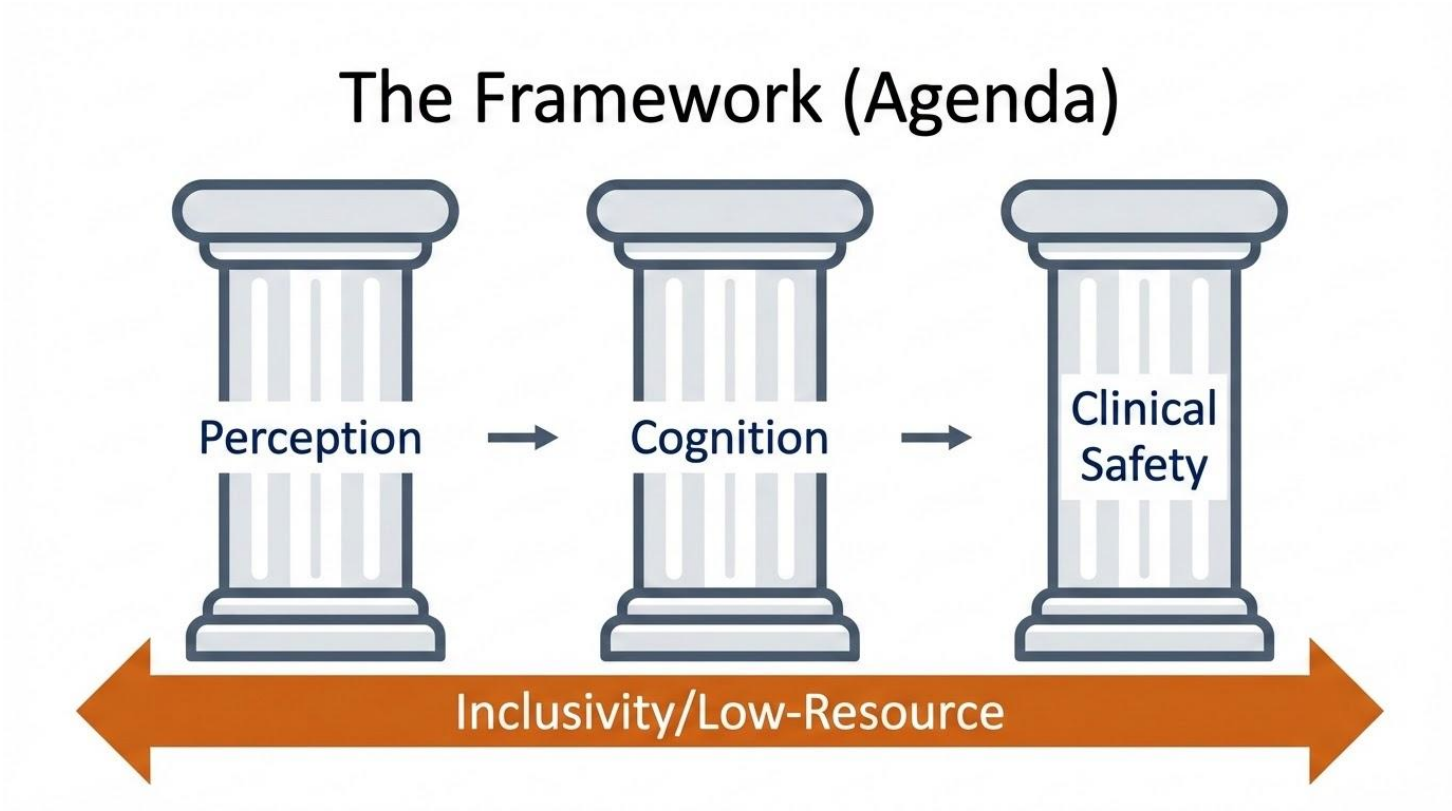
Psychologically Grounded AI:
Decoding Multimodal Affect, Structuring Cognition, and Building Safe Mental Health Tools

Sara (Ziwei) Gong
April 2026

Outline:


The Framework (Agenda)
Perception
Cognition
Clinical Safety
Inclusivity/Low-Resource

Part 1: Decoding Multimodal Emotions (The Signal)

1. The Baseline Challenge
   - Multimodal Fusion
2. The Acoustic Challenge
   - Amplifying Vocal Nuances
3. The Contextual Challenge
   - Synergistic Uncertainty (SURE)
4. The Demographic Challenge
   - Low-Resource & Cross-Cultural Perception

Multimodal Multi-loss Fusion Network for Sentiment Analysis

Zehui Wu∗, Ziwei Gong∗, Jaywon Koo, Julia Hirschberg

Department of Computer Science
Columbia University

Paper link: https://aclanthology.org/2024.naacl-long.197/

1. The Baseline Challenge – Multimodal Fusion

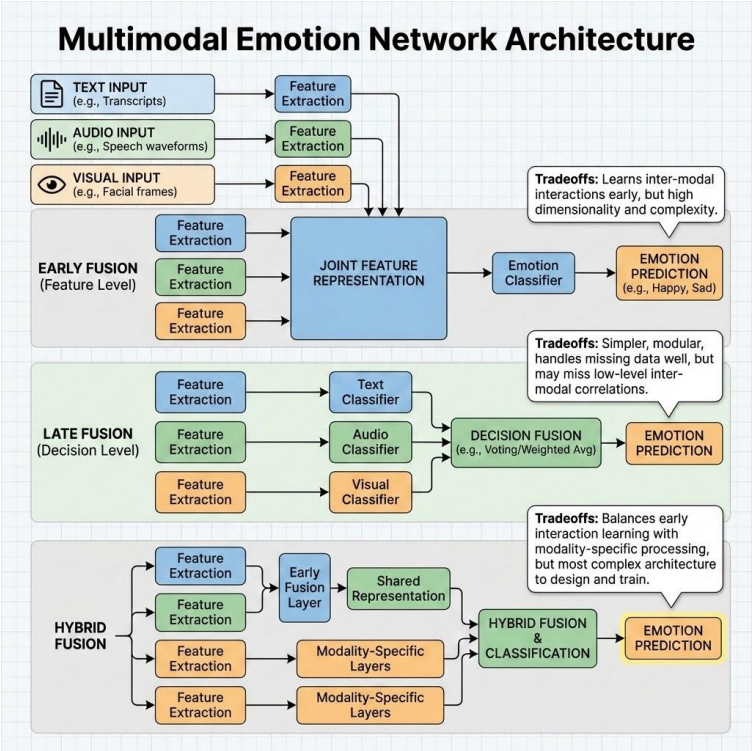
Goal: An optimized feature selection and fusion approach for enhancing sentiment detection in neural networks

Task: Multimodal sentiment analysis

- Sentiment analysis is one of the classification tasks that can really take advantage of signals other than text. The same sentence can be expressed very differently in emotion by using different pitch and intensity.

Achievements:

- Investigates the **optimal selection and fusion of feature encoders** across multiple modalities
- Examine the impact of **multi-loss training** within the multi-modality fusion network, identifying surprisingly important findings relating to subnet performance.
- Our best model achieves **state-of-the-art** performance for three sentiment datasets.



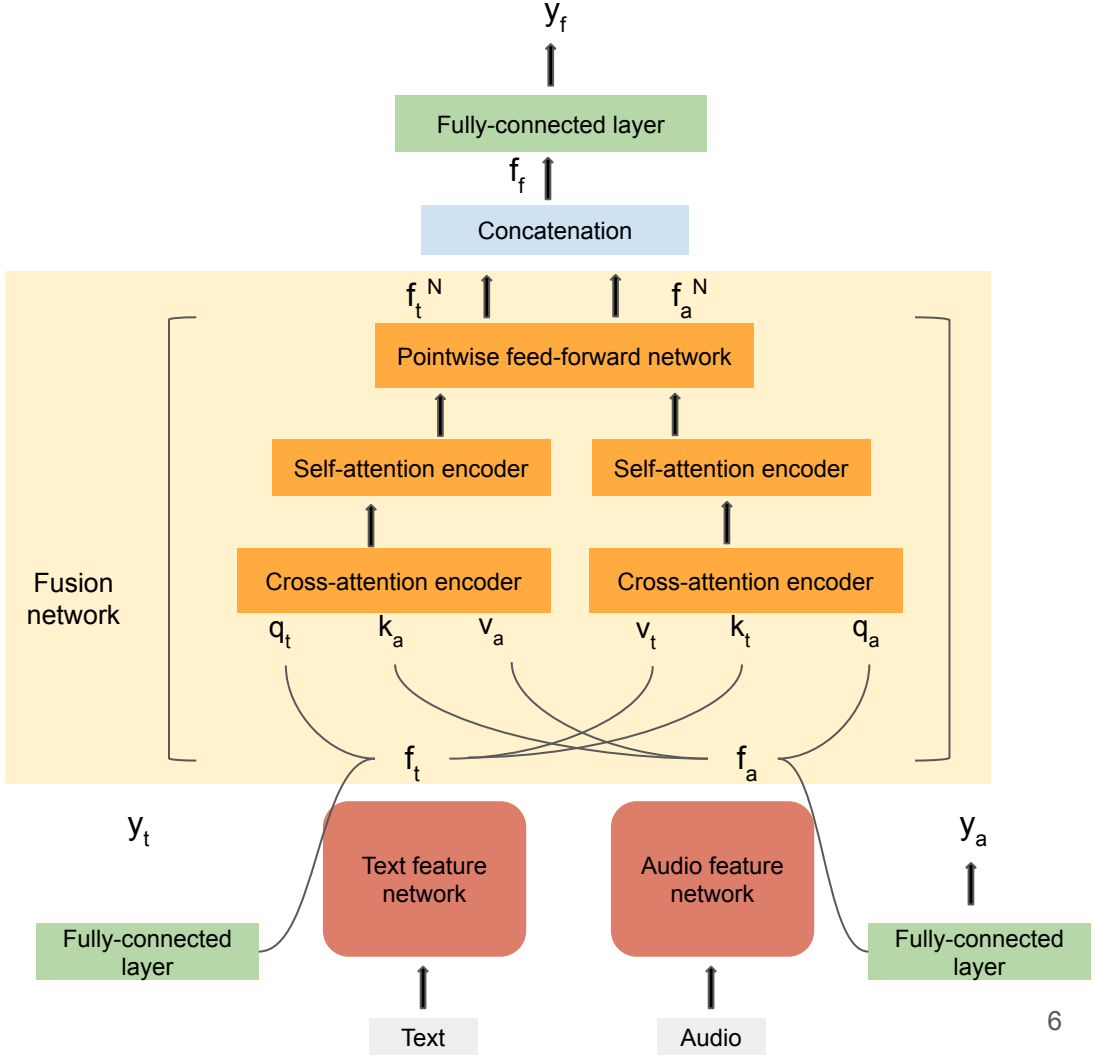
Model structure

Feature Network (red portion)
Pre-trained model for each modality

- Text: RoBerta
- Audio: pre-trained ASR model

Fusion Network (yellow portion)
For each modality:

- One Cross-attention encoder
- One Self-attention encoder
- Three fully connected layers



Sentiment Datasets

CMU-MOSI, CMU-MOSEI (English):

- Modality: audio, text, and video
- Size: 2199/23453 annotated video segments
- videos collected from YouTube

CH-SIMS (Mandarin)

- Modality: audio, text, and video
- Size: 2281 annotated video segments
- TV shows and movies
- multiple labels for the same utterance based on different modalities

Text features: Using fine-tuned RoBerta

CMU-MOSI (English):

	RoBerta(EN)
Class-2 ACC	0.8455

CH-SIMS (Mandarin):

	Translation + RoBerta(EN)	RoBerta(CH)
Class-2 ACC	0.7806	**0.7921**

(Class-2 ACC: positive vs negative valence accuracy)

Takeaway:

- Model performs well on both datasets.
- Directly using a Chinese pre-trained model is better than using a translated Chinese to English model.

Audio features: 3 Different Features

CMU-MOSI (English):

Feature name	Class-2 ACC
openSMILE	0.4606
Mel Spectrogram	0.4519
Fine-tuned Data2vec(EN)	**0.7099**

CH-SIMS (Mandarin):

Feature name	Class-2 ACC
openSMILE	0.6696
Mel Spectrogram	0.6805
Fine-tuned HuBert(CH)	**0.7465**

We use three types of **audio features**:

- openSMILE: low level descriptors extracted from raw audio signals
- Mel Spectrogram
- Large Pre-trained speech models

Takeaway:

- Pre-trained models work the best for both datasets

Model Structure Performance Comparison (Concat vs. Fusion)

MOSI	Has0_acc_2	Has0_F1	Non0_acc_2	Non0_F1	acc_5	acc_7	MAE	Corr
text-only	84.79	84.72	87.29	87.29	56.41	48.68	64.96	83.61
concat	85.77	85.74	87.6	87.62	**56.51**	**48.79**	**64.27**	**84.06**
fusion	**85.91**	**85.85**	**88.16**	**88.15**	56.08	48.25	64.29	83.8

MOSEI	Has0_acc_2	Has0_F1	Non0_acc_2	Non0_F1	acc_5	acc_7	MAE	Corr
text-only	84.81	84.95	86.34	86.19	54.99	52.7	53.31	78.6
concat	84.77	84.9	86.82	86.65	55.99	**53.94**	**51.63**	79.81
fusion	**85.61**	**85.67**	**87.01**	**86.81**	**56.85**	53.69	52.19	**79.94**

CH-SIMS	ACC_2	ACC_3	ACC_5	F1	MAE	CORR
text-only	79.21	65.06	42.02	79.14	42.65	59.4
concat	81.91	**70.68**	47.12	82.1	34.96	72.37
fusion	**82.93**	69.37	**49.38**	**82.9**	**33.2**	**73.26**

Takeaway:

- Adding audio feature is always better
- Transformer fusion works slightly better than concatenation overall.

Fusion Network Ablation

	Has0_ACC$_2$	Has0_F1	Non0_ACC$_2$	Non0_F1	ACC$_7$	MAE	Corr
fusion network	**85.91**	**85.85**	**88.16**	**88.15**	**48.25**	**64.29**	**83.8**
- self attention layers	85.13	85.12	87.35	87.38	46.94	65.81	81.53
- fully connected layers	85.13	85.09	87.2	87.2	46.65	66.48	83.07

Table 8: **Fusion Network Ablation Experiment Results**

Takeaway:

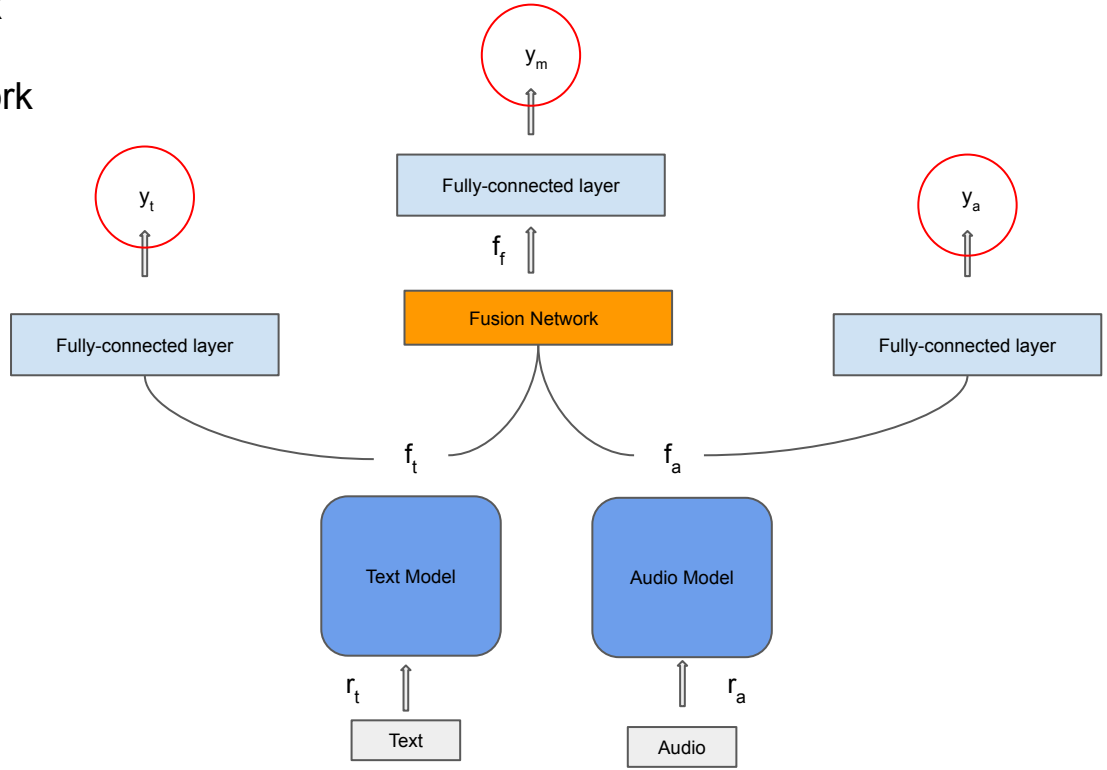
- Self-attention layers and fully connected layers are important in the Fusion Network

Multi-loss Training

Single loss: loss from the fusion network
vs.
Three losses: loss from the fusion network and two subnetworks

$$Loss = \sum_{m \in \{a,t,f\}} \alpha_m * loss\_fn(y_m, target_m)$$

(weighted sum of each loss)



Note:

- The Mandarin dataset (CH-sims) offers **distinct labels for different modalities**, so each loss can have **different targets**.
- The other two English datasets (CMU-MOSI, CMU-MOSEI) have only **one label**, so each loss has to use the **same target**.

