Multilingual Cognitive Impairment Detection in the Era of Foundation Models

Damar Hoogland[1] **Boshko Koloski**[1,2] **Jaya Caporusso**[1,2] **Tine Kolenik**[3]
Ana Zwitter Vitez[4] **Senja Pollak**[1] **Christina Manouilidou**[4] **Matthew Purver**[1,5]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] Institute of Synergetics and Psychotherapy Research, Paracelsus Medical University, Salzburg, Austria
[4] University of Ljubljana, Ljubljana, Slovenia
[5] Queen Mary University of London, London, UK

Abstract

We evaluate cognitive impairment (CI) classification from transcripts of speech in English, Slovene, and Korean. We compare zero-shot large language models (LLMs) used as direct classifiers under three input settings—transcript-only, linguistic-features-only, and combined—with supervised tabular approaches trained under a leave-one-out protocol. The tabular models operate on engineered linguistic features, transcript embeddings, and early or late fusion of both modalities. Across languages, zero-shot LLMs provide competitive no-training baselines, but supervised tabular models generally perform better, particularly when engineered linguistic features are included and combined with embeddings. Few-shot experiments focusing on embeddings indicate that the value of limited supervision is language-dependent, with some languages benefiting substantially from additional labelled examples while others remain constrained without richer feature representations. Overall, the results suggest that, in small-data CI detection, structured linguistic signals and simple fusion-based classifiers remain strong and reliable signals.

Keywords: cognitive decline detection, large language models, tabular foundation models, feature fusion

1. Introduction

Cognitive impairment (CI) refers to a state in which a person's cognitive functioning is below the expected level and is a diagnosable condition (Ray and Davidson, 2014). CI can involve varying degrees of deterioration of cognitive abilities such as memory, attention, executive functioning, and language, and it is often associated with neurodegenerative diseases like Alzheimer's disease (AD) and other conditions that cause dementia (Morley, 2018). Although relatively advanced CI associated with such diseases is often preceded by a mild cognitive impairment (MCI) phase, not all individuals with MCI progress to dementia (Petersen, 2016). Early identification of CI is essential to enable timely and appropriate clinical intervention, patient support, and participation in preventive or therapeutic programmes (Livingston et al., 2024).

Traditional diagnostic assessments for cognitive impairment include neurophysiological tests (Nasreddine et al., 2005), clinical and functional assessments (O'Bryant et al., 2008), neuroimaging and biomarker assessments (Hampel et al., 2018), and clinical interviews and observation (McKhann et al., 2011). Many of these assessments involve language, as individuals with CI frequently exhibit lexical retrieval difficulties, semantic degradation, syntactic simplification, and reduced discourse organisation, reflecting underlying deterioration in semantic memory and executive control (Taler and Phillips, 2008; Fraser et al., 2015; Boschi et al., 2017). For example, picture description tasks such as the Cookie Theft task (Goodglass and Kaplan, 1983) prompt individuals to describe a complex visual scene. The resulting descriptions enable qualitative and quantitative assessment of language production, including lexical retrieval, syntactic formulation, fluency, informativeness, and narrative organisation.

However, traditional diagnostic tools can be limited in their ability to detect early cognitive changes (Trzepacz et al., 2015) and often require in-person administration by trained professionals, making them resource-intensive, time-consuming, and impractical for frequent or large-scale screening (Slegers et al., 2018). Their outcomes can furthermore be affected by education and language background, introducing cultural and linguistic bias (Ramos-Henderson et al., 2025). Finally, because they are typically administered infrequently in clinical settings, they provide only snapshots of cognition rather than a continuous measure of change (Patnode et al., 2020).

Computational methods—particularly those leveraging Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL)—address several limitations of traditional assessments by enabling automated and fine-grained analysis of spontaneous speech. Such approaches can capture subtle linguistic and discourse changes that may precede clinical diagnosis, operate remotely and non-invasively, and allow for repeated or continuous monitoring over time (De la Fuente Garcia et al., 2020). While computational approaches risk inheriting or

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 1–12
12 May 2026.

amplifying linguistic and cultural biases present in existing assessments, they also have the potential to support diagnosis by trained clinicians when developed and evaluated responsibly.

With recent advances in pretrained foundation models and increased accessibility of computational resources, research on automatic CI detection from language has begun shifting from classical ML approaches that rely on expert-selected symbolic features towards methods built around pretrained language models and general-purpose tabular predictors (reviewed in Section 2). In this study, we compare inference-only foundation-model baselines (multilingual Large Language Models (LLMs) and tabular foundation models) against classical ML and fusion-based approaches across three languages, under unified zero-shot, few-shot, and leave-one-out evaluation protocols.

2. Related Work

Many studies investigating the detection and prediction of cognitive decline have employed classical ML approaches (Huang et al., 2024; Kaser et al., 2024). For example, Luz et al. (2021) employed a range of classical ML models—including linear discriminant analysis, decision trees, k-nearest neighbours, random forests, and support vector machines—to classify Alzheimer's vs. healthy speech and to predict scores obtained with the Mini-Mental State Examination test. These models were built on manually engineered acoustic and linguistic feature sets.

