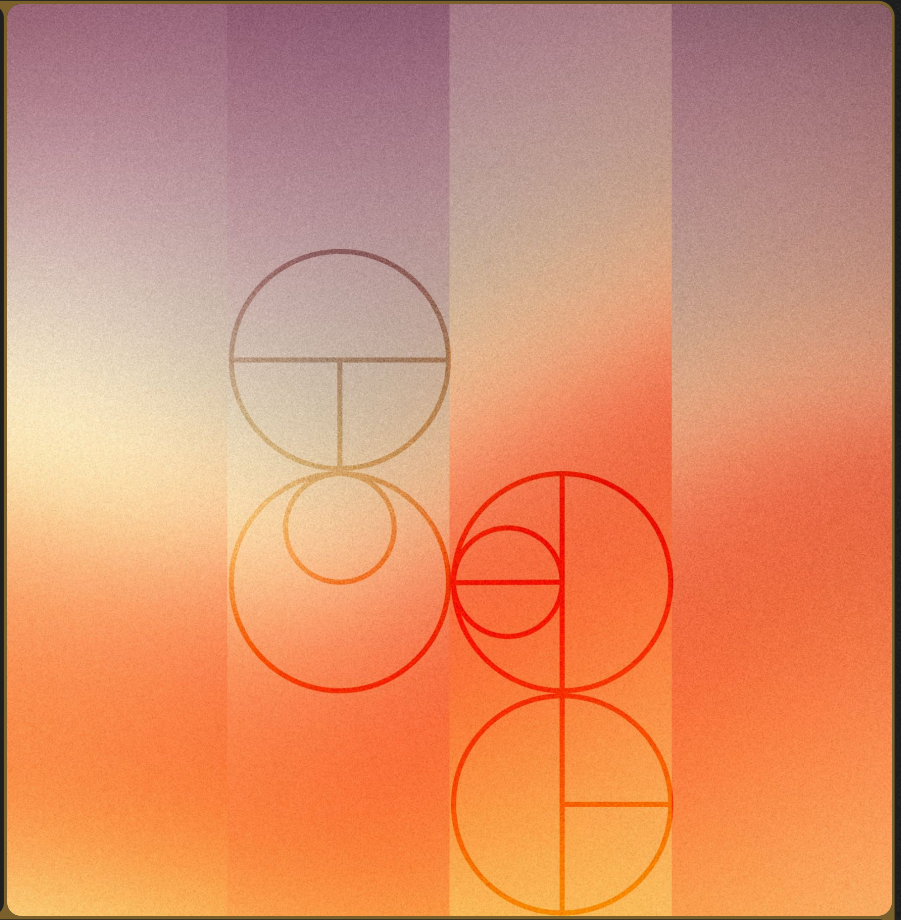
Voice-Based Psychiatric Assessment in the Real World

Agnes Norbury, George Fairs, Stefano Goria



scientific reports

Explore content | About the journal | Publish with us

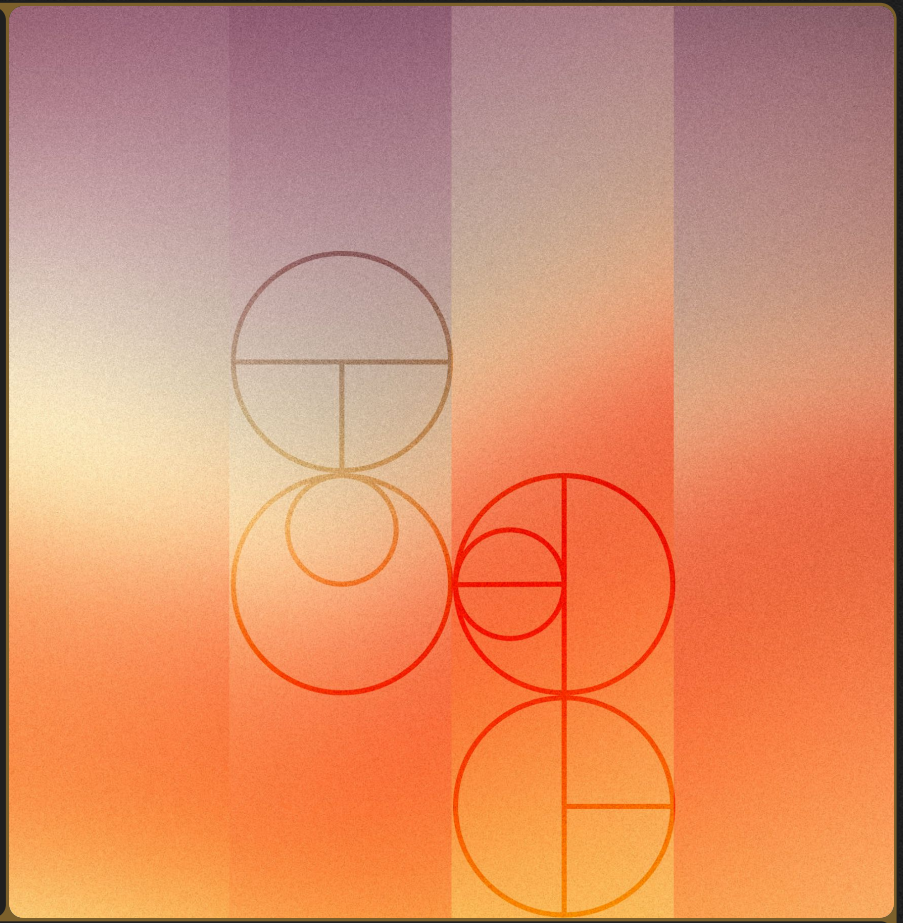
nature > scientific reports > articles > article

Article | Open access | Published: 06 February 2026

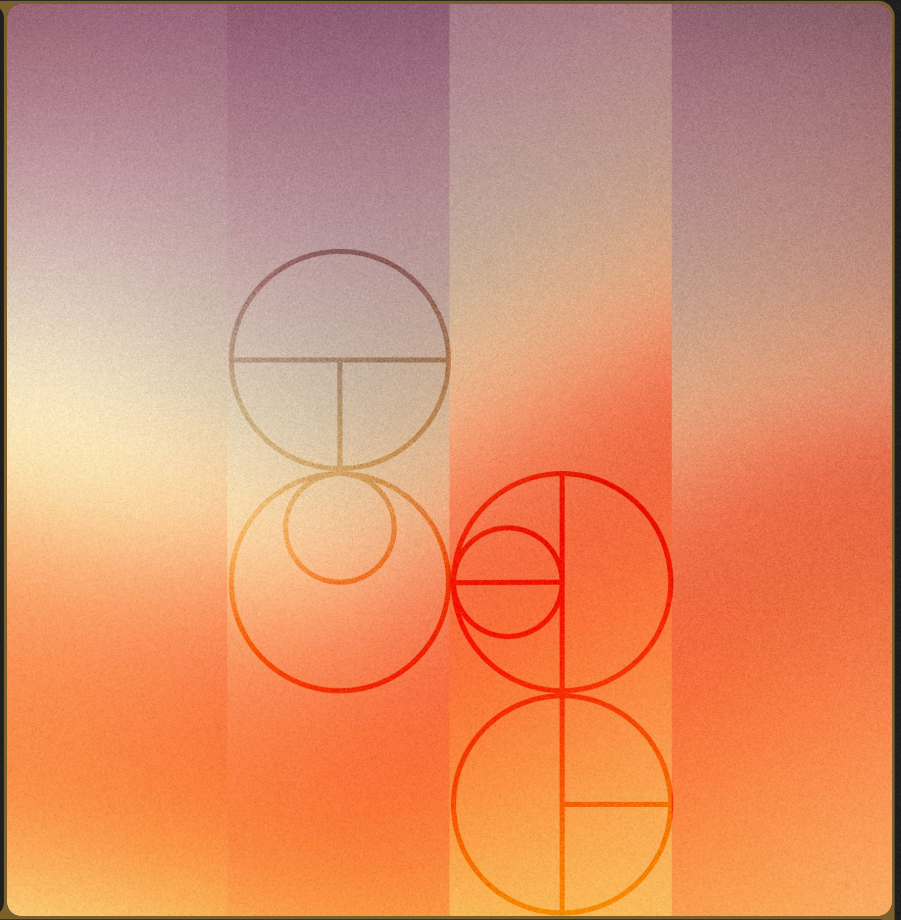
A multimodal Bayesian network for symptom-level depression and anxiety prediction from voice and speech data

