

Cross-Corpus Depression Detection in Semi-Structured Clinical Interviews

Misha Sadeghi^{1,2}, Phuc Truong Loc Nguyen¹, Andreas Triantafyllopoulos^{2,3},
Mahdis Habibpour^{1,2}, Robert Richer^{1,2}, Lydia Helene Rupp⁴,
Lena Schindler-Gmelch⁴, Marie Keinert⁴, Malin Hager⁴, Björn W. Schuller^{2,3,5,6},
Bernhard Egger⁷, Matthias Berking⁴, Bjoern M. Eskofier^{1,2,8,9}

¹ Machine Learning and Data Analytics Lab,
Friedrich-Alexander-Universität Erlangen-Nürnberg

² Munich Center for Machine Learning

³ Chair of Health Informatics, Department of Clinical Medicine, TUM University Hospital

⁴ Chair of Clinical Psychology and Psychotherapy,
Friedrich-Alexander-Universität Erlangen-Nürnberg

⁵ Munich Data Science Institute

⁶ Group on Language, Audio, & Music, Imperial College London

⁷ Chair of Visual Computing, Friedrich-Alexander-Universität Erlangen-Nürnberg

⁸ Chair of AI-supported Therapy Decisions, Ludwig-Maximilian University

⁹ Translational Digital Health Group, Institute of AI for Health, Helmholtz Munich

Correspondence: misha.sadeghi@fau.de

Abstract

Previous work links Major Depressive Disorder (MDD) to linguistic and paralinguistic speech patterns, highlighting speech analysis as a potential diagnostic tool for remote monitoring or clinical decision support (CDS) systems. Recent attempts to build MDD diagnostic models primarily rely on the English (E-)DAIC corpus. Despite promising results, these works rarely evaluate cross-corpus generalization and rely on self-reported questionnaires, limiting their clinical utility. We introduce a new, clinically validated German corpus, EmpkinS-EKSpresion, of semi-structured diagnostic interviews to test the reliability of spoken language-based measures of MDD. Our model achieves an F1-score of 0.86 on the EmpkinS-EKSpresion dataset and 0.74 on E-DAIC via zero-shot cross-corpus transfer, suggesting that MDD detection models can generalize across corpora and languages despite differences in interview format and ground truth definition, supporting further development of CDS systems.

1 Introduction

Major Depressive Disorder (MDD) affects approximately 5% of adults globally and leads to severe functional impairment when left untreated (World Health Organization, 2025; Woody et al., 2017). The clinical gold standards for diagnosis are semi-structured clinical interviews (American Psychiatric Association, 2013), in which clinicians sys-

tematically evaluate the presence and severity of core symptoms, such as persistent depressed mood, a loss of interest or pleasure, significant changes in sleep or appetite, and profound fatigue. At an earlier stage of the diagnostic process, validated self-report screening instruments such as the Patient Health Questionnaire-9 (PHQ-9), which covers all nine DSM-5 criteria, and its eight-item variant PHQ-8, which omits the suicidality assessment (Kroenke et al., 2009, 2001), are used to identify individuals who may warrant further clinical evaluation. While these methods are foundational to clinical practice, their manual administration strains overburdened healthcare professionals and limits diagnostic access in contexts where clinical expertise is scarce. Consequently, there is a clear need for scalable, automated tools to assist clinicians across the diagnostic pipeline, from early screening to clinical decision support (Insel, 2017; Torous et al., 2021).

Data from clinical interviews provide an ideal foundation for automated tools. While self-report questionnaires capture only semantic information, human experts intuitively integrate this with non-verbal cues such as pitch and prosody to form a holistic diagnosis. Motivated by this, automated depression detection has evolved from early systems relying on handcrafted acoustic features and manual transcripts (Williamson et al., 2016; Gong and Poellabauer, 2017; Lam et al., 2019) to modern architectures capturing complex cross-modal

dynamics. For instance, speech foundation models (e.g., HuBERT Hsu et al., 2021, Whisper Radford et al., 2023) have been combined with text embeddings (e.g., RoBERTa Liu et al., 2019) to integrate multimodal information (Gómez-Zaragoza et al., 2025; You et al., 2025; Maji et al., 2025). Researchers have further explored specialized audio architectures and novel training paradigms (Zhou et al., 2026; Fan et al., 2019; Dumpala et al., 2025; Liang et al., 2025; Zuo and Mak, 2025).

In parallel, advances in Large Language Models (LLMs) have enabled the high-fidelity automated analysis of clinical text, proving that psychiatric verbal data contain powerful predictive markers (Nazi and Peng, 2024; Yang et al., 2023; He et al., 2025; Wang et al., 2023; Sadeghi et al., 2023, 2024). However, the validity of automated depression assessment heavily depends on the quality of the ground truth. Standard benchmarks, such as the Extended DAIC-WOZ dataset (E-DAIC) (Gratch et al., 2014; DeVault et al., 2014), typically construct diagnostic labels by applying a fixed threshold to self-report screening tools (e.g., PHQ-8 ≥ 10). While the PHQ is a validated instrument, reducing it to a binary threshold discards severity information (Zhang et al., 2019), raising questions about its agreement with more reliable clinical markers of depression. Moreover, the over-reliance on a single benchmark dataset poses a significant risk of *overfitting*. E-DAIC was collected in a very specific context and features participants from a single data collection site. Differences in recording conditions, language, demographics, or depression phenotypes may hinder cross-corpus generalization, as also observed in facial-behaviour analysis (Sadeghi et al., 2026). While recent work has explored multilingual foundation models for psychiatric evaluation (Dong et al., 2025), bidirectional cross-corpus transfer of speech and language models with clinically validated ground truth remains understudied.

To address these gaps, we first introduce *EmpkinS-EKSpresion* (hereafter EKSpresion), a novel dataset of standardized German-language clinical interviews collected following a published study protocol (Keinert et al., 2024), with diagnostic ground truth anchored by the Structured Clinical Interview for DSM-5 (SCID-5-CV) (Beesdo-Baum et al., 2019). Together with E-DAIC, this dataset enables us to investigate three research questions: (RQ1) How accurately can textual, acoustic, and multimodal representations classify depres-

sion and estimate symptom severity within the two corpora? (RQ2) How well do these models generalize bidirectionally between EKSpresion and E-DAIC across domain and language boundaries? (RQ3) How does replacing EKSpresion’s SCID-5-CV labels with labels based on the PHQ-8 threshold used for E-DAIC affect cross-corpus performance?

To answer these questions, we evaluate binary classification and continuous PHQ-8 regression within both datasets, with additional PHQ-9 regression for EKSpresion, and benchmark our E-DAIC results against prior work. The cross-corpus evaluation is restricted to binary classification. Building upon our previous dual-stream text-based framework (Sadeghi et al., 2024), we adapt it to open-weight, locally deployable multilingual LLMs and extend it with auditory information, evaluating standard acoustic features, learned speech representations, and text-audio fusion.

2 Methods

2.1 Dataset Curation

We utilized EKSpresion, a novel, preregistered dataset of standardized German-language clinical interviews¹ (Keinert et al., 2024), conducted as face-to-face, semi-structured interviews by trained psychologists. The sample was a mixed community and help-seeking sample: participants were recruited via newspaper and social-media advertisements, mailing lists, and public flyers, and individuals with depression were additionally recruited through an outpatient psychotherapy clinic and by contacting psychiatric and psychosomatic clinics, counselling centres, psychotherapists, and general practitioners. From an initial enrollment of 256 participants, four were excluded due to technical failures or consent withdrawal, yielding a final cohort of 252 participants (126 MDD+, 126 MDD−) and 231.7 hours of recorded audio. The cohort spans a broad age range (18–75 years) and is predominantly female (69.0%), reflecting the typical demographic composition of depression studies (Albert, 2015). Diagnoses were established with the German version of the SCID-5-CV (Beesdo-Baum et al., 2019). Although the SCID-5-CV is a DSM-5 instrument, its German version provides a mapping from DSM-5 diagnoses to ICD-10 codes (Beesdo-Baum et al., 2019), as is standard clinical practice;

¹<https://osf.io/mswfg/>

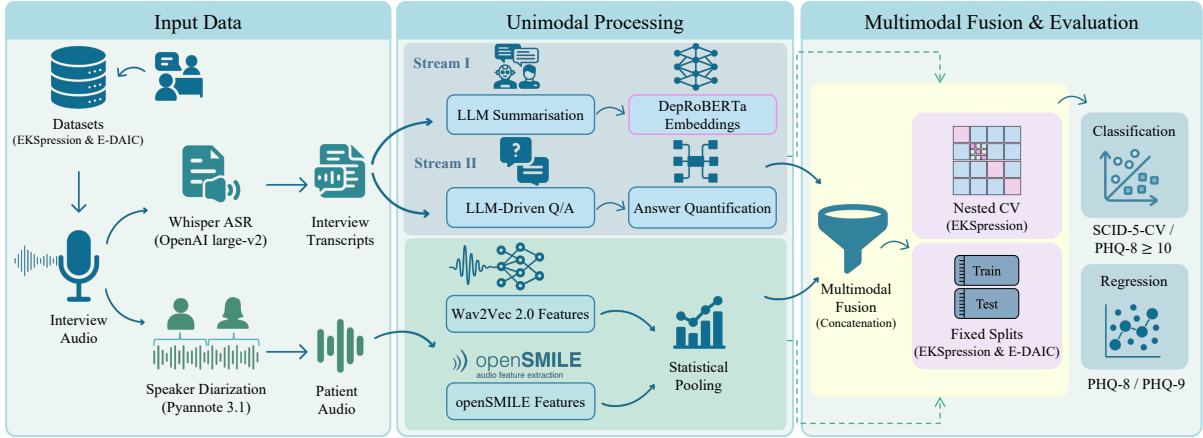


Figure 1: Overview of the proposed pipeline. (Left) Input data. (Middle) Unimodal processing of transcripts (via Llama 3 variants) and audio (Wav2Vec 2.0 Baevski et al., 2020; openSMILE Eyben et al., 2010). (Right) Fusion and model training. Pink borders indicate steps executed on training data (within CV folds or fixed training splits). Dashed green arrows show unimodal paths.

accordingly, the MDD+ group comprises participants fulfilling the criteria for an acute or recurrent depressive episode (ICD-10 F32 or F33). The MDD− group comprises individuals who did not meet criteria for a current depressive episode and reported a PHQ-9 score at or below 10. To maximise external validity, we deliberately did not apply further exclusion criteria; psychiatric comorbidities were common and systematically recorded. In the MDD− group ($n = 126$), 69 participants (55%) had no confirmed psychiatric diagnosis, 40 (32%) had a remitted or past depressive disorder with no current episode, and 29 (23%) had at least one non-depressive diagnosis, most frequently ADHD ($n = 12$) and anxiety disorders (phobic F40: $n = 8$; other anxiety F41: $n = 6$); these categories overlap, and no MDD− participant met criteria for a current depressive episode. In the MDD+ group ($n = 126$), 81 participants (64%) had at least one comorbid diagnosis, most frequently anxiety disorders (phobic F40: $n = 30$; other anxiety F41: $n = 18$; OCD F42: $n = 9$), stress-, trauma-, and adjustment-related disorders (F43: $n = 23$), persistent depressive disorder/dysthymia (F34: $n = 22$), and ADHD (F90: $n = 17$); 35 participants had two or more comorbid categories. The task is therefore depression detection within a realistic, clinically heterogeneous population rather than depression versus healthy controls.

