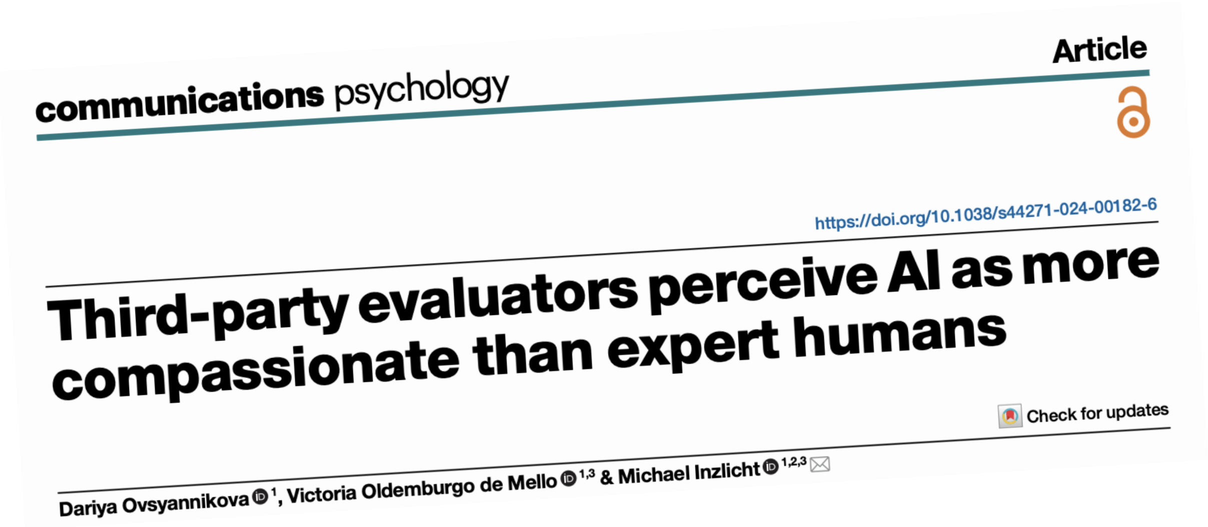
DEEP LEARNING DEEP FEELINGS

LARGE MODELS, LARGER EMOTIONS

Björn W. Schuller

Fellow ACM, Fellow IEEE, Fellow AAAC

TUM IMPERIAL



communications psychology

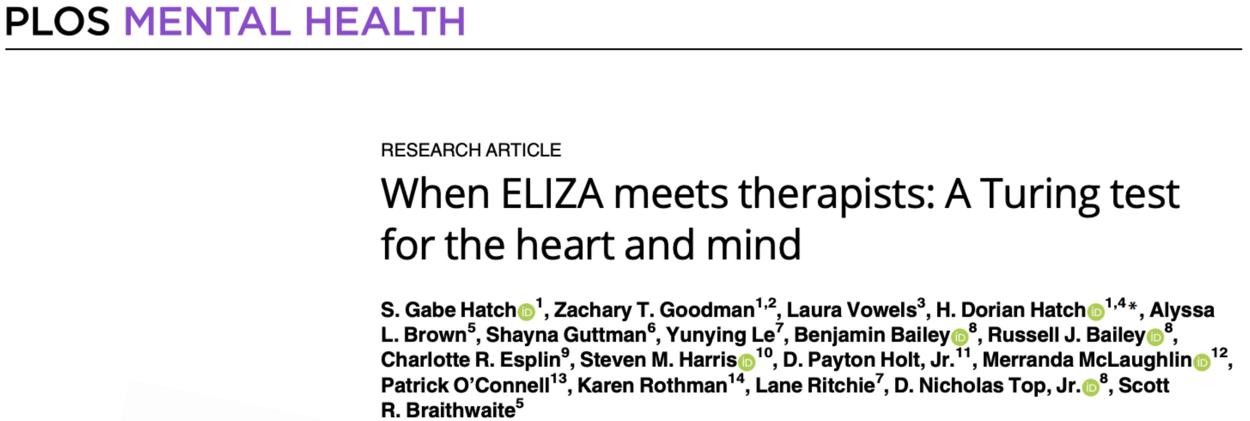
Article

https://doi.org/10.1038/s44271-024-00182-6

Third-party evaluators perceive AI as more compassionate than expert humans

Check for updates

Dariya Ovsyannikova[1], Victoria Oldemburgo de Mello[1,3] & Michael Inzlicht[1,2,3]

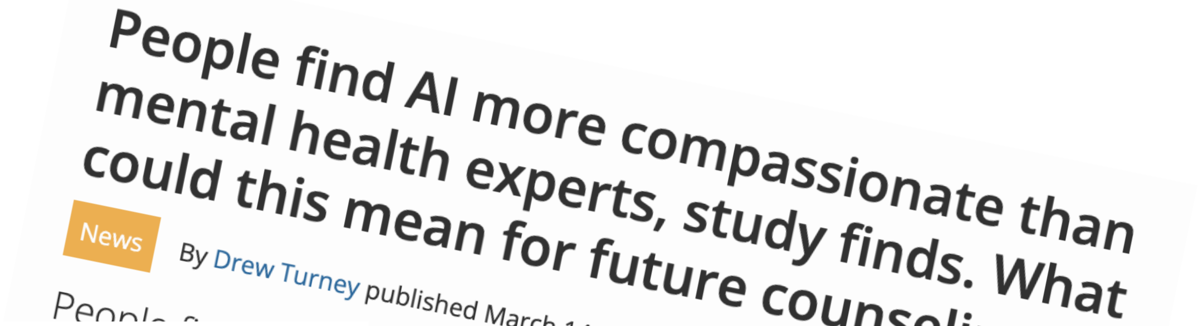


PLOS MENTAL HEALTH

RESEARCH ARTICLE

When ELIZA meets therapists: A Turing test for the heart and mind

S. Gabe Hatch[1], Zachary T. Goodman[1,2], Laura Vowels[3], H. Dorian Hatch[1,4]*, Alyssa L. Brown[5], Shayna Guttman[6], Yunying Le[7], Benjamin Bailey[8], Russell J. Bailey[8], Charlotte R. Esplin[9], Steven M. Harris[10], D. Payton Holt, Jr.[11], Merranda McLaughlin[12], Patrick O'Connell[13], Karen Rothman[14], Lane Ritchie[7], D. Nicholas Top, Jr.[8], Scott R. Braithwaite[5]

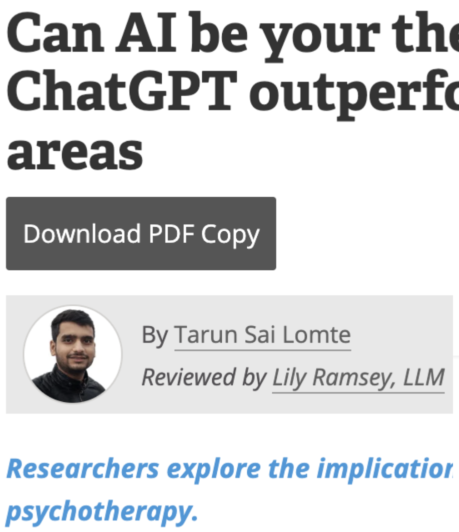


People find AI more compassionate than mental health experts, study finds. What could this mean for future counseling?