Agnes Norbury, George Fairs, Alexandra L. Georgescu, Matthew M. Nour, Emilia Molimpakis & Stefano Goria

Scientific Reports **16**, Article number: 5397 (2026) | Cite this article



Background:
The Clinical Problem



The challenge of psychiatric diagnosis

1

Complex inferential process

Psychiatrists integrate multiple information sources to arrive at diagnostic formulations that guide treatment decisions.

2

Reliance on subjective information

Unlike other medical specialties, psychiatric assessment depends primarily on clinical observation, patient self-report, and collateral information.

3

Absence of biological markers

No established “objective” diagnostic tests means diagnosis relies heavily on synthesizing complex, often ambiguous self-report and behavioural data.

A multimodal integration problem

Paralinguistic cues

Tone of voice, speech rate, fluency, and conversational responsiveness

Clinical constraints

Time pressures, capacity limitations, and cognitive demands of synthesizing uncertain information

Behaviour and cognition

Body language, psychomotor activity, and cognitive processing during clinical interview

Bias and inequity

Various biases can affect information gathering, weighting, and diagnostic conclusions

Potential barriers to adoption of digital phenotyping tools

Binary condition classification rather than symptom- or sign-level assessment

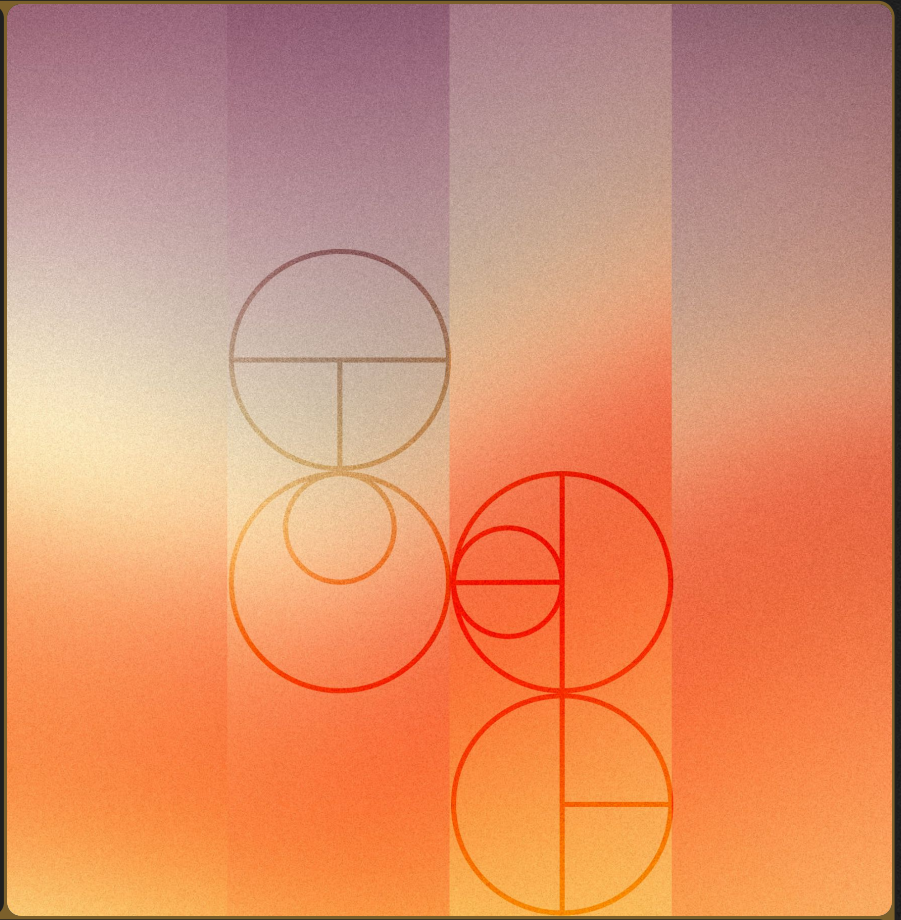
Small, homogeneous samples that compromise generalizability and fairness evaluation

Poor transparency or explainability of underlying models

Lack of integration or congruence with existing clinical workflows

Our Approach

Symptom and modality-specific surrogates + Bayesian Network modelling



Why did we choose this approach?

Preserves symptom-specific signal

Different symptoms are differentially reflected in different input modalities

Clinically useful

Granular outputs aid treatment planning, monitoring, and direct validation against clinical impressions

Principled combination

Models how clinicians may integrate multiple noisy information sources during assessment