More recently, others have moved to using Large Language Models (LLMs) for feature extraction. For example, de Arriba-Pérez et al. (2024) automatically extracted a large set of high-level, content-independent linguistic features from free dialogues using ChatGPT[1] prompts, alongside traditional n-gram features. These features were then analysed, selected, and used to train classical ML classifiers to detect cognitive decline. Other studies employed DL techniques, such as BERT-based transformer classifiers (Mao et al., 2022; Ilias and Askounis, 2022; Pahar et al., 2025; Zhu et al., 2022).

Large Language Models (LLMs) are increasingly employed as direct classifiers for cognitive decline detection (e.g., Jiang et al., 2026). Zheng et al. (2024) used pre-trained LLaMA-2[2] models with prompt engineering, Low-Rank Adaptation (LoRA) fine-tuning, and conditional learning to classify AD from speech transcripts of the ADReSS dataset (Luz et al., 2020). Guan et al. (2025) presented CD-Tron, a system built on the clinical LLM GatorTron[3], fine-tuned on labelled electronic health record note sections to detect early cognitive decline, and reported substantial improvements over smaller transformers and GPT-4. Botelho et al. (2024) used LLMs both as predictors and as feature extractors, prompting LLMs to produce interpretable macro-descriptors (e.g., coherence, lexical diversity, word-finding difficulty) which were then used as inputs to simple classifiers for AD detection.

Most studies reviewed above, both classical ML and LLM-based approaches, have focused on monolingual settings, with the majority using only English data. This makes it difficult to assess how well these paradigms transfer across languages and elicitation paradigms.

Tabular foundation models have recently been proposed as general-purpose predictors for structured data, enabling strong performance on small tabular datasets via in-context learning and reducing the need for task-specific training (Hollmann et al., 2025). In the cognitive decline domain, Ding et al. (2026) apply TabPFN to longitudinal AD modelling on the TADPOLE benchmark, predicting clinical diagnosis and cognitive scores from tabular patient records. Related work explores TabPFN-based approaches for dementia-related prediction in other neurodegenerative settings, such as predicting Parkinson's disease dementia using a hybrid LightGBM–TabPFN model with SHAP-based interpretability (Tran and Byeon, 2024). These studies do not use language data directly, and comparisons to language-based CI detection remain limited.

2.1. This study

Prior work has not systematically compared (i) classical ML models with expert-assisted linguistic features, (ii) embedding-based tabular models, (iii) tabular foundation models, and (iv) prompted LLM classifiers under a unified protocol and across multiple languages. In the present study, we conduct within-language experiments for three languages—English, Slovene, and Korean—and compare symbolic-feature-based models, embedding-based models, fusion strategies, and zero-shot LLM baselines under unified leave-one-out, zero-shot, and few-shot protocols. We further analyse representation alignment between symbolic features and embeddings to understand when and why multimodal fusion helps in small-data CI detection.

Our research questions (RQs) are as follows.

- **RQ1**: How well do LLMs (specifically, gpt-oss-20b and med-gemma-27b) discriminate CI vs. Healthy Control (HC) participants across three languages under zero-shot prompting?
- **RQ2**: How sensitive to the input modality (transcript-only, linguistic-features-only, or transcript+features) is the performance of LLMs?

[1] https://chat.openai.com/

[2] https://www.llama.com/llama2/

[3] https://huggingface.co/UFNLP/gatortron-base

- **RQ3**: Do expert-assisted symbolic linguistic features improve CI classification when paired with tabular models (TabPFN, RealMLP, and classical baselines) compared to (i) embedding-only representations and (ii) LLM-based classifiers, and are gains consistent across languages and evaluation paradigms?
- **RQ4**: Which integration strategy yields the best and most stable performance across languages among embeddings-only, symbolic-features-only, and multimodal fusion (normalised concatenation / feature reweighting, or late fusion), under both leave-one-out and few-shot evaluation?

3. Data and preprocessing

3.1. Datasets

To evaluate cross-linguistic generalisability and performance stability, we ran parallel experiments on three languages: English, Slovene, and Korean. The English and Slovene datasets were obtained from corpora of recordings of picture description tasks, while the Korean data came from a corpus of structured interviews. The English and Slovene datasets include participants with AD and HCs, while the Korean dataset includes participants with MCI and HCs. In all experiments we treat the positive class as PATIENT (AD or MCI, depending on the dataset) and the negative class as CONTROL (HC).

The included datasets are listed below, and Table 1 summarises the number of participants and diagnostic labels per dataset.

English: The English dataset consists of a subset of Cookie Theft Picture Descriptions from the Pitt Corpus (Becker et al., 1994; the original corpus is available on DementiaBank, MacWhinney et al., 2011), pre-processed for the ADReSS challenge (Luz et al., 2020). It includes participants with AD and Healthy Controls (HCs).