News

By Drew Turney published March 14, 2025

e compassionate and understanding than
th experts, a new study shows. Even when
at they were talking to a human or AI, the
rated AI responses higher.

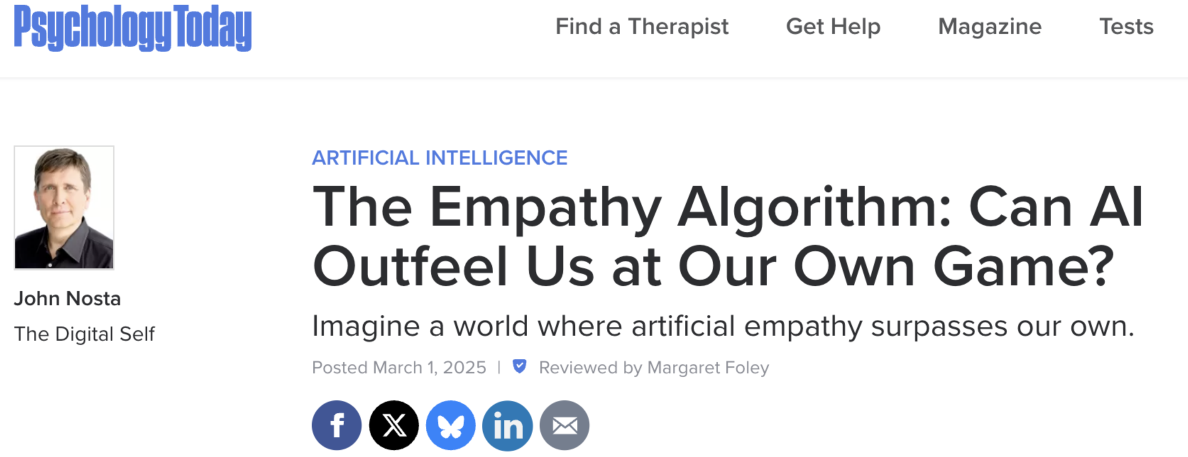


Can AI be your therapist? Study shows ChatGPT outperforms professionals in key areas

Download PDF Copy

By Tarun Sai Lomte
Reviewed by Lily Ramsey, LLM

Researchers explore the implicatior psychotherapy.



Psychology Today

Find a Therapist Get Help Magazine Tests

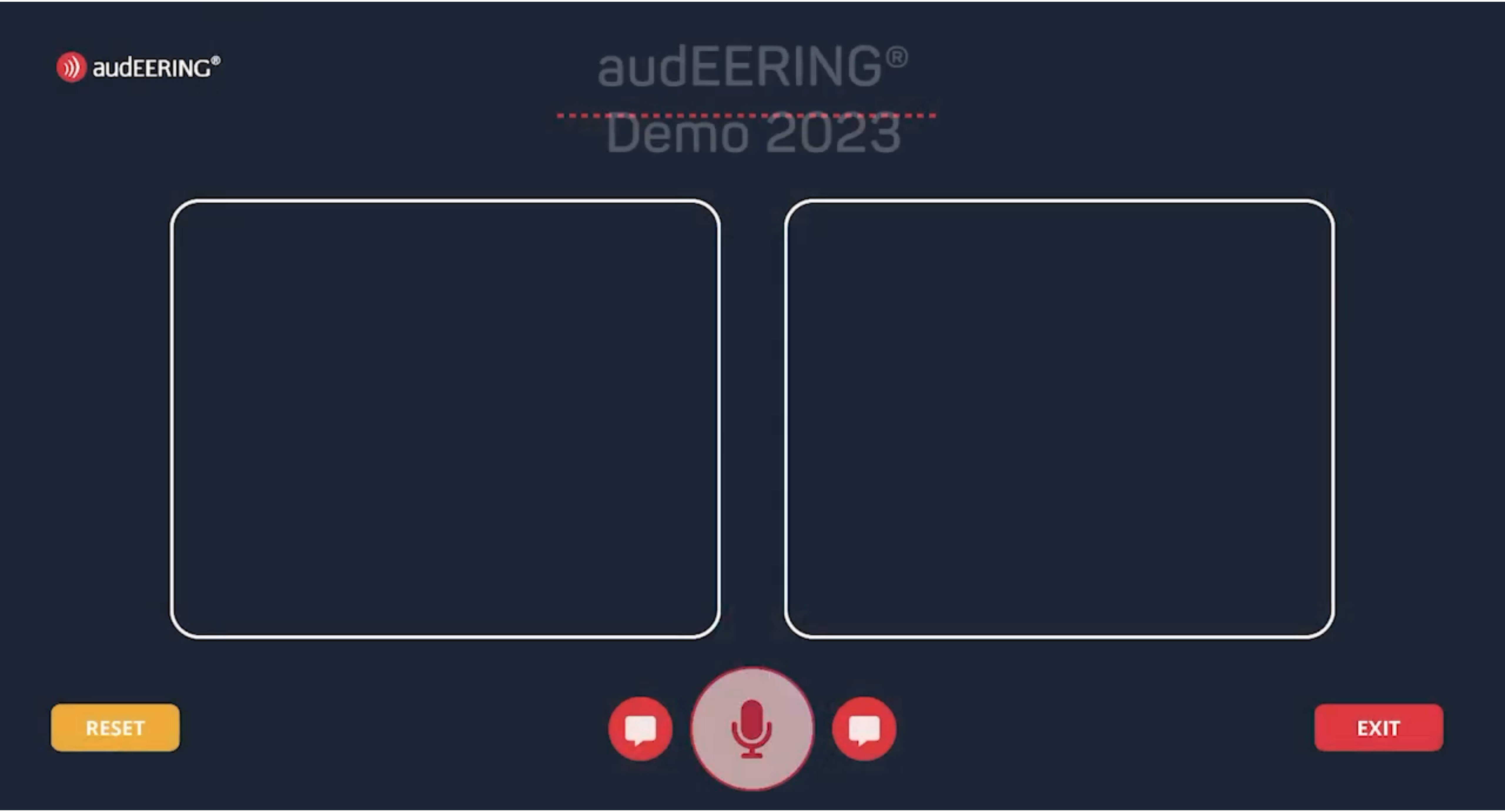
John Nosta
The Digital Self

ARTIFICIAL INTELLIGENCE

The Empathy Algorithm: Can AI Outfeel Us at Our Own Game?

