Speech and Language Foundation Models for Accurate and Interpretable Alzheimer's Dementia Recognition

Mengjie Qian, Katherine E. Jackson, Kate Knill
mq227@cam.ac.uk, kmk1001@cam.ac.uk

ALTA Institute / Department of Engineering, University of Cambridge

1. Introduction

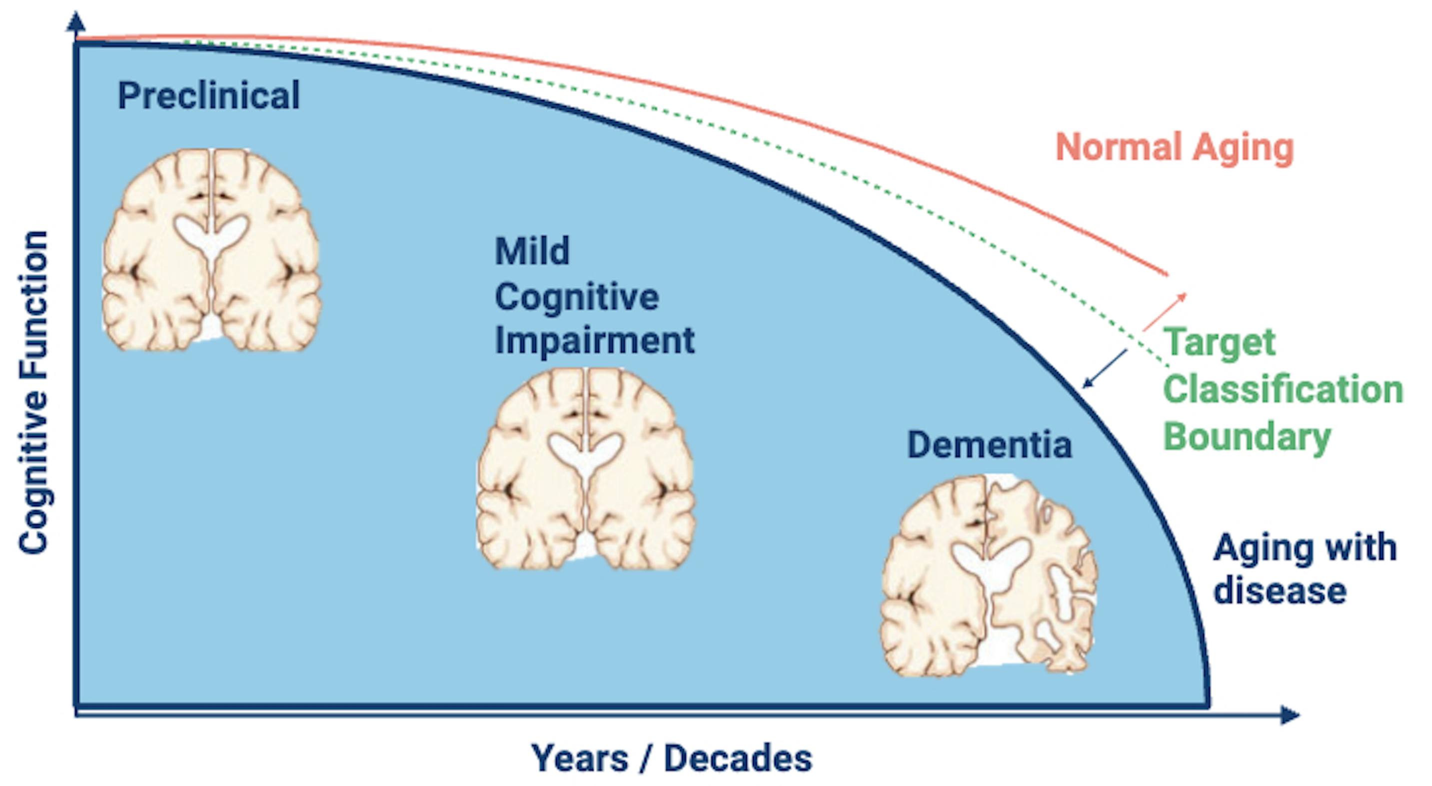


Figure: Typical progression of AD compared to normal cognitive ageing with potential target classification boundary.