Slovene: The Slovene dataset was collected as part of the CogLiTreat project, which investigated behavioural and transcranial magnetic stimulation interventions for language disorders. It includes recordings and transcripts of Slovene AD patients and control participants from the Ljubljana region who described the New Cookie Theft picture (Berube et al., 2019). The participants' responses were recorded and the detection of speech and silence was performed automatically using Praat (Boersma and Weenink, 2021). The recordings were orthographically transcribed by one of the interviewers and cross-checked by an independent native speaker of Slovene. Finally, each utterance was assigned to the participant or interviewer manually by the first author of the present study. We note two potential confounding factors. First, the patient recordings were delivered in a different file format (m4a) than the control group (WAV), which may have introduced confounds during preprocessing (e.g., silence detection may behave differently across formats). Second, the experimenter differed between the two groups, which may affect language use due to conversational alignment effects (Pickering and Garrod, 2004; Freud et al., 2018). We return to these issues in Section 8.

Korean: The Korean dataset was obtained from the Kang corpus, available on DementiaBank (MacWhinney et al., 2011). The Kang corpus includes participants with MCI and HCs. Each participant took part in a structured interview consisting of 16 questions. The corpus includes manual transcriptions.

3.2. Transcript pre-processing

For each dataset, we extracted the participant utterances and removed non-orthographic annotations (e.g., the use of '(.)' to indicate short pauses in the Pitt corpus).

For the English and Slovene datasets, each utterance was processed using that language's model from the Stanza NLP library (Qi et al., 2020). After tokenisation, tokens labelled as punctuation were excluded, and for each remaining token we extracted the surface form, lemma, universal part-of-speech (UPOS) tag, dependency relation, and syntactic head index.

For Korean, we used the MeCab morphological analyser for tokenisation and part-of-speech (POS) tagging (Kudo, 2005). Tokens labelled as punctuation were excluded (SF: sentence-final punctuation; SP: comma/pause; SS: brackets/quotation marks; SE: ellipsis; SO: other symbols), and the remaining POS tags were converted to their UPOS equivalents (Park and Tyers, 2019). Dependency-based features were not extracted for Korean, as MeCab does not provide dependency parsing.

3.3. Features

From the preprocessed participant-only transcripts we extracted eleven linguistic features that were reported as indicative of AD-related language change in a recent systematic review (Shankar et al., 2025). All features were calculated over all utterances per participant and per task. For Korean, two

Dataset	Condition	Patients	Controls	Total	Patient/Control (%)
English	AD	78	78	156	50.0 / 50.0
Slovene	AD	12	15	27	44.4 / 55.6
Korean	MCI	40	37	77	51.9 / 48.1
Total	–	**130**	**130**	**260**	**50.0 / 50.0**

Table 1: Participant statistics per dataset with within-dataset class proportions. AD: Alzheimer's Disease; MCI: Mild Cognitive Impairment.

dependency-based features (idea density and syntactic complexity) could not be computed (see Section 3.2); these values are treated as missing and handled by the imputation procedure described in Section 4.2.

Speech Rate The number of words uttered by the participant, divided by the total duration of the task in seconds. In English and Slovene, we divided the number of words by the duration from the start of the first participant utterance to the end of the last participant utterance, without excluding interviewer speech. In Korean, we excluded interviewer speech from the duration because the interviewer took a more active role due to the structured nature of the task. Interviewer speech could not be consistently removed from the total duration in English because accurate time-stamps were not available.

Type-Token Ratio The number of unique tokens in the participant's speech divided by the total number of words they uttered.

Repetitiveness The mean cosine distance between embeddings (produced by Sentence-BERT; Reimers and Gurevych, 2019) of each consecutive pair of participant utterances.

Coherence The mean cosine distance of embeddings (produced by Sentence-BERT; Reimers and Gurevych, 2019) between all possible pairs of different participant utterances.

Familiarity The mean familiarity per unique word used by the participant. Familiarity is expressed as the number of occurrences per million words in speech corpora, provided by frequency reference lists for each language (Dobrovoljc, 2018; Leech et al., 2014; Kim et al., 2024).

Idea Density The number of main verbs divided by the total number of tokens uttered by the participant. Not computed for Korean (Section 3.2).

Syntactic Complexity The mean maximal syntactic tree depth per participant utterance. Not computed for Korean (Section 3.2).

Verb Ratio The number of verbs divided by the total number of tokens.

Noun Ratio The number of nouns divided by the total number of tokens.

Pronoun Ratio The number of pronouns divided by the total number of tokens.

Pronoun to Noun Ratio The number of pronouns divided by the total number of nouns.

4. Modelling Methodology

4.1. Task and Data

We study binary classification of CI (AD or MCI, depending on the dataset) versus HC from speech-derived inputs. We evaluate performance separately for three languages: English, Slovene, and Korean (Section 3.1). All experiments are conducted within-language (i.e., training and evaluation never mix languages), and results are reported per language and aggregated across languages.

4.2. Input Representations

We compare three input configurations derived from each sample.

(1) Symbolic linguistic features. We use an 11-dimensional vector of expert-assisted textual features:

$$\mathbf{x}_{\text{feat}} \in \mathbb{R}^{11},$$

corresponding to the features listed in Section 3.3. For Korean, two feature dimensions are missing and are handled by imputation within each fold (see below).