To assess the reliability of the group assignment, a random 10% of the video-recorded interviews were independently re-rated by a second rater. Inter-rater agreement, quantified by Cohen’s κ (Cohen, 1960), was excellent ($\kappa = 1$;

all cases assigned to the same diagnostic group), where $\kappa \geq .80$ is considered satisfactory (Lan-dis and Koch, 1977). More broadly, the German SCID-5-CV has demonstrated satisfactory reliability (Beesdo-Baum et al., 2019), and large-scale studies confirm its diagnostic reliability and validity (Pettersson et al., 2015). As shown in Table 1, the two groups are well matched on age (Mann-Whitney U , $p = 0.654$) and sex (χ^2 , $p = 0.962$) and differ markedly in PHQ-8 and PHQ-9 ($p < 0.001$ each), consistent with the assigned diagnoses. Since group membership is defined by the clinician-administered SCID-5-CV rather than a self-report cut-off, the two labelling systems do not coincide: 17 MDD+ participants (13.5%) report a PHQ-9 below 10, while 32 MDD− participants (25.4%) score at or above the PHQ-8 ≥ 10 screening threshold. This divergence between clinical diagnosis and self-report screening is precisely what our two label settings are designed to probe. Note that the PHQ-8 was administered at online screening to guide recruitment, whereas the PHQ-9 was completed at the diagnostic assessment and contributed to the MDD− group definition; as the two were administered at different time points, although both cover the preceding two weeks, their scores can differ (see Limitations).

2.2 Textual Modality Processing

To approximate a clinician’s assessment, the textual processing stream (Figure 1) follows a three-stage logic: transcription via automatic speech recognition (ASR), consolidation of the dialogue into a readable narrative, and extraction of depression-

Demographic	MDD+ (<i>n</i> = 126)	MDD− (<i>n</i> = 126)
Age, mean ± SD	33.5 ± 13.2	36.5 ± 16.5
Sex (F / M / Div.)	88 / 36 / 2	86 / 38 / 2
PHQ-8, mean ± SD	14.1 ± 4.4	6.2 ± 4.1
Minimal (0–4)	2 (1.6%)	56 (44.4%)
Mild (5–9)	17 (13.5%)	38 (30.2%)
Moderate (10–14)	54 (42.9%)	27 (21.4%)
Mod. Severe (15–19)	37 (29.4%)	5 (4.0%)
Severe (20–24)	16 (12.7%)	0 (0.0%)
PHQ-9, mean ± SD	14.7 ± 4.7	4.4 ± 2.5
Minimal (0–4)	0 (0.0%)	67 (53.2%)
Mild (5–9)	17 (13.5%)	58 (46.0%)
Moderate (10–14)	46 (36.5%)	1 (0.8%)
Mod. Severe (15–19)	45 (35.7%)	0 (0.0%)
Severe (20–27)	18 (14.3%)	0 (0.0%)

Table 1: Cohort characteristics by group (SCID-5-CV; *N* = 252). Age: Mann-Whitney *U*, *p* = 0.654; Sex: χ^2 , *p* = 0.962; PHQ-8: *p* < 0.001; PHQ-9: *p* < 0.001. Severity categories follow (Kroenke et al., 2001).

relevant semantic representations. We bypass speaker diarization, relying on the contextual understanding of LLMs to distinguish patient responses within the full transcript, which also preserves the diagnostically informative patient–interviewer interaction (Agarwal et al., 2024). Both LLM streams are prompted to answer in English regardless of input language (Appendix A), so the same English DepRoBERTa applies across both corpora with no separate translation step.

2.2.1 ASR Pipeline

We utilized OpenAI’s Whisper “large-v2” (Radford et al., 2023) ASR model to generate high-fidelity transcripts from the raw audio. This architecture was chosen for its zero-shot capabilities; while maintaining a 2.7% Word Error Rate on LibriSpeech (Panayotov et al., 2015), it provides a 55.2% relative error reduction on out-of-distribution data compared to supervised models (Radford et al., 2023). We avoided manual transcript correction to retain the raw ASR output, challenging downstream models to handle real-world speech imperfections and providing a realistic assessment of fully automated clinical deployment. Building upon our previous work (Sadeghi et al., 2024), we implemented a dual-stream feature extraction architecture using locally deployable models. We integrated three instruction-tuned variants, Llama 3.1 8B, Llama 3.1 70B, and Llama 3.3 70B (Grattafiori et al., 2024), to evaluate the trade-offs between computational efficiency and diagnostic accuracy. These models drive the two subsequent feature extraction streams

described below.

2.2.2 Stream I: Global Context Embeddings

The first stream captures the global semantic context of the patient’s depression status. As raw ASR transcripts contain irrelevant chatter that dilutes the clinical signal, we employ a Transcript Transformation step. Following our previous work (Sadeghi et al., 2024), we augmented the baseline summary query with a System Message to enforce a defined persona and output boundaries detailed in Appendix A. This transformation distills the raw interview into a condensed summary, filtering out noise to isolate salient diagnostic features. These summaries are then processed by DepRoBERTa (Poświata and Perełkiewicz, 2022), a RoBERTa-based model (Liu et al., 2019) pre-trained on depression-related social media text. To adapt this model from its original English training data to our specific clinical domain, we fine-tuned it on our transformed transcripts. To align the continuous PHQ-8/9 scores with DepRoBERTa’s pre-training objective, we applied a three-class discretization scheme following (Sadeghi et al., 2024): *Severe* (≥ 14), *Moderate* ($7 - 13$), and *Not depression* (< 7). Although our cohort exhibits a well-distributed symptom range (Table 1), we retained this simplified mapping to maintain architectural compatibility with the model’s original classification objective. Therefore, this discretization serves as a technical adaptation for feature extraction, distinct from gold-standard clinical categories. The output of this stream is a 3-dimensional probability vector representing the likelihood of these diagnostic classes.

2.2.3 Stream II: LLM-Driven Q&A

While Stream I captured general contextual patterns, Stream II targets specific DSM-5 criteria via LLM-driven question-based feature extraction. This stream directly ingests the raw ASR transcripts to prevent the loss of granular symptomatic details. We leveraged Llama to simulate a structured assessment by querying 11 symptom-targeted questions covering somatic, cognitive, and affective domains (Sadeghi et al., 2024); these were derived in our previous work to discriminate depression and are not a clinically validated instrument. The complete prompt is detailed in Appendix A. To quantify symptoms, we mapped the constrained model outputs “YES”, “NO”, and “To Some Extent” to 1.0, 0.0, and 0.5, respectively. Instances of “Not Men-

tioned” were handled via mean imputation across that specific dimension for all participants. This process generated an 11-dimensional symptom vector. Finally, we concatenated this vector with the global probability vector from Stream I ($d = 3$) to yield a 14-dimensional textual feature set integrating both global context and specific diagnostic criteria.

2.3 Acoustic Modality Processing

The acoustic stream extracts paralinguistic features under two conditions: diarized (patient-isolated) and non-diarized (full recording), evaluated independently to assess the impact of speaker separation.

2.3.1 Preprocessing and Diarization

Audio recordings were resampled to 16 kHz mono. We used Pyannote.audio (Plaquet and Bredin, 2023; Bredin, 2023) to identify two speakers, isolating the patient as the speaker with the longest total vocal duration (Oesten et al., 2023). For the EKSpression dataset, features were extracted under two conditions: a) from the full recording without segmentation, to avoid diarization cascade errors and implicitly capture conversational dynamics, and b) from patient-only segments to compute patient-level statistics. For E-DAIC, audio was extracted between the first and last transcribed utterances to exclude non-interview speech.

2.3.2 Representation Extraction

We extracted embeddings using Wav2Vec 2.0 XLSR-53 to capture generalizable paralinguistics without task-specific supervision. For within-corpus cross-validation on our dataset, we used a German fine-tuned version (Grosman, 2021); for cross-corpus evaluations, the base multilingual model pre-trained on 53 languages (Baevski et al., 2020; Facebook AI, 2020). Audio was processed in 20-second segments, extracting 1024-dimensional latent representations from the Transformer hidden layers, aggregated using six statistical functionals (minimum, maximum, mean, standard deviation, skewness, kurtosis), yielding a 6144-dimensional vector.

2.3.3 Knowledge-Driven Features

To complement Wav2Vec 2.0 with interpretable baselines, we extracted eGeMAPS features (Eyben et al., 2015) using openSMILE (Eyben et al., 2010). As standard eGeMAPS functionals performed poorly in preliminary experiments, we man-

ually aggregated low-level descriptors using additional functionals: mean, standard deviation, minimum, maximum, median, interquartile range, skewness, and kurtosis, yielding a 200-dimensional vector per participant.