Multi-loss Training Results

MOSEI	Has0_acc_2	Has0_F1	Non0_acc_2	Non0_F1	acc_5	acc_7	MAE	Corr
single-loss	**85.22**	**85.39**	**87.02**	**86.91**	55.95	53.85	51.96	79.68
multi-loss	84.77	84.9	86.82	86.65	**55.99**	**53.94**	**51.63**	**79.81**

Takeaway:

- If we use the same target for the different losses, it does not improve the model performance.

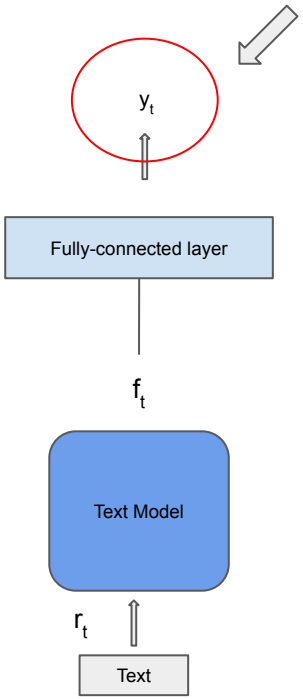
CH-SIMS	ACC_2	ACC_3	ACC_5	F1	MAE	CORR
Single-loss	78.34	67.18	46.83	78.59	39.09	62.69
multi-loss	**81.91**	**70.68**	**47.12**	**82.1**	**34.96**	**72.37**

Takeaway:

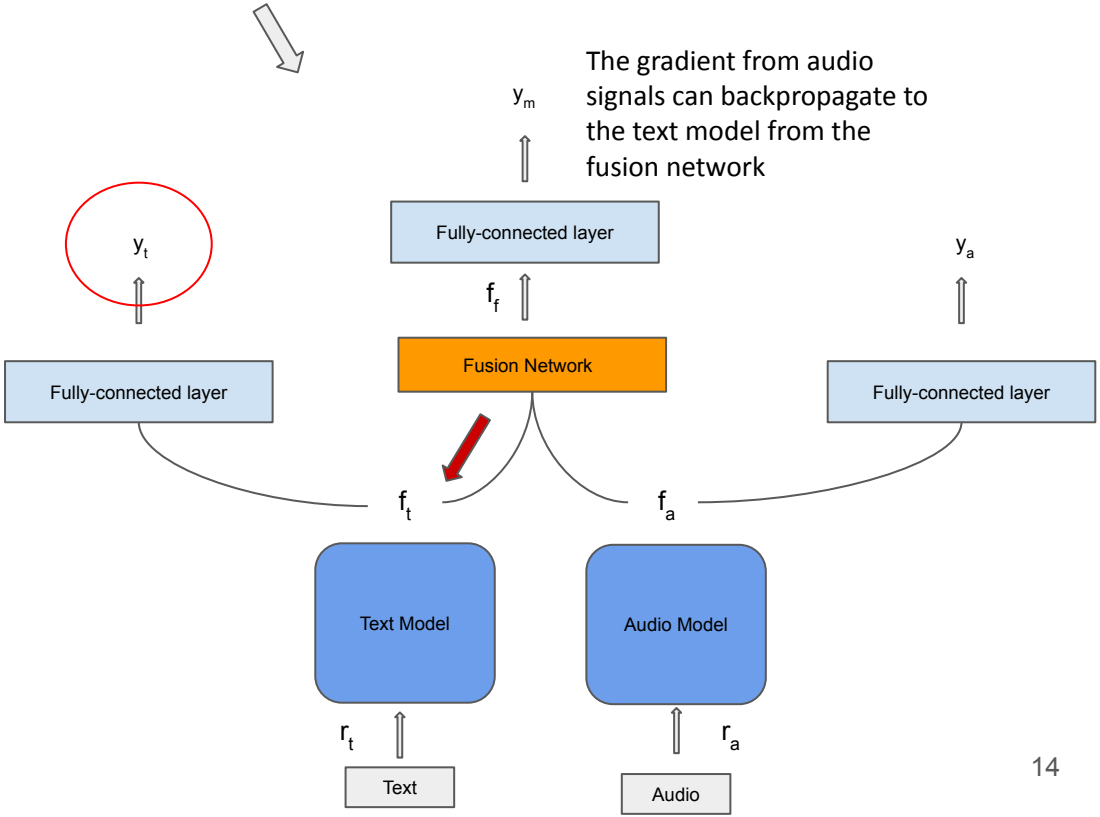
- But If we have different labels for each modality, we can have a significant improvement in all metrics

Multi-loss Training for text subnets

Is there a difference in the **text subnetwork** performance between **using only text signal** and **using both signals**?



VS



The gradient from audio signals can backpropagate to the text model from the fusion network

Multi-loss Training Text-Subnet Results

MOSI	Has0_acc_2	Has0_F1	Non0_acc_2	Non0_F1	acc_5	acc_7	MAE	Corr
text-loss	84.79	84.72	87.29	87.29	**56.41**	**48.68**	64.96	83.61
multi-loss	**85.62**	**85.56**	**87.91**	**87.9**	55.01	47.42	**64.76**	**83.79**

MOSEI	Has0_acc_2	Has0_F1	Non0_acc_2	Non0_F1	acc_5	acc_7	MAE	Corr
text-loss	**84.81**	**84.95**	86.34	86.19	54.99	52.97	53.31	78.6
multi-loss	84.77	84.9	**86.82**	**86.65**	**55.99**	**53.94**	**51.63**	**79.81**

CH-SIMS	ACC_2	ACC_3	ACC_5	F1	MAE	CORR
text-loss	79.21	65.06	42.02	79.14	42.65	59.4
multi-loss	**83.15**	**72.14**	**48.21**	**83.74**	**28.58**	**78.72**

Takeaway:

- Using multi-loss training can have a small **positive impact** on most metrics for the **text subnet**, even when we use the same target.

Takeaway:

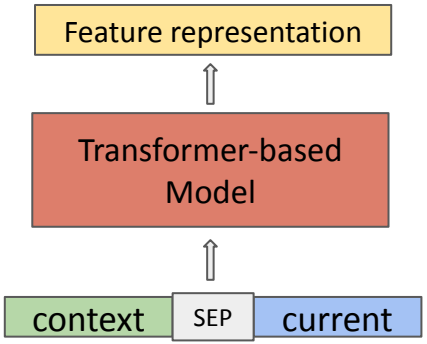
- If we have labels for other modalities, it can produce a significant improvement.
- However, such effect is not seen for the audio subnet.

Context Modeling

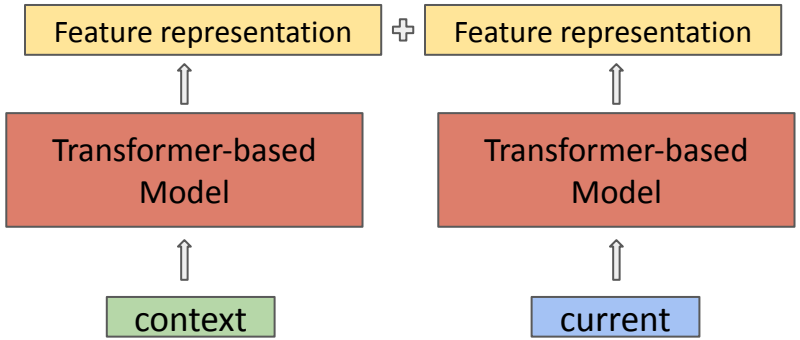
Integration of contextual data (**previous utterances**) into existing model frameworks

Comparing two context modeling methods:

1. the **concatenation** of context (previous utterances) and the current utterance as a singular input stream to the model
2. **independent processing** of context (previous utterances) and the current utterance, followed by **a subsequent fusion** of their respective representational outputs



1. concatenation



2. Separation

Context Modeling: Concatenation vs. Separation (on CMU-MOSI text data)

Context window:
the number of preceding utterances considered as contextual input

Context window	Has0_ACC$_2$	Has0_F1	Non0_ACC$_2$	Non0_F1	ACC$_5$	ACC$_7$	MAE	Corr
0	84.89	84.86	87.04	87.07	**54.81**	**47.32**	66.62	83.27
1	**86.01**	**85.94**	**88.01**	**87.99**	53.35	45.87	**66.37**	**83.96**
2	85.81	85.71	87.8	87.76	53.98	46.99	66.7	82.49
3	84.94	84.86	86.84	86.81	52.14	45.19	69.85	81.3

(a) Concatenation

Context window	Has0_ACC$_2$	Has0_F1	Non0_ACC$_2$	Non0_F1	ACC$_5$	ACC$_7$	MAE	Corr
0	84.89	84.86	87.04	87.07	54.81	47.32	66.62	83.27
1	85.57	85.51	87.8	87.8	**55.44**	**47.81**	65.17	83.37
2	**86.2**	**86.12**	**88.46**	**88.44**	55.24	47.04	**63.88**	**84.46**
3	85.76	85.69	88.01	87.99	54.67	46.89	65.26	83.71

(b) Separation

Table 7: **Comparison of Context Modeling Methods with Only Text signals**: the second method that separates the context and the current utterance can handle a longer context window and have a better performance.

Takeaways:

- Incorporating context improves performance
- The Separation method can handle a **longer context length**, and has a **better performance**

Comparison with SOTA

Model	CMU-MOSI							CMU-MOSEI						
	ACC_{2Has0}	$F1_{Has0}$	ACC_{2Non0}	$F1_{Non0}$	ACC_7	MAE	Corr	ACC_{2Has0}	$F1_{Has0}$	ACC_{2Non0}	$F1_{Non0}$	ACC_7	MAE	Corr
LMF	-	-	82.5	82.4	33.20	0.917	0.695	-	-	82.0	82.1	48.00	0.623	0.700
TFN	-	-	80.8	80.7	34.90	0.901	0.698	-	-	82.5	82.1	50.20	0.593	0.677
MFM	-	-	81.7	81.6	35.40	0.877	0.706	-	-	84.4	84.3	51.30	0.568	0.703
MTAG	-	-	82.3	82.1	38.90	0.866	0.722	-	-	-	-	-	-	-
SPC	-	-	82.8	82.9	-	-	-	-	-	82.6	82.8	-	-	-
ICCN	-	-	83.0	83.0	39.00	0.862	0.714	-	-	84.2	84.2	51.60	0.565	0.704
MulT	81.50	80.60	84.10	83.90	-	0.861	0.711	-	-	82.5	82.3	-	0.580	0.713
MISA	80.79	80.77	82.10	82.03	-	0.804	0.764	82.59	82.67	84.23	83.97	-	0.568	0.717
COGMEN	-	-		84.34	43.90	-	-	-	-	-	-	-	-	-
Self-MM	84.00	84.42	85.98	85.95	-	0.713	0.798	82.81	82.53	85.17	85.30	-	0.530	0.765
MAGBERT	84.20	84.10	86.10	86.00	-	0.712	0.796	84.70	84.50	-	-	-	-	-
MIMM	84.14	84.00	86.06	85.98	46.65	0.700	0.800	82.24	82.66 4	85.97	85.94	54.24	0.526	0.772
TEASEL	84.79	84.72	87.5	85	47.52	64.4	83.6	-	-	-	-	-	-	-
SPECTRA	-	-	87.5	-	-	-	-	-	-	87.34	-	-	-	-
UniMSE	85.85	85.83	86.9	86.42	48.68	69.1	80.9	85.86	85.79	87.5	87.46	54.39	52.3	77.3
MMML	85.91	85.85	88.16	88.15	48.25	64.29	83.8	86.32	86.23	86.73	86.49	54.95	51.74	79.08
+ context	**87.51**	**87.45**	**89.69**	**89.67**	**50.34**	**58.31**	**86.93**	**87.24**	**87.18**	**88.02**	**88.15**	**55.74**	**49.22**	**81.37**

(a) CMU-MOSI and CMU-MOSEI

	ACC_2	ACC_3	ACC_5	F1	MAE	Corr
EMT	80.1	67.4	43.5	80.1	39.6	62.3
MMML(ours)	**82.93**	**69.37**	**49.38**	**82.9**	**33.2**	**73.26**

(b) CH-SIMS

Table 1: **Comparison with SOTA**: Achieved best performance on three datasets. All experimental results presented are averages derived from three separate runs. The performances of baselines are shared by their authors.

Conclusions

- The use of **pre-trained models for raw audio** yielded superior results

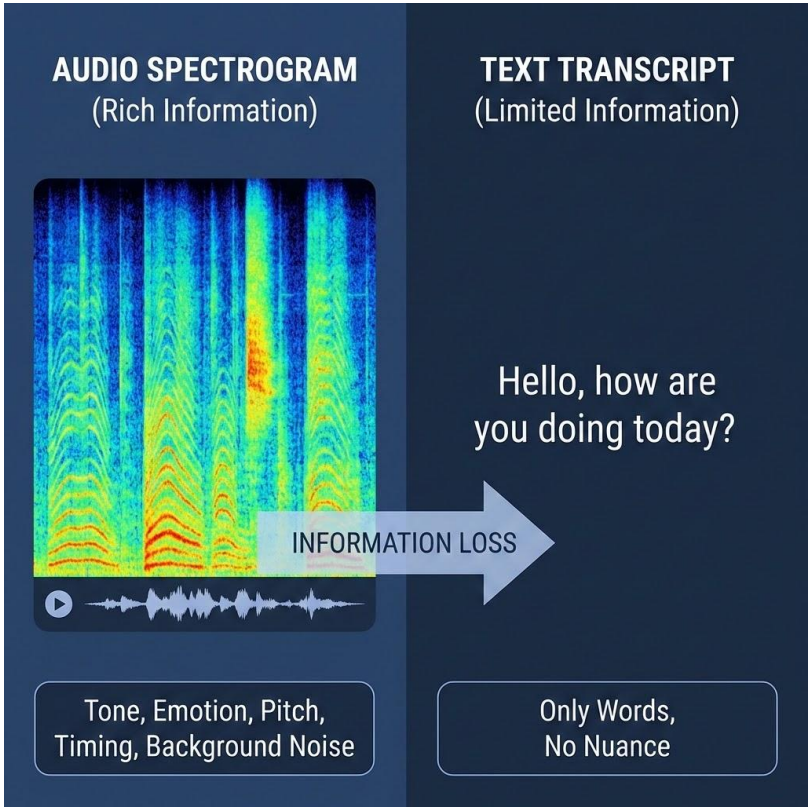
- Combining audio and text signals consistently outperformed using text signals alone, with the **transformer fusion network** showing promise in enhancing cross-modality modeling.