Maintains expert primacy

Supports direct clinical intervention

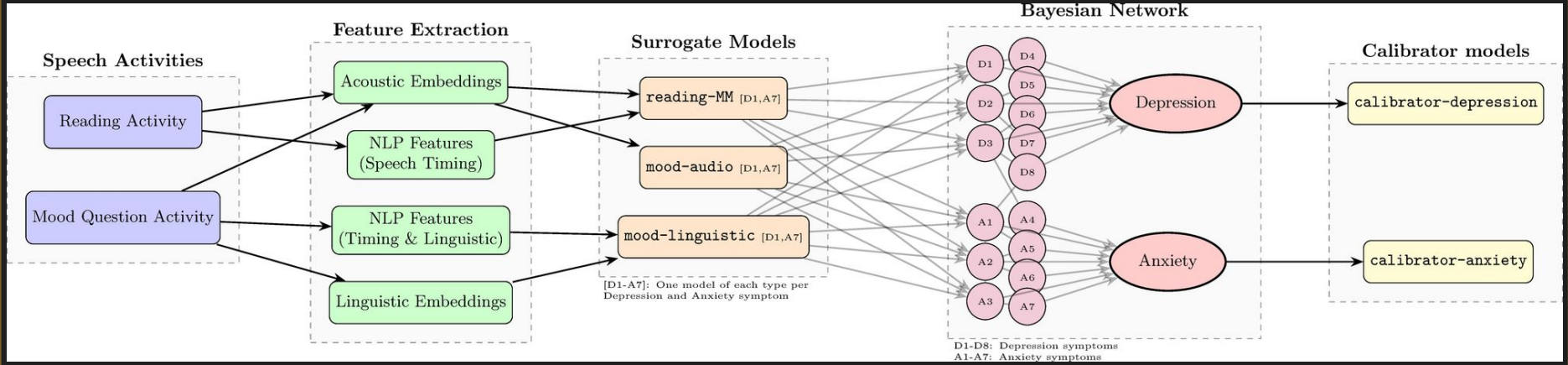
Handles comorbidity

Naturally accounts for symptom co-occurrence patterns common in mental health

Transparent & explainable

Probabilistic reasoning with inspectable inter-symptom dependencies

Model overview

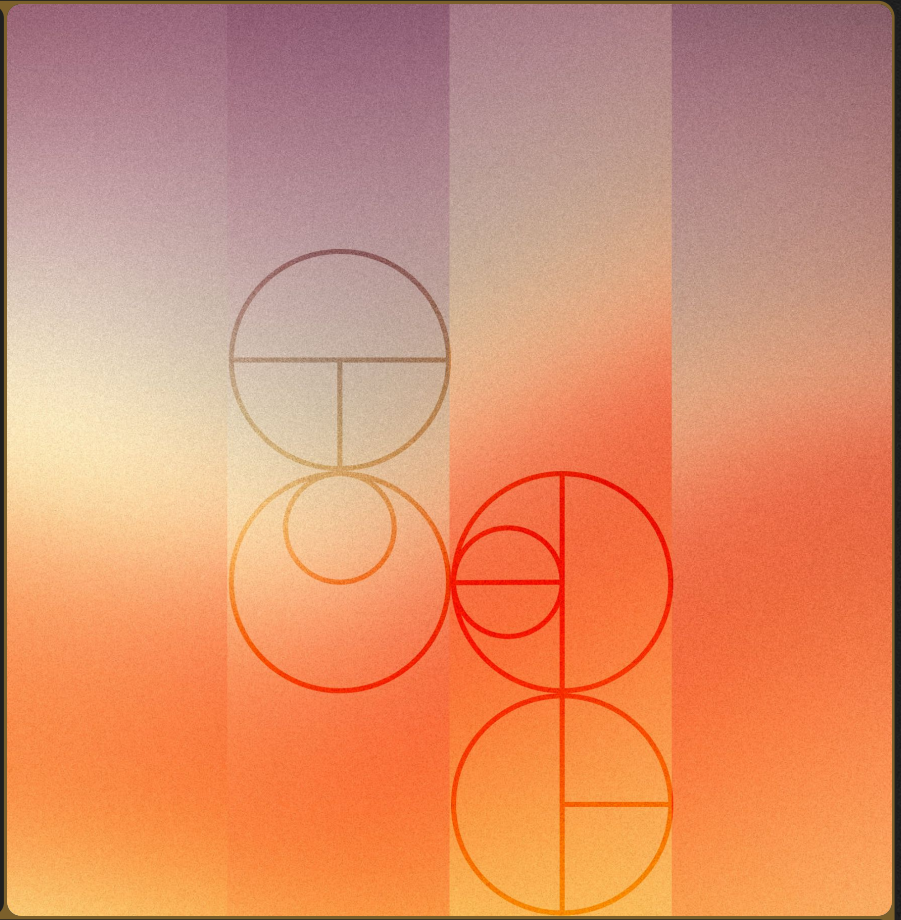


Two speech activities: reading aloud a standardised passage + answering a question about recent mood (~1 min each)

Three surrogate types: reading-MM (acoustic + speech timing), mood-audio (acoustic), mood-linguistic (linguistic + NLP) = 45 models

15 symptom nodes: 8 depression (PHQ-8) + 7 anxiety (GAD-7) symptoms, each with learned inter-symptom relationships

Data & Implementation



Dataset

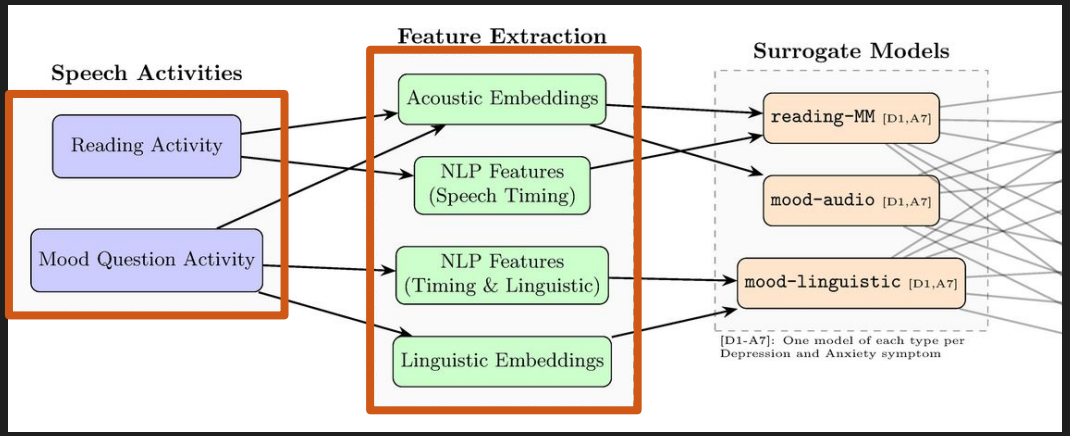
~30,000

unique speakers

	Development	Calibration	Test
N	21,379	6,325	2,431
Age (mean)	37.6 (SD 12.9)	37.4 (SD 13.2)	37.1 (SD 13.0)
Age range	18–89	18–88	18–80
% Female	61%	66%	67%

- No user overlap between splits
- Test set completely unseen during model development
- Separate calibration set for training output calibrator models
- Recruitment quotas for mental health history + race/ethnicity diversity - other confounds (e.g. physical health) also controlled for

Observables: activities & feature extraction



Representation embeddings -> rich representations designed to preserve signal

Acoustic & linguistic embeddings -> large pre-trained encoders

NLP features -> “handcrafted”