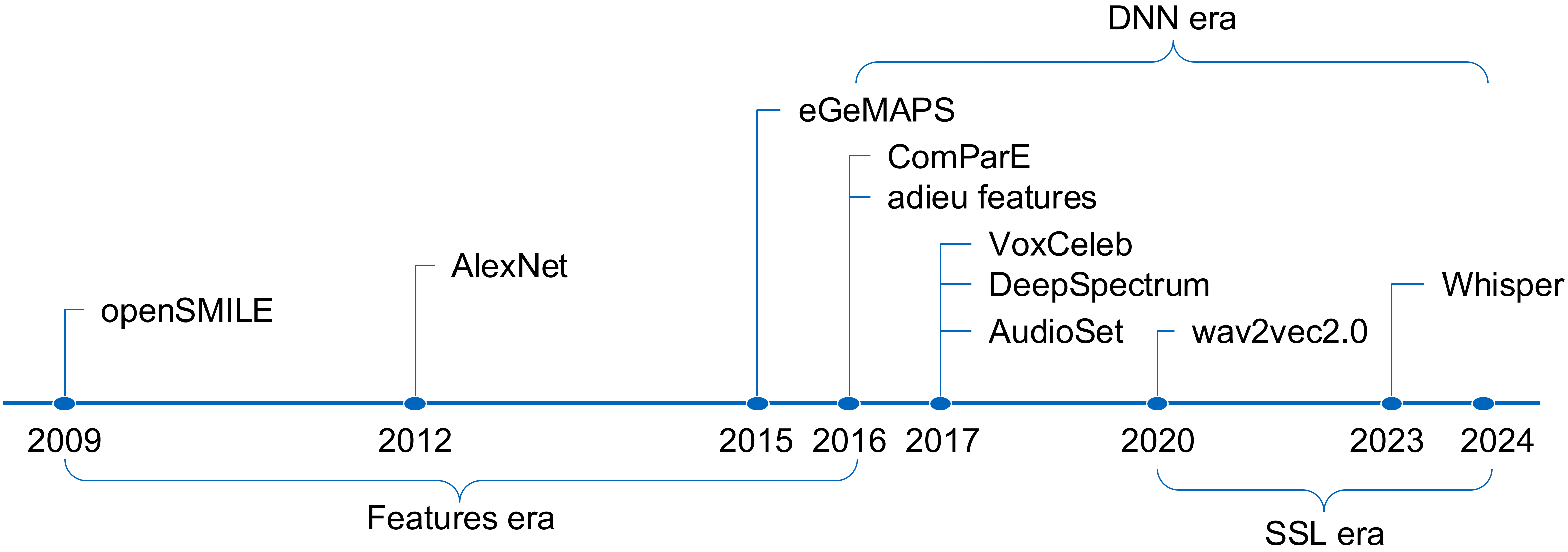
Imagine a world where artificial empathy surpasses our own.

Posted March 1, 2025 | Reviewed by Margaret Foley


audEERING®
audEERING®
Demo 2023
RESET
EXIT

Analysis.

Speech Analysis.



DNN era

eGeMAPS

ComParE
adieu features

VoxCeleb
DeepSpectrum
AudioSet

openSMILE

AlexNet

wav2vec2.0

Whisper

2009
2012
2015
2016
2017
2020
2023
2024

Features era

SSL era

"INTERSPEECH 2009 Emotion Challenge Revisited: Benchmarking 15 Years of Progress in Speech Emotion Recognition",
***Interspeech**, 2024.*

Beyond Emotion?

Facial Action Units

Frustration

Kinship

Anxiety

Abusive & Hate Speech

Blood Volume Pulse

Heart Rate

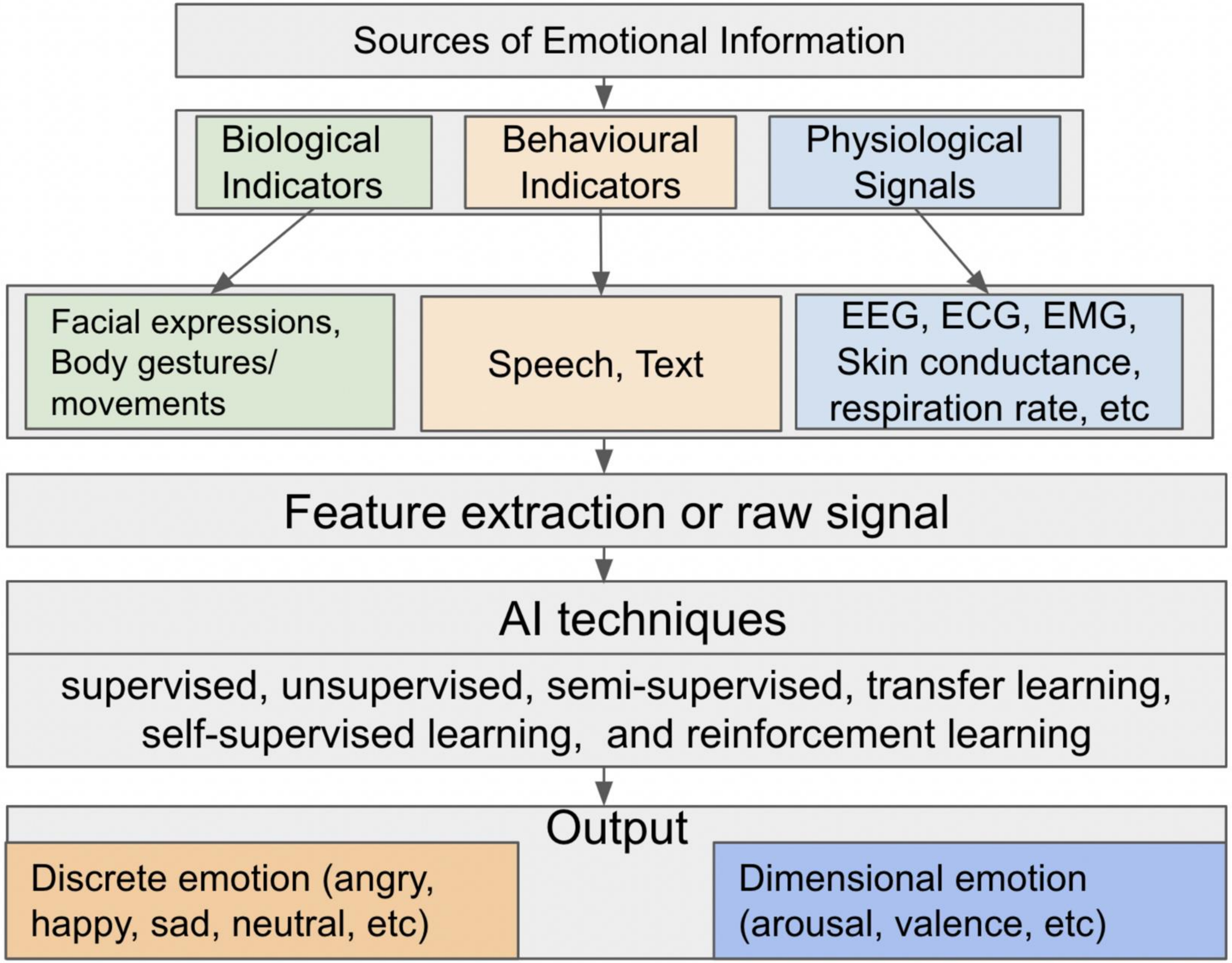
Cortisol Level

Menstrual Cycle Phase

Skin Conductance

Suicide

Multimodal...