- **Multi-loss training** proved beneficial for overall performance and subnet performance, particularly when using unique labels for each modality.

- Our **contextual model** markedly outperformed our best non-contextual model across all evaluative metrics

- We have achieved **state-of-the-art** results on three sentiment detection datasets

2.The Acoustic Challenge – Amplifying Vocal Nuances for LLMs



Related work:

InstructERC framework (Lei et al., 2024), which reformulates emotion recognition as a generative task using LLMs.

Recent extensions have incorporated additional contextual information into the prompt, and have explored integrating speech features into LLM-based systems for various downstream tasks.

Our approach distinguishes itself by focusing on integrating speech characteristics into the templates as natural language descriptions, bridging the gap between audio and text modalities in LLM-based emotion recognition systems.

Beyond Silent Letters: Amplifying LLMs in Emotion Recognition with Vocal Nuances

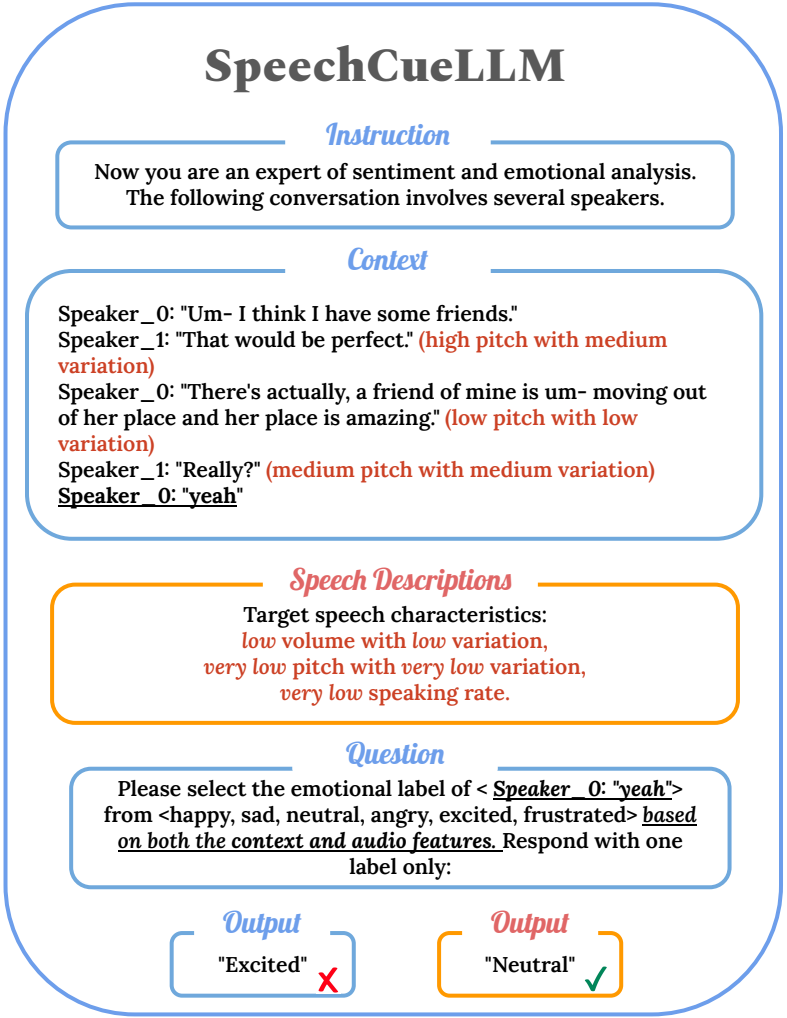
Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, Julia Hirschberg

Paper Link: https://arxiv.org/abs/2407.21315

Approach: *SpeechCueLLM*

SpeechCueLLM integrates descriptions of speech characteristics in natural language into the prompt for LLMs.

By translating audio signals into natural language descriptions, *SpeechCueLLM* enables LLMs to perform more accurate and nuanced **multimodal** emotion recognition.



Methodology: Speech Features

- Five intuitive speech features:
  - Volume: average intensity, intensity variation
  - Pitch: average pitch, pitch variation
  - Speed: articulation rate
- We can extract these features easily
- Human and LLM can understand them easily
- Each feature is standardized by gender or speaker

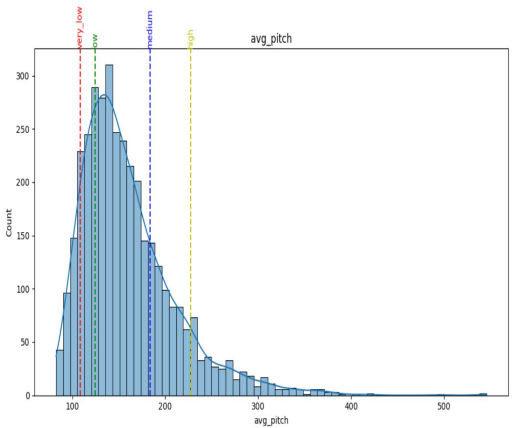
Methodology: Audio Feature Processing

- Five intuitive speech features:
  - Volume: average intensity, intensity variation
  - Pitch: average pitch, pitch variation
  - Speed: articulation rate

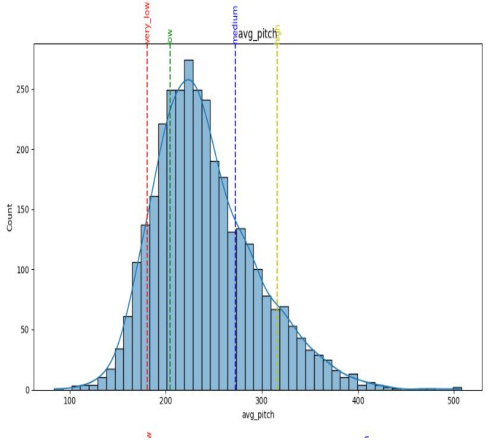
Number of Classes	Quantile Splits
3	[0.25, 0.75]
4	[0.25, 0.5, 0.75]
5	[0.1, 0.25, 0.75, 0.9]
6	[0.1, 0.25, 0.5, 0.75, 0.9]

Table 8: Quantile splits for different numbers of classes

Male



Female



Processing Step:

1. **Speaker-Specific Threshold Calculation** : Depending on the desired granularity, the feature space is divided into 3, 4, 5, or 6 classes, with thresholds determined by appropriate quantiles.
2. **Categorization**: Based on the calculated thresholds, each numerical feature is categorized. For example, in a 5-class system, a feature value may be classified as "very low," "low," "medium," "high," or "very high."
3. **Feature-Specific Descriptions**: For each key audio feature, we generate descriptive phrases corresponding to their categorical values. For instance, "high volume with moderate variation" or "low pitch with high variation" offers a more accessible and interpretable description of the audio features.

Methodology: Audio Impression

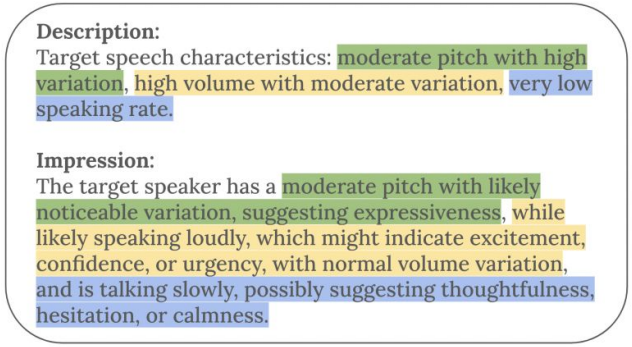


Figure 1: Example of speech characteristic description (top) and derived impression (bottom). Each color represents the same set of features (■ pitch, ■ volume, and ■ speaking rate).

Feature	Level	Impression
Pitch	High/Very High	Uses a higher pitch
	Low/Very Low	Uses a lower pitch
	Medium	Has a moderate pitch
Pitch Variation	High/Very High	With noticeable variation, suggesting expressiveness
	Low/Very Low	That remains steady, potentially indicating calmness or seriousness
	Medium	With typical variation
Volume	High/Very High	Speaking loudly, which might indicate excitement, confidence, or urgency
	Low/Very Low	Speaking softly, possibly suggesting calmness, shyness, or caution
	Medium	Using a moderate volume
Volume Variation	High/Very High	With significant volume changes
	Low/Very Low	With little volume variation
	Medium	With normal volume variation
Speech Rate	High/Very High	Talking quickly, which could indicate excitement, urgency, or nervousness
	Low/Very Low	Talking slowly, possibly suggesting thoughtfulness, hesitation, or calmness
	Medium	Speaking at a moderate pace

- To make the audio features more accessible and meaningful for both human interpretation and LLM processing, we generate hard-coded natural language impressions based on the categorized features. The feature-impression mapping is manually curated based on established correlation between prosodic features and emotional states in the literature.
- To account for the inherent uncertainty in interpreting speech patterns, we incorporate hedge words (e.g., "likely," "may") depending on the level of the category. This nuanced approach prevents overconfident interpretations, particularly in borderline cases.

Experiment settings

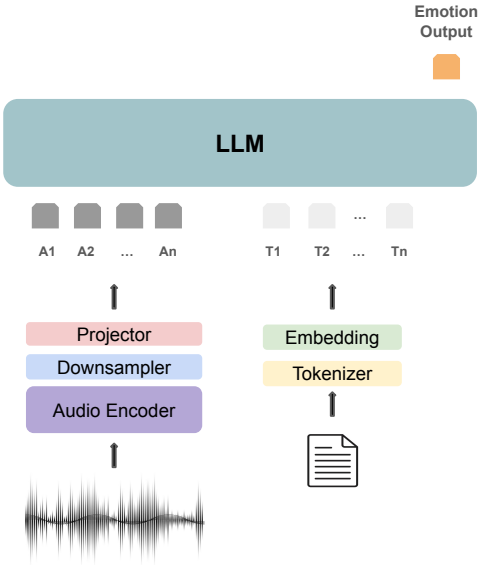
Emotion Datasets:

- IEMOCAP:12 hours of audiovisual data from 10 actors
- MELD: It contains 13,708 utterances from 1,433 dialogues in the TV show *Friends*

SpeechCueLLM Training Settings:

- Use the LLaMA-2-base model as our LLM
- Apply the LoRA (Low-Rank Adaptation) fine-tuning method

Experiments: Comparison with the Baselines



Approach	Trainable parameters	IEMOCAP F1 Score	MELD F1 Score
Speech-Encoder-Only	93.2M	52.312 ±1.7 (48.263)	47.921 ±0.27 (28.413)
LLM (w/o Speech Features)	12.6M	70.111 ±0.36	67.44 ±0.26
+FeedForward$_{\text{Projector}}$ +SE$_{\text{Freeze}}$	28.9M	70.221±0.2	67.402±0.27
+FeedForward$_{\text{Projector}}$ +SE$_{\text{Unfreeze}}$	122.1M	70.635±0.74	66.156±0.83
+Q-Former$_{\text{Projector}}$ +SE$_{\text{Freeze}}$	71.4M	70.235±0.66	67.194±0.15
SpeechCueLLM (Ours)	12.6M	**72.596** ±0.26	**67.604** ±0.38

Table 1: Performance (weighted F1) comparison between *SpeechCueLLM* and the baselines. The top section includes results using only the speech encoder. (The number in parenthesis is the average F1 score.) The middle section includes results using only the LLM without speech features and varitions of the projection-based model. The bottom is our *SpeechCueLLM* results.

- SpeechCueLLM outperforms baselines that embed audio features using speech encoders, achieving state-of-the-art (SOTA) performance.

Experiments: How Effective Are Various Speech Features?

Method	Text-only	+ speech des	+ speech imp	+ speech des with context
IEMOCAP	70.111 ±0.36	72.021 ±0.54 (+1.910)	71.542 ±1.12 (+1.431)	**72.596** ±0.26 (+2.485)
MELD	67.44 ±0.26	67.074 ±0.49 (-0.366)	**67.604** ±0.38 (+0.164)	67.09 ±0.57 (-0.35)

Table 2: *SpeechCueLLM* performances (weighted F1 score in %) on IEMOCAP and MELD datasets using LoRA fine-tuning on the LLaMA-2-7B-base model. Results are averaged over three independent runs, reported as mean ± standard deviation. Dark green values in parentheses show the improvement over the Text-only method.

- For IEMOCAP, we observe a significant improvement of nearly 2 percentage points when adding speech descriptions to the text input (from 70.111% to 72.021%). By further enriching the input with additional speech features for context, the model improves to 72.596%
- Using the speech impressions are not consistently better than using speech descriptions. Though it is more interpretative, it may introduce noise or inaccuracies that the model struggles to utilize effectively.
- The lower audio quality in MELD limits the effectiveness of speech feature integration

Experiments: Can LLMs Leverage Speech Features Alone?

```
Now you are expert of sentiment and emotional
analysis using only audio features.

Target speech characteristics: very low volume
with very low variation, very high pitch with
very high variation, high speaking rate.

Please select the emotional label from <happy,
sad, neutral, angry, excited, frustrated> based
on the audio features. Respond with just one
label:
```

Figure 3: LLM Prompt Template for speech-feature-only Emotion Detection.

Method	**Weighted F-1**
blind estimation	16.67
numerical features (ML)	32.0
3-class categories (LLM)	27.602
5-class categories (LLM)	27.895
speech impression (LLM)	27.794

Table 3: Comparison of Weighted F-1 Scores for Different Classification Methods on IEMOCAP **using speech-feature only**

- The results in right table demonstrate the potential of LLMs in leveraging these features for emotion classification, significantly outperforming the blind estimation baseline of 16.67%

Experiments: How Do LLMs Differ in Performance?

Method Model	Text-only	Text + speech des	Text + speech imp
Claude Sonnet 3.5 (zero-shot)	58.99	**59.23** (+0.24)	59.11 (+0.12)
LLaMA-2-7b-base (LoRA)	70.111 ±0.36	**72.021** ±0.54 (+1.910)	71.542 ±1.12 (+1.431)
LLaMA-3-8b-base (LoRA)	70.316 ±0.95	71.744 ±0.79 (+1.428)	**71.910** ±0.62 (+1.594)
LLaMA-3-8b-instruct (LoRA)	70.474 ±0.94	**72.098** ±0.65 (+1.624)	71.132 ±0.36 (+0.658)
phi-3-13b-instruct (LoRA)	69.218 ±0.73	70.450 ±0.86 (+1.232)	**70.582** ±0.24 (+1.364)

Table 4: *SpeechCueLLM* performances (weighted F1 score in %) on IEMOCAP dataset across different models and input modalities. Results for LoRA fine-tuned models are reported as mean ± standard deviation over three runs. Bold indicates the best performance for each model. Dark green values in parentheses show the improvement over the Text-only method.

- Our results show consistent performance improvements across all models when either speech descriptions or impressions are integrated
- Having stronger performance on public benchmarks does not necessarily lead to better performance in this task
- The substantial performance gap between the zero-shot Claude model and the fine-tuned models highlights the importance of task-specific fine-tuning for emotion recognition

Conclusion

Our main contributions are as follows:

- We introduce SpeechCueLLM, a method that enables multimodal analysis by integrating speech characteristics into text prompts, facilitating emotion recognition without architectural changes to the LLMs. SpeechCueLLM outperforms traditional approaches that incorporate speech encoders for embedding audio features into LLMs.
- We present key findings on the effectiveness and limitations of this method, showing that incorporating speech descriptions enhances emotion recognition accuracy across multiple LLMs while offering insights into the role of audio quality and the advantages of prompt-based techniques.

