



BACKGROUND

The lack of non-invasive, accessible, and scalable biomarkers for Alzheimer's disease and related dementias (AD/ADRD) remains a major challenge. While blood-based biomarkers are actively studied, alternative modalities show promise. Audio biomarkers, easily collected via standardized tasks or natural conversation, can capture speech features associated with AD/ADRD pathology¹⁻³.

METHODS

We extracted multilingual audio recordings (English, Spanish, Mandarin, and German) from the DementiaBank subrepository of TalkBank. Figure 1 presents distributions of the data for training and testing, and shows the diagnosis distribution and the type of tasks available in each language. The diagnoses found in this dataset consist of healthy controls (HC), mild cognitive impairment (MCI), Alzheimer's disease (AD), and other dementias (OD). Time-aligned transcripts were used to isolate participant speech. All recordings were processed through a voice enhancement pipeline to improve clarity and reduce acoustic variability, including background noise reduction, dynamic range compression, frequency equalization, and loudness normalization.

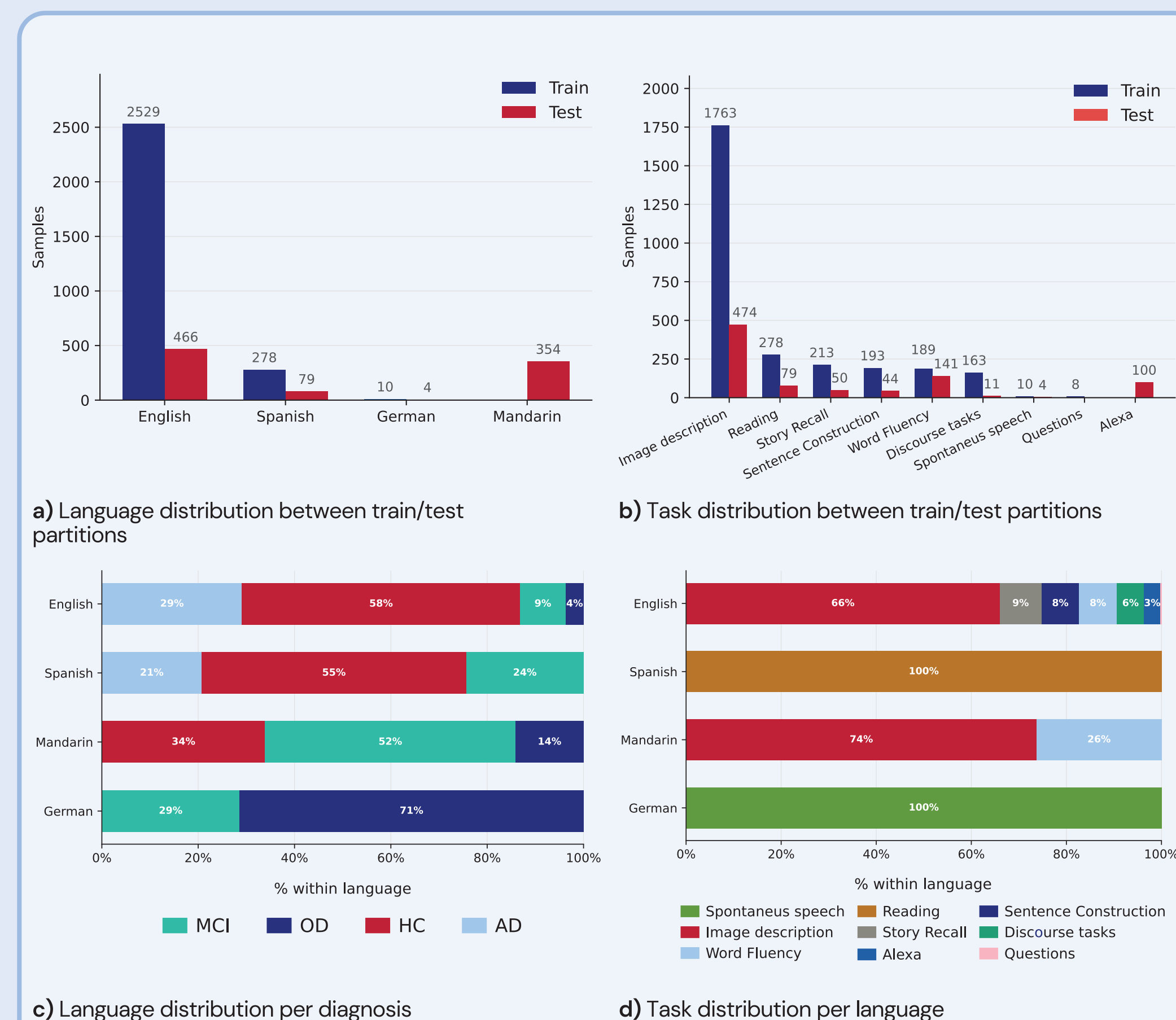


Figure 1. Dataset language, task and diagnosis distributions

We evaluated several models, including distilled and non-distilled Whisper⁴ variants and a hybrid transformer architecture, for temporal modeling, with optional early/late fusion of demographic metadata and audio features. Figure 2 shows a schema of the original Whisper architecture, and the audio segmentation and feature fusion approaches. Recordings were segmented into 30-second chunks to match Whisper encoder requirements. However, the audios were also split into 10-second chunks for the transformer architecture. Participant-level predictions were computed by averaging segment-level class probabilities for Whisper models, while the hybrid architecture produced participant-level predictions directly. To address language imbalance and improve generalizability, we introduced a speech-to-speech translation pipeline as shown in Figure 3.