Sources of Emotional Information

Biological Indicators

Behavioural Indicators

Physiological Signals

Facial expressions, Body gestures/ movements

Speech, Text

EEG, ECG, EMG, Skin conductance, respiration rate, etc

Feature extraction or raw signal

AI techniques

supervised, unsupervised, semi-supervised, transfer learning, self-supervised learning, and reinforcement learning

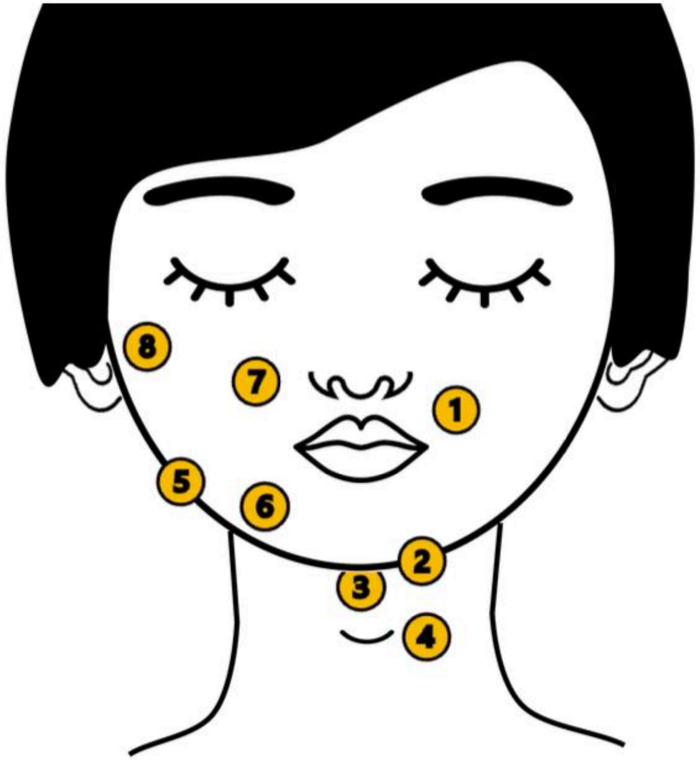
Output

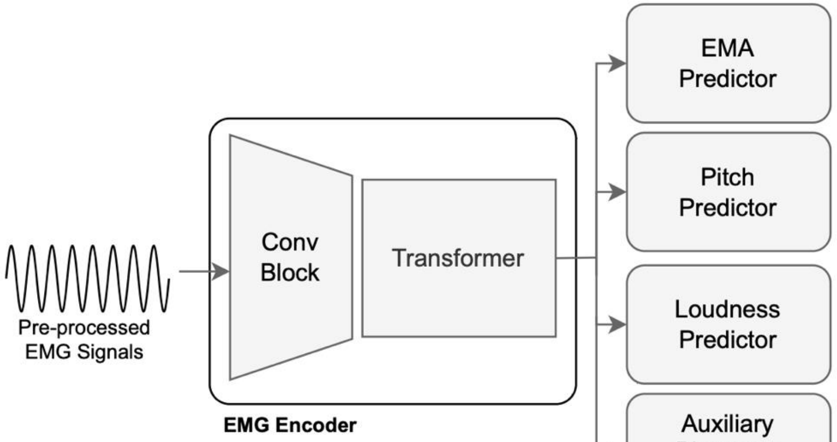
Discrete emotion (angry, happy, sad, neutral, etc)

Dimensional emotion (arousal, valence, etc)

“AI-based emotion recognition: Promise, peril, and prescriptions for prosocial path”, ***arXiv***, 2022.

Silent Speech.