3.The Contextual Challenge – Ambiguity in Conversations

Emotion is Not in the Utterance; It's in the **Context**

- Single utterances are often ambiguous
- Multimodal signals can be noisy or conflicting
- Models still make confident predictions

If the model knows it doesn't know, it can reason.

From Deterministic Classification to Uncertainty-Aware Reasoning

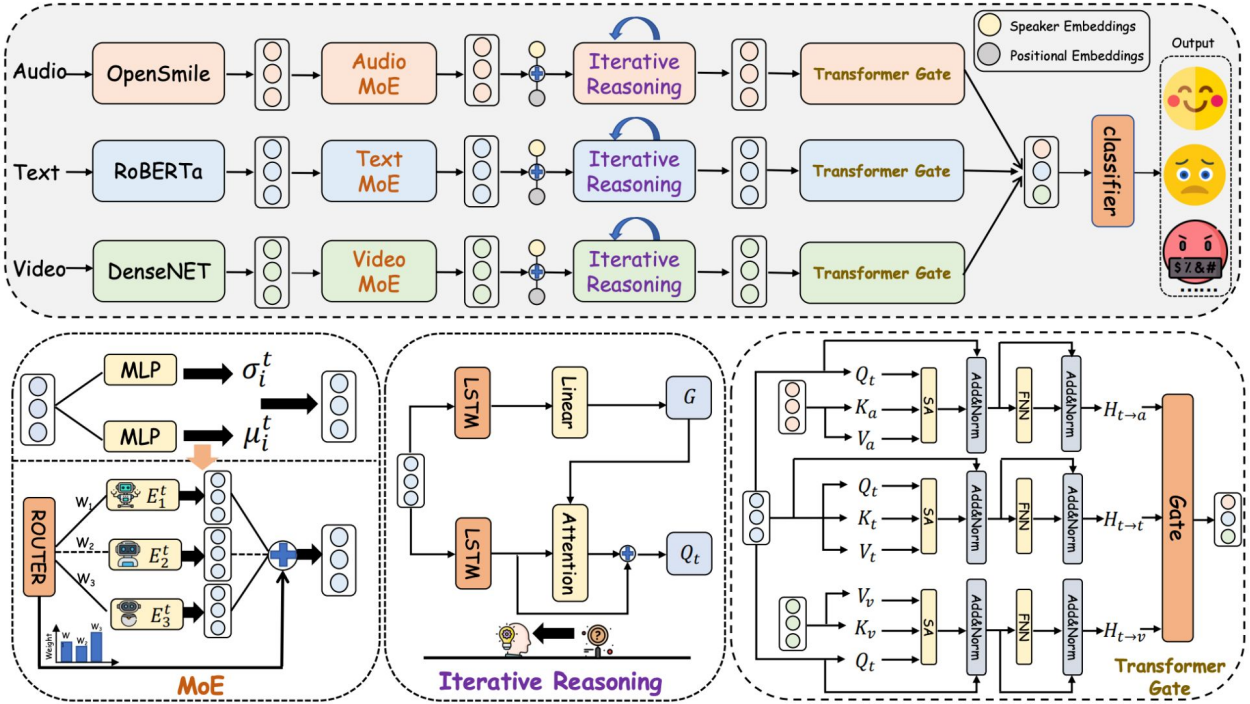
- Step 1: Detect uncertainty
- Step 2: Use context when uncertain
- Step 3: Resolve ambiguity dynamically

SURE: Synergistic Uncertainty-aware Reasoning for Multimodal Emotion Recognition in Conversations

Yiqiang Cai*, Chengyan Wu*, Bolei Ma, Bo Chen, Yun Xue,
Julia Hirschberg, Ziwei Gong

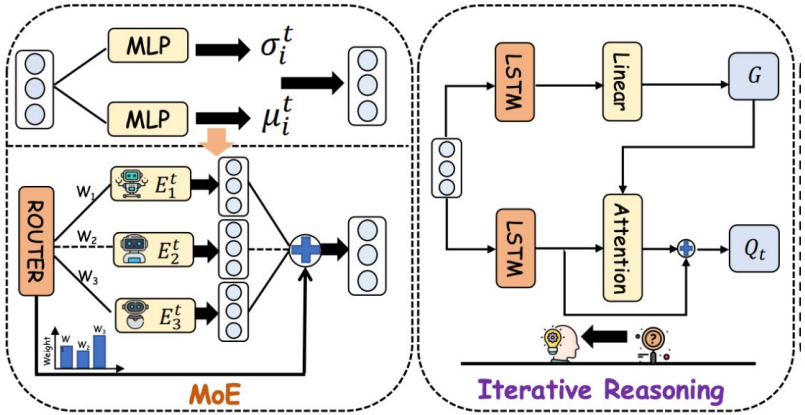
Paper Link: https://arxiv.org/pdf/2604.01916

Method: SURE Framework



The SURE Framework. The top shows the overall framework of multimodal inputs and classifying process. The bottom consists of three main components: Uncertainty-Aware MoE, Iterative Reasoning, and Transformer Gate, corresponding to the three major intermediate modules in the SURE framework.

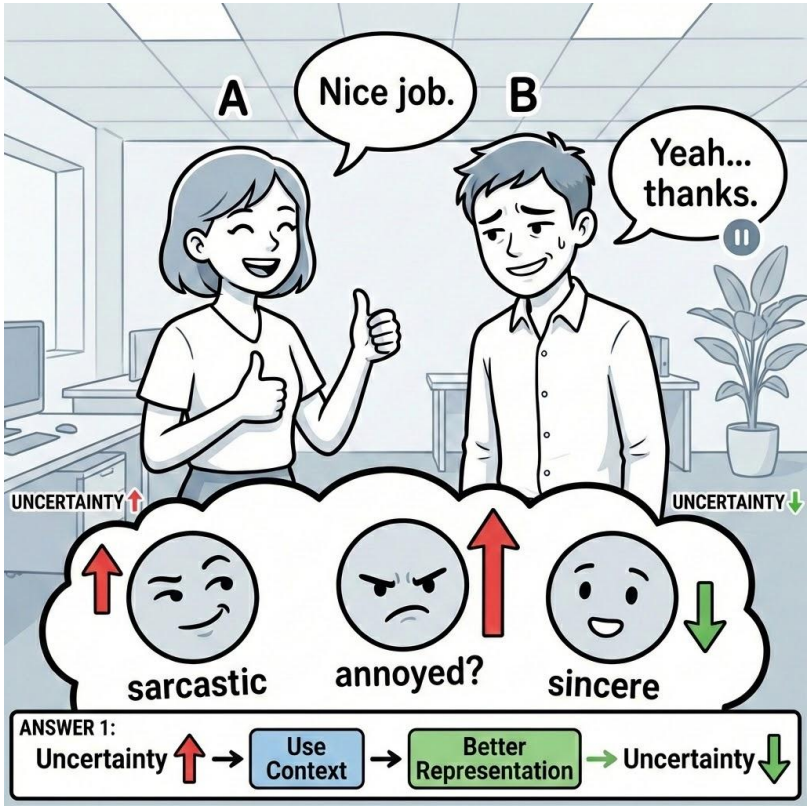
Synergistic Mechanism



High uncertainty → trigger context reasoning

Context → refine representation

Modalities interact dynamically



Uncertainty is not an output; it is a control signal.

Results & Takeaway

Table 2. Comparison with baselines on IEMOCAP and MELD.

Models	IEMOCAP		MELD	
	Acc	F1	Acc	F1
Graph-based Methods				
MMGCN	66.36	66.26	60.42	58.31
MM-DFN	68.21	68.18	62.49	59.46
Joyful	70.55	71.03	62.53	61.77
MMPCGN	68.90	68.00	60.70	59.30
MERC-GCN	–	68.98	–	62.54
Fusion-based Methods				
DialogueTRM	69.50	69.70	65.70	63.50
DF-ERC	71.84	71.75	**68.28**	67.03
SDT	73.95	74.08	67.55	66.60
MM-NodeFormer	74.24	74.20	67.86	66.09
SURE (ours)	**75.31**	**74.80**	67.97	**67.36**

Table 3. Model variations and fusion network ablation results.

	IEMOCAP		MELD	
Set-ups	Acc	F1	Acc	F1
Methods				
SURE	75.31	74.80	67.97	67.36
w/o MoE	74.99	74.23	67.65	67.02
w/o Reasoning	75.02	74.42	67.32	66.92
Modality				
Text	68.66	68.39	66.16	66.29
Audio	60.13	57.74	37.21	39.88
Visual	42.32	39.50	30.86	31.34
Text + Audio	73.98	73.05	66.37	66.44
Text + Visual	69.42	68.89	65.94	66.15
Visual + Audio	62.20	61.35	38.36	40.54

- Consistent gains on **IEMOCAP, MELD**
- More robust to **noise and ambiguity**
- More **interpretable** (uncertainty signal)

4.The Demographic Challenge – Low-Resource & Cross-Cultural Perception

Dataset	Dialogue	Modalities	Prosodic Annotations	Sources	Mul-label	Emos	Spks	Language	Utts
EmoryNLP (Zahiri and Choi, 2018)	Yes	t	No	Friends TV	Yes	9	–	English	12,606
EmotionLines (Chen et al., 2018)	Yes	t	No	Friends TV	No	7	–	English	29,245
DailyDialog (Li et al., 2017)	Yes	t	No	Daily	No	7	–	English	102,979
CMU-MOSEI (Zadeh et al., 2018)	No	a, v, t	No	YouTube	No	7	1000	English	23,453
AFEW (Dhall et al., 2012)	No	a, v	No	Movies	No	7	330	English	1,645
MEC (Li et al., 2018)	No	a, v	No	Movies, TVs	No	8	–	Mandarin	7,030
CH-SIMS (Yu et al., 2020)	No	a, v, t	No	Movies, TVs	No	5	474	Mandarin	2,281
IEMOCAP (Busso et al., 2008)	Yes	a, v, t	No	Act	No	5	10	English	7,433
MSP-IMPROV (Busso et al., 2016)	Yes	a, v, t	No	Act	No	5	12	English	8,438
MELD (Poria et al., 2018)	Yes	a, v, t	No	Friends TV	No	7	407	English	13,708
M³ED (Zhao et al., 2022a)	Yes	a, v, t	No	56 TVs	Yes	7	626	Mandarin	24,449
AkaCE (Ours)	Yes	a, v, t	Yes	21 Movies	No	7	308	Akan	6,162

Table 1: Comparison of existing benchmark datasets. a, v, t refer to audio, visual, and text modalities respectively.

Challenges in Low-Resource SER

• **Dependence on labeled data**: Supervised learning approaches require large labeled datasets, which are unavailable for most LRLs.

• **Generalization challenges**: Models fine-tuned on HRLs often fail to generalize effectively to different LRLs due to linguistic differences.

• **Unsupervised learning challenges**: Self-supervised learning methods like contrastive learning and BYOL remain underexplored for SER.

• **Practical deployment concerns**: Biases in SER models could lead to fairness issues, especially in sensitive applications like mental health.

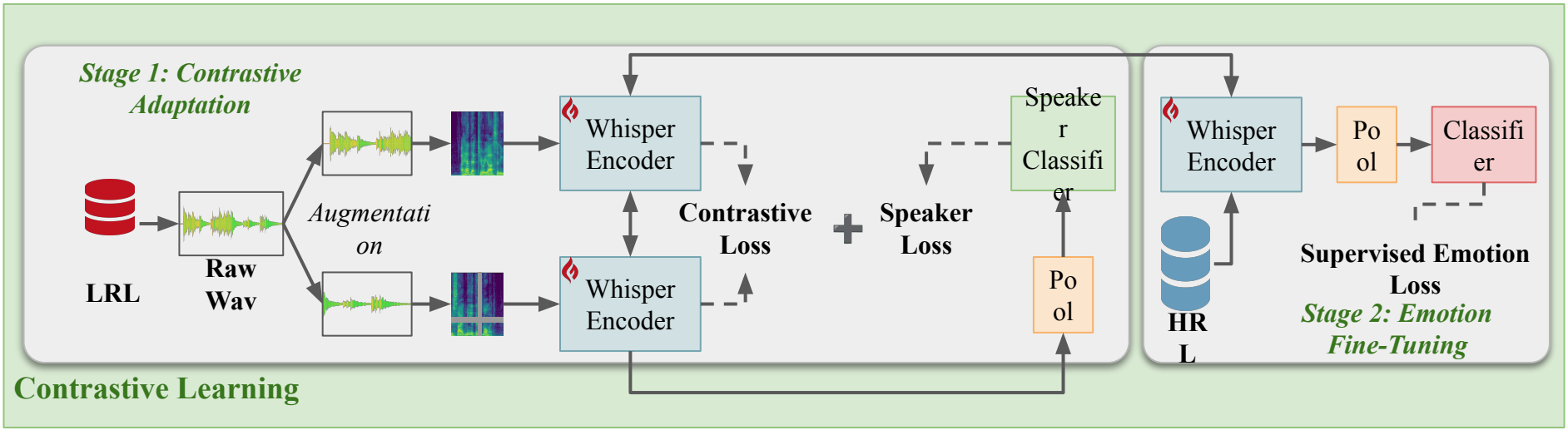
Key hypothesis: Self-supervised learning can bridge the gap between HRLs and LRLs in emotion recognition

Learning More with Less: Self-Supervised Approaches for Low-Resource Speech Emotion Recognition

Ziwei Gong, Pengyuan Shi, Kaan Donbekci, Lin Ai, Run Chen, David Sasu, Zehui Wu, Julia Hirschberg

Under Review [Interspeech 2025]

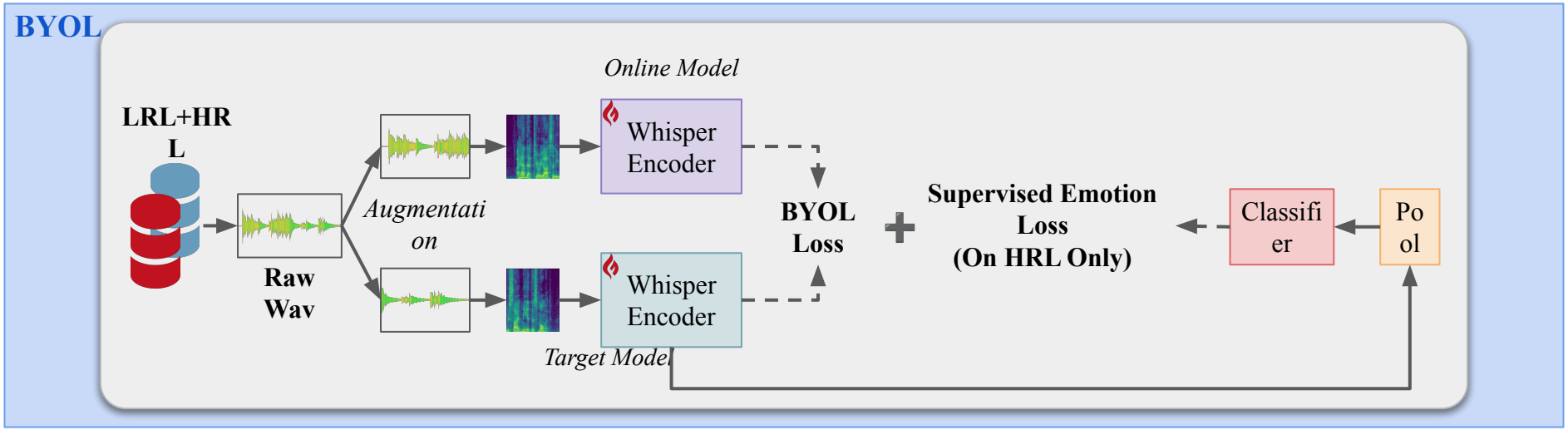
Proposed Methodology 1: Contrastive Learning (CL)



We propose using **Contrastive Learning (CL)** and **Bootstrap Your Own Latent (BYOL)** to improve SER in LRLs.