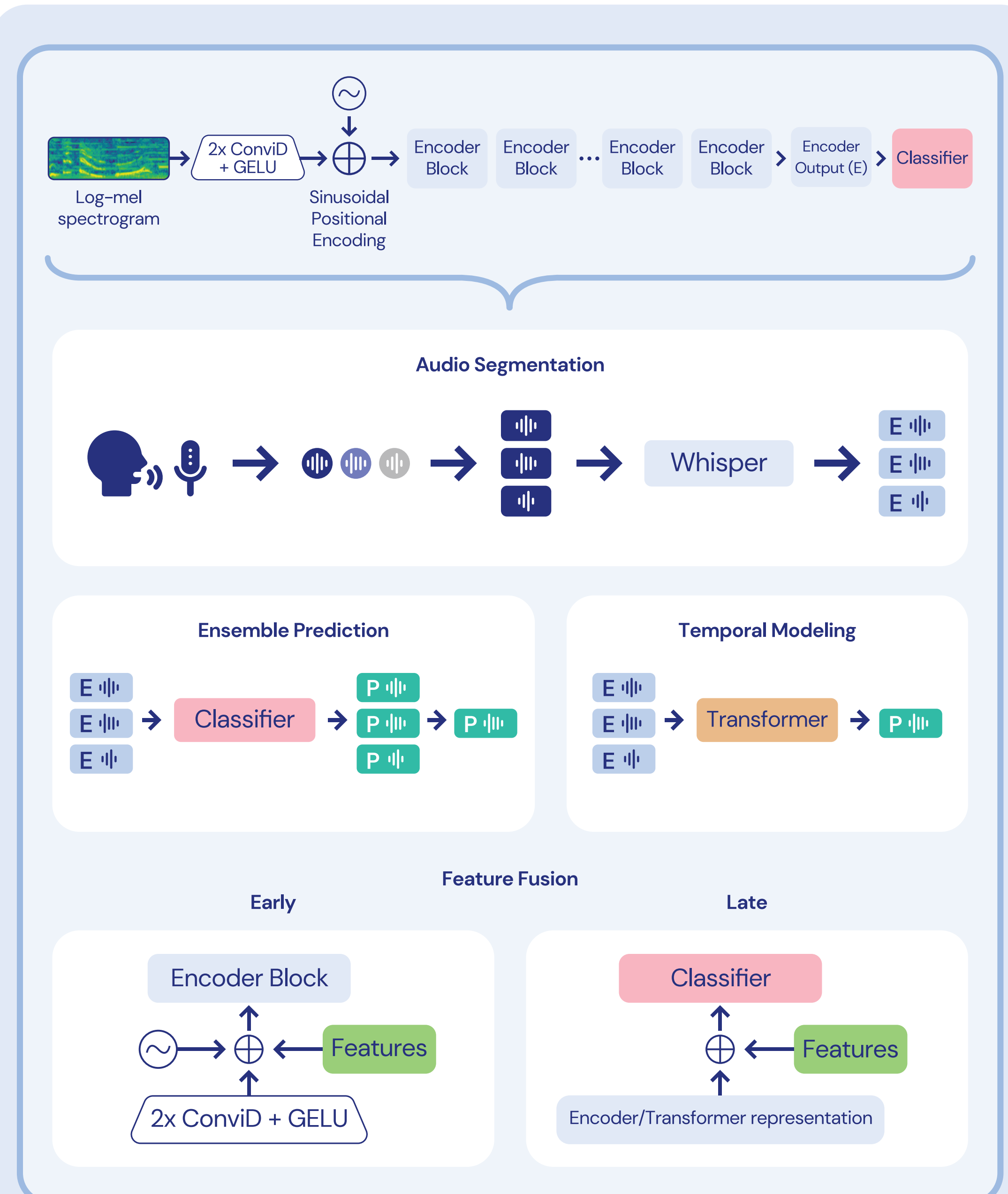


Figure 2. From top to bottom: Whisper architecture, audio segmentation encoding, ensemble and transformer architecture prediction, and early/late feature fusion techniques.

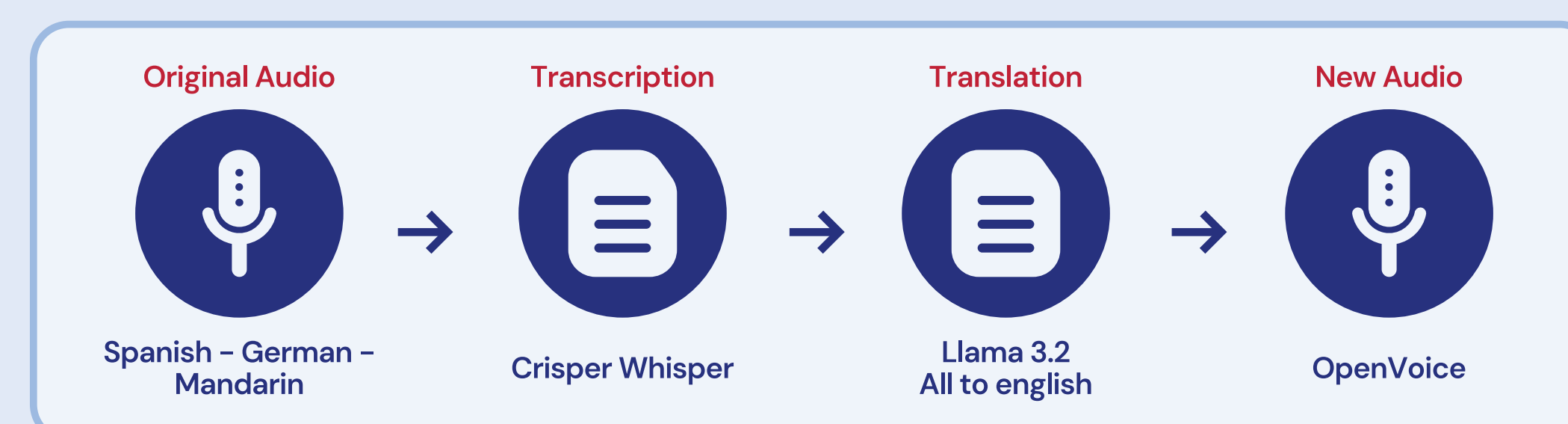


Figure 3. Diagram of the speech-to-speech pipeline

RESULTS

The most relevant experiments are displayed in Table 1. Whisper Large achieved the best overall performance across diagnostic classes (HC, MCI, AD, and OD), particularly when intermediate layers were frozen. AUPRC and accuracy were 0.799/0.757 (HC), 0.440/0.677 (MCI), 0.798/0.728 (AD), and 0.399/0.899 (OD), respectively. Incorporating demographic metadata improved OD classification, suggesting utility for certain clinical subtypes. Our hybrid architecture effectively captured temporal speech dynamics and achieved the strongest OD performance, potentially due to longer recordings in this group. The speech-to-speech translation approach improved predictions for most non-healthy classes in Mandarin and Spanish, with notable gains for AD, as shown in Table 2.