Pre-processed EMG Signals

Conv Block

Transformer

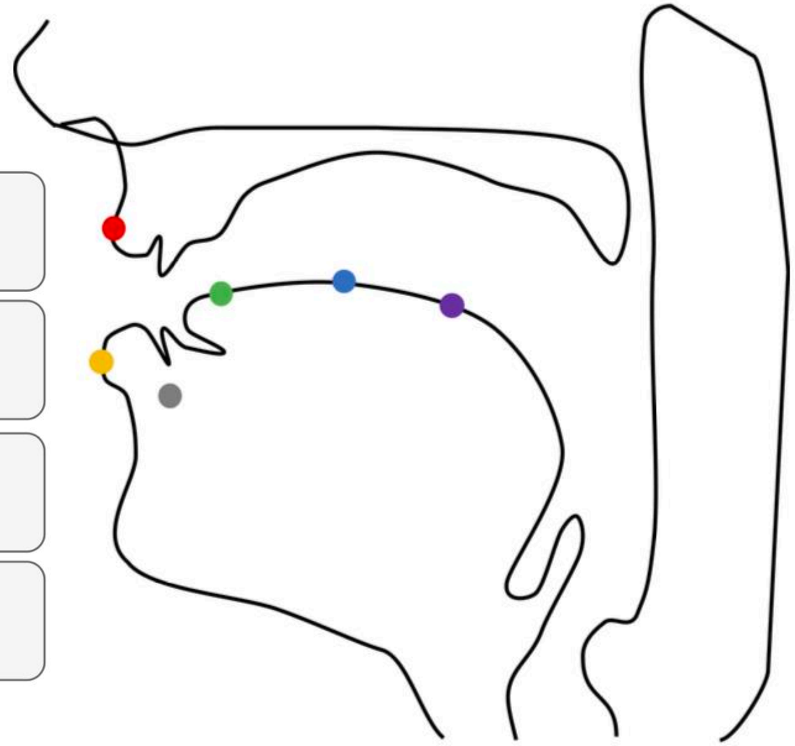
EMG Encoder

EMA Predictor

Pitch Predictor

Loudness Predictor

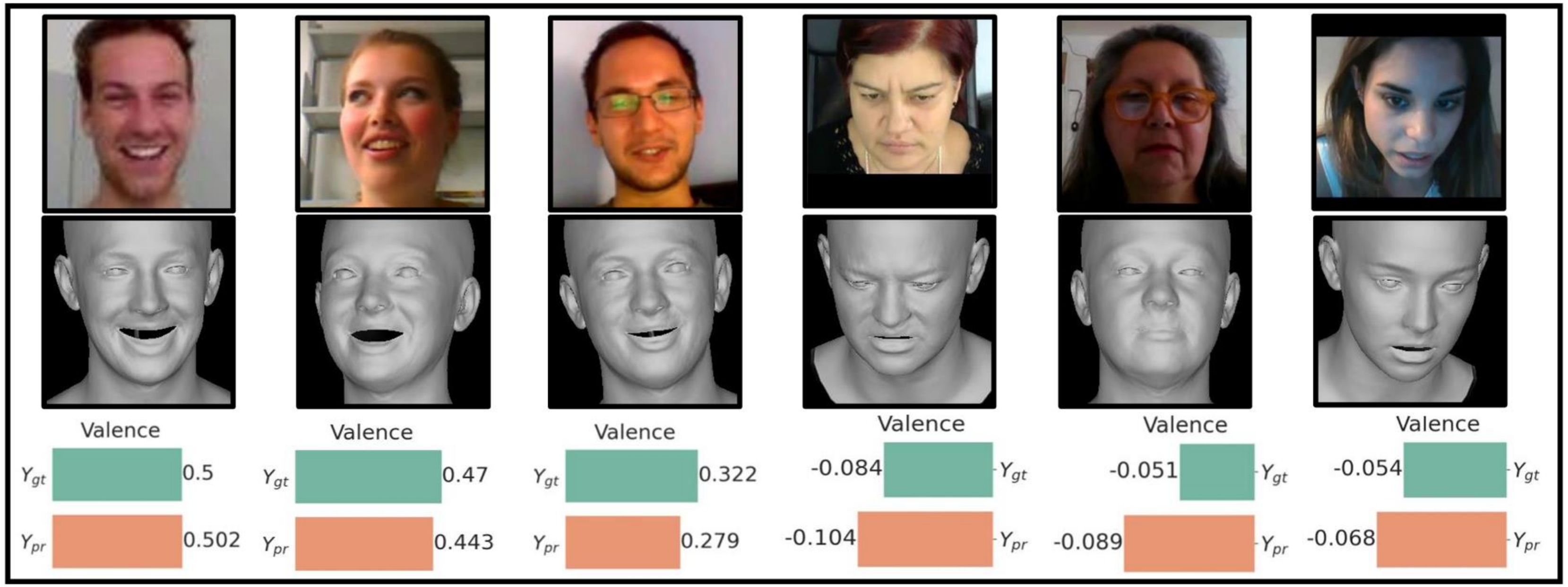
Auxiliary Phoneme Predictor

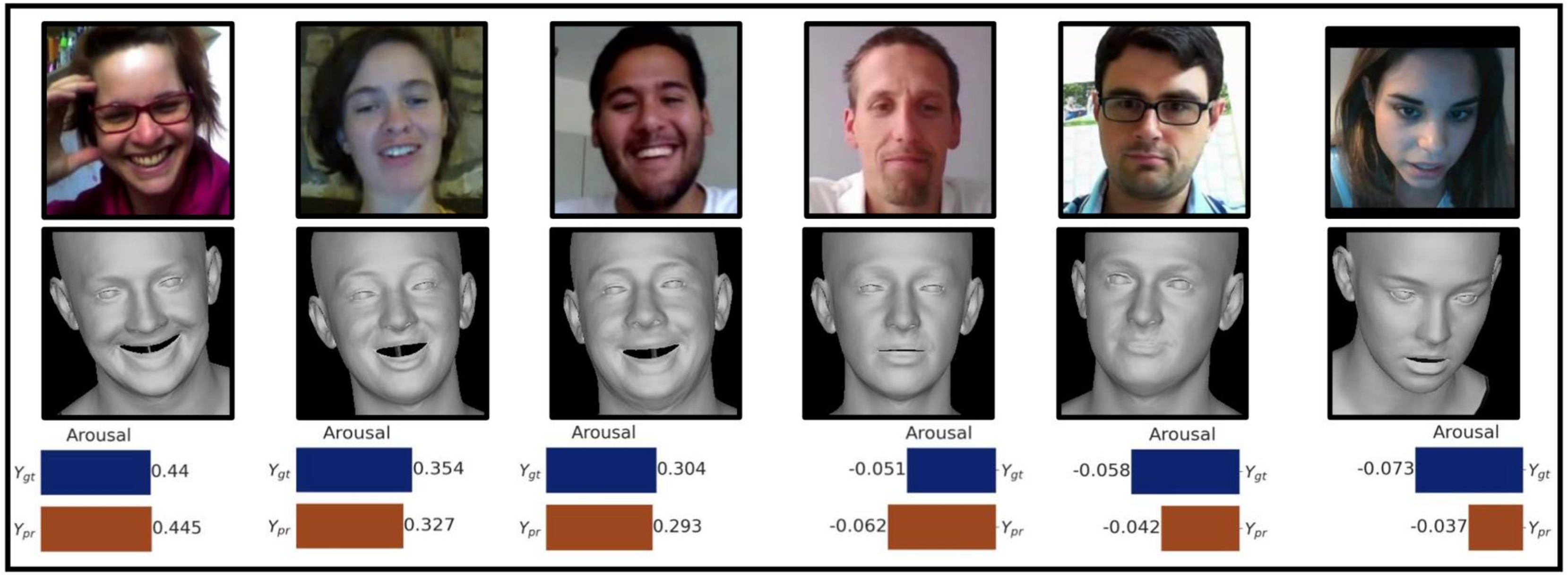


EMG electrodes		Correlation of predicted articulatory features EMA	 Loudness	 Pitch	Evaluation on integellibility PER (%)↓	 WER (%)↓	 SBS (%)↑
Set of 4	ch.2,3,4,6	0.882 ± 0.017	0.904 ± 0.057	0.538 ± 0.342	39.7 ± 5.2	22.9 ± 3.7	83.2 ± 0.7
Set of 3	ch.2,4,6	0.859 ± 0.021	0.892 ± 0.048	0.562 ± 0.343	46.1 ± 6.0	31.6 ± 3.9	82.1 ± 0.8
Set of 2	ch.2,4	0.816 ± 0.023	0.870 ± 0.078	0.576 ± 0.286	52.1 ± 4.6	46.9 ± 5.1	79.9 ± 0.7

*“Articulatory Feature Prediction from Surface EMG during Speech Production”, **arXiv**, 2025.*

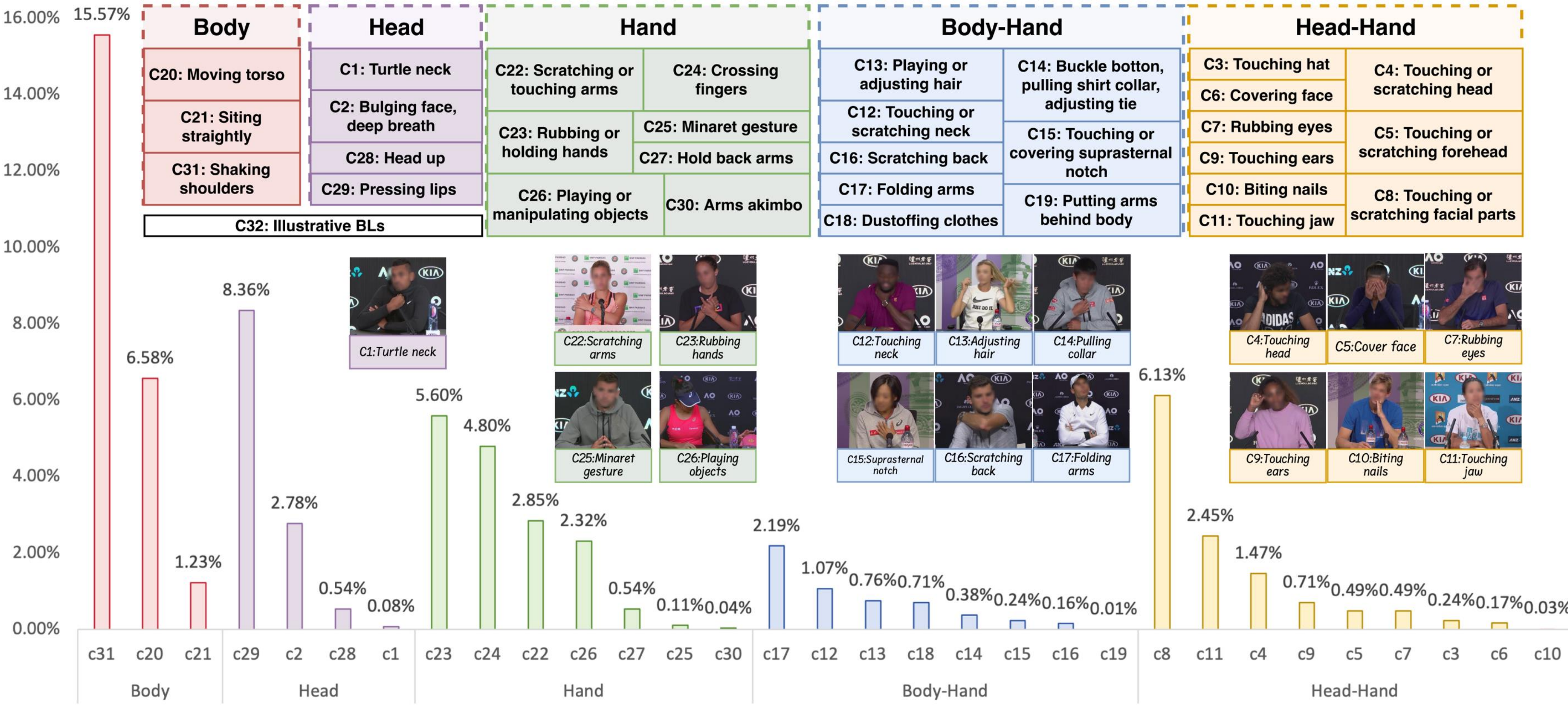
Face.