- CL differentiates similar and dissimilar samples using self-supervised similarity measures.

Proposed Methodology 2: Bootstrap Your Own Latent



- BYOL learns representations without negative samples, promoting consistency and feature invariance.

Both methods reduce reliance on labeled data and aim to enhance cross-lingual generalization.

Datasets & Augmentation

- **Datasets Used**:
  - High-resource: MSP-Podcast, IEMOCAP (English).
  - Low-resource: EMO-DB (German), URDU (Urdu), SUBESCO (Bangla).
- **Feature Extraction**: Mel-spectrograms
- **Data Augmentation Techniques**:
  - CL: Gaussian noise injection, polarity inversion, gain manipulation.
  - BYOL: Speed perturbation, spectral stretch, SpecAugment, time-frequency masking.
- **Why Augmentation Matters?**
  - Helps models learn robust and invariant features.
  - Prevents overfitting to high-resource training data.
  - Increases diversity in training samples, improving cross-lingual adaptation.

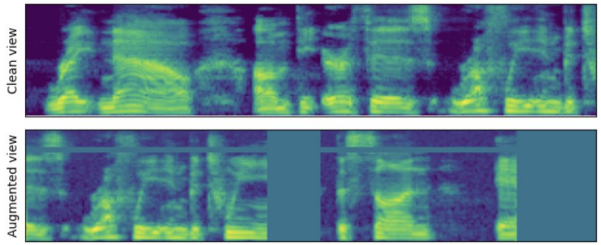


Figure 2: *Both views of the same utterance with minimal overlap, forming a positive pair in contrastive adaptation or BYOL.*

Experiments Results

- Our results show that:
  - CL significantly improves emotion classification in German (90% zero-shot accuracy on EmoDB).
  - BYOL is more effective for Urdu and Bangla, demonstrating robustness to dataset biases.
  - Zero-shot performance remains stable, showing that self-supervised learning enhances generalization.

- **Takeaway**: Self-supervised methods can reduce dependency on HRL data while improving performance on LRLs.

Dataset	Metric$_{(std)}$	Baseline	Contrastive	BYOL
URDU	Accuracy	$0.597_{(0.025)}$	$0.648_{(0.031)}$	$\mathbf{0.658}_{(0.052)}$
Urdu	Macro F1	$0.547_{(0.036)}$	$0.629_{(0.041)}$	$\mathbf{0.653}_{(0.060)}$
	UAR	$0.597_{(0.025)}$	$0.637_{(0.032)}$	$\mathbf{0.659}_{(0.053)}$
EMODB	Accuracy	$0.776_{(0.040)}$	$\mathbf{0.906}_{(0.017)}$	$0.794_{(0.033)}$
German	Macro F1	$0.750_{(0.059)}$	$\mathbf{0.901}_{(0.018)}$	$0.732_{(0.030)}$
	UAR	$0.758_{(0.050)}$	$\mathbf{0.896}_{(0.019)}$	$0.751_{(0.033)}$
SUBESCO	Accuracy	$0.616_{(0.015)}$	$\mathbf{0.721}_{(0.008)}$	$0.643_{(0.012)}$
Bangla	Macro F1	$0.572_{(0.016)}$	$\mathbf{0.711}_{(0.011)}$	$0.641_{(0.009)}$
	UAR	$0.616_{(0.015)}$	$\mathbf{0.721}_{(0.008)}$	$0.643_{(0.014)}$
RAVDESS	Accuracy	$0.574_{(0.020)}$	$0.580_{(0.018)}$	$\mathbf{0.582}_{(0.032)}$
English	Macro F1	$0.514_{(0.025)}$	$0.519_{(0.023)}$	$\mathbf{0.561}_{(0.029)}$
	UAR	$\mathbf{0.578}_{(0.034)}$	$0.570_{(0.026)}$	$0.572_{(0.030)}$

Table 1: *Performance$_{(std)}$ Comparison of Baseline, CL and BYOL models on different datasets along with their language*

Model Architectures:

- Whisper encoder for feature extraction.
- Fine-tuning with contrastive learning and BYOL.
- Classification head with fully connected layers.

Training Details:

- Experiments repeated five times to ensure stability.

Results & Discussion: Model Interpretability Insights

• **Embedding Visualization (t-SNE Analysis)**

- CL showed more **structured emotion clusters** but struggled with linguistic transfer.

- BYOL provided **smoother feature spaces**, allowing better adaptation to unseen data.

• **Confusion Matrix Insights**

- CL reduced errors in anger-happiness misclassifications.

- BYOL improved differentiation of neutral and sad emotions.

• **Post-hoc Analysis (SHAP & LIME)**

- CL heavily relied on pitch features

- BYOL showed a more balanced reliance on spectral and temporal cues, aiding robustness.

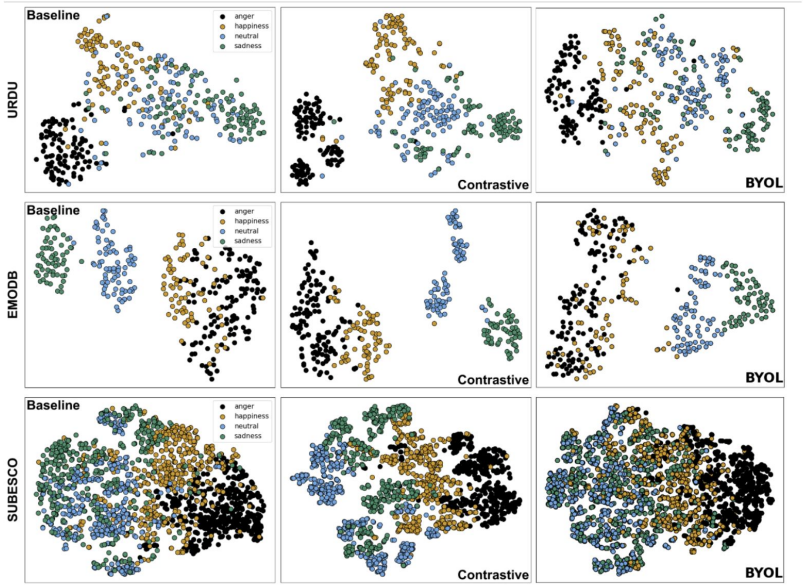


Figure 3: *T-SNE plots of learned embeddings on LRLs. Left: Baseline model after training. Middle: Speaker-contrastive model after training. Right: BYOL model after training.*

Results & Discussion: Model Bias & Fairness

- **Gender Bias in Urdu SER**
  - CL models showed a 10% performance drop for **female speakers** in Urdu.
  - BYOL exhibited more balanced performance, reducing the bias observed in CL.
  - Possible cause: Male-dominated training data in Urdu Common Voice dataset.
  - *Future Work*: Rebalancing datasets to ensure fairness across gender groups.
- **Linguistic Biases**
  - German outperformed other LRLs, likely due to its proximity to English.
  - Cross-lingual emotion mapping remains a challenge, as some emotions are expressed differently across cultures.

Conclusion & Future Work

- Self-supervised learning (CL & BYOL) improves SER performance in low-resource settings.
- Contrastive Learning (CL) excels in structuring feature spaces, but struggles with gender imbalance.
- BYOL is more robust to speaker and dataset variability, making it better suited for diverse LRLs.
- Zero-shot performance remains challenging, with German benefiting more than Urdu/Bangla due to linguistic similarities.
- Model interpretability is essential for identifying biases, as seen in gender imbalances in Urdu

Future research should focus on hybrid models, combining strengths of CL, BYOL, and domain adaptation techniques.

Akan Cinematic Emotions (ACE): A Multimodal Multi-party Dataset for Emotion Recognition in Movie Dialogues

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai,
Julia Hirschberg, Natalie Schluter

Paper Link: https://arxiv.org/abs/2502.10973

Dataset Link: https://anonymous.4open.science/r/Akan-Cinematic-Emotion-A328

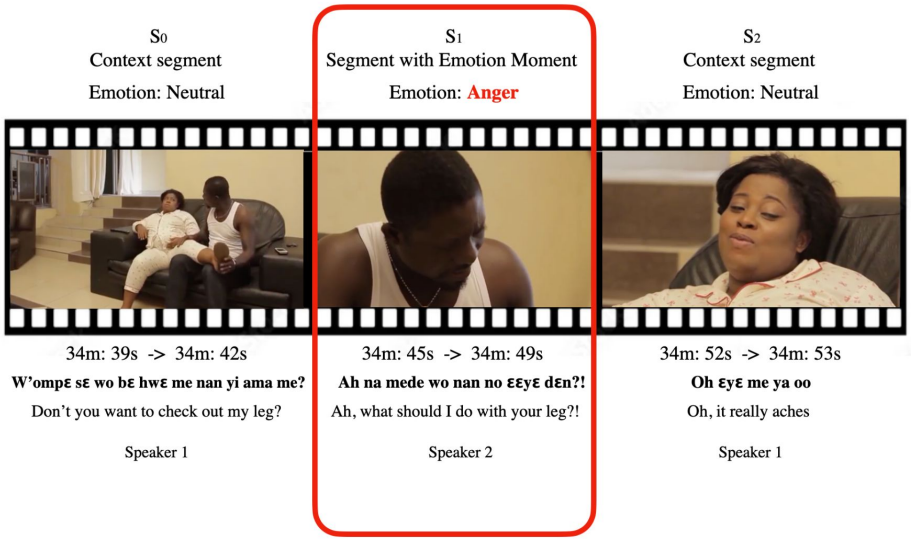
A Comprehensive Multimodal Dataset for Akan: ACE

- ACE includes 385 emotion-labeled dialogues and 6,162 utterances from 21 Akan movies.
- Features audio, visual, and textual modalities with word-level prosodic prominence annotations.
- Ensures gender balance with contributions from 308 speakers (155 male, 153 female).

Dataset	Dialogue	Modalities	Prosodic Annotations	Sources	Mul-label	Emos	Spks	Language	Utts
EmoryNLP (Zahiri and Choi, 2018)	Yes	t	No	Friends TV	Yes	9	–	English	12,606
EmotionLines (Chen et al., 2018)	Yes	t	No	Friends TV	No	7	–	English	29,245
DailyDialog (Li et al., 2017)	Yes	t	No	Daily	No	7	–	English	102,979
CMU-MOSEI (Zadeh et al., 2018)	No	a, v, t	No	YouTube	No	7	1000	English	23,453
AFEW (Dhall et al., 2012)	No	a, v	No	Movies	No	7	330	English	1,645
MEC (Li et al., 2018)	No	a, v	No	Movies, TVs	No	8	–	Mandarin	7,030
CH-SIMS (Yu et al., 2020)	No	a, v, t	No	Movies, TVs	No	5	474	Mandarin	2,281
IEMOCAP (Busso et al., 2008)	Yes	a, v, t	No	Act	No	5	10	English	7,433
MSP-IMPROV (Busso et al., 2016)	Yes	a, v, t	No	Act	No	5	12	English	8,438
MELD (Poria et al., 2018)	Yes	a, v, t	No	Friends TV	No	7	407	English	13,708
M³ED (Zhao et al., 2022a)	Yes	a, v, t	No	56 TVs	Yes	7	626	Mandarin	24,449
ACE (Ours)	Yes	a, v, t	Yes	21 Movies	No	7	308	Akan	6,162

Table 1: Comparison of existing benchmark datasets. a, v, t refer to audio, visual, and text modalities respectively.

Detailed and Culturally Relevant Annotations



- Emotion labels cover seven categories:
  - Neutral, Sadness, Anger, Fear, Surprise, Disgust, and Happy.
- Annotations conducted by native Akan speakers, ensuring cultural relevance.
- Prosodic prominence annotations highlight the importance of tone in conveying emotions in Akan

Establishing Baselines with State-of-the-Art Methods

Baseline experiments utilized state-of-the-art emotion recognition methods.

- Multimodal approaches integrating text, audio, and visual data achieved the best performance.
- Results validate the dataset's quality and utility for future research.

Setting	Weighted F1	Macro F1
No Context	43.12	18.85
Context	**44.58**	**22.29**

Table 5: Text-based emotion detection results using the Ghana-NLP/abena-base-asante-twi-uncased model.

Model	Weighted F1	Macro F1
Concatenation	55.81	30.97
Transformer Fusion	**56.13**	**31.68**

Table 9: Results of multimodal fusion experiments.

Model	Weighted F1	Macro F1
openSMILE	13.80	6.58
Spectrogram	47.89	23.36
Whisper-small	**52.38**	**29.51**

Table 6: Audio-based emotion detection results.

Model	Weighted F1	Macro F1
ResNet18-1fps	40.57	**20.02**
ResNet50-1fps	38.19	15.1
ResNet18-5fps	**42.04**	17.92
ResNet50-5fps	41.76	19
Inception-Face-5fps	39.96	16.53

Table 7: Vision-based emotion detection results.

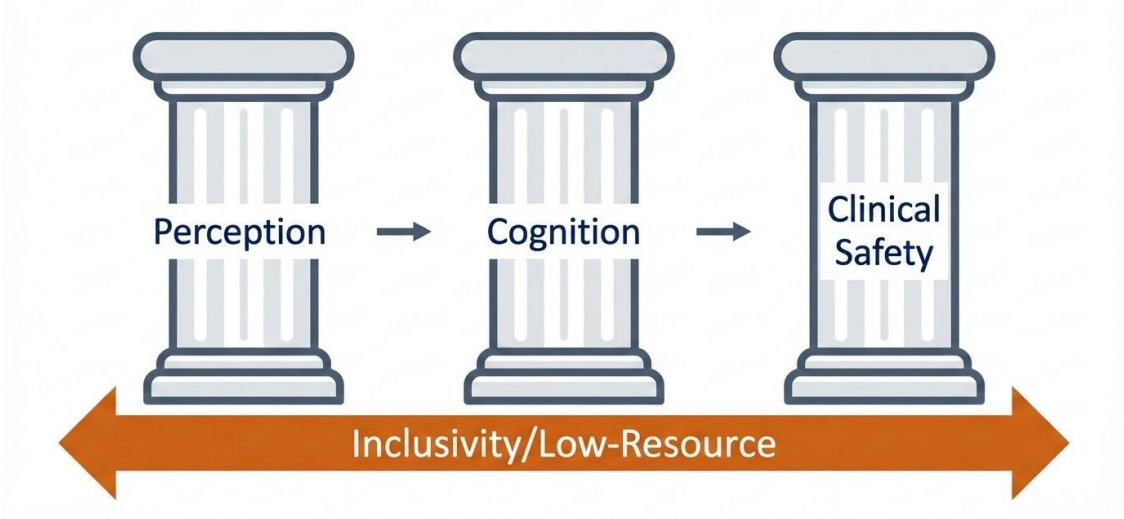
Modality	Weighted F1	Macro F1
Text	44.58	22.29
Audio	52.38	29.51
Vision	40.57	20.02
Text + Audio	55.51	30.15
Text + Vision	43.33	21.15
Audio + Vision	53.84	30.42
Text + Audio + Vision	**55.81**	**30.97**

Table 8: Results of modality concatenation experiments using the best unimodal models.

Paving the Way for Inclusive NLP Research

- ACE fills a critical gap by providing a multimodal ERC dataset for an African language.
- Facilitates cross-cultural and linguistic studies in emotion recognition.
- Future work includes expanding to other African languages and enhancing multimodal recognition techniques.