2.4 Model Evaluation and Multimodal Fusion

For unimodal text experiments, we used a 14-dimensional feature vector consisting of 3 DepRoBERTa depression probability scores and 11 LLM-extracted symptom features. Unimodal audio experiments evaluated eGeMAPS ($d = 200$) and Wav2Vec 2.0 representations ($d = 6144$) independently. For multimodal evaluation, we employed an early fusion scikit-learn (Pedregosa et al., 2011) pipeline. To prevent high-dimensional audio representations from overwhelming the compact text features, audio vectors were standardized (StandardScaler) and subjected to dynamic dimensionality reduction, while text features were passed through unaltered. We implemented a custom feature selection module evaluating four strategies: Principal Component Analysis (PCA), ANOVA F-values, Mutual Information (MI), and Complementary PCA (cPCA), in which audio features are pre-filtered via ANOVA, reduced via PCA, and pruned of components highly correlated with text features ($|r| > 0.70$). Each strategy optimized target dimension $k \in \{50, 100, 200\}$.

Within-corpus generalization was estimated via nested cross-validation (CV): a 5-fold Stratified K-Fold outer loop for classification and standard K-Fold for regression, with an inner 3-fold CV optimizing feature selection and model hyperparameters via GridSearchCV. Within each outer fold, DepRoBERTa was fine-tuned on the training data using layer-specific learning rates (Appendix B). All dimensionality reduction, feature scaling, and DepRoBERTa fine-tuning occur strictly within the training partition. For cross-corpus evaluations and comparability with state-of-the-art, we replicated this pipeline using the official E-DAIC partitions (train/development/test: 163/56/56). To avoid dependence on any single split of the EKSpression dataset, we repeated each transfer over all five cross-validation folds and report the mean $\pm$ SD across folds. We evaluated 16 models for both binary classification and regression: (Logistic) Regression, Ridge, Lasso, ElasticNet, Support Vector Classifier and Regressor (SVC/SVR), Random Forest, XGBoost, AdaBoost, Decision Tree, k -Nearest Neighbors (kNN), Extra Trees, Gradient Boost-

ing, Multi-Layer Perceptron (MLP), LightGBM, and CatBoost. Hyperparameter grids were optimized via inner 3-fold CV (EKSpresion dataset) or on the development partition (E-DAIC); details in Appendix B. Classification performance was evaluated using the F1-score of the depressed class and Area Under the Curve (AUC); regression used Mean Absolute Error (MAE) and the coefficient of determination (R^2).²

2.5 Cross-Corpus Evaluation

We assessed out-of-distribution transferability via a bidirectional zero-shot cross-corpus analysis between the EKSpresion dataset and E-DAIC (Gratch et al., 2014; DeVault et al., 2014), probing robustness against domain shift (psychologist-administered SCID-5-CV vs. virtual-agent interviews) and language shift (German vs. English). We evaluated both directions: (1) EKSpresion $\rightarrow$ E-DAIC, where a model is trained on the training partition of EKSpresion and evaluated on the test partition of E-DAIC; (2) E-DAIC $\rightarrow$ EKSpresion, which represents the reverse direction. To isolate the effect of label mismatch, we used two configurations: Setting 1 paired our SCID-5-CV diagnoses with E-DAIC’s PHQ-8 ≥ 10 threshold; Setting 2 redefined our positive class using the same PHQ-8 ≥ 10 threshold to harmonize labels. For the textual modality, we bypassed E-DAIC’s official transcripts and applied our Whisper ASR pipeline directly to the raw audio. Transcripts underwent dual-stream feature extraction (Sections 2.2.2 and 2.2.3) across all three Llama variants, yielding the 14-dimensional textual vector per participant. DepRoBERTa was fine-tuned on the source training partition with identical optimization to Section 2.4; features were extracted from the target test partition without domain adaptation. For the acoustic modality, we recomputed eGeMAPS and Wav2Vec 2.0 features directly from raw E-DAIC audio to ensure an aligned feature space across languages, bypassing official feature releases. All dimensionality-reduction strategies were fitted on the source training partition to prevent target-domain leakage. Hyperparameters were selected using only source-domain data: the E-DAIC development set when E-DAIC was the source and the EKSpresion training partition when EKSpresion was the source, without access

to the target corpus. Finally, we applied SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to selected interpretable cross-corpus fusion models.

3 Results and Discussion

We first report 5-fold CV in-domain results for EKSpresion (Sections 3.1–3.3), followed by cross-corpus generalization (Section 3.4), averaged over all five folds.

3.1 Textual Modality

Llama 3.3-70B consistently outperformed earlier variants across both tasks (Table 2). For classification, it improved F1 from 0.84 for the best Llama 3.1 model to 0.86 and AUC from 0.86 to 0.92, indicating stronger discriminative power. For regression, it achieved MAE=3.48 and R^2 =0.40, compared with R^2 =0.30 and 0.31 for Llama 3.1-8B and Llama 3.1-70B, respectively. Thus, while the smaller models detect general depressive markers, Llama 3.3-70B provides better severity estimation.

3.2 Acoustic Modality

Acoustic models showed limited predictive power compared to text (Table 2). Using non-diarized Wav2Vec 2.0 features, the best classification reached F1=0.68 (AUC=0.65) and the best regression MAE=4.41. Diarization produced mixed effects: F1 decreased slightly to 0.67 and MAE increased to 4.49, whereas AUC improved to 0.68. Our main hypothesis is that diarization errors contribute to the small losses in F1 and MAE; alternatively, models may benefit from the overall conversational context. A thorough investigation is left for future work. eGeMAPS features showed a similar ceiling: non-diarized classification reached F1=0.59 (AUC=0.59), increasing slightly to F1=0.60 (AUC=0.60) with diarization. For regression, diarized eGeMAPS achieved MAE=4.75 (R^2 =0.04), versus MAE=4.62 (R^2 =0.08) without diarization. While our statistical pooling approach, incorporating skewness and kurtosis, was designed to capture global prosodic shifts such as reduced vocal variability and monotone speech associated with depression (Menne et al., 2024; Bauer et al., 2024), the performance gap versus text likely reflects the nature of the SCID-5-CV assessment, where diagnosis is driven primarily by explicit symptom reports rather than paralinguistic cues, limiting the standalone predictive power of acoustic features in this setting.

²Source code is publicly available:
<https://github.com/MishaSadeghi/cross-corpus-llm-speech-depression-detection>

Feature Configuration	Classification (SCID-5-CV)				Regression (PHQ-8)	
	Pipeline	F1 $\uparrow$	AUC $\uparrow$	Pipeline	MAE $\downarrow$	R^2 $\uparrow$
<i>Text</i>						
Unimodal (L3.1-8B)	CatBoost	0.83 $\pm$ 0.04	0.85 $\pm$ 0.07	kNN	3.73 $\pm$ 0.56	0.30 $\pm$ 0.09
Unimodal (L3.1-70B)	Extra Trees	0.84 $\pm$ 0.03	0.86 $\pm$ 0.03	Ridge	3.88 $\pm$ 0.43	0.31 $\pm$ 0.10
Unimodal (L3.3-70B)	Extra Trees	0.86 $\pm$ 0.05	0.92 $\pm$ 0.03	CatBoost	3.48 $\pm$ 0.43	0.40 $\pm$ 0.15
<i>Audio</i>						
Unimodal (W2V, ND)	Extra Trees (PCA)	0.68 $\pm$ 0.02	0.65 $\pm$ 0.06	Random Forest (PCA)	4.41 $\pm$ 0.44	0.15 $\pm$ 0.10
Unimodal (W2V, D)	Extra Trees (PCA)	0.67 $\pm$ 0.03	0.68 $\pm$ 0.07	Ridge (PCA)	4.49 $\pm$ 0.32	0.11 $\pm$ 0.17
Unimodal (eGeMAPS, ND)	MLP (ANOVA)	0.59 $\pm$ 0.08	0.59 $\pm$ 0.12	Ridge (cPCA)	4.62 $\pm$ 0.54	0.08 $\pm$ 0.09
Unimodal (eGeMAPS, D)	Random Forest (MI)	0.60 $\pm$ 0.08	0.60 $\pm$ 0.07	SVR (ANOVA)	4.75 $\pm$ 0.45	0.04 $\pm$ 0.11
<i>Text + Audio</i>						
Fusion (L3.1-8B + W2V, D)	Random Forest (PCA)	0.84 $\pm$ 0.05	0.85 $\pm$ 0.06	Extra Trees (cPCA)	3.75 $\pm$ 0.46	0.33 $\pm$ 0.05
Fusion (L3.1-8B + eGeMAPS, ND)	AdaBoost (ANOVA)	0.81 $\pm$ 0.03	0.84 $\pm$ 0.06	CatBoost (ANOVA)	3.82 $\pm$ 0.32	0.31 $\pm$ 0.05
Fusion (L3.1-70B + W2V, D)	CatBoost (PCA)	0.83 $\pm$ 0.04	0.85 $\pm$ 0.04	CatBoost (PCA)	3.82 $\pm$ 0.47	0.32 $\pm$ 0.13
Fusion (L3.1-70B + eGeMAPS, D)	CatBoost (PCA)	0.82 $\pm$ 0.04	0.84 $\pm$ 0.04	CatBoost (ANOVA)	3.97 $\pm$ 0.29	0.26 $\pm$ 0.07
Fusion (L3.3-70B + W2V, ND)	AdaBoost (PCA)	0.86 $\pm$ 0.03	0.91 $\pm$ 0.04	AdaBoost (PCA)	3.40 $\pm$ 0.53	0.42 $\pm$ 0.19
Fusion (L3.3-70B + eGeMAPS, D)	LightGBM (ANOVA)	0.86 $\pm$ 0.07	0.90 $\pm$ 0.06	CatBoost (MI)	3.51 $\pm$ 0.38	0.40 $\pm$ 0.15

Table 2: EKSpresion classification and PHQ-8 regression (mean $\pm$ SD; 5-fold CV). D/ND: diarized/non-diarized; cPCA: Complementary PCA; MI: Mutual Information; L3.1/L3.3: Llama 3.1/3.3; W2V: Wav2Vec 2.0. Bold: best per modality group, with lower SD breaking ties after rounding. Recall, RMSE, and PHQ-9 results are reported in Appendix D.