(2) Embedding-based representation. We compute a fixed-dimensional dense embedding from the transcript of the participant's utterances using a frozen multilingual embedding model (`google/embedding-gemma-300m`):

$$\mathbf{x}_{\text{emb}} = f_{\text{emb}}(t), \qquad \mathbf{x}_{\text{emb}} \in \mathbb{R}^{d}.$$

Embeddings are computed once and reused across all evaluation runs.

(3) Fusion of embeddings and features. We consider two fusion strategies that combine the embedding and symbolic feature modalities.

For *early fusion*, we preprocess each modality independently, re-weight the symbolic features to compensate for the dimensionality imbalance ($d \gg 11$), and concatenate into a single vector passed to one classifier:

$$w = \sqrt{\frac{d}{11}}, \qquad \mathbf{x}_{\text{early}} = \left[\, \tilde{\mathbf{x}}_{\text{emb}} \;;\; w \cdot \tilde{\mathbf{x}}_{\text{feat}} \,\right].$$

For *late fusion*, we train two independent classifiers of the same family—one on $\tilde{\mathbf{x}}_{\text{emb}}$ and one on $\tilde{\mathbf{x}}_{\text{feat}}$—and combine their outputs by averaging the predicted class probabilities:

$$\hat{p}_{\text{late}} = \frac{1}{2}\Big(\hat{p}_{\text{emb}} + \hat{p}_{\text{feat}}\Big), \qquad \hat{y}_{\text{late}} = \mathbf{1}\big[\hat{p}_{\text{late}} \geq 0.5\big].$$

This decision-level combination allows each modality to be preprocessed and modelled independently before fusion.

Preprocessing. To prevent data leakage, all preprocessing steps are fit using training data only within each evaluation fold or episode. We apply median imputation per feature dimension using a `SimpleImputer` fit on the training split. We standardise each feature dimension using z-score normalisation (`StandardScaler`) fit on the training split. For both fusion variants, embeddings and symbolic features are imputed and standardised separately.

4.3. Models

We compare tabular foundation models, classical ML baselines, and prompted LLMs.

4.3.1. Tabular and Foundational Models

We train the following classifiers per language and representation (embeddings, features, early fusion, and late fusion): TabPFN (foundational in-context tabular classifier), RealMLP (tabular deep learning baseline), Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM) with both linear (SVM-Linear) and radial basis function (SVM-RBF) kernels, LightGBM (LGBM), and k-nearest neighbours (k-NN) with $k \in \{3, 5, 7\}$. For late fusion, each model family is instantiated twice per fold or episode (once per modality) and their predicted probabilities are averaged as described in Section 4.2.

4.3.2. LLM-based Classification

We evaluate two LLMs as direct classifiers: gpt-oss-20b and med-gemma-27b. Both are used in an inference-only setting via an OpenAI-compatible API endpoint served through vLLM.

We test three prompt variants: (1) transcript-only, (2) linguistic-only (the 11 symbolic features rendered as a numeric list), and (3) full-data (the transcript concatenated with the symbolic feature list). Full prompt templates are provided in Appendix 1.

Models are instructed to output exactly one token: CONTROL or PATIENT. Where supported by the inference server, we enforce a guided-choice constraint restricting outputs to {CONTROL, PATIENT}. Temperature is set to 0 for deterministic decoding unless otherwise stated.

4.4. Evaluation Protocols

All evaluations are conducted *within-language*. We report per-language performance and aggregated averages across languages. To ensure direct comparability between tabular models and LLMs across all evaluation settings, both model families share the same outer leave-one-out loop and the same episodic sampling scheme for few-shot conditions.

4.4.1. Full-Data Evaluation (Leave-One-Out)

For each language L, we perform leave-one-out cross-validation over all n_L samples. Each fold holds out exactly one sample i for testing and uses the remaining $n_L - 1$ samples as the training pool. For tabular models, preprocessing is fit on the training pool only and applied to both train and test. For LLMs, no demonstrations are used in the zero-shot condition. We refer to this setting as the *full-data* evaluation. For each language, we also report a majority-class predictor trained and evaluated under the same LOO protocol.

4.4.2. Few-Shot Episodic Evaluation

To study few-shot behaviour under a unified and directly comparable protocol, we apply the same outer LOO loop and inner episodic sampling scheme to tabular models.

Table 2: LOO Macro-F1 across all models and input configurations. Classical ML, tabular foundation models (TFMs), gradient boosting, and LLM zero-shot baselines. Best per language in **bold**. † = zero-shot (no training data).