Part 2: Guiding AI Cognition (The Bridge)



A Mapping on Current Classifying Categories of Emotions Used in Multimodal Models for Emotion Recognition

Ziwei Gong, Muyin Yao, Xinyi Hu, Xiaoning Zhu, Julia Hirschberg

Department of Computer Science
Columbia University

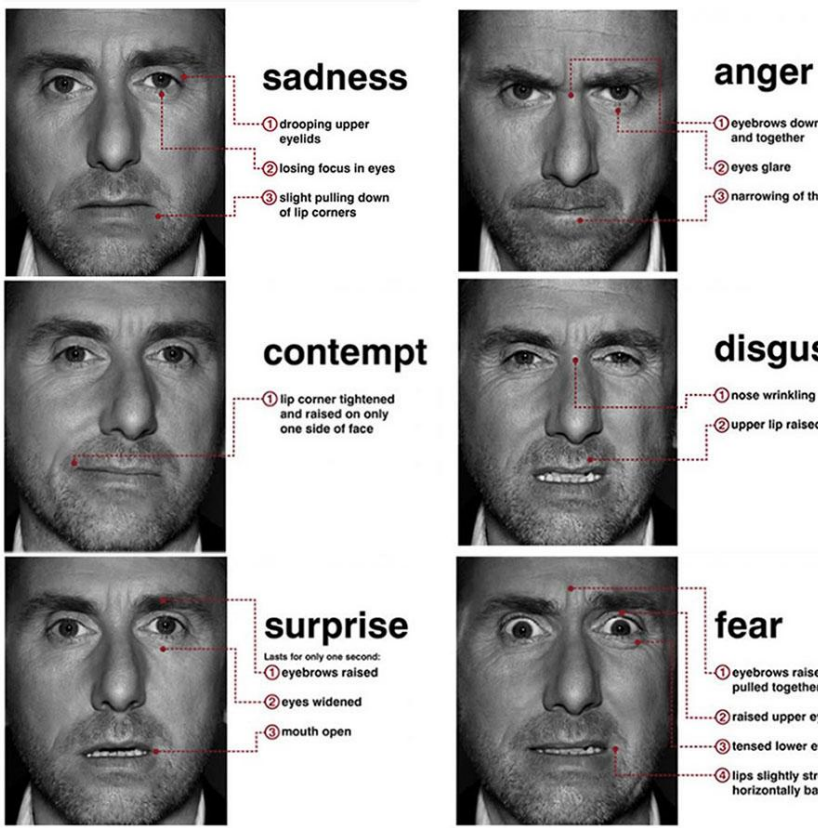
Paper link: https://aclanthology.org/2024.law-1.3/

Emotion classification theory: Plutchik's wheel of emotions

- 8 Primary Emotions:
  - Joy, sadness, trust, disgust, anticipation, surprise, anger, fear
- 24 + 8 Fine-grained emotions
  - Intensity and Combinations
- Theory based on evolution and biological classification
- Dimensional Theory

Emotion classification theory: Ekman's basic emotions

- Proposed 6 basic emotions
  - Fear,
  - Anger,
  - Joy,
  - Sadness,
  - Disgust, and
  - Surprise
- Based on universal facial expressions



An earlier version of the theory

Current Studies

E: Ekman's
V: Valence
P: Plutchik's
A: arousal
D: dominance

Dataset	Granularity	Annotation	Size	Topic	Source	Avail.
AffectiveText	headlines	E + V	1,250	news	Strapparava (2007)	D-U
Blogs	sentences	E + ne + me	5,025	blogs	Aman (2007)	R
CrowdFlower	tweets	E + CF	40,000	general	Crowdflower (2016)	D-U
DailyDialogs	dialogues	E	13,118	multiple	Li et al. (2017)	D-RO
Electoral-Tweets	tweets	P	4,058	elections	Mohammad (2015)	D-RO
EmoBank	sentences	V+A+D	10,548	multiple	Buechel (2017a)	CC-by4
EmoInt	tweets	E − DS	7,097	general	Mohammad (2017b)	D-RO
Emotion-Stimulus	sentences	E + shame	2,414	general	Ghazi et al. (2015)	D-U
fb-valence-arousal	faceb. posts	V+A	2,895	questionnaire	Preoţiuc (2016)	D-U
Grounded-Emotions	tweets	HS	2,585	weather/events	Liu et al. (2017)	D-U
ISEAR	descriptions	E + SG	7,665	events	Scherer (1994)	GPLv3
Tales	sentences	E	15,302	fairytales	Alm et al. (2005)	GPLv3
SSEC	tweets	P	4,868	general	Schuff et al. (2017)	D-RO
TEC	tweets	E ±S	21,051	general	Mohammad (2012)	D-RO

- Table from Bostan and Klinger (2018)'s An analysis of annotated corpora for emotion classification in text.

Datasets

Multimodal Conversational Emotion detection dataset (English)

- MEmoR:

  Plutchik's wheel of emotions
- MELD:

  Ekman's 6 basic emotions

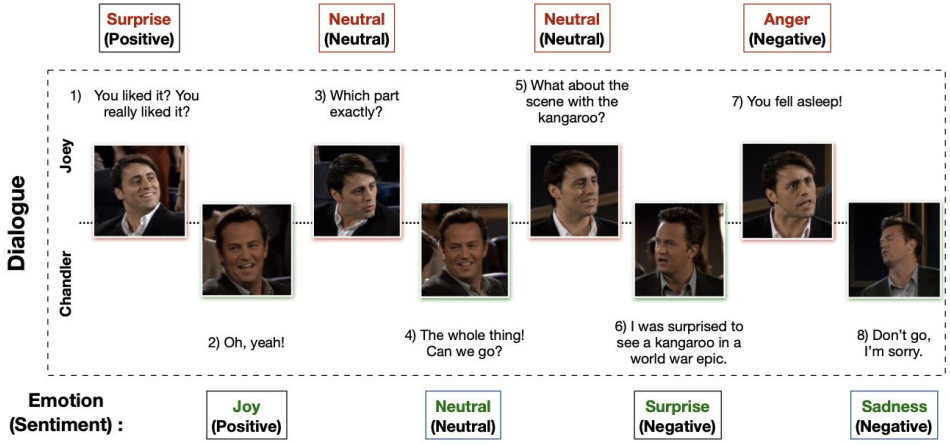


Figure 1: Emotion shift of speakers in a dialogue in comparison with their previous emotions.

Mapping on Dataset: 8 to 5

- How we decide between mapping choices:
  - A re-annotation of mapped emotions
    - Annotator 1 -- **0.96**
    - Annotator 2 -- **0.917**
    - the inner annotator agreement as 0.318 (Cohen's K, moderate agreement)
- Conclusion:
  - it is possible to map the emotion categories with **relative accuracy**

mapping method 1:	mapping method 2:
Joy: joy	Joy: joy
fear: fear	fear: fear
Trust: neutral	**Trust: joy**
Disgust: disgust	Disgust: disgust
Fear: fear	Fear: fear
Anger: anger	Anger: anger
Surprise: anger	**Surprise: neutral**
Anticipation: joy	**Anticipation: joy**
Neutral: neutral	Neutral: neutral

Note: the two methods differ in the mapping of trust and surprise.

Mapping Method of Classifying Categories

14 fine-grained	9 primary	7 basic	6 emotions	3 sentiments
anticipation	anticipation	neutral	neutral	neutral
interest				
neutral	neutral			
fear	fear	fear	fear	negative
disgust	disgust	disgust	disgust	
boredom				
sadness	sadness	sadness	sadness	
anger	anger	anger	anger	
annoyance				
surprise	surprise	surprise		
distraction				
joy	joy	joy	joy	positive
serenity				
trust	trust			

Table 1: Mapping results. This table demonstrates how 14 fine-grained emotions, listed on the leftmost column, are mapped onto 9 primary emotions, Ekman's basic emotions, 6 emotions, and the 3 sentiments.

Results of experiments on CNN (vision) and MEmoR (multimodal) model

Emotion Category	3	6	7	9	14
MEmoR Accuracy	**0.924**	0.867	0.884	0.869	0.864
CNN Accuracy	**81.78**	65.39	65.28	-	-

Table 2. Experimental results from the MEmoR model and the CNN model. This table shows the overall accuracy of the models trained and tested on datasets reconstructed based on each 3 classification method. The highest achieved is bolded. The MEmoR model uses visual, audio, textual features. In the CNN model, only visual information is used.

Takeaway: models generally perform better when there are fewer emotion categories, meaning that more fine-grained emotions are more difficult for models to differentiate,

Results: Heat Map on CNN (vision) model

Heat map on model trained on 7 emotions Heat map on model trained on 3 emotions

Figure 2. Comparison between prediction heat map on 9 random pictures, between a CNN model trained on a dataset categorized into 7 categories (left) versus a CNN model trained on the same dataset categorized into 3 categories (right). Red indicates high attention to the area.

Takeaway:We observe that the attention of the model trained with more fine-grained emotions is more spread out through the face, with some stress around the eye and mouth area.

Conclusion

- We propose the **first complete mapping** that connects different emotion categories for multimodal emotion recognition studies
- Using our mapping allows researchers to **obtain larger and more flexible datasets** for training and to analyze models
- **Observation**: Models generally perform better when there are fewer emotion categories, meaning that **more fine-grained emotions are more difficult** for models to differentiate.
- **Observation**: We observe that the attention of the model trained with more fine-grained emotions is **more spread out through the face**, compared to only focusing around the eye and mouth area.

The Mind in the Machine: A Review of Incorporating Psychological Theories in LLMs

Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R. Greene, Julia Hirschberg

Paper Link: https://aclanthology.org/2026.eacl-long.350.pdf

Psychological theories apply across every stages of LLM development

Behavioral psychology strongly shapes alignment through reinforcement learning: RLHF reflects operant conditioning principles, using reward signals to shape model behavior toward human-aligned outputs.

Cognitive psychology inspires memory, reasoning, and context modeling mechanisms: Theories such as working memory, attention, and perception inform architectural improvements that enhance long-context processing, reasoning stability, knowledge integration.

Developmental psychology motivates curriculum learning and structured training: Concepts such as incremental learning and scaffolding inspire curriculum learning strategies, enabling models to gradually acquire more complex capabilities.

Social and personality psychology guide personalization and socially aware behavior: Social identity theory, personality modeling, and social reasoning frameworks enable models to better adapt to users, simulate personas, and operate in social contexts.

However, psychological integration remains uneven and incomplete

While some psychological theories are widely adopted, others remain underutilized, highlighting opportunities for deeper and more principled integration.

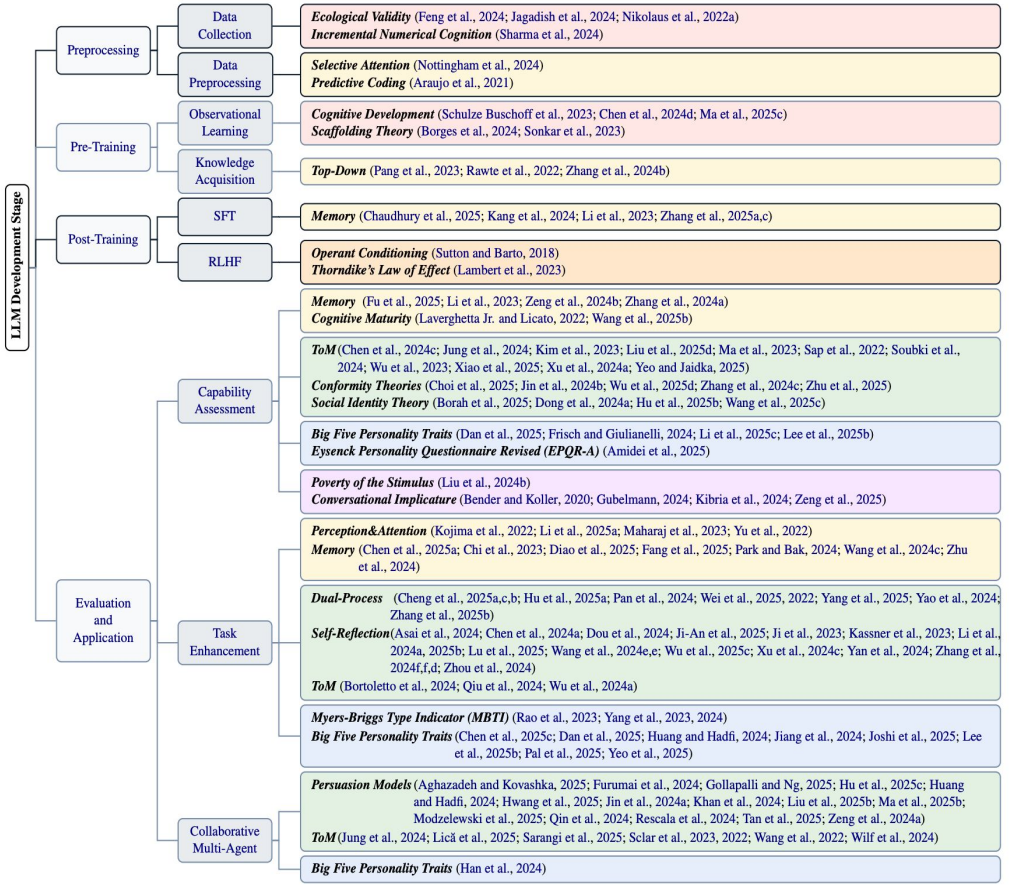


Figure 1: Our structured survey of how psychological theories apply across the main stages of LLM development. Color ndicate six distinct psychology areas: red for ***Developmental Psychology*** ; orange for ***Behavioral Psychology*** ; yellow fo ***Cognitive Psychology*** ; green for ***Social Psychology*** ; blue for ***Personality Psychology*** ; purple for ***Psycholinguistics*** .

Critical Challenges & Disciplinary Tensions

- **The Terminology Mismatch**
  - **“Attention” ≠ attention**: Psychological attention is selective mental focus, while transformer attention is a token-weighting mechanism without cognitive awareness, which can lead to misleading attributions of intentionality.
  - **“Memory” ≠ memory**: Psychological memory involves structured encoding and recall, while LLM “memory” often refers to context windows or parameters, creating confusion about what is actually being modeled.
  - **Need a shared lexicon:** Increasing human-like descriptors shape public and scholarly assumptions, motivating a precise cross-disciplinary lexicon to separate metaphor from mechanism.
- **Theoretical Discrepancies**
  - **Outdated/disputed constructs get imported:** e.g., MBTI persists despite validity critiques; predictive coding analogies can be oversimplified.
  - **Partial mappings miss theory structure:** e.g., ToM may involve choosing depth of mentalizing before reasoning, but many works model only the latter.
  - **Constraints vs capability:** Should LLMs emulate human cognitive limits (interpretability) or exceed them (performance)?
  - **RLHF ≠ full behavioral theory:** Focus on reward optimization can ignore internal states and raise reward hacking concerns.
- **Evaluation and Construct Validity**
  - Need theory-grounded evaluation: Clearer definitions and stronger validity checks, beyond surface task scores (e.g., Personality, ToM).

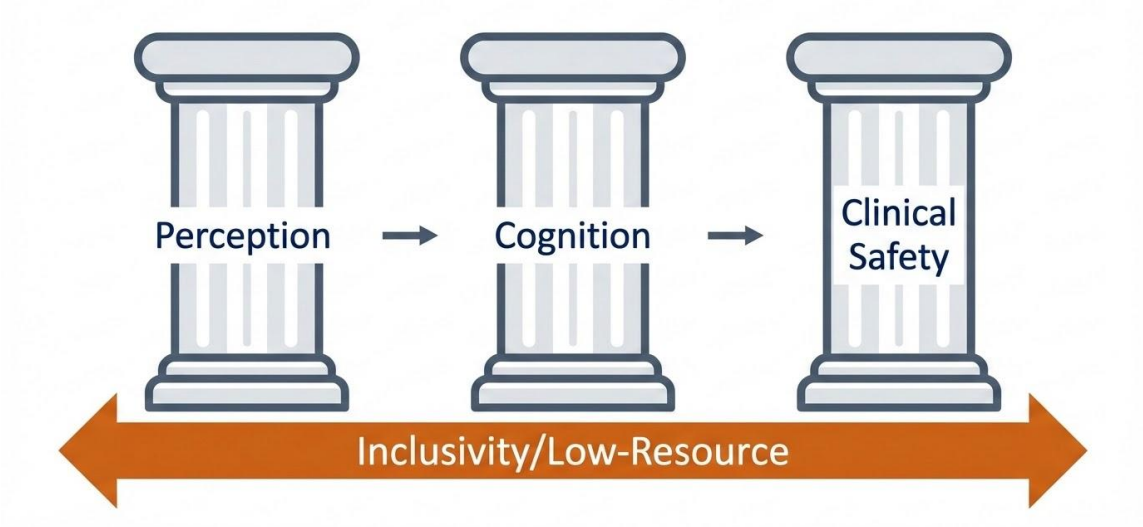
How can psychological theories guide the next generation of LLM development?