3.3 Multimodal Fusion

Contrary to our expectation that acoustic features would provide complementary cues, early fusion produced no classification gain and only marginal regression improvements (Table 2). The Llama 3.3-70B fusions with non-diarized Wav2Vec 2.0 and diarized eGeMAPS both matched the best text-only F1 of 0.86, with AUCs of 0.91 and 0.90 versus 0.92 for text alone. The best fusion regression marginally improved MAE from 3.48 to 3.40 and R^2 from 0.40 to 0.42, but adding acoustic features provided no consistent benefit. Previous work on the (E-)DAIC corpus has also consistently found higher performance for text-based models, with little-to-no improvement or even degradation from including auditory information (Qureshi et al., 2019; Ray et al., 2019; Ringeval et al., 2019; Williamson et al., 2016). Dong et al. (2025) further identifies modality imbalance as a prevalent issue across depression datasets, where fusion frequently fails to surpass the best unimodal model. This is explained by the nature of the spoken content: semi-structured interviews intentionally discuss depressive symptoms, often mentioning an existing diagnosis, providing rich cues that current language models capture well. In contrast, prosodic and acoustic markers appear weaker and more individual-dependent (Lewis et al., 2025). These patterns extend to our corpus, given its similar semantic content to E-DAIC.

3.4 Benchmarking and Cross-Corpus Generalization

Tables 3 and 4 present the within-corpus performance of our proposed framework on the standard

Method	Mod.	MAE $\downarrow$	RMSE $\downarrow$
State-of-the-Art (Evaluated on E-DAIC)			
Rodrigues Makiuchi et al. (2019)	T	-	6.88
Ray et al. (2019)	T	4.02	4.73
Sadeghi et al. (2023)	T	4.26	5.36
Sadeghi et al. (2024)	T	3.86	4.66
Tank et al. (2024) [†]	T	3.16	3.98
Tank et al. (2024)	T	4.89	6.29
Mandal et al. (2025)	T	4.02	5.22
Schmidt et al. (2025)	T	3.85	4.52
Teng et al. (2025) [†]	T	3.37	-
Ringeval et al. (2019)	A	-	8.19
Zhang et al. (2019)	A	5.77	6.78
Rodrigues Makiuchi et al. (2019)	A	-	7.02
Tank et al. (2024)	A	-	5.39
Mandal et al. (2025)	A	5.14	6.35
Rodrigues Makiuchi et al. (2019)	A+T	-	6.11
Mandal et al. (2025)	A+T	4.12	5.27
Our Approach (In-Domain E-DAIC)			
Unimodal (eGeMAPS)	A	5.24	6.44
Unimodal (W2V)	A	4.79	5.91
Unimodal (L3.1-8B)	T	3.36	<u>4.16</u>
Unimodal (L3.1-70B)	T	3.86	4.92
Unimodal (L3.3-70B)	T	3.68	4.75
Fusion (L3.1-8B + W2V)	A+T	3.41	4.43
Fusion (L3.1-70B + W2V)	A+T	3.65	4.54
Fusion (L3.3-70B + W2V)	A+T	3.69	4.74
Fusion (L3.3-70B + eGeMAPS)	A+T	3.64	4.71

Table 3: E-DAIC Regression Benchmark (PHQ-8). Bold/underlined: best/second-best. [†] Proprietary models (GPT-4/GPT-4o). Mod.: Modality (A: Acoustic, T: Text); L3.1/L3.3: Llama 3.1/3.3; W2V: Wav2Vec 2.0; -: Not Reported.

E-DAIC benchmark to ensure contextual comparability with existing literature. For the regression task (Table 3), our locally deployable unimodal L3.1-8B approach achieved an MAE of 3.36 and an RMSE of 4.16. We note that the current state-of-the-art of Tank et al. (Tank et al., 2024) is achieved utilizing a closed-source, cloud-based GPT-4 model (MAE 3.16), whereas our locally deployable alternative outperforms their own open-source implementation (MAE 4.89). This

Method	Train	Mod.	E-DAIC	EKS
Prior Work (Evaluated on E-DAIC)				
Ali et al. (2024) [†]	– (ZS)	T	0.65	-
Guo et al. (2025)	E-DAIC	T	0.64	-
Zhang et al. (2025)	E-DAIC	T	0.86	-
Cheong et al. (2024)	E-DAIC	A	0.56	-
Ali et al. (2024) [†]	– (ZS)	A	0.56	-
Ali et al. (2024) [†]	– (ZS)	A+T	0.60	-
Our Approach (Within-Corpus & Cross-Corpus)				
<i>Setting 1: EKS uses SCID labels, E-DAIC uses PHQ-8 ≥ 10 labels</i>				
eGeMAPS	E-DAIC	A	0.55	0.67 $\pm$ 0.03
eGeMAPS	EKS	A	0.48 $\pm$ 0.02	0.60 $\pm$ 0.08
W2V	E-DAIC	A	0.52	0.68 $\pm$ 0.02
W2V	EKS	A	0.48 $\pm$ 0.02	0.68 $\pm$ 0.02
L3.1-8B	E-DAIC	T	0.78	0.78 $\pm$ 0.02
L3.1-8B	EKS	T	0.66 $\pm$ 0.05	0.83 $\pm$ 0.04
L3.1-70B	E-DAIC	T	0.80	0.78 $\pm$ 0.04
L3.1-70B	EKS	T	0.74 $\pm$ 0.06	0.84 $\pm$ 0.03
L3.3-70B	E-DAIC	T	0.79	0.82 $\pm$ 0.02
L3.3-70B	EKS	T	0.62 $\pm$ 0.12	0.86 $\pm$ 0.05
L3.3-70B+eGeMAPS	E-DAIC	A+T	0.68	0.80 $\pm$ 0.03
L3.3-70B+eGeMAPS	EKS	A+T	0.49 $\pm$ 0.04	0.86 $\pm$ 0.07
L3.1-8B+W2V	E-DAIC	A+T	0.73	0.81 $\pm$ 0.04
L3.1-8B+W2V	EKS	A+T	0.55 $\pm$ 0.13	0.84 $\pm$ 0.05
L3.1-70B+W2V	E-DAIC	A+T	0.70	0.78 $\pm$ 0.03
L3.1-70B+W2V	EKS	A+T	0.72 $\pm$ 0.06	0.82 $\pm$ 0.07
L3.3-70B+W2V	E-DAIC	A+T	0.71	<u>0.83 $\pm$ 0.02</u>
L3.3-70B+W2V	EKS	A+T	0.62 $\pm$ 0.09	0.86 $\pm$ 0.03
Avg. (all)	E-DAIC	A, T, A+T	0.70	0.77
Avg. (all)	EKS	A, T, A+T	0.60	0.80
<i>Setting 2: Both EKS and E-DAIC datasets use PHQ-8 ≥ 10 labels</i>				
L3.1-70B	E-DAIC	T	0.80	0.79 $\pm$ 0.05
L3.1-70B	EKS	T	<u>0.75 $\pm$ 0.02</u>	0.79 $\pm$ 0.08
L3.3-70B	E-DAIC	T	0.79	0.84 $\pm$ 0.06
L3.3-70B	EKS	T	0.68 $\pm$ 0.03	0.85 $\pm$ 0.05
L3.1-8B+W2V	E-DAIC	A+T	0.73	0.83 $\pm$ 0.04
L3.1-8B+W2V	EKS	A+T	0.55 $\pm$ 0.11	0.82 $\pm$ 0.04
L3.1-70B+W2V	E-DAIC	A+T	0.70	0.80 $\pm$ 0.06
L3.1-70B+W2V	EKS	A+T	0.68 $\pm$ 0.09	0.81 $\pm$ 0.08
Avg. (all)	E-DAIC	A, T, A+T	0.70	0.79
Avg. (all)	EKS	A, T, A+T	0.60	0.80

Table 4: Prior work and our within- and cross-corpus depression classification results. Values are depressed-class F1, except for Cheong et al., who report weighted F1. Mod.: A/T/A+T = acoustic/text/both; ‘+’ in our model names marks early fusion. Our results are mean $\pm$ SD over five folds; E-DAIC within-corpus values use the official fixed split. Bold/underlined: best within-/cross-corpus per section (ties broken by lower SD). [†]: E-DAIC development set (other prior work: test set); ZS: zero-shot. EKS: EKSprespression; eGeMAPS: acoustic feature set; W2V: Wav2Vec 2.0; L3.1/L3.3: Llama 3.1/3.3. Setting 2 shows selected models; its averages include all models reported in Appendix E.

demonstrates the high efficacy of our specific LLM prompting and extraction pipeline.

Similarly, for binary classification (Table 4), our best text-only pipeline (L3.1-70B) achieved an F1-score of 0.80. This closely approaches the state-of-the-art F1-score of 0.86 reported by Zhang et al. (Zhang et al., 2025). While their Multi-Instance Learning framework processes individual sentences independently, our approach achieves competitive results by distilling the entire conversational context into a holistic clinical summary via

locally deployable LLMs.

Importantly, Table 4 also contains our main results – the cross-corpus evaluation between EKSprespression and E-DAIC, using two separate labeling schemes for our data. We first focus on the original labels provided by each dataset (*Setting 1*). While the clinical construct differed (SCID-5-CV vs. PHQ-8 ≥ 10), the overall *goal* was the same: to classify individuals as suffering from depression or not.

In general, we observe that cross-corpus generalization was highly *asymmetric*; models trained on E-DAIC generalized much better on the EKSprespression dataset than the other way around. Training on E-DAIC and evaluating on EKSprespression yielded an average F1-score of 0.77 (best: 0.83), whereas the opposite led to an average of 0.60 (best: 0.74). Interestingly, the average cross-corpus performance when trained on E-DAIC and evaluated on our dataset was higher than the in-domain performance on E-DAIC (0.77 vs 0.70). We hypothesize that this difference is partially caused by the severe label imbalance in E-DAIC, where about 75% of participants fall below the PHQ-8 cutoff; in contrast, our proposed dataset was intentionally constructed to be balanced. Nevertheless, using an identical labeling scheme for both datasets (PHQ-8 ≥ 10 ; *Setting 2*) produced only modest changes: the average E-DAIC $\rightarrow$ EKSprespression performance increased to 0.79 (best: 0.84), whereas the average EKSprespression $\rightarrow$ E-DAIC performance remained 0.60 (best: 0.75). Thus, the asymmetric cross-corpus performance persisted, suggesting that the difference in labels alone cannot explain it. Finally, we note that text-based models show strong generalization performance despite the difference in language (English vs. German). This was explicitly built into our pipeline, as our summarization models were instructed to output their summaries in English; this allowed us to apply the same DepRoBERTa model across both datasets.