Method	Input	English	Slovene	Korean
Non-learning baseline				
Majority	—	0.333	0.357	0.342
LLMs (zero-shot†)				
MedGemma-27B†	Full	0.361	0.518	0.595
	Ling.	0.333	0.357	0.342
	Trans.	0.413	0.492	0.579
GPT-OSS-20B†	Full	0.413	0.617	0.609
	Ling.	0.500	0.555	0.349
	Trans.	0.556	0.603	0.621
Tabular foundation models				
TabPFN	Emb	0.343	**0.852**	0.523
	Feat	0.814	0.454	0.740
	Early	0.784	**0.852**	0.792
	Late	**0.816**	0.727	0.753
RealMLP	Emb	0.493	0.802	0.566
	Feat	0.746	0.682	0.692
	Early	0.743	0.701	0.737
	Late	0.738	0.754	0.680
Gradient boosting				
LGBM	Emb	0.549	0.357	0.342
	Feat	0.782	0.357	0.621
	Early	0.763	0.357	0.621
	Late	0.776	0.357	0.633
Classical ML				
LR	Emb	0.549	**0.852**	0.523
	Feat	0.807	0.540	0.739
	Early	0.801	0.735	0.792
	Late	0.788	0.727	**0.805**
RF	Emb	0.549	**0.852**	0.523
	Feat	0.781	0.635	0.675
	Early	0.665	**0.852**	0.714
	Late	0.737	**0.852**	0.714
SVM (linear)	Emb	0.549	**0.852**	0.523
	Feat	0.813	0.540	0.727
	Early	0.750	0.659	0.778
	Late	0.769	0.814	0.789
SVM (RBF)	Emb	0.549	**0.852**	0.523
	Feat	0.806	0.442	0.714
	Early	0.738	**0.852**	0.766
	Late	0.807	**0.852**	0.712
kNN-3	Emb	0.333	0.357	0.523
	Feat	0.712	0.508	0.726
	Early	0.654	**0.852**	0.778
	Late	0.698	0.250	0.476
kNN-5	Emb	0.325	0.308	0.523
	Feat	0.750	0.463	0.674
	Early	0.652	0.814	0.779
	Late	0.582	0.583	0.476
kNN-7	Emb	0.410	**0.852**	0.523
	Feat	0.762	0.365	0.673
	Early	0.665	**0.852**	0.752
	Late	0.543	0.735	0.500

For each held-out test sample i, we sample k examples per class uniformly at random from the remaining $n_L - 1$ samples (the support set), using only these $2k$ samples for training. We repeat this sampling for $E = 3$ episodes with different random seeds and aggregate predictions across episodes by majority vote; for models producing calibrated probabilities, we additionally average scores before thresholding. We evaluate $k \in \{1, 2, 3, 5\}$. For tabular models, the $2k$ support samples are used to fit the classifier (with preprocessing fit on the same $2k$ samples). For late fusion, two classifiers are fit on the $2k$ samples (one per modality) and their probabilities are averaged.

4.4.3. Zero-Shot Evaluation for LLMs

For each language and prompt modality, we classify each sample independently with no labeled demonstrations.

Evaluation For each language, model, and evaluation mode we report Macro-F1. To estimate variability due to episodic sampling and any stochasticity in model training, we repeat the full tabular evaluation pipeline three times with different random seeds. Results are aggregated by Language, Model, Evaluation Mode and reported as mean. LLM evaluations are deterministic at temperature 0 and are therefore reported as single-run results.

5. Results and Discussion

Leave-one-out (LOO) Macro-F1 results are reported in Table 2. Few-shot results (embeddings-only; k shots per class) are reported in Table 3. We discuss the findings by research question.

Table 3: Few-shot Macro-F1 (embeddings-only, k shots/class, 3 seeds) vs. LLM zero-shot reference. Best overall per language in **bold**.

Method	k	English	Slovene	Korean
LLM zero-shot reference (best across models & modalities)				
Best LLM†	0	**0.556**	0.617	0.621
Tabular foundation models				
TabPFN	1	0.439	0.846	0.538
	2	0.442	0.815	0.504
	3	0.436	**0.852**	0.560
	5	0.472	**0.852**	0.532
RealMLP	1	0.404	**0.852**	0.497
	2	0.461	0.815	0.522
	3	0.448	0.839	0.468
	5	0.455	**0.852**	0.570
Classical ML				
LR	1	0.400	**0.852**	0.497
	2	0.431	**0.852**	0.568
	3	0.444	**0.852**	0.552
	5	0.449	**0.852**	**0.667**
RF	1	0.382	**0.852**	0.497
	2	0.462	**0.852**	0.568
	3	0.458	**0.852**	0.552
	5	0.482	**0.852**	0.452

RQ1: LLM zero-shot CI detection. Zero-shot LLM performance is generally limited relative to supervised tabular models. The best zero-shot LLM result is GPT-OSS-20B on Korean (0.621 Macro-F1), which is +0.279 above the majority baseline (0.342). On English and Slovene, LLM performance varies substantially by prompt modality; for example, MedGemma-27B collapses to majority-class behaviour in the linguistic-only setting on English (0.333 Macro-F1). Despite being a medical-domain model, MedGemma-27B underperforms GPT-OSS-20B across all three languages, suggesting general instruction-following behaviour and robustness to prompt formatting may matter more than domain specialisation in this setting.

RQ2: Modality sensitivity. No single input modality consistently dominates across languages and models. Transcript-only input works best for English and Korean with GPT-OSS-20B, while full-data prompts (transcript + features) do not reliably outperform single-modality prompts. This suggests that, in an inference-only setting, LLMs may struggle to integrate heterogeneous numeric and textual evidence as reliably as simpler tabular fusion approaches.