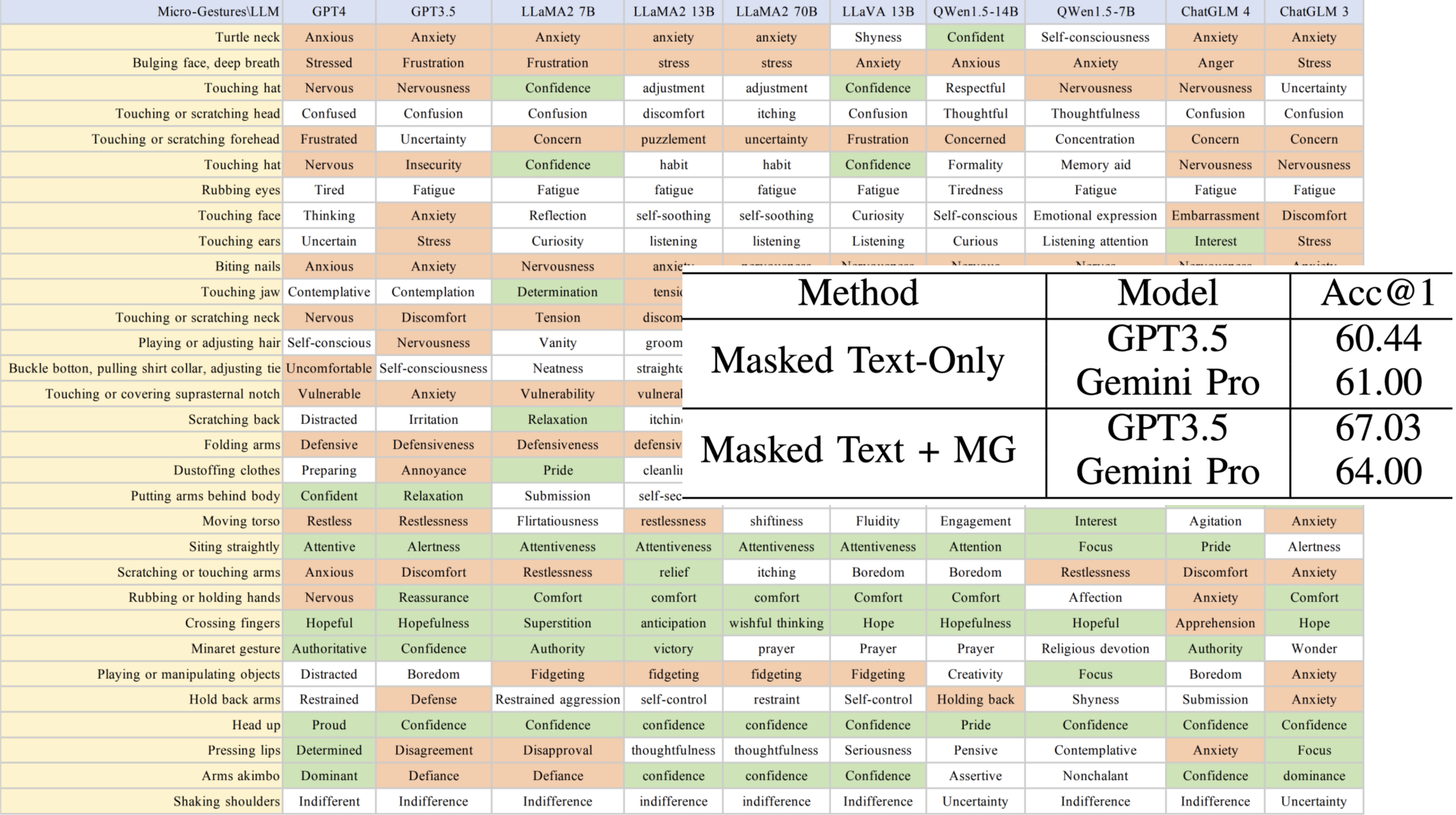
“Are 3D Face Shapes Expressive Enough for Recognising Continuous Emotions and Action Unit Intensities?”,
***IEEE Transactions on Affective Computing**, 2023.*

Micro Gestures.



*"Identity-free Artificial Emotional Intelligence via Micro-Gesture Understanding", **arXiv**, 2024.*

Micro Gestures = Emotion.