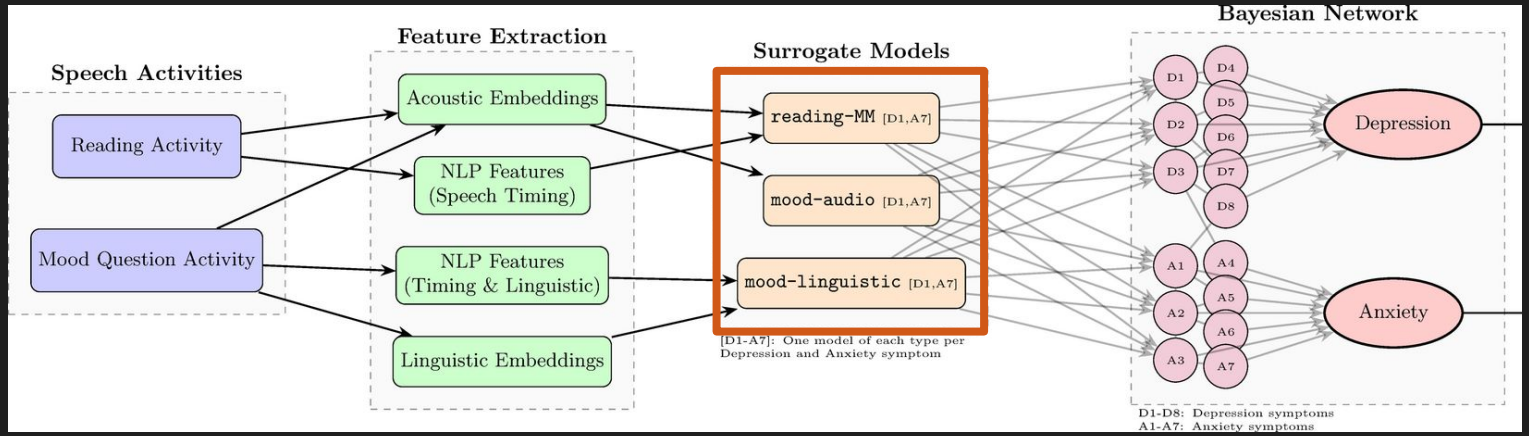
Reading Activity

Standardised passage (Aesop's "The North Wind and the Sun"). Controlled, paralinguistic focus - captures "how" it's said.

Mood Question Activity

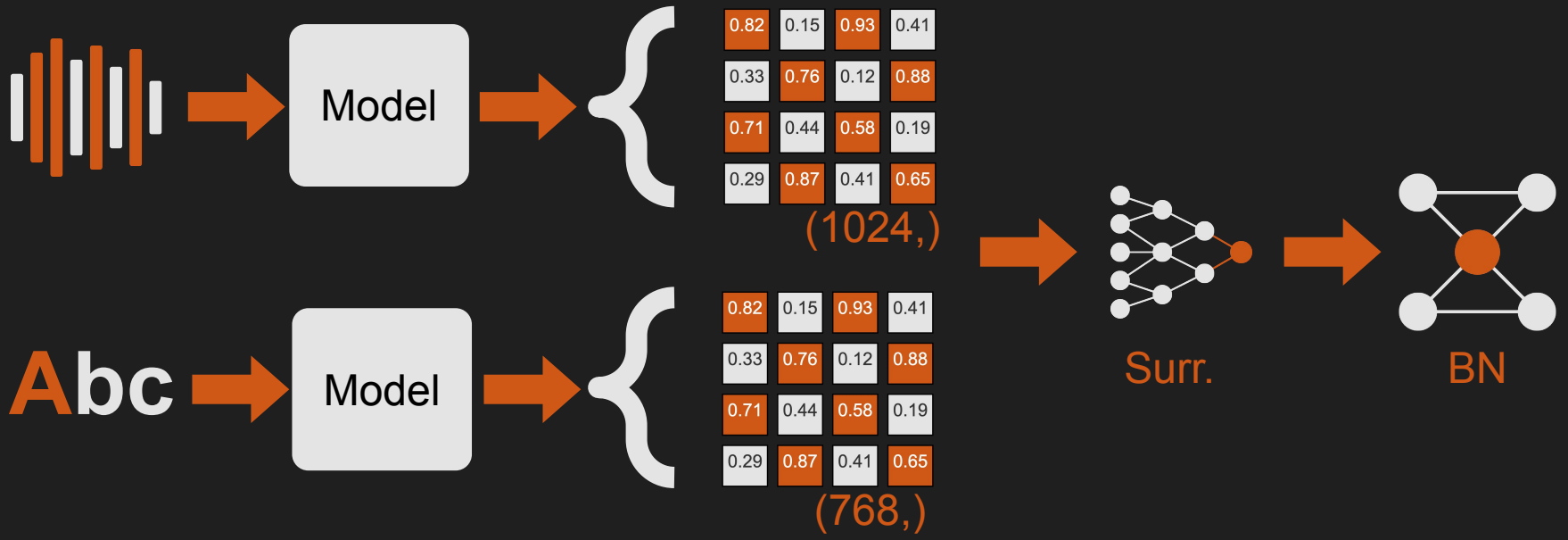
Free-form description of recent mood. Captures both semantic ("what") and paralinguistic ("how") content.

The surrogate architectures

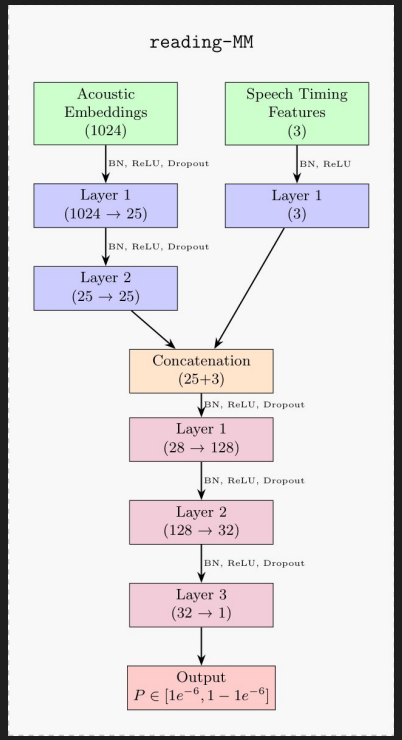


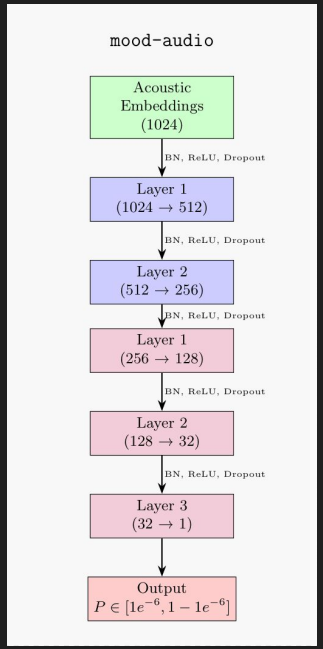
Why use surrogates? A problem of dimensionality...

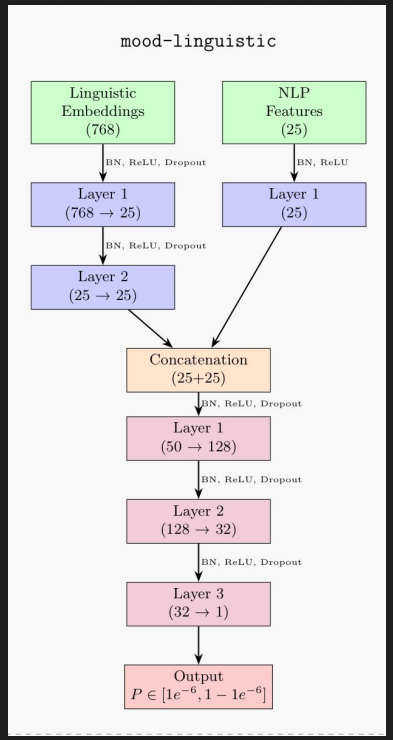
Dense embeddings produced from large pre-trained models are largely intractable for direct Bayesian Network inference. How can we compress and effectively combine these representations?