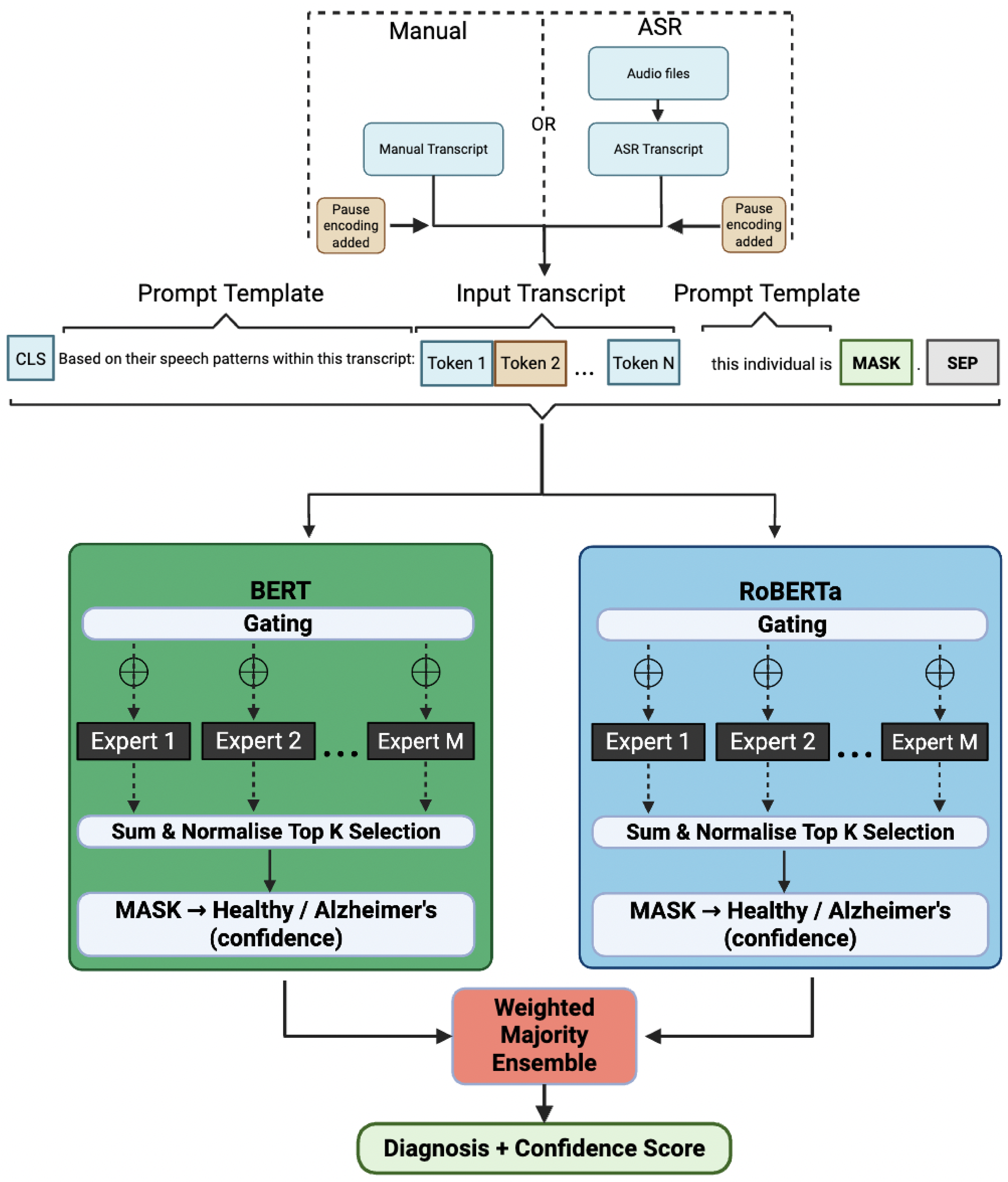
Alzheimer's Dementia (AD) is a neurodegenerative disorder defined by the accumulation of amyloid plaques and neurofibrillary tangles in the brain which cause cognitive impairment.

- Motivation
  - Early and accurate diagnosis is increasingly important as new clinical interventions emerge and recruitment to clinical trials remains challenging.
  - Subtle linguistic and prosodic changes in speech can emerge early and serve as accessible biomarkers for detection.
- Objective
  - Improve AD detection accuracy on manual and ASR (Automatic Speech Recognition) transcripts derived from spontaneous speech samples.
  - Increase interpretability of the model, addressing a key barrier to the field's adoption in real-world settings.

2. Datasets

- **DementiaBank Pitt Corpus:** 33 hours of audio and transcripts from 397 participants (104 controls, 208 AD, 85 unknown), collected via the TalkBank project [1]. Recordings are primarily Cookie Theft picture descriptions, widely used for speech-based dementia research.
- **ADReSS Challenge 2020 dataset:** A balanced subset of DementiaBank (78 AD, 78 controls) with demographic matching [2]. Provides standardised audio and transcripts designed for benchmarking AD detection models.