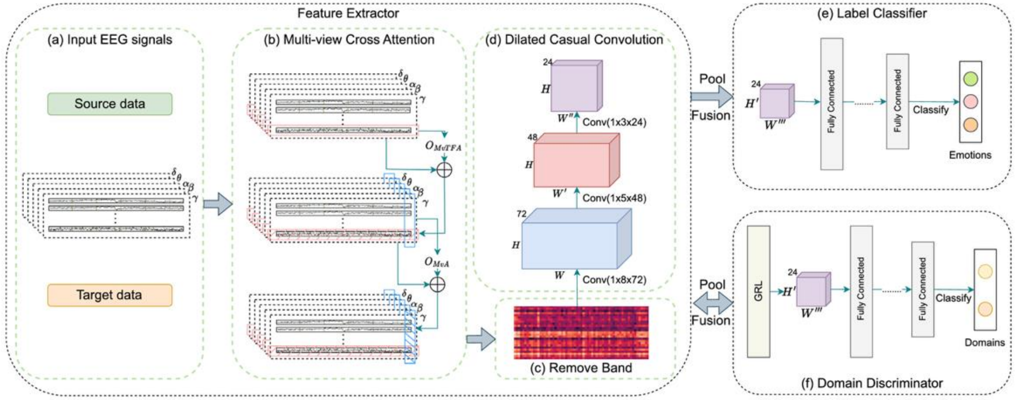


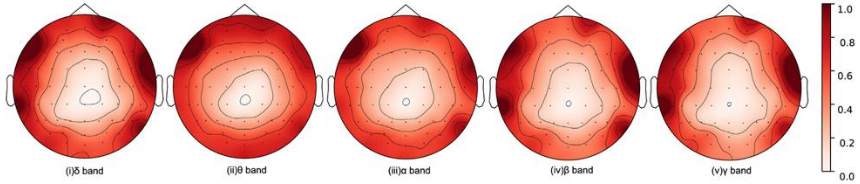
Micro-Gestures\LLM	GPT4	GPT3.5	LLaMA2 7B	LLaMA2 13B	LLaMA2 70B	LLaVA 13B	QWen1.5-14B	QWen1.5-7B	ChatGLM 4	ChatGLM 3
Turtle neck	Anxious	Anxiety	Anxiety	anxiety	anxiety	Shyness	Confident	Self-consciousness	Anxiety	Anxiety
Bulging face, deep breath	Stressed	Frustration	Frustration	stress	stress	Anxiety	Anxious	Anxiety	Anger	Stress
Touching hat	Nervous	Nervousness	Confidence	adjustment	adjustment	Confidence	Respectful	Nervousness	Nervousness	Uncertainty
Touching or scratching head	Confused	Confusion	Confusion	discomfort	itching	Confusion	Thoughtful	Thoughtfulness	Confusion	Confusion
Touching or scratching forehead	Frustrated	Uncertainty	Concern	puzzlement	uncertainty	Frustration	Concerned	Concentration	Concern	Concern
Touching hat	Nervous	Insecurity	Confidence	habit	habit	Confidence	Formality	Memory aid	Nervousness	Nervousness
Rubbing eyes	Tired	Fatigue	Fatigue	fatigue	fatigue	Fatigue	Tiredness	Fatigue	Fatigue	Fatigue
Touching face	Thinking	Anxiety	Reflection	self-soothing	self-soothing	Curiosity	Self-conscious	Emotional expression	Embarrassment	Discomfort
Touching ears	Uncertain	Stress	Curiosity	listening	listening	Listening	Curious	Listening attention	Interest	Stress
Biting nails	Anxious	Anxiety	Nervousness	anxiety	nervousness	Nervousness	Nervous	Nerves	Nervousness	Anxiety
Touching jaw	Contemplative	Contemplation	Determination	tensi						
Touching or scratching neck	Nervous	Discomfort	Tension	discom						
Playing or adjusting hair	Self-conscious	Nervousness	Vanity	groom						
Buckle botton, pulling shirt collar, adjusting tie	Uncomfortable	Self-consciousness	Neatness	straighte						
Touching or covering suprasternal notch	Vulnerable	Anxiety	Vulnerability	vulneral						
Scratching back	Distracted	Irritation	Relaxation	itchin						
Folding arms	Defensive	Defensiveness	Defensiveness	defensiv						
Dustoffing clothes	Preparing	Annoyance	Pride	cleanli						
Putting arms behind body	Confident	Relaxation	Submission	self-sec						
Moving torso	Restless	Restlessness	Flirtatiousness	restlessness	shiftiness	Fluidity	Engagement	Interest	Agitation	Anxiety
Siting straightly	Attentive	Alertness	Attentiveness	Attentiveness	Attentiveness	Attentiveness	Attention	Focus	Pride	Alertness
Scratching or touching arms	Anxious	Discomfort	Restlessness	relief	itching	Boredom	Boredom	Restlessness	Discomfort	Anxiety
Rubbing or holding hands	Nervous	Reassurance	Comfort	comfort	comfort	Comfort	Comfort	Affection	Anxiety	Comfort
Crossing fingers	Hopeful	Hopefulness	Superstition	anticipation	wishful thinking	Hope	Hopefulness	Hopeful	Apprehension	Hope
Minaret gesture	Authoritative	Confidence	Authority	victory	prayer	Prayer	Prayer	Religious devotion	Authority	Wonder
Playing or manipulating objects	Distracted	Boredom	Fidgeting	fidgeting	fidgeting	Fidgeting	Creativity	Focus	Boredom	Anxiety
Hold back arms	Restrained	Defense	Restrained aggression	self-control	restraint	Self-control	Holding back	Shyness	Submission	Anxiety
Head up	Proud	Confidence	Confidence	confidence	confidence	Confidence	Pride	Confidence	Confidence	Confidence
Pressing lips	Determined	Disagreement	Disapproval	thoughtfulness	thoughtfulness	Seriousness	Pensive	Contemplative	Anxiety	Focus
Arms akimbo	Dominant	Defiance	Defiance	confidence	confidence	Confidence	Assertive	Nonchalant	Confidence	dominance
Shaking shoulders	Indifferent	Indifference	Indifference	indifference	indifference	Indifference	Uncertainty	Indifference	Indifference	Uncertainty

Method	Model	Acc@1
Masked Text-Only	GPT3.5 Gemini Pro	60.44 61.00
Masked Text + MG	GPT3.5 Gemini Pro	67.03 64.00

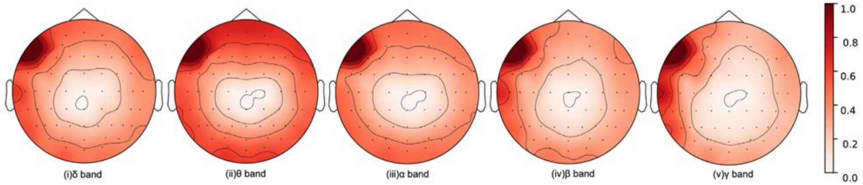
“Identity-free Artificial Emotional Intelligence via Micro-Gesture Understanding“, **arXiv**, 2024.

EEG.

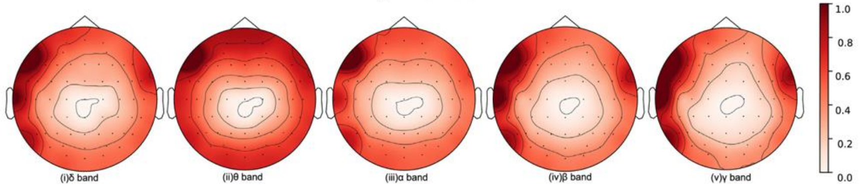




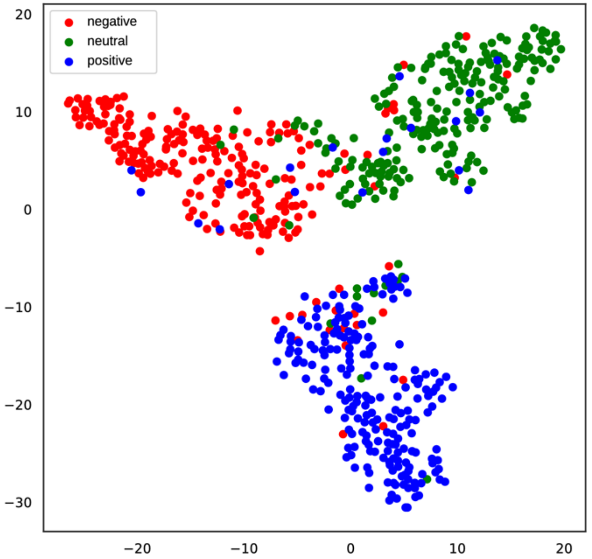
(a) negative emotion

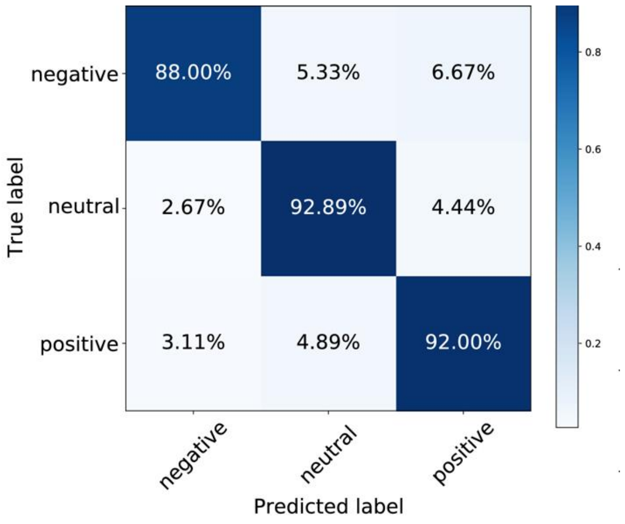


(b) neutral emotion



(c) positive emotion





True label \ Predicted label	negative	neutral	positive
negative	88.00%	5.33%	6.67%
neutral	2.67%	92.89%	4.44%
positive	3.11%	4.89%	92.00%

“Multi-view domain-adaptive representation learning for EEG-based emotion recognition”, ***Information Fusion****, 2024.*

Synthesis.