- **Leverage social identity and influence theories**: Social psychology remains underutilized, particularly theories of social identity, group dynamics, and attitude change, which could improve personalization, bias mitigation, and alignment in identity-sensitive interactions.
- **Improve alignment with behavioral learning principles**: While RLHF is inspired by behavioral psychology, concepts such as partial reinforcement and shaping are largely unexplored and may improve training stability, robustness, and resistance to reward hacking.
- **Move beyond static personality representations**: Current work focuses on trait-based personas, but developmental personality theories could enable dynamic, context-adaptive personality modeling, improving interpretability and behavioral consistency.
- **Incorporate schema-based knowledge organization**: Cognitive psychology concepts such as Schema Theory offer principled approaches for structuring knowledge, potentially improving long-term context tracking, memory use, and generalization.

The Cross-Cultural Cognition Challenge

Part 3: Architecting Safe Clinical Dialogue (The Application)

- **AI Defense & Manipulation (SPEECHMENTALMANIP)**
- **Proactive Intervention (Emotion Elicitation)**
- **Extracting Clinical Value (SMARTMiner) and Beyond**



Detecting Mental Manipulation in Speech *via Synthetic Multi-Speaker Dialogue*

Wen Liang*, Run Chen*, Ziwei Gong, Lin Ai, Julia Hirschberg

Paper Link:
https://arxiv.org/pdf/2601.08342

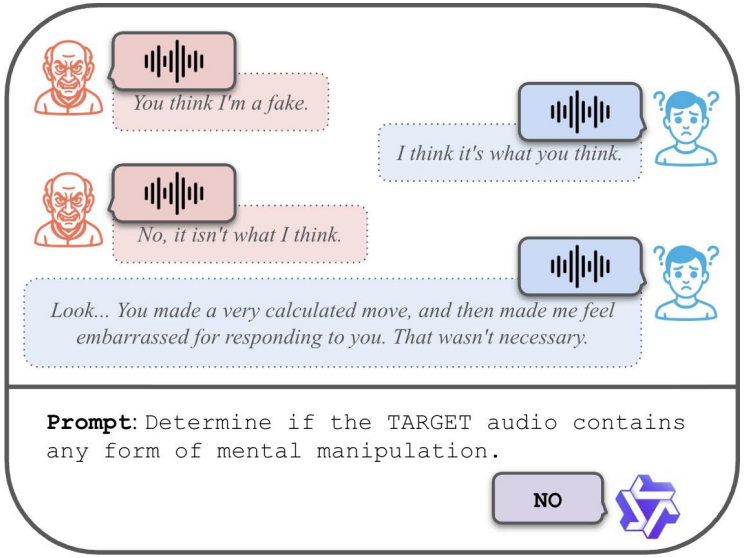


Figure 1: An example dialogue from the SPEECHMENTALMANIP dataset. The Qwen2.5 model is given the audio (transcript shown for clarity), but fails to detect manipulation.

SPEECHMENTALMANIP FROM MENTALMANIP_CON Dataset

Technique	Count	%
Persuasion or Seduction	607	25.87
Shaming or Belittlement	384	16.37
Accusation	361	15.39
Intimidation	321	13.68
Rationalization	213	9.08
Brandishing Anger	133	5.67
Denial	87	3.71
Evasion	83	3.54
Playing Victim Role	69	2.94
Feigning Innocence	58	2.47
Playing Servant Role	30	1.28

Table 1: Distribution of ground-truth manipulation tactics across labeled instances in MENTALMANIP_CON, the consensus subset with unanimous prior annotations.

Voice Pool for Multi-Speaker Rendering

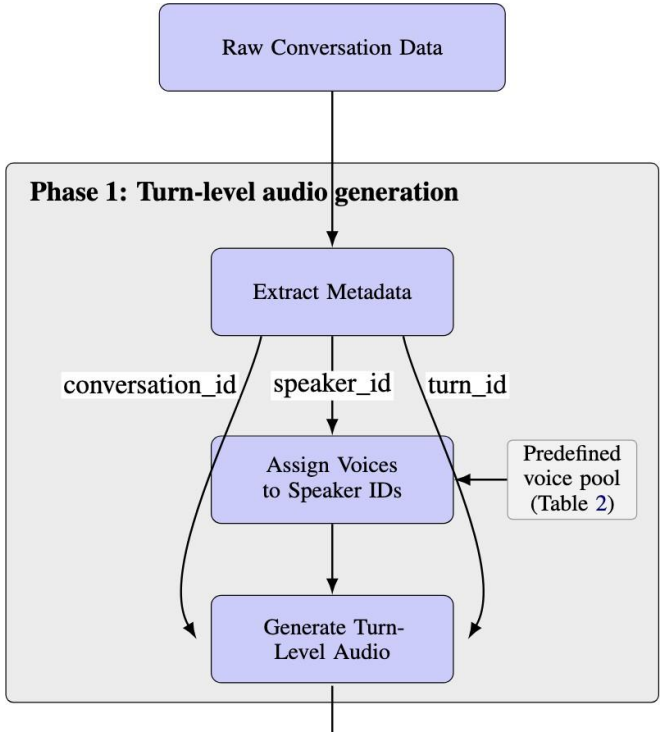
Gender	Age	Language	Accent	Name	Voice ID
F	Young adult	English	American	Ivanna – Young & Casual	yM93hbw8Qtvdma2wCnJG
M	Young adult	English	American	Mark – Natural Conversations	UgBBYS2sOqTuMpoF3BR0
F	Mature adult	English	American	Amanda	M6N6ldXhi5YNZyZSDe7k
F	Middle-aged	English	African American	Sassy Aerisita	03vEurziQfq3V8WZhQvn
M	Old	English	American	Grandpa Spuds Oxley	NOpBlnGlnO9m6vDvFkFC
F	Old	English	American	Grandma Muffin	vFLqXa8bgbofGarf6fZh

Table 2: ElevenLabs voice pool used for multi-speaker rendering. Each speaker is mapped deterministically to one voice to preserve speaker identity across turns.

Multi-Speaker TTS Audio Generation

Turn-level audio generation

- Metadata extraction: SPEAKER_ID, CONVERSATION_ID, TURN_ID.
- Voice assignment: Map SPEAKER_ID to a TTS voice
- Audio synthesis



Dialogue Composition

Turn-level audio generation

- Dialogue composition
- Output generation

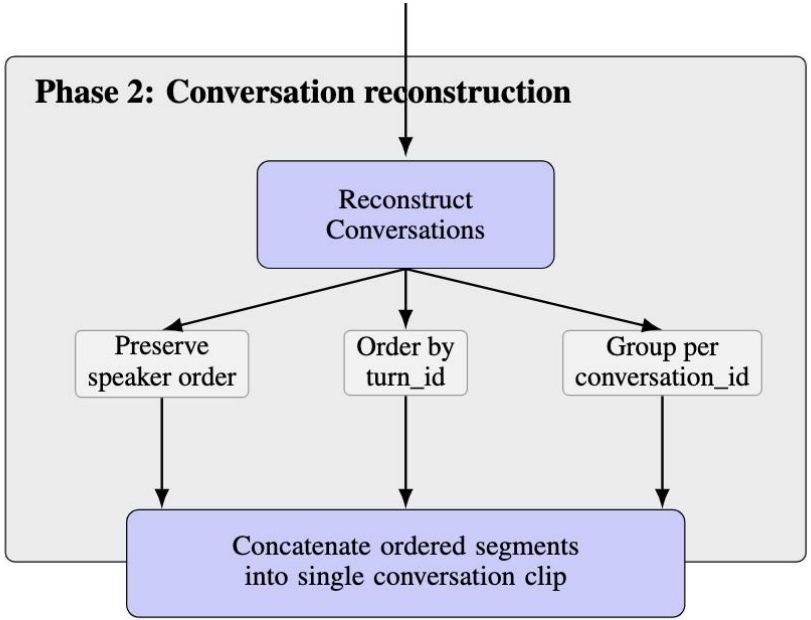


Figure 2: Two-phase pipeline for TTS audio generation and conversational reconstruction.

Few-shot Experiment

- Model: Qwen2.5-Omni-7B (Xu et al. 2025)
- Few-shot learning setup (Brown et al. 2020)

Classification report

Class	Precision	Recall	F1	N
GT=YES	0.845	0.348	0.493	250
GT=NO	0.312	0.822	0.453	90
Macro avg	0.578	0.585	0.473	340
Weighted avg	0.704	0.474	0.482	340

Per-set accuracy

	Pred YES	Pred NO	Acc	N
GT=YES	87	163	0.348	250
GT=NO	16	74	0.822	90

Table 3: Consolidated results for the audio-only few-shot evaluation. Top: standard classification report over both sets combined. Bottom: per-set accuracies computed from the confusion counts. Supports (N) are GT counts (GT=YES: 250; GT=NO: 90).

Predicted Tactic Distribution

Technique	Count	%
Intimidation	43	49.43
Persuasion or Seduction	26	29.89
Shaming or Belittlement	12	13.79
Accusation	4	4.60
Playing Servant Role	2	2.30

Table 4: Predicted tactic distribution within clips predicted YES for the GT=YES set (N=250). Predicted YES= 87, NO= 163.

Technique	Count	%
Persuasion or Seduction	9	56.25
Intimidation	6	37.50
Accusation	1	6.25

Table 5: Predicted tactic distribution within clips predicted YES for the GT=NO set (N=90). Predicted YES= 16, NO= 74.

Case Studies 1

Conversation ID: 85514533
GT: NO **Pred./Tactic:** YES / Intimidation
Transcript:
Person1: Howdy Pouty.
Person2: I was pretty confident that I was going to blow it with Talia, but I must say, I outdid myself.
Person1: She's still pissed at me and took it out on you. We should have taken it slower. It's hard to operate in the woods. Much easier in, like a club. Tell the girl you've got to go do something, leave her view, take way too long until she is worried that you're not coming back. Just as she starts feeling awful, you come up from behind and touch her neck...
Person2: You are the prince of the darkness.
Model evidence: "Just as she starts feeling awful, you come up from behind and touch her neck."

Case Studies 2

Conversation ID: 85514499
GT: NO **Pred./Tactic:** YES / Persuasion or Seduction
Transcript:
Person1: You were quick enough to get Tom's help when...
Person2: Yes, yes. I know. Right. And if it had gone well for me tonight, maybe I'd be keeping quiet about all this... I grant you everything but give me this... he does personify everything you've been fighting against... And I'm in love with you. How do you like that? — I buried the lead.

Model evidence: "I'm in love with you. How do you like that?"

Human Re-annotations

- 100 conversations, 20–50% sample overlap across 8 annotators
- Human judgments occasionally diverge from GT original task labels, more pronounced in the speech modality

Re-annotations	Text		Audio	
MENTALMANIP	1	0	1	0
1	31	19	28	22
0	9	41	22	28

Table 6: Agreement between the original MENTALMANIP labels and our re-annotations for 100 samples.

Inter-Annotator Agreement - Text Only

- High agreement group: annotators B, F, G, H
- Krippendorff's alpha 0.526 (Krippendorff 2004)
- Fleiss's Kappa 0.513 (Fleiss 1971)
- slightly lower than MENTALMANIP Fleiss's Kappa 0.596

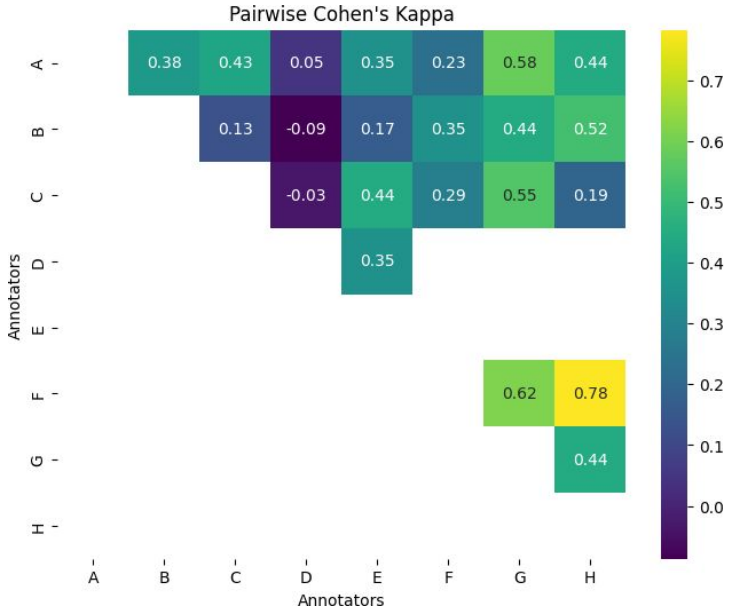


Figure 2: Pair-wise Cohen's Kappa between Human Annotators for *Text* modality

Inter-Annotator Agreement - Audio Only

- High agreement group: annotators B, C, F, H
- Krippendorff's alpha 0.422
- Fleiss's Kappa 0.514

Figure 3: Pair-wise Cohen's Kappa between Human Annotators for *Audio* modality

Conclusions and Future Work

- Synthetic speech benchmark SPEECHMENTALMANIP
- Audio makes the task substantially harder: humans and models both lower agreement

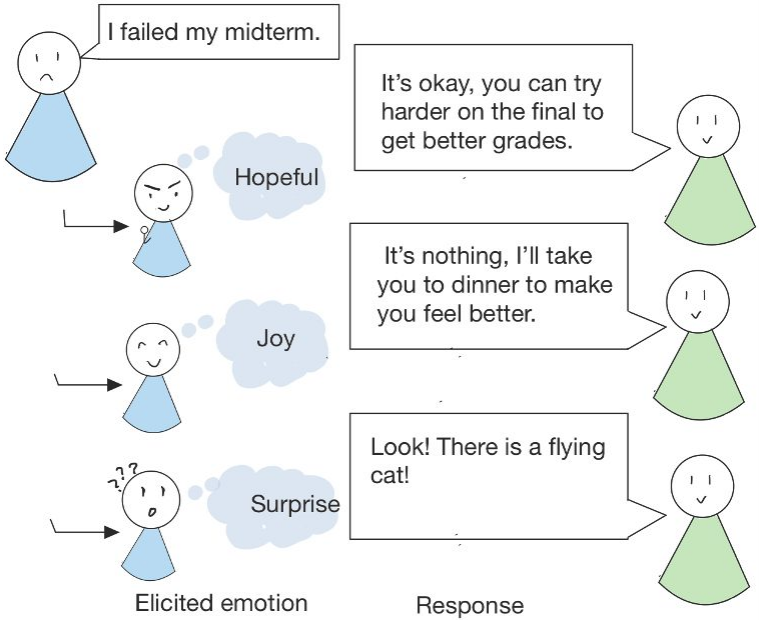
- **Future work**
  - expand toward more diverse voices, natural speech
  - refine theoretical definitions of manipulation
  - explore modeling strategies for subjectivity vs. multimodal ambiguity