RQ3: Symbolic features + tabular models vs. LLMs. Tabular models with symbolic features decisively outperform zero-shot LLM baselines (+0.18 to +0.26 Macro-F1 across languages when comparing best results per language). The 11-feature vector is highly informative: on English, TabPFN achieves 0.814 with features alone, a +0.471 improvement over embeddings-only (0.343). Classical ML baselines remain competitive, with LR achieving the best Korean result (0.805, late fusion). Slovene is an exception where embeddings dominate features (0.852 vs. 0.454 for TabPFN), though potential confounds (different recording formats and experimenters across groups) may inflate embedding-based separability (Sections 3.1 and 8).

RQ4: Fusion strategies. Early fusion generally performs best for Slovene and yields strong performance for Korean, indicating that combining complementary information sources can improve stability across datasets. Features-only is strongest for English, where the engineered linguistic signal is particularly predictive and adding high-dimensional embeddings can dilute that signal for some models. To better understand when fusion helps, we analyse alignment between feature space and embedding space (Table 4). English and Korean show near-orthogonal representations (CKA 0.024 and 0.016; low Overlap@5), consistent with fusion providing complementary information. Slovene exhibits higher alignment (CKA 0.181; Overlap@5 0.200), suggesting that in this dataset embeddings may already encode much of the feature-level signal (or reflect dataset-specific confounds).

Few-shot vs. zero-shot LLMs. Even with minimal labels, tabular models can match or exceed LLM zero-shot performance for some languages. For Slovene, embedding-based few-shot models exceed the best LLM from just k=1 example per class (0.846–0.852 vs. 0.617). For Korean, LR reaches 0.667 at k=5, exceeding the best LLM (0.621). In English, the best LLM zero-shot result (0.556) remains higher than embedding-only few-shot baselines, highlighting the importance of incorporating symbolic features and/or stronger representations for small-data supervision.

Table 4: Feature–embedding space alignment. Low CKA and Spearman ρ indicate weak alignment between representations; Procrustes disparity (Williams et al., 2021) near 2.0 indicates geometric dissimilarity. Overlap@5 = shared kNN neighbours; Purity@5 = same-class neighbours.

Language	CKA	Spearman ρ	Procrustes	Overlap@5	$\text{Purity}_{\text{feat}}$@5	$\text{Purity}_{\text{emb}}$@5
English	0.024	0.001	1.758	0.032	0.658	0.501
Slovene	0.181	0.116	1.432	0.200	0.511	0.563
Korean	0.016	0.005	1.835	0.049	0.610	0.610

6. Conclusion

We evaluated CI detection across three languages (English, Slovene, and Korean) comparing LLM

zero-shot prompting, tabular foundation models, and classical ML. LLMs provide usable no-training baselines (best Macro-F1: 0.621), but supervised tabular models—especially those using expert-assisted symbolic features and fusion—achieve substantially higher performance (+0.18 to +0.26 Macro-F1 over the best LLM per language). Across datasets, lightweight classical models remain highly competitive, with TabPFN reaching 0.816 on English (late fusion) and LR reaching 0.805 on Korean (late fusion). Alignment analysis indicates that feature and embedding representations are weakly aligned in English and Korean, providing principled support for fusion; Slovene shows higher alignment, consistent with embeddings dominating in that dataset. For practical deployment in small-data CI detection, tabular models operating on transparent linguistic markers offer a strong and interpretable alternative to inference-only LLM classification, while LLM prompting remains a useful reference point when training data are unavailable.

7. Code Availability

The source code is publicly available at https://github.com/bkolosk1/foundational-ci-detection.

8. Limitations

The Slovene dataset (n=27) exhibits potential confounds: different recording formats and experimenters for patient and control groups may inflate embedding-based performance. Two symbolic features (idea density and syntactic complexity) are unavailable for Korean due to parser limitations, and are therefore treated as missing and imputed; this reduces the amount of available symbolic information for that language. While the symbolic features are individually interpretable, fusion with 768-dimensional embeddings reduces transparency; future work should investigate explanation methods (e.g., SHAP) for fusion models and assess robustness across elicitation paradigms. The relatively small dataset size, even if it is typical for this field, restricts the strength and generalisability of our conclusions. Moreover, interviewer speech was not excluded from the speech-rate calculation in the English and Slovene data, which may have affected these measures. As a result, the reported speech-rate values may not fully reflect participant speech alone and should be interpreted with caution. A further limitation is that the study does not examine potential sources of bias related to participant characteristics such as age, education, and linguistic background, all of which may affect linguistic performance and, in turn, influence classification outcomes. It also does not account for differences in cognitive or brain reserve. Assessing such demographic and individual differences is essential before any clinical deployment. Finally, two of the three datasets use picture description tasks; generalisation to other elicitation paradigms (spontaneous speech, narrative recall) remains to be investigated.

Acknowledgments

We acknowledge the financial support from the Slovenian Research Agency ARIS via the projects Cross-Lingual Analysis for Detection of Cognitive Impairment in Less-Resourced Languages (CroDeCo; J6-60109), and Natural Language Processing for Corpus Analysis in the Medical Humanities (BI-VB/25-27-021), and research core funding for the programme Knowledge Technologies (P2-0103).

BK is funded by the Young Researcher Grant PR-12394, and JC by the Young Researcher Grant PR-13409.