3.5 Cross-Corpus Explainability

We applied SHAP (Lundberg and Lee, 2017) to an interpretable Extra Trees classifier combining the 14 textual features with eGeMAPS acoustic features, fitted on the source training data of each direction: the official E-DAIC split and the median-performing EKSprespression cross-validation fold, respectively (model details in Appendix H). SHAP values were computed over the source training data to characterize the learned representations and iden-

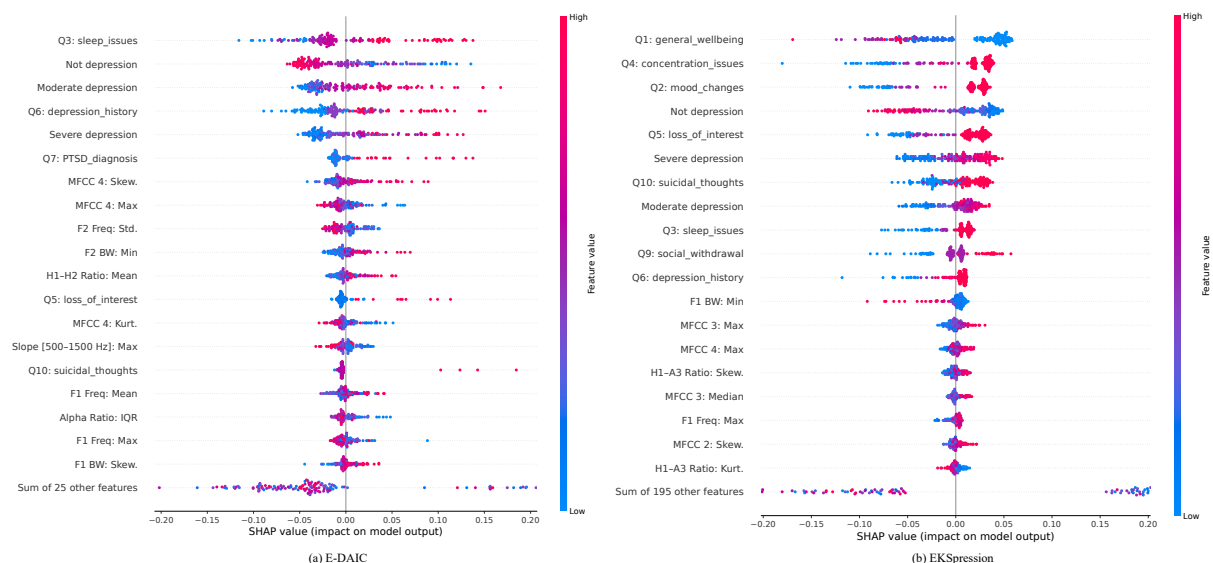


Figure 2: SHAP beeswarm plots of the top features for the Extra Trees eGeMAPS+text fusion, trained on (a) E-DAIC (official split) and (b) the EKSpresion dataset (median-performing CV fold). ‘Q’ denotes LLM-driven question features; ‘Not/Moderate/Severe depression’ indicates DepRoBERTa probability scores. Acoustic feature suffixes denote standard statistical functionals (e.g., skewness, standard deviation, kurtosis, and interquartile range).

tify source-specific biases. Figure 2 presents features ranked by mean absolute SHAP value for the Llama 3.3-70B integration in both directions.

In the E-DAIC $\rightarrow$ EKSpresion direction (Plot a), sleep issues (Q3) are most influential, followed by the DepRoBERTa probabilities, depression history (Q6), and PTSD (Q7); the latter likely reflects E-DAIC’s military-weighted sample (Gratch et al., 2014). The leading acoustic cues are complementary: MFCC 4 skewness and maximum (spectral-envelope and voice-quality variation), F2 frequency variability and bandwidth (vocal-tract resonance and articulation), and H1-H2 (breathiness of phonation), together indicating the reduced articulatory dynamics and altered phonation of psychomotor slowing (Cummins et al., 2015; Scherer et al., 2013). In the EKSpresion $\rightarrow$ E-DAIC direction (Plot b), general wellbeing (Q1), concentration (Q4), and mood changes (Q2) dominate, spanning functional, cognitive, and affective symptoms, followed by the DepRoBERTa probabilities, loss of interest (Q5), and suicidal thoughts (Q10). The recurrence of Q3, Q5, Q6, Q10, and the DepRoBERTa probabilities across directions indicates transferable semantic markers, whereas acoustic features rank lower with smaller SHAP magnitudes (panel b). Overall, in this setting, LLM-extracted symptoms transfer more robustly across interview paradigms than acoustic patterns, which

remain highly sensitive to recording environment and corpus-specific demographics.

4 Conclusion

This study introduced a clinically validated German dataset of semi-structured diagnostic interviews and investigated the cross-corpus generalizability of depression detection models. By anchoring diagnostic labels to the SCID-5-CV rather than self-report proxy thresholds, our results indicate that automated systems can approximate formal clinical judgments. Text-based models vastly outperformed audio-based ones, with fusion providing only marginal improvements, if any. We attribute this result to the abundant linguistic information in semi-structured interviews, where conversations explicitly target diagnostically relevant topics. The evaluated models also maintained their predictive power in zero-shot cross-corpus transfer to E-DAIC, suggesting that multilingual language models can support depression detection across the two corpora and languages studied. This supports the development of clinical decision support (CDS) systems that assist clinicians throughout the diagnostic process, from early screening to formal assessment. Future work could evaluate generalizability across additional corpora and improve the predictive power of audio-based models.

Limitations

This study has several limitations. First, while participants were recruited from varied backgrounds, all were fluent German speakers and interviews were conducted at a single clinical site, which may limit generalizability to other languages, healthcare systems, and cultural contexts where depression phenotypes or interview dynamics differ substantially. Second, four missing PHQ-9 scores were imputed using the cohort mean, a simplification that could introduce a small bias in the regression task. Relatedly, the PHQ-8 and PHQ-9 were administered at different stages of the study: the PHQ-8 at online screening to guide recruitment and the PHQ-9 at the later diagnostic session contributed to group assignment. Although both cover the preceding two weeks, the two self-report scores reflect different time points and should not be read as a single cross-sectional snapshot. Third, our acoustic models show limited predictive power and limited cross-corpus generalization, likely due to differences in recording conditions, microphone types, ambient noise, and language between datasets; developing acoustic representations robust to such domain and cross-lingual shifts remains an open challenge. Fourth, our pipeline relies on Whisper ASR without manual correction, meaning transcription errors could propagate to downstream models; the impact of ASR quality on diagnostic accuracy warrants further investigation. Finally, while our models are evaluated on two datasets spanning two languages, broader validation across additional corpora, languages, and clinical settings is needed before deployment in real-world clinical workflows.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and received approval from the ethics committee of Friedrich-Alexander-Universität Erlangen-Nürnberg (443_20 B). The study protocol was preregistered on the Open Science Framework (<https://osf.io/mswfg/>) and retrospectively registered in the German Clinical Trials Register on November 25, 2024 (Clinical Trial Number: DRKS00035577). Any protocol modifications were documented in the preregistration and submitted to the ethics committee as amendments. All participants provided written informed consent in two stages: initially via UniPark for the screening phase, and subsequently before the first diagnostic session for the main study.

To ensure data pseudonymization, enrolled participants were assigned a unique personal code. The password-protected key linking these codes to identifying information was stored securely and independently from the study data, accessible only to authorized research personnel.

To protect participant privacy, all processing was performed on institutional infrastructure using locally deployable models, ensuring that no sensitive clinical audio or transcripts were transmitted to external servers. This deliberate design choice aligns directly with strict healthcare data regulations.

The proposed system is intended solely as a CDS tool to assist trained clinicians and is not designed to replace human diagnostic judgment. Researchers and clinicians deploying such systems should be aware that model performance may vary across demographic groups, languages, and clinical settings not represented in the training data, and should validate system outputs accordingly before clinical use.

To balance reproducibility with participant privacy, the raw EKSpresion dataset will not be released. Instead, derived acoustic and textual features for EKSpresion, together with the feature-extraction code, will be made publicly available upon publication. E-DAIC-derived features will be released only if permitted under the E-DAIC license. This will allow authorized researchers to reproduce our pipeline without releasing sensitive clinical recordings or personal identifiers.

Acknowledgments

We thank the students who assisted with study execution and data collection. This study is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) - SFB 1483 – Project-ID 442419336. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU); the hardware is partially funded by the DFG. In addition, M.S. acknowledges travel support from the Munich Center for Machine Learning (MCML) to present this work at EMNLP 2026. During the preparation of this work, the authors used AI-based assistants for proofreading and code debugging; all scientific content, analysis, and conclusions are the authors' own.