3. Approach



- A hybrid approach combining prompt-based fine-tuning and Mixture-of-Experts (MoE) using BERT, RoBERTa, and Whisper for ASR
  - **Pause Encoding:** Inject medium/long pause features into prompts.
  - **Prompt Engineering:** Frame diagnosis as a masked LM with dynamic templates and prompt positions.
  - **Ensemble Strategies:** Fuse model and prompt variants via weighted majority voting.
  - **ASR analysis:** Compare results of Whisper vs. manual transcriptions.
  - **MoE Attention:** Gated attention selects specialised expert projections per segment.

4. Feature Analysis

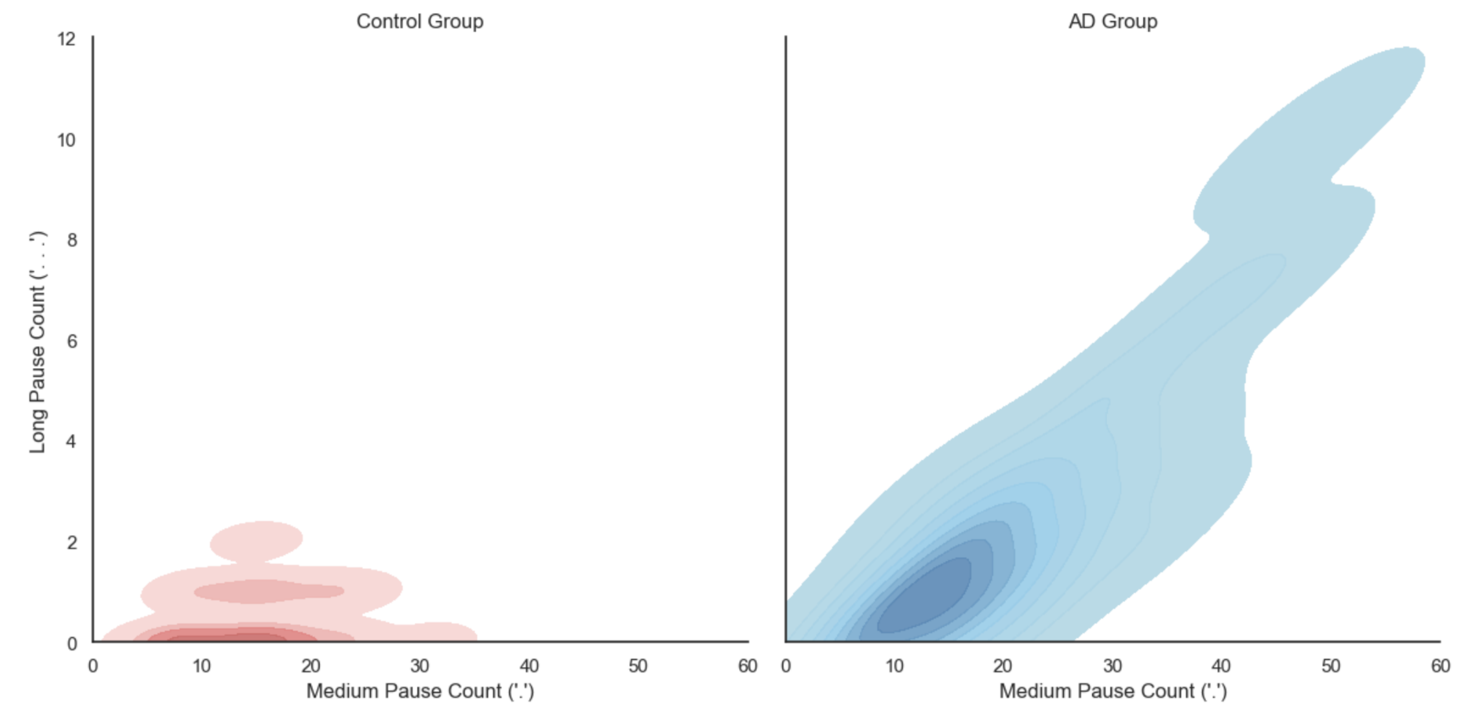


Figure: Relationship between frequency of medium and long Pauses present per transcript. Modelled by a GMMs for Control (red) and AD (blue) groups from ADDreSS Dataset [1].

- Other features (e.g., filler words, speech rate, syntax) showed subtler patterns, making deep learning approaches effective for modelling the complex classification boundary.