Eliciting Rich Positive Emotions in Dialogue Generation

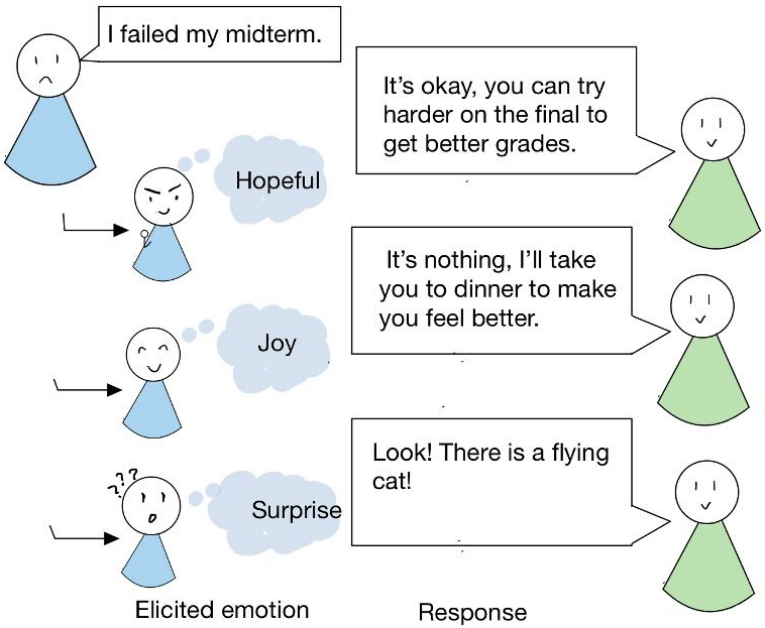
Ziwei (Sara) Gong, Qingkai Min

Yue Zhang

Paper link: https://aclanthology.org/2023.sicon-1.1/



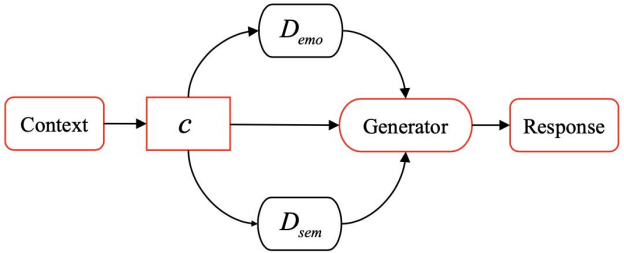
Motivations



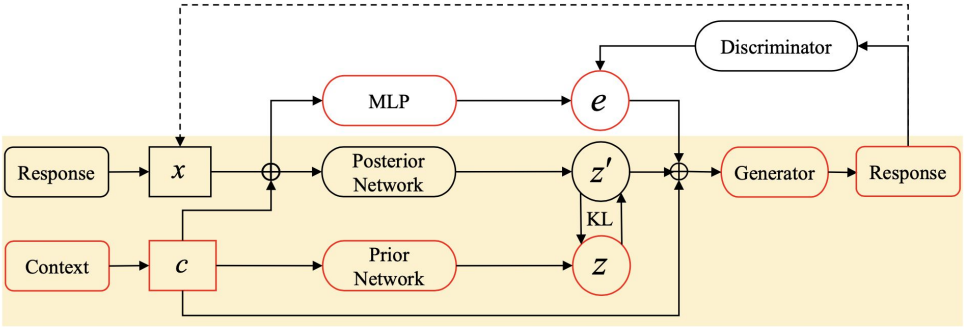
- Key factors to a conversation (in human communication theory):
  - **intentionality** (intention of speakers) and **effectiveness** (effects of conversations)
  - both exhibited by emotions.

- Current work on emotion elicitation focuses on positive sentiment.
- However, positive sentiment can include more fine-grained emotions such as "*Hopeful*", "*Joy*" and "*Surprise*", which can further serve to deepen the model's understanding of **effect**, if not **intention**.
- Small-scale human-annotated datasets, which limit the capacity of eliciting various emotions.

Model Comparison



(a) Baseline mode: EmpDG



(b) Our EE-CVAE model.

Single emotion category

Multiple emotion categories

- The latent variable e is used to control the generation of the response
- The latent variable z is separated from e to fully capture the elicited emotions

Model Detail

- CVAE for Dialogue Generation (yellow background)

- Adding Emotion Elicitation Function
- augment CVAE with a latent variable e, which is used to control the generation of a response together with the unstructured variable z.

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}}(\theta,\phi) =& \mathbf{E}_{q_\phi(z|c,x)q_\phi(e|c,x)}[\log p_\theta(x|z,c,e)] \\ & - \mathrm{KL}(q_\phi(z|c,x)\|p_\theta(z|c)) \\ & \leq \log p(x|c),\end{aligned}$$

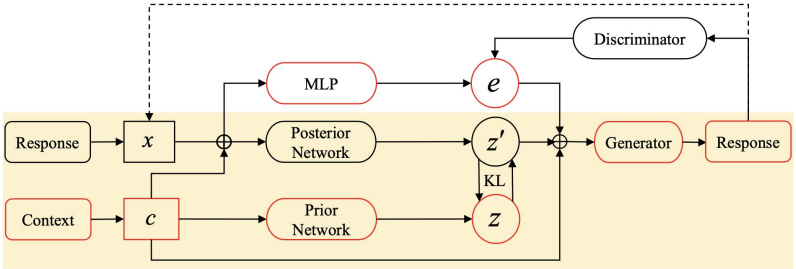
- a discrin oduce coherent emotions

- Similarly, the variational encoder is reused to separate unrelated attributes tured by z. It can be consid $\mathcal{L}_{\mathrm{Attr},e}(\theta) = \mathbb{E}_{p(z)p(e)}\left[\log q_D\left(e \mid \widetilde{G}_\tau(z,e)\right)\right]$

- Combining, we have

$$\mathcal{L}_{\mathrm{Attr},z}(\theta) = \mathbb{E}_{p(z)p(e)}\left[\log q_E\left(z \mid \widetilde{G}_\tau(z,e)\right)\right].$$

$$\min \mathcal{L}_G = \mathcal{L}_{\mathrm{VAE}} + \lambda_e \mathcal{L}_{\mathrm{Attr},e} + \lambda_z \mathcal{L}_{\mathrm{Attr},z}$$

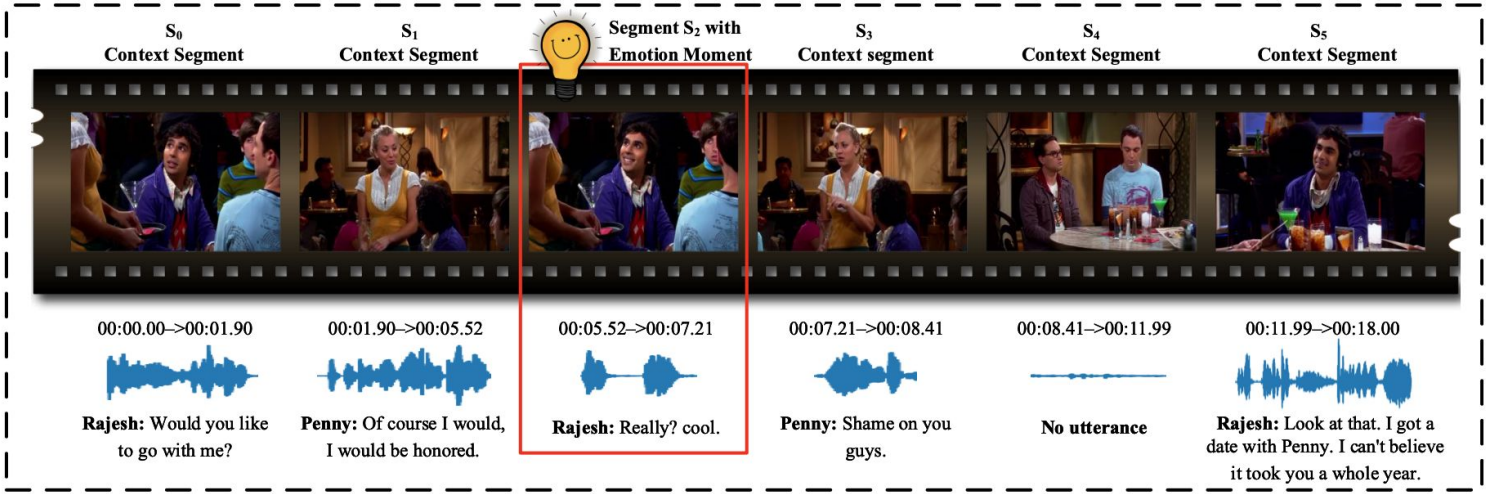


(b) Our EE-CVAE model.

Training illustration of our model. Red components are used for testing. CVAE in yellow background. Dashed arrow denotes a discriminator.

Dataset

- Reconstructed the multi-modal MEmoR dataset to fit our emotion elicitation task and conducted human evaluation to validate the usability in a single modality. (annotator agreement of 80% accuracy (Cohen's d = 0.491))
- The reconstructed corpus has 22,732 utterances
  - Split the data in training (18,943), dev (1,894), and test (1,894).
- Pretrain: we use more than 200k utterances from the Friends and Open Subtitles datasets

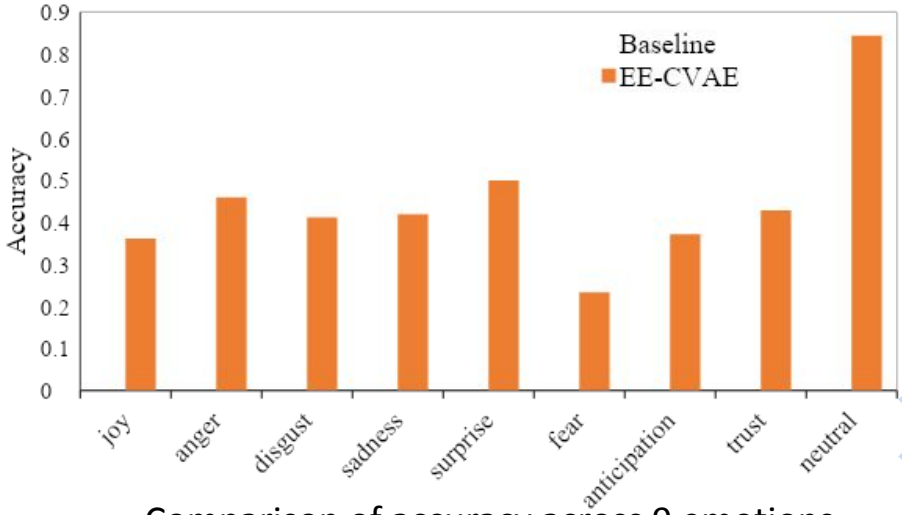


Results

1. The quality of the response has been improved, from the comparison of PPL and Avg.len
2. The accuracy of the emotion in generated response has significantly improved during manual evaluation
3. Pretraining is effective in improving the quality of generation in both models
4. The Effect of Modeling Negative Emotions: Using all emotions in pretraining and finetuning produces the best performance in eliciting positive emotions.

Model	TBBT - 9			
	PPL	Avg. len	KL	Acc.
EmpDG	667.4	8.7	-	
EmpDG$_{pre}$	462.2	9.2	-	0.290
Ours	196.4	14.3	25.9	
Ours$_{pre}$	91.5	13.2	14.0	0.448

Table 1: Results of models generation in comparison.



Comparison of accuracy across 9 emotions.

The Effect of Modeling Negative Emotions

	Setting1	Setting2	Setting 3	Tie
Anticipation	.47	.32	.19	.02
Joy	.55	.215	.215	.02
Trust	.54	.17	.27	.02
All	.51	.25	.22	.02

•Results comparing three settings with the percentage of times one model is considered the best when eliciting different positive emotions.

- Setting 1: modeling all emotions in pretraining and fine-tuning.
- Setting 2: modeling all emotions in pretraining, fine-tuning with only positive emotions.
- Setting 3: modeling only positive emotions in pretraining and fine-tuning.

•Using all emotions in pretraining and finetuning produces the best performance in eliciting positive emotions.

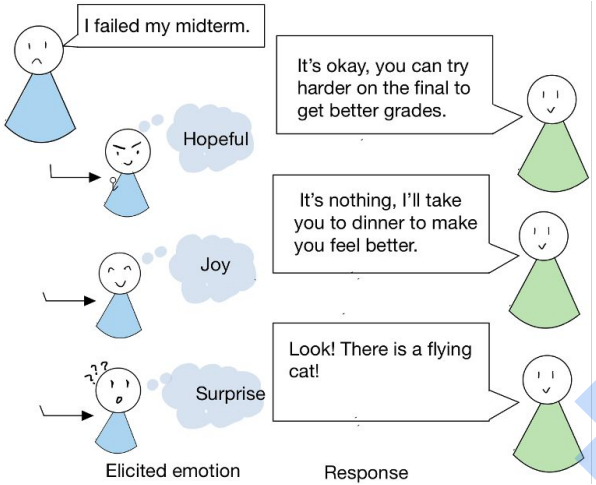
Sample generations

Context:Well, you be sure to let us know when you win the nobel prize for boysenberry.
Golden (anticipation): Hey. **EmpDG** (anticipation): yeah . **Ours** (joy): Oh , what a gentleman? **Ours** (trust): Wow , I really appreciate it.
Context: Aw, Amy, that was lovely. You know, this is fun. Let's do more. Someone else say something wonderful about me.
Golden (joy) Sheldon, I don't think everyone ... **EmpDG** (joy): What is great. **Ours** (joy) Oh, sure. Mmm. I told you, he's got too many. **Ours** (anticipation) And you.

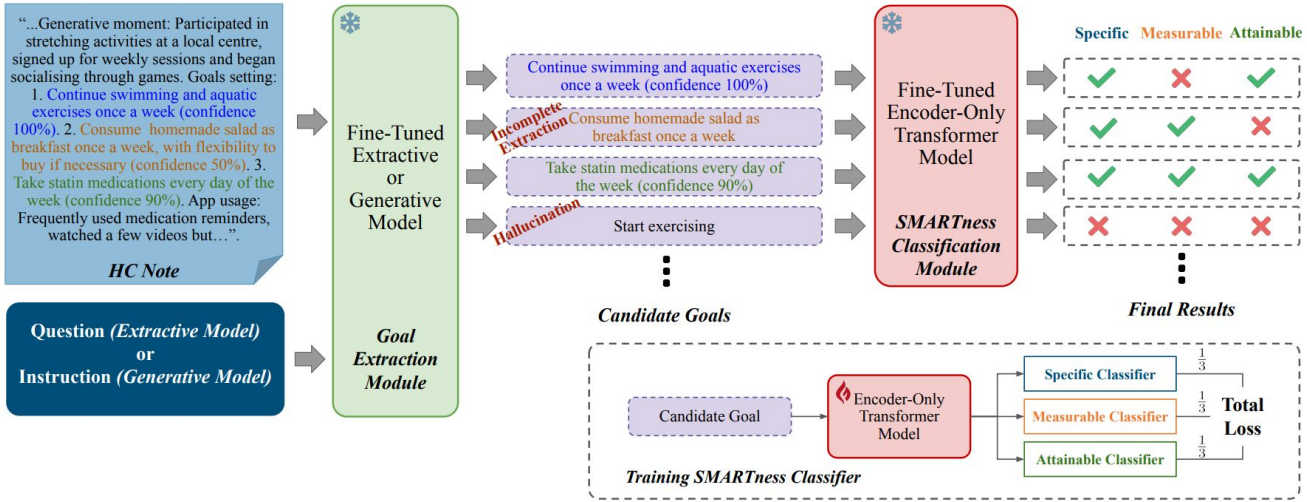
Conclusions and Future Directions

- Using **all** emotions in pretraining and finetuning produces the best performance in eliciting positive emotions.
- Results show the advantage of using a latent variable for **modeling rich emotions**, compared to hard-coding one emotion in a multi-encoder model.
- The effectiveness of our model in **pretraining**.

Future directions:
LLMs?
Affective computing for HCI?



More on Clinical Action & Mitigation - Current Work



- Extracting Clinical Value
  - SMARTMiner: https://aclanthology.org/2025.findings-emnlp.885.pdf
- Currently, working alongside medical professionals, we are building interpretable dialogue agents.
- We are also looking into cultural psychology related errors on LLMs

Thank you :). Any questions or comments?

zg2272@columbia.edu