References

- Navneet Agarwal, Gaël Dias, and Sonia Dollfus. 2024. Multi-view graph-based interview representation to improve depression level estimation. *Brain Informatics*, 11(1):14.
- Paul R Albert. 2015. Why is depression more prevalent in women? *Journal of Psychiatry and Neuroscience*, 40(4):219–221.
- Abdelrahman A Ali, Aya E Fouda, Radwa J Hanafy, and Mohammed E Fouda. 2024. Leveraging audio and text modalities in mental health: A study of llms performance. *arXiv preprint arXiv:2412.10417*.
- American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5 edition. American Psychiatric Association, Arlington, VA.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Jonathan F Bauer, Maurice Gerczuk, Lena Schindler-Gmelch, Shahin Amiriparian, David Daniel Ebert, Jarek Krajewski, Björn Schuller, and Matthias Berking. 2024. Validation of machine learning-based assessment of major depressive disorder from paralinguistic speech characteristics in routine care. *Depression and Anxiety*, 2024(1):9667377.
- Katja Beesdo-Baum, Michael Zaudig, and Hans-Ulrich Wittchen. 2019. *SCID-5-CV: strukturiertes klinisches Interview für DSM-5-Störungen-Klinische Version: deutsche Bearbeitung des Structured Clinical Interview for DSM-5 Disorders-Clinician version von Michael B. First, Janet BW Williams, Rhonda S. Karg, Robert L. Spitzer*. Hogrefe.
- Hervé Bredin. 2023. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th Interspeech Conference (INTERSPEECH 2023)*, pages 1983–1987. ISCA.
- Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. 2024. Fairrefuse: Referee-guided fusion for multi-modal causal fairness in depression detection. In *IJCAI*, pages 7224–7232.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49.
- David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhomme, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and 3 others. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.
- Zhongren Dong, Haotian Guo, Weixiang Xu, Huan Zhao, and Zixing Zhang. 2025. Foundation model-based evaluation of neuropsychiatric disorders: A lifespan-inclusive, multi-modal, and multi-lingual study. *IEEE Journal of Selected Topics in Signal Processing*.
- Sri Harsha Dumpala, Chandramouli S Sastry, Rudolf Uher, and Sageev Oore. 2025. Test-time training for speech-based depression detection. In *Proc. Interspeech 2025*, pages 479–483.
- Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and Khiet P. Truong. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.
- Facebook AI. 2020. Wav2vec2-large-xlsr-53. <https://huggingface.co/facebook/wav2vec2-large-xlsr-53>.
- Weiquan Fan, Zhiwei He, Xiaofen Xing, Bolun Cai, and Weirui Lu. 2019. Multi-modality depression detection via multi-scale temporal dilated cnns. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 73–80.
- Lucía Gómez-Zaragozá, Javier Marín-Morales, Mariano Alcañiz, and Mohammad Soleymani. 2025. Speech and text foundation models for depression detection: Cross-task and cross-language evaluation. In *Proc. Interspeech 2025*, pages 5253–5257.
- Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76.
- Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Lrec*, volume 14, pages 3123–3128. Reykjavik.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Amy Yang. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in German. <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german>.
- Yanrong Guo, Youwei Guo, Bingxin Yang, Jingjing Wu, Shijie Hao, and Richang Hong. 2025. Subthreshold depression detection with text-guided multimodal learning. *IEEE Transactions on Computational Social Systems*.
- Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.
- Thomas R Insel. 2017. Digital phenotyping: technology for a new science of behavior. *Jama*, 318(13):1215–1216.
- Marie Keinert, Lena Schindler-Gmelch, Lydia Helene Rupp, Misha Sadeghi, Klara Capito, Malin Hager, Farnaz Rahimi, Robert Richer, Bernhard Egger, Bjoern M Eskofier, and Matthias Berking. 2024. Facing depression: Evaluating the efficacy of the EmpkinS-EKSpression reappraisal training augmented with facial expressions—protocol of a randomized controlled trial. *BMC Psychiatry*, 24(1):896.
- Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.
- Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.
- Genevieve Lam, Huang Dongyan, and Weisi Lin. 2019. Context-aware deep learning for multi-modal depression detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Robert Lewis, Szymon Fedor, Nelson Hidalgo Julia, Joshua Curtiss, Jiyeon Kim, Noah Jones, David Mischoulon, Thomas F Quatieri, Nicholas Cummins, Paola Pedrelli, and Rosalind Picard. 2025. Towards the objective characterisation of major depressive disorder using speech data from a 12-week observational study with daily measurements. In *Proc. Interspeech 2025*, pages 494–498.
- Wenrui Liang, Rong Zhang, Xuezhen Zhang, Ying Ma, and Wei-Qiang Zhang. 2025. Depressgen: Synthetic data generation framework for depression detection. In *Proc. Interspeech*, pages 464–468.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Bubai Maji, Monorama Swain, Shazia Nasreen, Debabrata Majumdar, Rajlakshmi Guha, Aurobinda Routray, and Anders Sjøgaard. 2025. A study on the impact of foundation models on automatic depression detection from speech signals. In *Proc. Interspeech 2025*, pages 5258–5262.
- Aishik Mandal, Dana Atzil-Slonim, Tamar Solorio, and Iryna Gurevych. 2025. Enhancing depression detection via question-wise modality fusion. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 44–61.
- Felix Menne, Felix Dörr, Julia Schröder, Johannes Tröger, Ute Habel, Alexandra König, and Lisa Wagens. 2024. The voice of depression: speech features as biomarkers for major depressive disorder. *BMC psychiatry*, 24(1):794.
- Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.
- Marie Oesten, Robert Richer, Luca Abel, Nicolas Rohleder, and Bjoern M Eskofier. 2023. Vostress—voice-based detection of acute psychosocial stress. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Agneta Pettersson, Kristina Bengtsson Boström, Petter Gustavsson, and Lisa Ekselius. 2015. Which instruments to support diagnosis of depression have sufficient accuracy? a systematic review. *Nordic journal of psychiatry*, 69(7):497–508.

- Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. *arXiv preprint arXiv:2310.13025*.
- Rafał Poświata and Michał Perelkiewicz. 2022. Opi@lt-edl-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282.
- Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Anupama Ray, Siddharth Kumar, Rutvik Reddy, Pre-rana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88.
- Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12.
- Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63.
- Misha Sadeghi, Bernhard Egger, Reza Agahi, Robert Richer, Klara Capito, Lydia Helene Rupp, Lena Schindler-Gmelch, Matthias Berking, and Bjoern M Eskofier. 2023. Exploring the capabilities of a language model-only approach for depression detection in text data. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE.
- Misha Sadeghi, Robert Richer, Bernhard Egger, Lena Schindler-Gmelch, Lydia Helene Rupp, Farnaz Rahimi, Matthias Berking, and Bjoern M Eskofier. 2024. Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3(1):66.
- Misha Sadeghi, Robert Richer, Lydia Helene Rupp, Lena Schindler-Gmelch, Marie Keinert, Farnaz Rahimi, Malin Hager, Bernhard Egger, Matthias Berking, and Bjoern M. Eskofier. 2026. Do depressive facial patterns transfer across cultures and contexts? evidence from a german RCT and E-DAIC. In *Proceedings of the ACM KDD Workshop on AI for Cognitive and Mental Health Support (AI4Mental)*, Jeju, South Korea.
- Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. 2013. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Interspeech*, pages 847–851.
- Fabian Schmidt, Seyedehmoniba Ravan, and Vladimir Vlassov. 2025. Probabilistic textual time series depression detection. *arXiv preprint arXiv:2511.04476*.
- Chayan Tank, Sarthak Pol, Vinayak Katoch, Shaina Mehta, Avinash Anand, and Rajiv Ratn Shah. 2024. Depression detection and analysis using large language models on textual and audio-visual modalities. *arXiv preprint arXiv:2407.06125*.
- Shiyu Teng, Jiaqing Liu, Rahul Kumar Jain, Shurong Chai, Ruibo Hou, Tomoko Tateyama, Lanfen Lin, and Yen-Wei Chen. 2025. Enhancing depression detection with chain-of-thought prompting: From emotion to reasoning using large language models. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–6. IEEE.
- John Torous, Sandra Bucci, Imogen H Bell, Lars V Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World psychiatry*, 20(3):318–335.
- Yuqing Wang, Yun Zhao, and Linda Petzold. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine learning for healthcare conference*, pages 804–823. PMLR.
- James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrienne Schwarzentruher, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F. Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 11–18.
- CA Woody, AJ Ferrari, DJ Siskind, HA Whiteford, and MG Harris. 2017. A systematic review and meta-regression of the prevalence and incidence of perinatal depression. *Journal of affective disorders*, 219:86–92.
- World Health Organization. 2025. [Depressive disorder \(depression\)](#). Accessed 21 May 2026.
- Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan

- Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.
- Jiajun You, Shuai Wang, Xun Gong, and Xiang Wan. 2025. M3I: A multi-modal and multi-lingual depression detection framework. In *Proc. Interspeech 2025*, pages 5268–5272.
- Larry Zhang, Joshua Driscoll, Xiaotong Chen, and Reza Hosseini Ghomi. 2019. Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 47–53.
- Xu Zhang, Chenlong Li, Weisi Chen, Jiaxin Zheng, and Feihong Li. 2025. Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework. *Scientific Reports*, 15(1):6637.
- Longhui Zhou, Bin Hu, and Zhi-Hong Guan. 2026. Gaborformer: A method for depression detection through hybrid acoustic feature extraction and fusion. *Journal of Affective Disorders*, page 121172.
- Lishi Zuo and Man-Wai Mak. 2025. Leveraging ordinal information for speech-based depression classification. In *Proc. Interspeech 2025*, pages 484–488.

A LLM Prompt Architecture

Figures 3 and 4 detail the exact prompts used for the two textual feature extraction streams described in Sections 2.2.2 and 2.2.3.

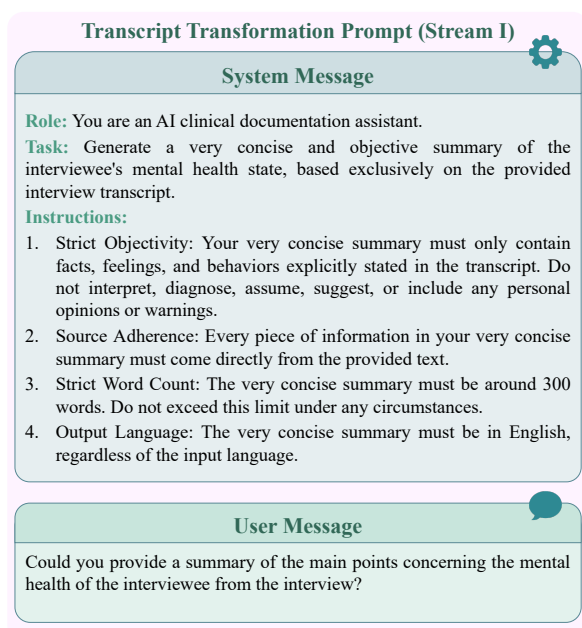


Figure 3: Stream I Transcript Transformation prompt. The System Message defines a clinical AI role, while the User Message requests a clinical summary following our previous work (Sadeghi et al., 2024).

