Tomoeda, C. K., Bayles, K. A., Trosset, M. W., Azuma, T., and McGeagh, A. (1996). Cross-sectional analysis of Alzheimer disease effects on oral discourse in a picture description task. *Alzheimer Dis. Assoc. Disord.* 10, 204–215. doi: 10.1097/00002093-199601040-00006

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* 15, 130–138. doi: 10.2174/1567205014666171121114930

U.S. Food and Drug Administration. FDA Clears First Blood Test Used in Diagnosing Alzheimer's Disease. Available online at: https://www.fda.gov/news-events/press-announcements/fda-clears-first-blood-test-used-diagnosing-alzheimers-disease (2025).

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Woolsey, C. R., Bisht, P., Rothman, J., and Leroy, G. (2024). Utilizing large language models to generate synthetic data to increase the performance of BERT-based neural networks. *AMIA Sum. Transl. Sci. Proc.* 2024:429. doi: 10.48550/arXiv.2405.06695

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. C-pack: Packaged resources to advance general Chinese embedding. Red Hook, NY, USA: ICLR 2024 conference proceedings. (2023).

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., et al. Qwen2.5-Omni Technical Report. Available online at: https://arxiv.org/abs/2503.20215 (2025).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). XLNet: generalized autoregressive Pretraining for language understanding. *Adv Neural Inf Process Syst* 32, 5753–5763. doi: 10.48550/arXiv.1906.08237

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT. Available online at: https://arxiv.org/abs/1904.09675 (2020).

Zhang, Z., Nezhad, M. J. M., Hosseini, S. M. B., Zolnour, A., Zonour, Z., Hosseini, S. M., et al. (2025). A scoping review of large language model applications in healthcare. *Stud. Health Technol. Inform.* 329, 1966–1967. doi: 10.3233/SHTI251302

Zhu, Y., Liang, X., Batsis, J. A., and Roth, R. M. (2021). Exploring deep transfer learning techniques for alzheimer's dementia detection. *Front. Comput. Sci.* 3:Article number 624683. doi: 10.3389/fcomp.2021.624683

Zolnoori, M., Vergez, S., Xu, Z., Esmaeili, E., Zolnour, A., Anne Briggs, K., et al. (2024b). Decoding disparities: evaluating automatic speech recognition system performance in transcribing black and white patient verbal communication with nurses in home healthcare. *JAMIA Open* 7:ooae130. doi: 10.1093/jamiaopen/ooae130

Zolnoori, M., Zolnour, A., and Topaz, M. (2023). ADscreen: a speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia. *Artif. Intell. Med.* 143:102624. doi: 10.1016/j.artmed.2023.102624

Zolnoori, M., Zolnour, A., Vergez, S., Sridharan, S., Spens, I., Topaz, M., et al. (2024a). Beyond electronic health record data: leveraging natural language processing and machine learning to uncover cognitive insights from patient-nurse verbal communications. *J. Am. Med. Inform. Assoc.* 32:ocae300. doi: 10.1093/jamia/ocae300