5. Key Results

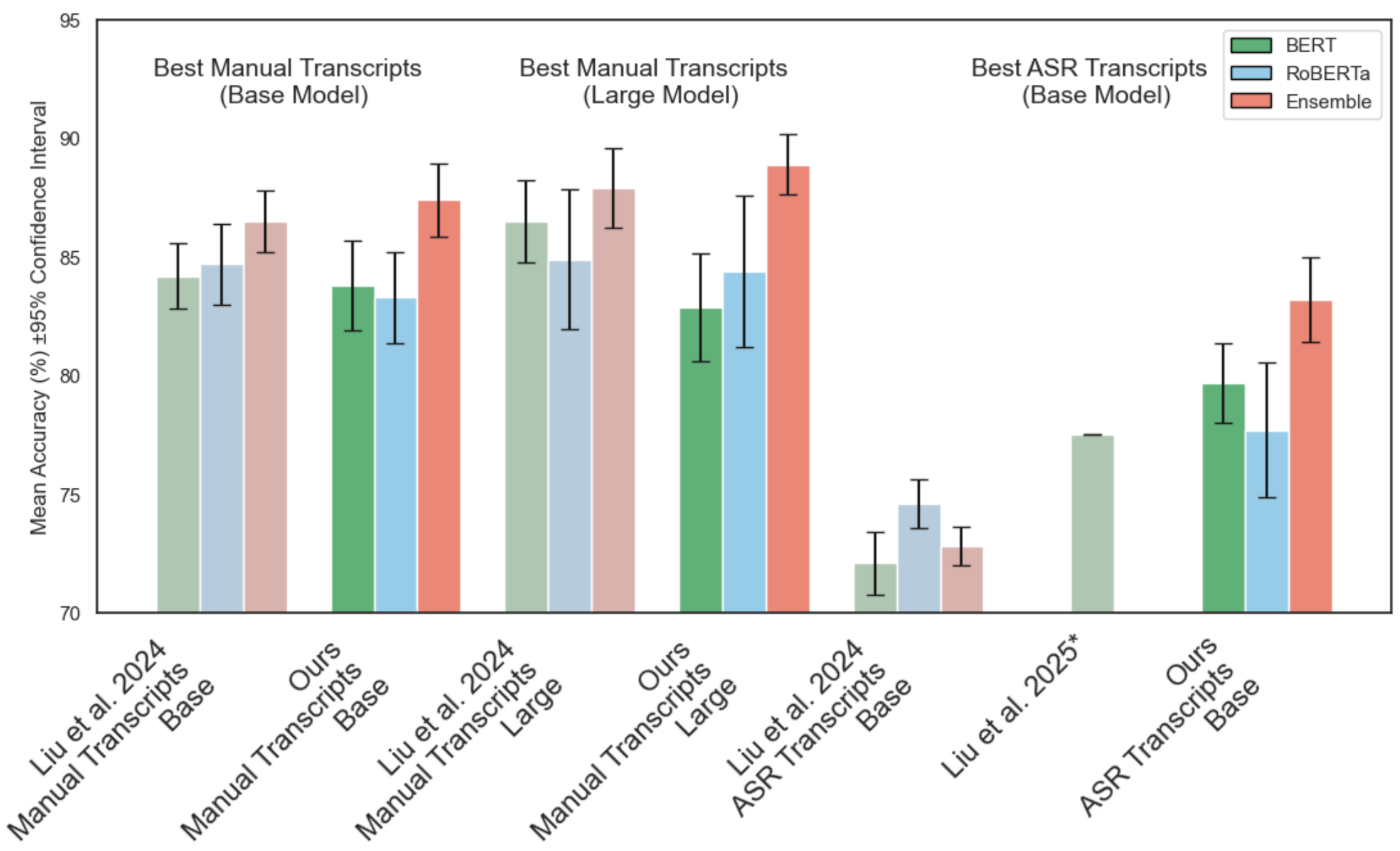


Figure: Comparing mean model accuracies ±95% confidence intervals obtained and reported in [3] and [4].

- Improved accuracy
- Interpretability: The regularised MoE model shows clear expert specialisation, making its decisions more interpretable.
  - Expert 1: syntactic structure
  - Expert 2: tokens relevant to AD: long pauses, grammatical connectors
  - Expert 3: spatially linked words
  - Expert 4: descriptive nouns and verbs: more common in control

6. Conclusions

- Improved AD classification accuracy with a Whisper + LLM architecture.
- MoE adds interpretability via specialised expert pathways.

7. References

[1] J. T. Becker, et al. "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis." Archives of neurology 51.6 (1994): 585-594.
[2] S. Luz, et al. "Alzheimer's dementia recognition through spontaneous speech." Frontiers in computer science 3 (2021): 780169.
[3] T. Liu, et al. "Leveraging Prompt Learning and Pause Encoding for Alzheimer's Disease Detection." 2024 IEEE 14th ISCSLP.
[4] T. Liu, et al. "Beyond Manual Transcripts: The Potential of Automated Speech Recognition Errors in Improving Alzheimer's Disease Detection." arXiv preprint arXiv:2505.19448 (2025).