The English dataset was based on the Pitt Corpus (Becker et al., 1994), which was produced with the support of grants NIA AG03705 and AG05133 to the original authors of the corpus.

9. Bibliographical References

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Shauna Berube, Jodi Nonnemacher, Cornelia Demsky, Shenly Glenn, Sadhvi Saxena, Amy Wright, Donna C Tippett, and Argye E Hillis. 2019. Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with aphasia. *American Journal of Speech-Language Pathology*, 28(1S):321–329.

Paul Boersma and David Weenink. 2021. Praat: Doing phonetics by computer (version 6.4.06) [computer software]. http://www.praat.org/.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in Psychology*, 8:269.

Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2024. Macro-descriptors for alzheimer's

disease detection using large language models. In *Interspeech*, pages 1975–1979.

Francisco de Arriba-Pérez, Silvia García-Méndez, Javier Otero-Mosquera, and Francisco J González-Castaño. 2024. Explainable cognitive decline detection in free dialogues with a machine learning approach based on pre-trained large language models. *arXiv preprint arXiv:2411.02036*.

Sofia De la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. 2020. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574.

Yilang Ding, Jiawen Ren, Jiaying Lu, Gloria Hyunjung Kwak, Armin Iraji, Shengpu Tang, and Alex Fedorov. 2026. Longitudinal progression prediction of alzheimer's disease with tabular foundation model.

Kaja Dobrovoljc. 2018. Gos corpus n-grams 2.0. https://www.clarin.si/repository/xmlui/handle/11356/1195.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Debora Freud, Ruth Ezrati-Vinacour, and Ofer Amir. 2018. Speech rate adjustment of adults during conversation. *Journal of fluency disorders*, 57:1–10.

Harold Goodglass and Edith Kaplan. 1983. *The assessment of aphasia and related disorders*.

Hao Guan, John Novoa-Laurentiev, and Li Zhou. 2025. Cd-tron: Leveraging large clinical language model for early detection of cognitive decline from electronic health records. *Journal of Biomedical Informatics*, page 104830.

Harald Hampel, Sid E O'Bryant, José L Molinuevo, Henrik Zetterberg, Colin L Masters, Simone Lista, Steven J Kiddle, Richard Batrla, and Kaj Blennow. 2018. Blood-based biomarkers for alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology*, 14(11):639–652.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

Lihe Huang, Hao Yang, Yiran Che, and Jingjing Yang. 2024. Automatic speech analysis for detecting cognitive decline of older adults. *Frontiers in Public Health*, 12:1417966.

Loukas Ilias and Dimitris Askounis. 2022. Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4153–4164.

Lei Jiang, Yue Zhou, and Natalie Parde. 2026. What do llms know about alzheimer's disease? fine-tuning, probing, and data synthesis for ad detection. *arXiv preprint arXiv:2602.11177*.

Alyssa N Kaser, Laura H Lacritz, Holly R Winiarski, Peru Gabirondo, Jeff Schaffert, Alberto J Coca, Javier Jiménez-Raboso, Tomas Rojo, Carla Zaldua, Iker Honorato, et al. 2024. A novel speech analysis algorithm to detect cognitive impairment in a spanish population. *Frontiers in Neurology*, 15:1342907.

Jin-seo Kim, Anna Seo Gyeong Choi, and Sunghye Cho. 2024. Kofren: Comprehensive korean word frequency norms derived from large scale free speech corpora. In *Proceedings of the Joint Conference on Language Resources and Evaluation*.

Takumitsu Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Routledge.

Gill Livingston, Jonathan Huntley, Kathy Y Liu, Sergi G Costafreda, Geir Selbæk, Suvarna Alladi, David Ames, Sube Banerjee, Alistair Burns, Carol Brayne, et al. 2024. The lancet commissions. *Lancet*, 404:572–628.

Saturnino Luz, Farhana Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Proceedings of Interspeech 2020*, pages 2172–2176.

Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. 2021. Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3:780169.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25:1286–1307.

Chengsheng Mao, Jie Xu, Luke Rasmussen, Yikuan Li, Prakash Adekkanattu, Jennifer Pacheco, Borna Bonakdarpour, Robert Vassar, Guoqian Jiang, Fei Wang, et al. 2022. Ad-bert: using pre-trained contextualized embeddings to predict the progression from mild cognitive impairment to alzheimer's disease. *arXiv preprint arXiv:2212.06042*.

Guy M McKhann, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux, et al. 2011. The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269.

John E Morley. 2018. An overview of cognitive impairment. *Clinics in Geriatric Medicine*, 34(4):505–513.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Sid E O'Bryant, Stephen C Waring, C Munro Cullum, James Hall, Laura Lacritz, Paul J Massman, Philip J Lupo, Joan S Reisch, Rachelle Doody, Texas Alzheimer's Research Consortium, et al. 2008. Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study. *Archives of Neurology*, 65(8):1091–1095.

Madhurananda Pahar, Fuxiang Tao, Bahman Mirheidari, Nathan Pevy, Rebecca Bright, Swapnil Gadgil, Lise Sproson, Dorota Braun, Caitlin Illingworth, Daniel Blackburn, et al. 2025. Cognospeak: an automatic, remote assessment of early cognitive decline in real-world conversational speech. In *2025 IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*, pages 1–7. IEEE.