The surrogate architectures



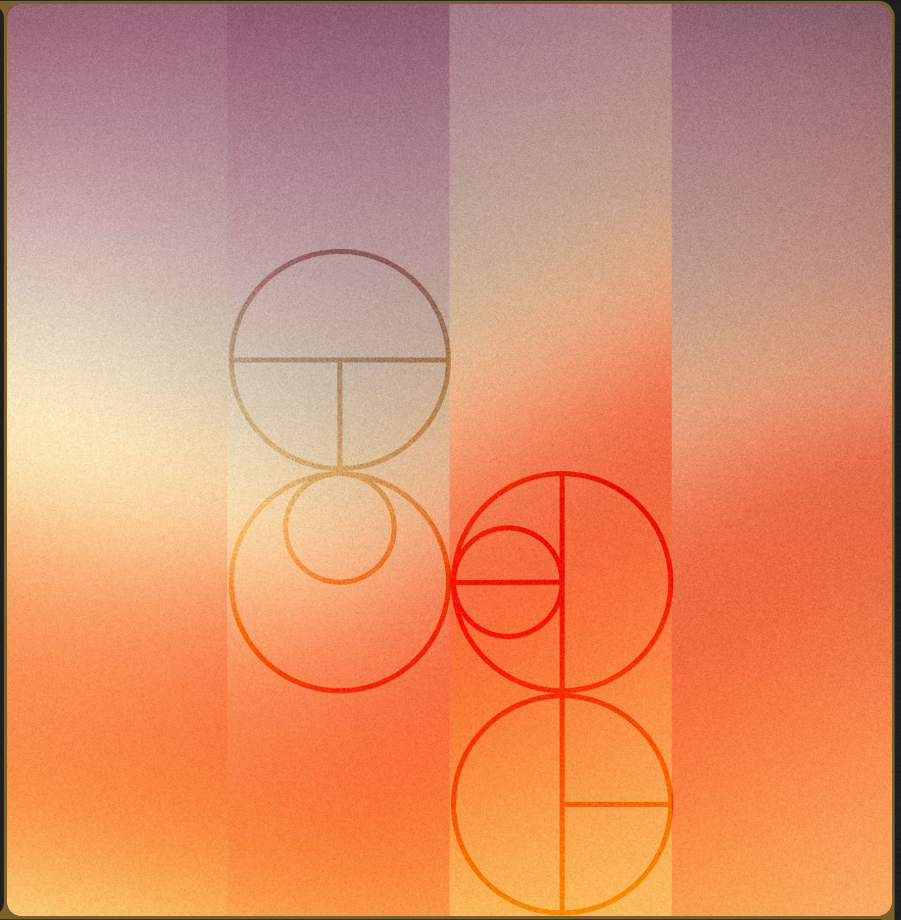




- Binary classification targets from PHQ-8/GAD-7 (symptom present ≥ half of days vs not)
- Binary cross-entropy loss + AdamW optimiser + LR scheduling + early stopping + gradient clipping
- Model selection via stratified nested cross-validation with Optuna Bayesian optimisation

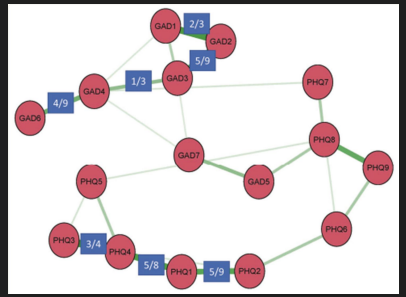
Bayesian Network

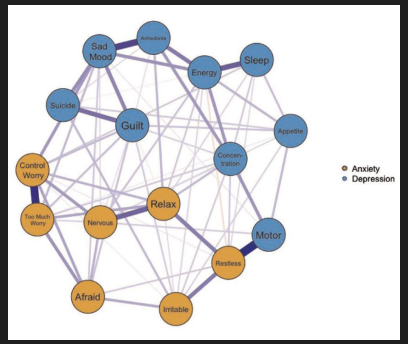
Structure, parameterisation, and inference

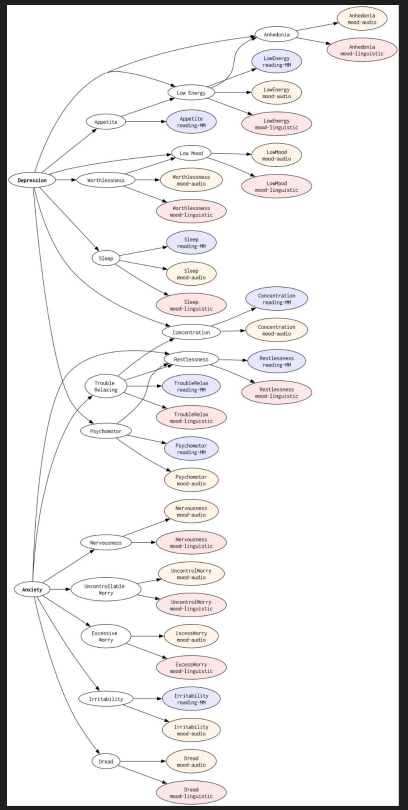


Network structure

- Hybrid structure combining published literature on symptom relationships with data-driven insights
- Inter-symptom edge directions informed by temporal analyses identifying which symptom changes tend to precede others over time
- Surrogate-to-symptom connections pruned from a fully connected network based on informativeness and specificity
- **Design constraint:** every symptom required to be informed by at least one paralinguistic model in order to prevent over-reliance on linguistic features
- Final structure selected by optimising ROC-AUC at symptom and condition levels while preserving realistic cross-symptom prediction correlations