B Hyperparameter Search Grids

Table 5 details the hyperparameter grids searched during inner cross-validation (or on the development partition for fixed-split evaluations) for all evaluated classifiers and regressors.

C Best-Found Hyperparameters (Fixed Splits)

Table 6 reports the best-found hyperparameters for the E-DAIC fixed-split experiments (selected via PredefinedSplit on the development partition). For nested cross-validation experiments on the EKSpresion dataset, best hyperparameters vary across folds; the search grids are reported in Table 5.

D Full In-Domain Results

Tables 7 and 8 report results omitted from Table 2 for space. Table 7 presents classification Recall and PHQ-8 regression RMSE. Since the EKSpresion dataset additionally includes PHQ-9 scores, unlike

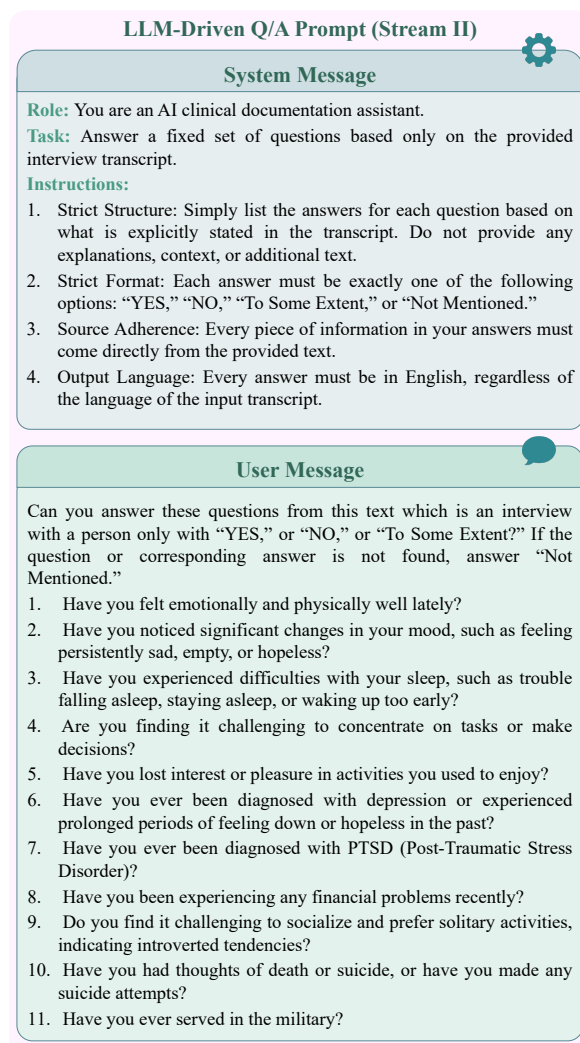


Figure 4: Stream II LLM-driven Q/A prompt. The System Message constrains the output format for 11 symptom-targeted questions adapted from our previous work (Sadeghi et al., 2024) and presented in the User Message.

E-DAIC which only provides PHQ-8, Table 8 further reports full PHQ-9 regression results (MAE, RMSE, R^2), offering a complementary view of depressive symptom severity estimation across all nine DSM-5 criteria. All results are from 5-fold cross-validation on the EKSpresion dataset.

E Full Cross-Corpus Results

Table 9 reports the Setting 2 model configurations omitted from Table 4.

F Per-Fold and Worst-Case Cross-Corpus Results

To assess the robustness of our cross-corpus results beyond the average, Table 10 reports

Model Family	Hyperparameter Grid
Tree & Ensemble	Estimators: $\{10, 50, 100, 200\}$ Depth: $\{10, 20, 30, \text{None}\}$ (standard); $\{3, 5, 7\}$ (boosting) Min. split samples: $\{2, 5, 10\}$ Learning rate: $\{0.01, 0.05, 0.1, 0.2\}$ <i>Covers: Random Forest, XGBoost, AdaBoost, Extra Trees, Gradient Boosting, LightGBM, CatBoost</i>
Linear & Distance	Regularization C/α : $\{10^{-3}, 10^{-2}, \dots, 10^2\}$ Kernel: {RBF, polynomial, sigmoid} Neighbors k : $\{3, 5, 7, 9\}$ <i>Covers: Logistic Regression, Ridge, Lasso, ElasticNet, SVC/SVR, kNN</i>
Multi-Layer Perceptron (MLP)	Layers/neurons: 1–2 layers, up to 100 neurons Activation: {rectified linear unit (ReLU), tanh} L_2 penalty: $\{10^{-4}, 10^{-3}, 10^{-2}\}$
Feature Selection	Method: {PCA, ANOVA F-values, Mutual Information, cPCA} <i>cPCA: audio features pre-filtered via ANOVA, reduced via PCA, then pruned of components correlated with text features ($r > 0.70$)</i> Target dimension k : $\{50, 100, 200\}$
DepRoBERTa Fine-tuning	Optimizer: AdamW Batch size: 32; Max token length: 512 Max epochs: 10 Encoder learning rate: 10^{-5} Classification head learning rate: 5×10^{-6}

Table 5: Hyperparameter search grids for all evaluated models and the DepRoBERTa fine-tuning configuration. cPCA: Complementary PCA; RBF: Radial Basis Function; SVC/SVR: Support Vector Classifier/Regressor; kNN: k -Nearest Neighbors.

Task	Feature	Model	Best Hyperparameters
<i>Classification</i>			
	Text (L3.1-8B)	LogReg	$C=0.001$, solver=liblinear
	Text (L3.1-70B)	MLP	hidden=(50,50), act=tanh, $\alpha=10^{-4}$
	Text (L3.3-70B)	CatBoost	depth=3, iter=200, lr=0.1, $\ell_2=3$
	eGeMAPS	LogReg	$C=0.001$, PCA $k=10$
	Wav2Vec	LogReg	$C=0.1$, PCA $k=100$
<i>Regression</i>			
	Text (L3.1-8B)	AdaBoost	lr=1.0, loss=square, $n=100$
	Text (L3.1-70B)	CatBoost	depth=7, iter=200, lr=0.05, $\ell_2=1$
	Text (L3.3-70B)	LightGBM	lr=0.1, depth=3, $n=50$, leaves=15
	eGeMAPS	Random Forest	depth=10, $n=50$, PCA $k=10$
	Wav2Vec	Random Forest	depth=10, $n=100$, PCA $k=50$

Table 6: Best-found hyperparameters for E-DAIC fixed-split experiments, selected on the development partition. LogReg: Logistic Regression; act: activation function; lr: learning rate; iter: iterations; ℓ_2 : L2 regularisation weight; n : number of estimators.

the F1-score (depressed class) for each of the five outer cross-validation folds, together with the mean $\pm$ SD and the worst-performing (minimum) fold. Even in the worst fold, the best text and fusion pipelines remain strong in the E-DAIC $\rightarrow$ EKSpresion direction (worst-fold

$F1 \geq 0.75$ for all text and W2V-fusion configurations), and the strongest EKSpresion $\rightarrow$ E-DAIC fusion configuration (L3.1-70B + W2V) does not drop below $F1=0.64$. This confirms that the reported cross-corpus performance is not driven by a single favourable split.

Feature Configuration	Mod.	Recall $\uparrow$	RMSE $\downarrow$
<i>Text</i>			
Unimodal (L3.1-8B)	T	0.88 ± 0.03	4.85 ± 0.60
Unimodal (L3.1-70B)	T	0.94 ± 0.06	4.79 ± 0.45
Unimodal (L3.3-70B)	T	0.91 ± 0.13	4.46 ± 0.62
<i>Audio</i>			
Unimodal (W2V, ND)	A	0.83 ± 0.11	5.33 ± 0.45
Unimodal (W2V, D)	A	0.76 ± 0.06	5.43 ± 0.44
Unimodal (eGeMAPS, ND)	A	0.62 ± 0.19	5.54 ± 0.52
Unimodal (eGeMAPS, D)	A	0.61 ± 0.10	5.68 ± 0.59
<i>Text + Audio</i>			
Fusion (L3.1-8B + W2V, D)	A+T	0.91 ± 0.05	4.76 ± 0.43
Fusion (L3.1-8B + eGeMAPS, ND)	A+T	0.85 ± 0.09	4.81 ± 0.34
Fusion (L3.1-70B + W2V, D)	A+T	0.91 ± 0.05	4.77 ± 0.57
Fusion (L3.1-70B + eGeMAPS, D)	A+T	0.93 ± 0.08	4.96 ± 0.25
Fusion (L3.3-70B + W2V, ND)	A+T	0.93 ± 0.03	4.34 ± 0.63
Fusion (L3.3-70B + eGeMAPS, D)	A+T	0.90 ± 0.07	4.47 ± 0.52

Table 7: Classification Recall and PHQ-8 in-domain regression RMSE omitted from Table 2 (Mean $\pm$ SD, 5-fold CV). Mod.: Modality. (D)/(ND): Diarized/Non-Diarized. L3.1/L3.3: Llama 3.1/3.3; W2V: Wav2Vec 2.0. Bold: best per modality group.