Jungyeul Park and Francis Tyers. 2019. A new annotation scheme for the sejong part-of-speech tagged corpus. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 195–202.

Carrie D Patnode, Leslie A Perdue, Rebecca C Rossom, Megan C Rushkin, Nadia Redmond, Rachel G Thomas, and Jennifer S Lin. 2020. Screening for cognitive impairment in older adults: updated evidence report and systematic review for the us preventive services task force. *JAMA*, 323(8):764–785.

Ronald C. Petersen. 2016. Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, 22(2):404–418.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. 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*.

Miguel Ramos-Henderson, Carlos Calderón, and Marcos Domic-Siede. 2025. Education bias in typical brief cognitive tests used for the detection of dementia in elderly population with low educational level: a critical review. *Applied Neuropsychology: Adult*, 32(1):253–261.

Sujata Ray and Susan Davidson. 2014. Dementia and cognitive decline. a review of the evidence. *Age UK*, 27:10–12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ravi Shankar, Anjali Bundele, and Amartya Mukhopadhyay. 2025. A systematic review of natural language processing techniques for early detection of cognitive impairment. *Mayo Clinic Proceedings: Digital Health*, 3(2):100205.

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556.

Vinh Quang Tran and Haewon Byeon. 2024. Predicting dementia in parkinson's disease on a small tabular dataset using hybrid lightgbm–tabpfn and shap. *Digital Health*, 10:20552076241272585.

Paula T Trzepacz, Helen Hochstetler, Shufang Wang, Brett Walker, Andrew J Saykin, and

Alzheimer's Disease Neuroimaging Initiative. 2015. Relationship between the montreal cognitive assessment and mini-mental state examination for assessment of mild cognitive impairment in older adults. *BMC Geriatrics*, 15(1):107.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott W Linderman. 2021. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pages 4738–4750.

Tian Zheng, Xurong Xie, Xiaolan Peng, Hui Chen, and Feng Tian. 2024. Alzheimer's disease detection based on large language model prompt engineering. In *International Conference on Social Robotics*, pages 207–216. Springer Nature Singapore.

Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. 2022. Domain-aware intermediate pretraining for dementia detection with limited data. In *Interspeech*, volume 2022, page 2183.

Appendix 1: Prompt templates

All LLM-based classifications use the same system instruction and output constraint, differing only in the data fields provided to the model. The system prompt and output format are shared across all three variants:

System Prompt (all variants)

You are a binary classifier for a research dataset (non-diagnostic). Use only the provided transcript and/or linguistic metrics. Inputs may be English, Slovene, or Korean; treat multilingualism, accent, dialect, and topical content as neutral. Ignore demographic/identity attributes and stereotypes. Assume no class base-rate. Do not reveal reasoning.

Output: Exactly one word — `Control` or `Patient`.

The three prompt variants differ in which data fields are included in the query block:

Variant 1 — Transcript-only

`[DATA INPUT FOR ID: {id} Language: {language}]`

`[TRANSCRIPT]`
{transcript_patient}

`[INSTRUCTIONS]`
Classify strictly as "Control" or "Patient" using only evidence present in the provided fields. Output exactly one word.

`[OUTPUT FORMAT]`
Return exactly one word: Control or Patient.

Variant 2 — Linguistic-only

`[DATA INPUT FOR ID: {id} Language: {language}]`

`[LINGUISTIC METRICS]`
- Speech Rate: *{value}*
- Ttr: *{value}*
- Noun Ratio: *{value}*
- Verb Ratio: *{value}*
- Pronoun Ratio: *{value}*
- Pronoun To Noun Ratio: *{value}*
- Mean Frequency: *{value}*
- Coherence: *{value}*
- Repetitiveness: *{value}*
- Idea Density: *{value}*
- Syntactic Complexity: *{value}*

`[INSTRUCTIONS]`
Classify strictly as "Control" or "Patient" using only evidence present in the provided fields. Output exactly one word.

`[OUTPUT FORMAT]`
Return exactly one word: Control or Patient.

Variant 3 — Full-data (Transcript + Linguistic Metrics)

`[DATA INPUT FOR ID: {id} Language: {language}]`

`[TRANSCRIPT]`
{transcript_patient}

`[LINGUISTIC METRICS]`
- Speech Rate: *{value}*
- Ttr: *{value}*
- Noun Ratio: *{value}*
- Verb Ratio: *{value}*
- Pronoun Ratio: *{value}*
- Pronoun To Noun Ratio: *{value}*
- Mean Frequency: *{value}*
- Coherence: *{value}*
- Repetitiveness: *{value}*
- Idea Density: *{value}*
- Syntactic Complexity: *{value}*

`[INSTRUCTIONS]`
Classify strictly as "Control" or "Patient" using only evidence present in the provided fields. Output exactly one word.

`[OUTPUT FORMAT]`
Return exactly one word: Control or Patient.

For few-shot variants, a `[EXAMPLES (labeled)]` block is prepended before the query, containing $2k$ demonstration cases (one per support sample), each formatted identically to the query block above and annotated with their ground-truth label (`Label: Control` or `Label: Patient`).