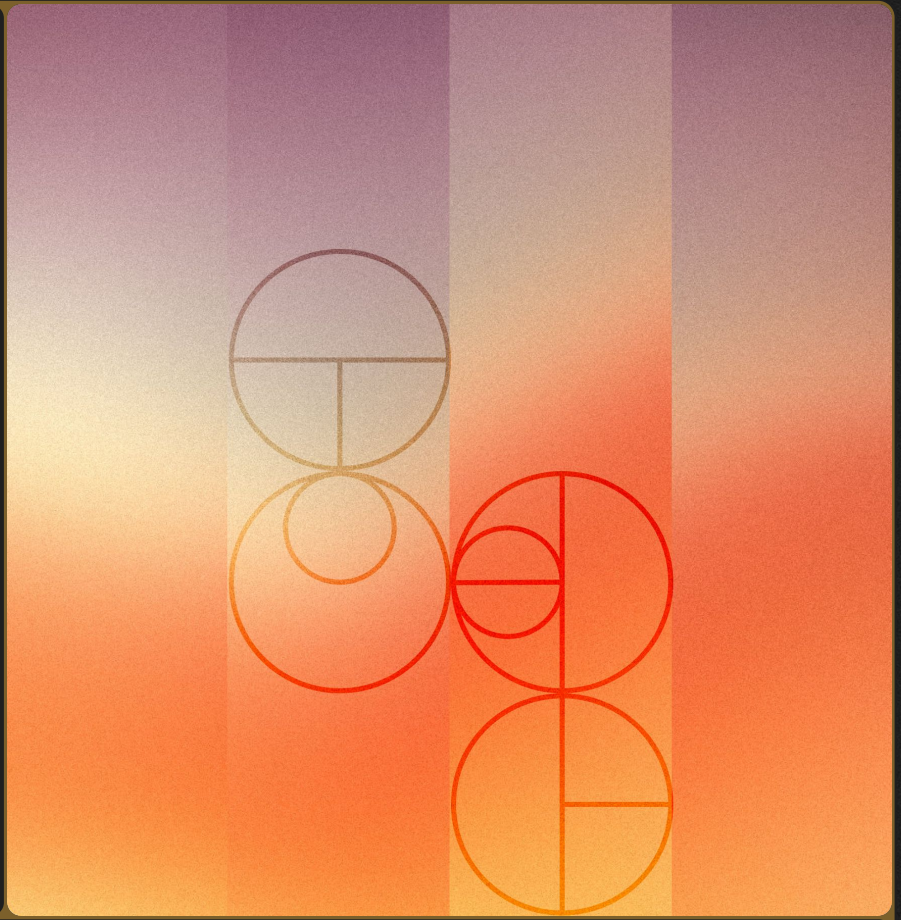
Parameterisation and inference

Parameter estimation: conditional probability distributions estimated using Bayesian estimation with Dirichlet equivalent uniform priors - this regularises distributions and prevents overfitting by assigning weight to rare or unseen state combinations

Regularisation tuning: prior strength tuned during development to balance overall performance with condition-specific sensitivity, i.e. ensuring depression and anxiety predictions are not overly correlated

Inference & calibration: exact inference via Variable Elimination (no approximation needed for this network structure); output calibrators (trained on separate calibration set) ensure condition probabilities align with observed case rates

Results



Condition-level performance & calibration

	ROC-AUC	ECE
Depression	**0.842**	0.018
Anxiety	**0.831**	0.015

ROC-AUC > 0.80: generally considered clinically useful discrimination

ECE < 0.05: excellent calibration, where predicted probabilities closely match observed case rates

Condition severity correlates with independent outcomes:

Quality of Life (r ≈ -0.50)
Psychosocial functioning (r ≈ 0.47)
Healthy days (r ≈ 0.44)

Symptom-level performance

Depression symptoms (ROC-AUC)

	Anhedonia	Low Mood	Sleep	Low Energy	Appetite	Worthless	Concentr.	Psychom.
Test	0.741	**0.801**	0.706	0.734	0.690	0.757	0.667	0.714

Anxiety symptoms (ROC-AUC)

	Nervous	Uncontrol. Worry	Excessive Worry	Trouble Relaxing	Restless	Irritable	Dread
Test	**0.749**	0.746	0.736	0.739	0.680	0.706	0.721

- Core symptoms (anhedonia, low mood, nervousness, uncontrollable worry) approach or exceed 0.75 ROC-AUC
- All individual symptoms > 0.66 ROC-AUC -> generally in fair-to-good range
- Some symptoms (concentration, restlessness) may benefit from additional data sources e.g. cognitive tasks

Multimodal integration properties

Successful integration

- Network outperforms every individual surrogate for every symptom
- Confirms effective weighting and integration across noisy inputs

Built-in redundancy

- Good signal exists across both paralinguistic and linguistic inputs
- Combining paralinguistic surrogates matches or improves on linguistic-only performance

Example: sleep symptom integration (see Fig. 3 in paper)

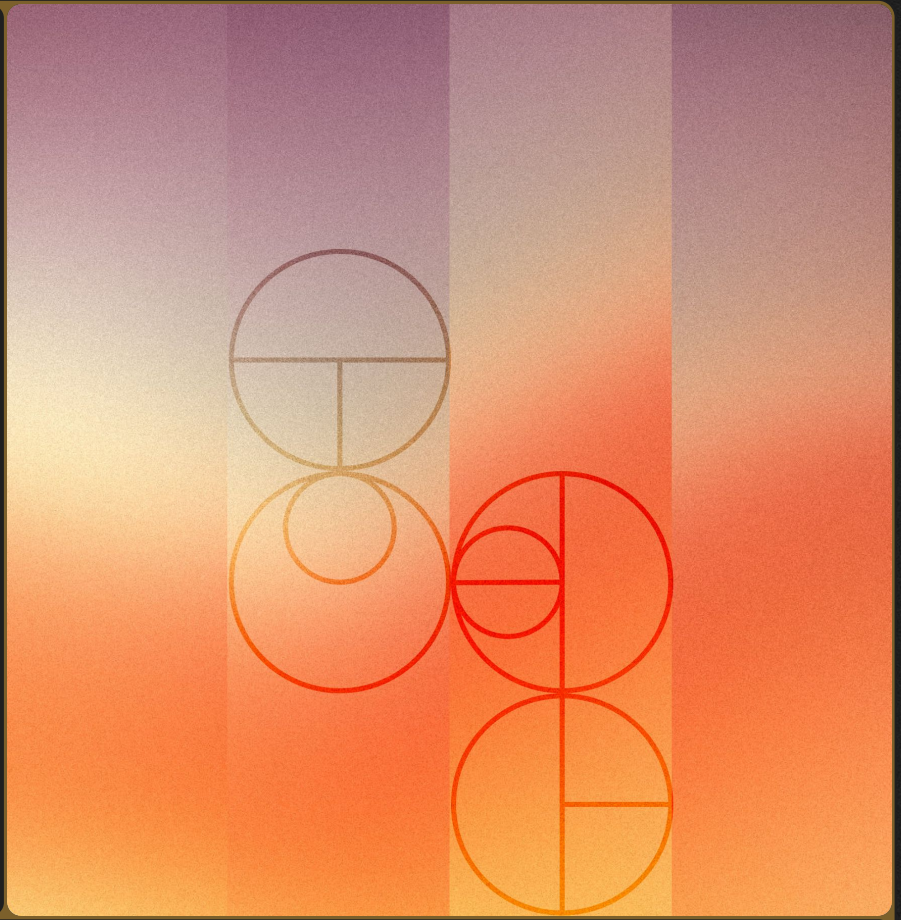
Individual surrogate AUCs: reading-MM = 0.62, mood-audio = 0.66, mood-linguistic = 0.68. After BN integration: 0.71 ROC-AUC. The network learns different CPD weights for each surrogate, assigning higher posterior weights to more discriminative models for this symptom.

Fairness across demographic groups

Group	Dep. ROC-AUC	Anx. ROC-AUC	ECE Diff	Eq. Odds Diff (Dep)	Eq. Odds Diff (Anx)
Age (<35 / ≥35)	0.826 / 0.848	0.814 / 0.836	0.006	0.099	0.047
Birth Sex (F / M)	0.838 / 0.828	0.825 / 0.800	0.036	0.111	0.251
Race (White / Other)	0.850 / 0.840	0.843 / 0.814	0.055	0.024	0.060
Device (Laptop / Mobile)	0.831 / 0.856	0.839 / 0.819	0.006	0.054	0.106

- ROC-AUC ≥ 0.80 across all examined subgroups for both conditions
- Most group differences in equalized odds ratios in excellent (<0.05) or good (<0.10) ranges
- **One area warranting attention:** sex-based differences for anxiety (equalized odds ratio 0.251), potentially driven by higher base rates in women - so may benefit from sex-specific calibration or more broadly multi-calibration techniques
- Performance robust across device types and chronic health conditions

Clinical Translation



Clinical usefulness: screening tool metrics

	PPV	NPV	LR+	LR–
Depression	0.69	0.83	4.7	0.46
Anxiety	0.71	0.80	5.4	0.55

LR+ ≈ 5: a user is ~5x more likely to have the condition after a positive result

NPV ≈ 0.80: model is reliable at ruling out conditions when negative

PPV ≈ 0.70: some false positives - acceptable for screening where the cost of missing cases outweighs false alarms

Good properties for a screening / triage / monitoring tool?