Feature Configuration	Mod.	Pipeline	MAE $\downarrow$	RMSE $\downarrow$	R^2 $\uparrow$
<i>Text</i>					
Unimodal (L3.1-8B)	T	Random Forest	3.72 ± 0.63	4.79 ± 0.73	0.42 ± 0.12
Unimodal (L3.1-70B)	T	LightGBM	3.70 ± 0.60	4.69 ± 0.62	0.48 ± 0.05
Unimodal (L3.3-70B)	T	Extra Trees	2.98 ± 0.27	3.91 ± 0.32	0.61 ± 0.05
<i>Audio</i>					
Unimodal (W2V, ND)	A	Ridge (PCA)	4.63 ± 0.72	5.64 ± 0.76	0.19 ± 0.15
Unimodal (W2V, D)	A	Extra Trees (MI)	4.87 ± 0.60	5.79 ± 0.69	0.16 ± 0.06
Unimodal (eGeMAPS, ND)	A	SVR (MI)	5.18 ± 0.50	6.23 ± 0.58	0.02 ± 0.08
Unimodal (eGeMAPS, D)	A	XGBoost (cPCA)	5.00 ± 0.60	6.06 ± 0.66	0.10 ± 0.11
<i>Text + Audio</i>					
Fusion (L3.1-8B + W2V, D)	A+T	Random Forest (cPCA)	3.70 ± 0.27	4.84 ± 0.31	0.42 ± 0.07
Fusion (L3.1-8B + eGeMAPS, D)	A+T	LightGBM (ANOVA)	3.69 ± 0.38	4.80 ± 0.42	0.43 ± 0.06
Fusion (L3.1-70B + W2V, ND)	A+T	CatBoost (PCA)	3.86 ± 0.68	4.94 ± 0.71	0.38 ± 0.11
Fusion (L3.1-70B + eGeMAPS, ND)	A+T	Gradient Boosting (PCA)	4.00 ± 0.53	5.07 ± 0.59	0.35 ± 0.10
Fusion (L3.3-70B + W2V, ND)	A+T	Extra Trees (MI)	3.12 ± 0.36	4.03 ± 0.39	0.60 ± 0.02
Fusion (L3.3-70B + eGeMAPS, D)	A+T	Extra Trees (ANOVA)	3.12 ± 0.21	4.05 ± 0.34	0.60 ± 0.06

Table 8: Full in-domain PHQ-9 regression results (MAE, RMSE, R^2 ; Mean $\pm$ SD, 5-fold CV) omitted from Table 2. Mod.: Modality. (D)/(ND): Diarized/Non-Diarized. cPCA: Complementary PCA. MI: Mutual Information; L3.1/L3.3: Llama 3.1/3.3; W2V: Wav2Vec 2.0. Bold: best per modality group.

G Computational Infrastructure

All experiments were conducted on a high-performance computing infrastructure. CPU-intensive tasks, including eGeMAPS extraction and classical machine learning pipelines, utilized nodes with up to 32 cores and 8 GB RAM per core. GPU-intensive operations, specifically Wav2Vec 2.0 feature extraction and LLM inference, were performed on NVIDIA A100-SXM4-40GB GPUs (40 GB VRAM, CUDA 13). Average runtimes for the 5-fold nested CV framework varied by modality. Unimodal text-based experiments typically completed

within 3.5 hours, while unimodal acoustic pipelines required between 11 and 20 hours. Multimodal fusion experiments represented the most computationally expensive configurations, requiring an average of 14 hours per GPU-based job.

H SHAP Model Configuration Details

Table 11 details the optimal model configurations used for the SHAP explainability analysis in Section 3.5. Both directions use an interpretable Extra Trees fusion of the 14 textual features with eGeMAPS acoustic features. Each classifier is fit-

Method	Train	Mod.	E-DAIC	EKS
<i>Setting 2: Both EKS and E-DAIC datasets use PHQ-8 ≥ 10 labels</i>				
eGeMAPS	E-DAIC	A	0.55	0.71 ± 0.07
eGeMAPS	EKS	A	0.49 ± 0.01	0.71 ± 0.06
W2V	E-DAIC	A	0.52	0.72 ± 0.03
W2V	EKS	A	0.48 ± 0.03	0.73 ± 0.08
L3.1-8B	E-DAIC	T	0.78	0.80 ± 0.04
L3.1-8B	EKS	T	0.65 ± 0.15	0.80 ± 0.04
L3.1-70B	E-DAIC	T	0.80	0.79 ± 0.05
L3.1-70B	EKS	T	<u>0.75 ± 0.02</u>	0.79 ± 0.08
L3.3-70B	E-DAIC	T	0.79	<u>0.84 ± 0.06</u>
L3.3-70B	EKS	T	0.68 ± 0.03	0.85 ± 0.05
L3.3-70B+eGeMAPS	E-DAIC	A+T	0.68	0.81 ± 0.05
L3.3-70B+eGeMAPS	EKS	A+T	0.60 ± 0.12	0.83 ± 0.05
L3.1-8B+W2V	E-DAIC	A+T	0.73	0.83 ± 0.04
L3.1-8B+W2V	EKS	A+T	0.55 ± 0.11	0.82 ± 0.04
L3.1-70B+W2V	E-DAIC	A+T	0.70	0.80 ± 0.06
L3.1-70B+W2V	EKS	A+T	0.68 ± 0.09	0.81 ± 0.08
L3.3-70B+W2V	E-DAIC	A+T	0.71	0.83 ± 0.03
L3.3-70B+W2V	EKS	A+T	0.52 ± 0.06	0.85 ± 0.06
Avg. (all)	E-DAIC	A, T, A+T	0.70	0.79
Avg. (all)	EKS	A, T, A+T	0.60	0.80

Table 9: Full within- and cross-corpus results for Setting 2, with PHQ-8 ≥ 10 labels in both EKS and E-DAIC. Values are depressed-class F1. Model results are mean $\pm$ SD over five folds, except E-DAIC within-corpus results, which use the official split. Mod.: A/T/A+T denote acoustic/text/both; ‘+’ denotes early fusion. Bold/underline: best within-/cross-corpus result, respectively, with lower SD breaking ties. EKS: EKSpresion; eGeMAPS: acoustic feature set; W2V: Wav2Vec 2.0; L3.1/L3.3: Llama 3.1/3.3.

Method	Mod.	F0	F1	F2	F3	F4	Mean $\pm$ SD	Worst fold
<i>E-DAIC $\rightarrow$ EKSpresion</i>								
Unimodal (eGeMAPS)	A	0.65	0.71	0.68	0.67	0.65	0.67 ± 0.03	0.65
Unimodal (W2V)	A	0.66	0.68	0.70	0.68	0.69	0.68 ± 0.02	0.66
Unimodal (L3.1-8B)	T	0.81	0.76	0.78	0.79	0.76	0.78 ± 0.02	0.76
Unimodal (L3.1-70B)	T	0.75	0.83	0.81	0.75	0.76	0.78 ± 0.04	0.75
Unimodal (L3.3-70B)	T	0.79	0.83	0.85	0.79	0.82	0.82 ± 0.02	0.79
Fusion (L3.3-70B + eGeMAPS)	A+T	0.77	0.79	0.85	0.79	0.78	0.80 ± 0.03	0.77
Fusion (L3.1-8B + W2V)	A+T	0.85	0.79	0.87	0.78	0.78	0.81 ± 0.04	0.78
Fusion (L3.1-70B + W2V)	A+T	0.79	0.77	0.81	0.80	0.75	0.78 ± 0.03	0.75
Fusion (L3.3-70B + W2V)	A+T	0.80	0.85	0.85	0.83	0.81	0.83 ± 0.02	0.80
<i>EKSpresion $\rightarrow$ E-DAIC</i>								
Unimodal (eGeMAPS)	A	0.47	0.47	0.46	0.48	0.52	0.48 ± 0.02	0.46
Unimodal (W2V)	A	0.49	0.47	0.52	0.49	0.47	0.48 ± 0.02	0.47
Unimodal (L3.1-8B)	T	0.67	0.71	0.58	0.69	0.65	0.66 ± 0.05	0.58
Unimodal (L3.1-70B)	T	0.85	0.69	0.69	0.72	0.73	0.74 ± 0.06	0.69
Unimodal (L3.3-70B)	T	0.70	0.51	0.48	0.74	0.67	0.62 ± 0.12	0.48
Fusion (L3.3-70B + eGeMAPS)	A+T	0.52	0.42	0.50	0.52	0.50	0.49 ± 0.04	0.42
Fusion (L3.1-8B + W2V)	A+T	0.62	0.47	0.47	0.75	0.47	0.55 ± 0.13	0.47
Fusion (L3.1-70B + W2V)	A+T	0.64	0.69	0.79	0.79	0.69	0.72 ± 0.06	0.64
Fusion (L3.3-70B + W2V)	A+T	0.70	0.47	0.63	0.68	0.60	0.62 ± 0.09	0.47

Table 10: Per-fold cross-corpus classification F1 (depressed class) across all five outer CV folds, with mean $\pm$ SD and the worst (minimum) fold, under Setting 1 (EKSpresion uses SCID labels; E-DAIC uses PHQ-8 ≥ 10). F0–F4 denote the five folds. Mod.: Modality (A: Acoustic, T: Text); L3.1/L3.3: Llama 3.1/3.3; W2V: Wav2Vec 2.0. Complements Table 4.

ted on its source training data: E-DAIC’s official split for E-DAIC $\rightarrow$ EKSpresion, and the median-performing cross-validation fold of the EKSpresion dataset for EKSpresion $\rightarrow$ E-DAIC; this median is defined with respect to the fixed Extra Trees

explainability model. E-DAIC $\rightarrow$ EKSpresion selects the ANOVA top-30 eGeMAPS features pruned of those correlated with the text stream ($|r| > 0.70$); EKSpresion $\rightarrow$ E-DAIC retains all 200 eGeMAPS features.

Property	Value
<i>E-DAIC → EKSprespression (Plot a in Figure 2)</i>	
Model	Extra Trees
<i>n_estimators</i>	200
<i>max_depth</i>	None
Total features	44 (14 text + 30 eGeMAPS)
Feature selection	ANOVA ($k = 30$) + cPCA ($ r > 0.70$)
LLM	Llama 3.3-70B
Training data	E-DAIC official split
<i>EKSprespression → E-DAIC (Plot b in Figure 2)</i>	
Model	Extra Trees
<i>n_estimators</i>	200
<i>max_depth</i>	None
Total features	214 (14 text + 200 eGeMAPS)
Feature selection	None
LLM	Llama 3.3-70B
Training data	EKSprespression median fold (fold 3)

Table 11: Model configuration for the SHAP explainability analysis. Both directions use an interpretable Extra Trees fusion of the Llama 3.3-70B textual pipeline (14 features) with eGeMAPS acoustic features, each fitted on its source training data (E-DAIC’s official split, and the median-performing EKSprespression cross-validation fold, respectively). E-DAIC→EKSprespression applies ANOVA ($k=30$) feature selection with cPCA pruning of text-correlated features; EKSprespression→E-DAIC retains all 200 eGeMAPS features. cPCA: Complementary PCA.