	General	HC	MCI	AD	OD
Whisper Ensemble	0.611	0.799	0.440	0.798	0.399
Whisper First 30s	0.541	0.786	0.428	0.748	0.194
Transformer All chunks (30s length)	0.596	0.802	0.446	0.769	0.360
Transformer All chunks (10s length)	0.587	0.802	0.391	0.725	0.424
Whisper + Metadata (early)	0.610	0.800	0.436	0.798	0.398
Transformer + Metadata (early)	0.595	0.808	0.447	0.763	0.354
Transformer + Audio Features (late)	0.337	0.428	0.333	0.377	0.200

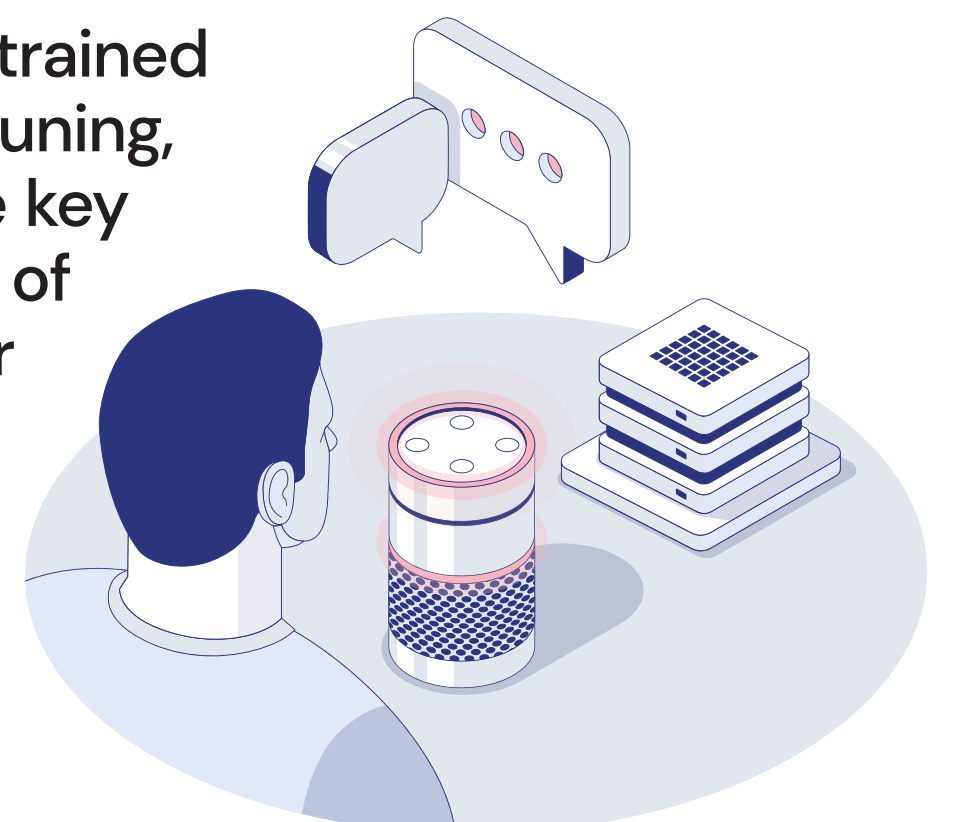
Table 1. AUPRC [0,1] – higher is better – for the most relevant experiments.

		HC	MCI	AD	OD
Whisper Ensemble	Spanish	0.397	0.666	0.679	1
	Mandarin	0.533	0.502	0.504	0.839
Speech-to-speech	Spanish	0.564	0.679	0.884	1
	Mandarin	0.483	0.516	0.921	0.860

Table 2. Accuracy [0,1] – higher is better – for the translation approach.

CONCLUSIONS

In summary, our results show that larger pretrained models, careful adaptation via partial fine-tuning, and modeling speech patterns over time are key to improving the accuracy and reliability of speech-based AD/ADRD prediction. Higher capacity models showed gains from large audios which points to two viable strategies for future data collection: maximizing variability through short recordings when using small models or leveraging longer recordings to enrich sequential speech modeling. Finally, translating non-English speech, while preserving speaker characteristics, can effectively reduce the language domain gap and improve model performance.



DISCLOSURE OF FUNDING

This research was funded by IGC Pharma, LLC.

REFERENCES

- Pastoriza-Domínguez, P. et al. Speech pause distribution as an early marker for Alzheimer's disease. *Speech Commun.* 136, 107–117 (2022).
- Pulido, M. L. B. et al. Alzheimer's disease and automatic speech analysis: A review. *Expert Syst. Appl.* 150, 113213 (2020).
- l, H. et al. Temporal parameters of spontaneous speech in Alzheimer's disease. *Int. J. Speech Lang. Pathol.* doi:https://doi.org/10.3109/17549500903137256.
- Radford, A. et al. Robust Speech Recognition via Large-Scale Weak Supervision. Preprint at <https://doi.org/10.48550/arXiv.2212.04356> (2022).