- E.g. positive results flag for clinical follow-up and confirmation
- Well-calibrated probabilities allow stakeholders to choose thresholds for their specific context (e.g. via decision curve analysis)

Clinician-in-the-loop: explainability & intervention

The Bayesian Network architecture enables two key clinical properties:

- **Explainability -** explicit reporting of individual symptom severity estimates and their contributions to overall condition probabilities. Clinicians can inspect which symptoms are driving predictions.

- **Modifiability (do-operations) -** clinicians can intervene directly in model predictions, e.g. by isolating symptom nodes based on clinical discussion with the client, then updating predictions in real time.

Key takeaways & limitations

Takeaways

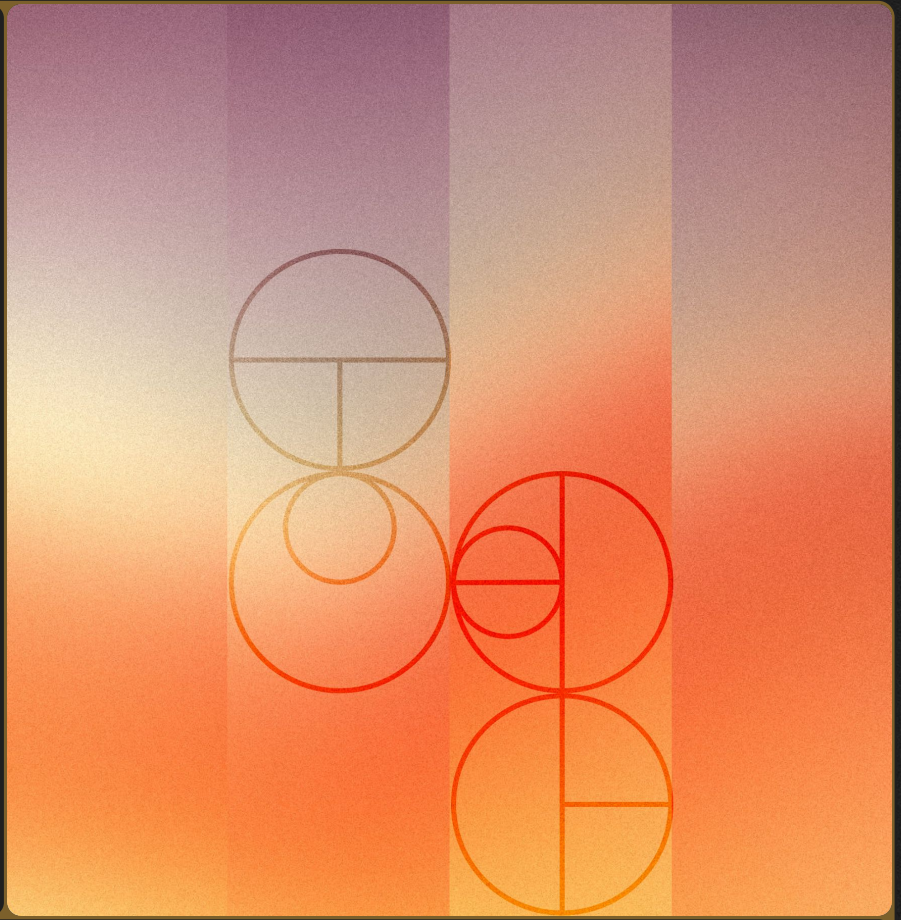
- Bayesian Networks offer a principled, explainable, intervenable approach for multimodal information integration
- Scale (30k+ speakers) enables robust performance, calibration, and fairness evaluation
- Symptom-level outputs align with clinical practice - symptoms, not diagnoses, guide treatment
- Multimodal integration exceeds any individual source; built-in redundancy protects against single-modality failure

Limitations

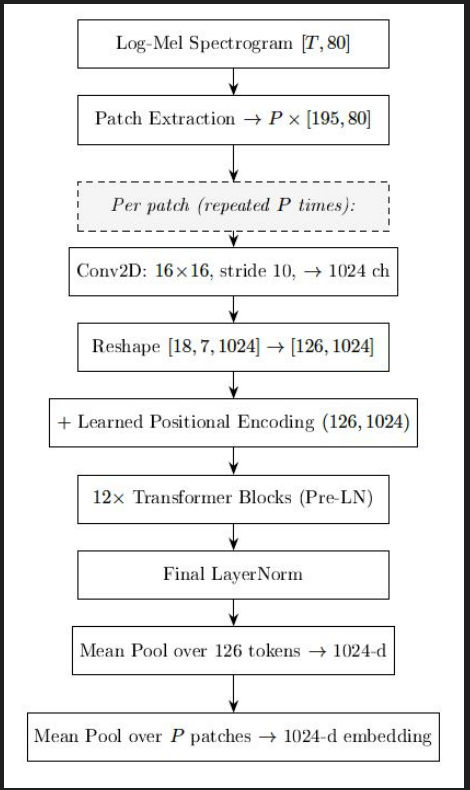
- Self-report labels (PHQ-8, GAD-7); English-speaking US/UK sample; certain symptoms may benefit from non-speech sources; *some* fairness metrics warrant further investigation

Real-World Deployment

Real-world constraints don't just limit you — they sometimes drive fundamental research.



From lab to production: new constraints



- Production demands real-time inference — latency directly limits throughput and user experience
- Privacy-sensitive health data favours on-device (edge) deployment — but edge devices have limited compute and memory
- First tool: quantisation — compress TRILLsson5 from 338 MiB (float32) → 85 MiB (int8), making edge deployment viable

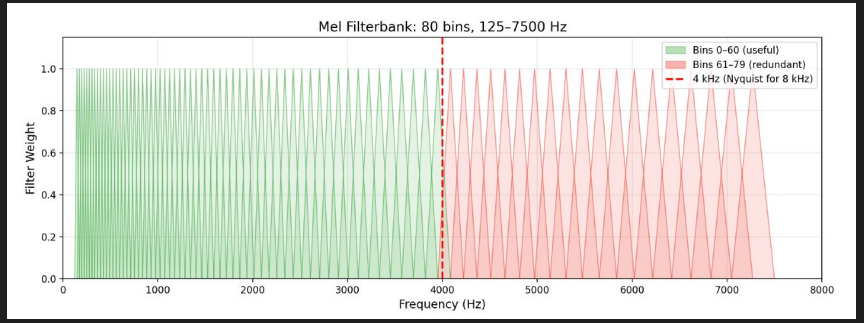
These are all standard challenges; it requires work, but not new research

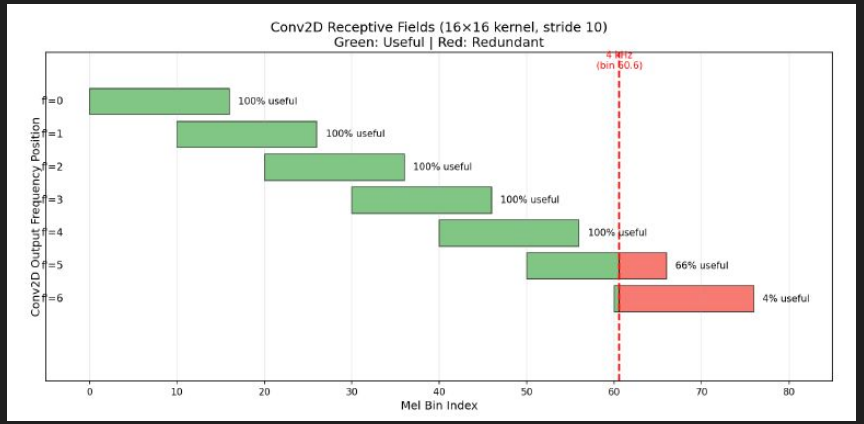
Business reality: phones use 8 kHz audio

- Key client requirement: system must work over phone calls
- Standard telephony (PSTN) transmits at 8 kHz sample rate — unchanged since the 1960s
- Our entire pipeline — TRILLsson5, mel spectrograms, all training — assumes 16 kHz
- Naïve fix: upsample 8 kHz → 16 kHz. **It works — equivalent downstream performance**
- This raises a question: what is the model actually doing with those frequencies?

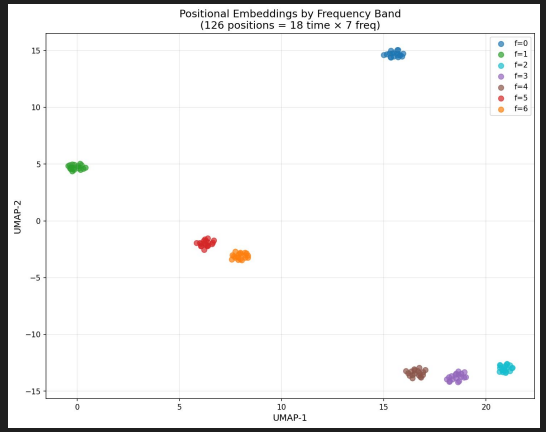
Tracing the 4 kHz boundary through the model

- 8 kHz audio has a hard Nyquist limit at 4 kHz — no real content exists above this
- Tracing through TRILLsson5's mel frontend: 19 of 80 mel bins (bins 61–79) carry zero useful energy
- These empty bins flow into Conv2D patches, **creating 36 of 126 transformer tokens that process nothing**
- The model was carrying 29% redundant tokens without us knowing





The model had learned to segregate frequency bands



Block	L→L	L→H	H→L	H→H	Cross/Within	Interpretation
0	0.0089	0.0056	0.0055	0.0140	0.49	Strong segregation
1	0.0085	0.0060	0.0058	0.0120	0.57	Strong segregation
2	0.0086	0.0062	0.0062	0.0105	0.65	Moderate segregation
⋮	⋮	⋮	⋮	⋮	⋮	
5	0.0082	0.0074	0.0080	0.0079	0.95	Near-uniform
⋮	⋮	⋮	⋮	⋮	⋮	
11	0.0084	0.0067	0.0088	0.0058	1.09	Slight integration

- Positional embeddings for high-freq tokens (f=5,6) form a distinct cluster — cross-group cosine similarity: 0.03 (nearly orthogonal to low-freq tokens)
- Attention patterns in early layers show 40–50% less cross-attention between low-freq and high-freq tokens
- The model had already partially learned to ignore these tokens
- Removing them isn't a hack — **it aligns with the model's own learned representations**

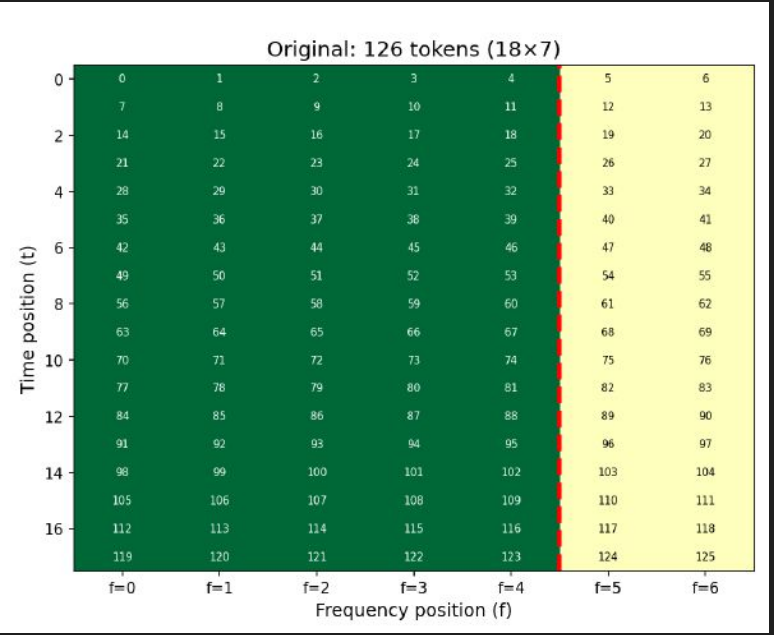
Closed-form model reduction — no retraining required

Token reduction: truncate mel filterbank 80 → 64 bins, tokens 126 → 90 (29% fewer)

- Embedding similarity with original: 0.989 (near-identical)
- Attention computation: 49% reduction (quadratic in token count)
- Measured speedup: 1.32×

Combined with layer pruning (12 → 9 layers):

- Speedup: 1.64×
- 25% fewer parameters
- All via closed-form weight surgery — zero retraining



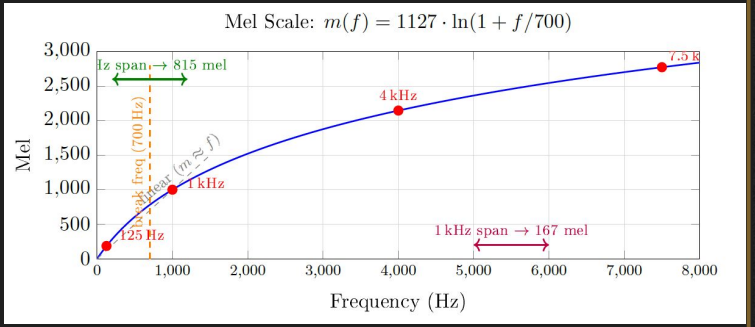
Real-world constraints drive fundamental research

Business constraint: phones use 8 kHz

Better model: faster, smaller, principled — no retraining

- Opens new questions: should paralinguistic models use perceptual (mel) frontends at all, or production-aligned frontends that preserve what doctors can't hear?

Production constraints aren't just obstacles — they're a lens that reveals what your model is actually doing. Sometimes that understanding drives the next research question.



Thank you! Questions?