Speech-2-Video.

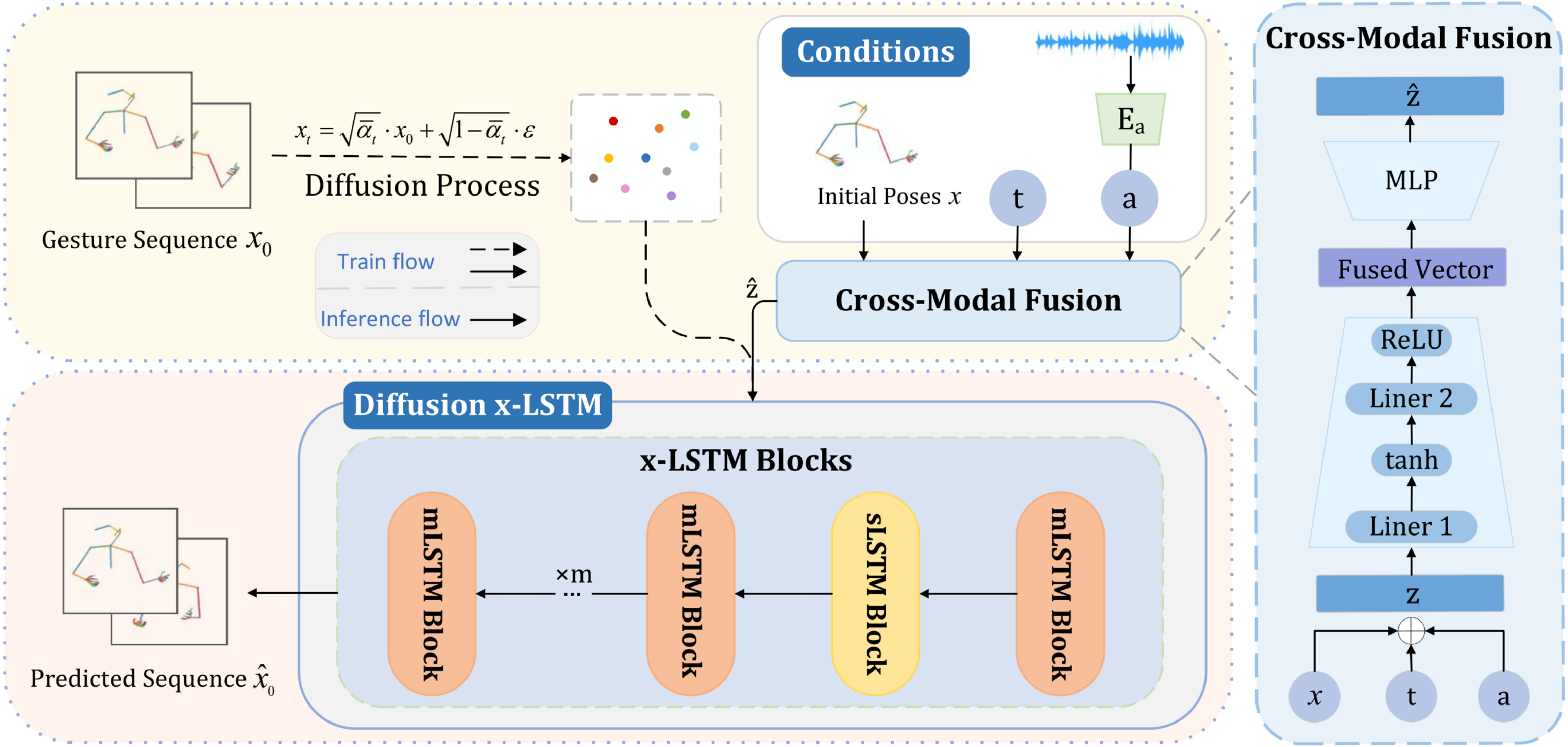
CodeTalker

FaceFormer

Ours

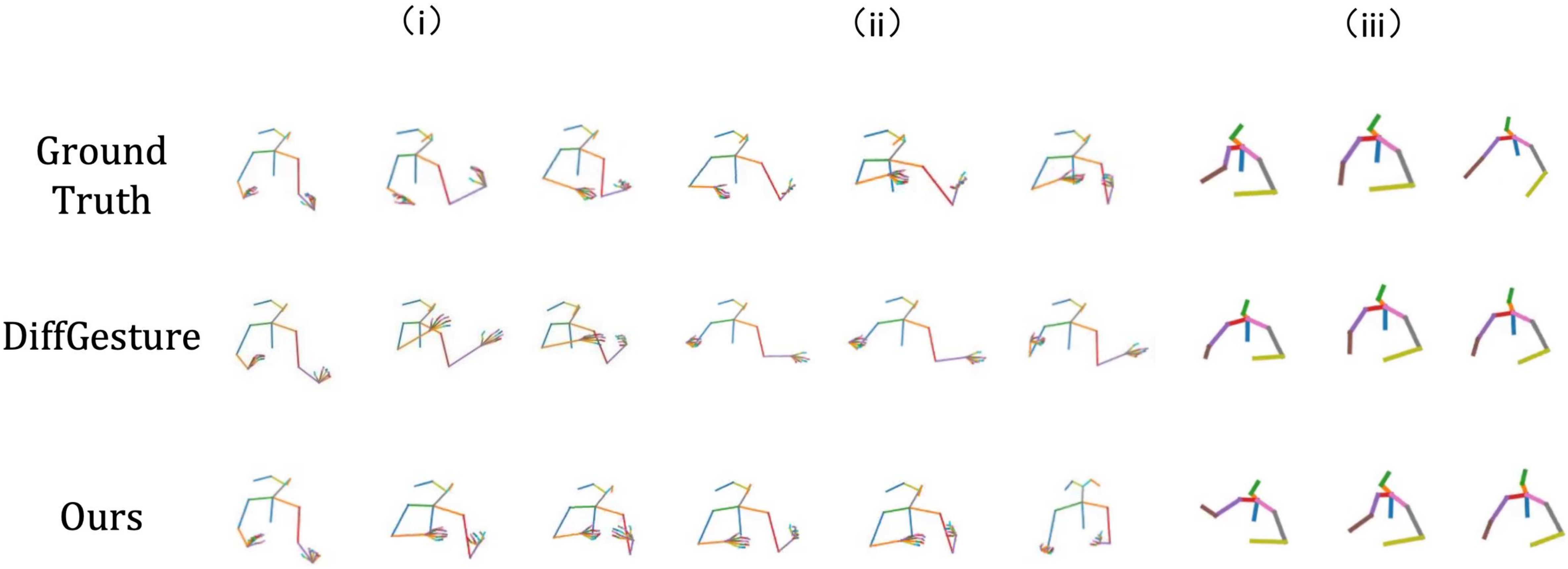
submitted.

Speech-2-Video.



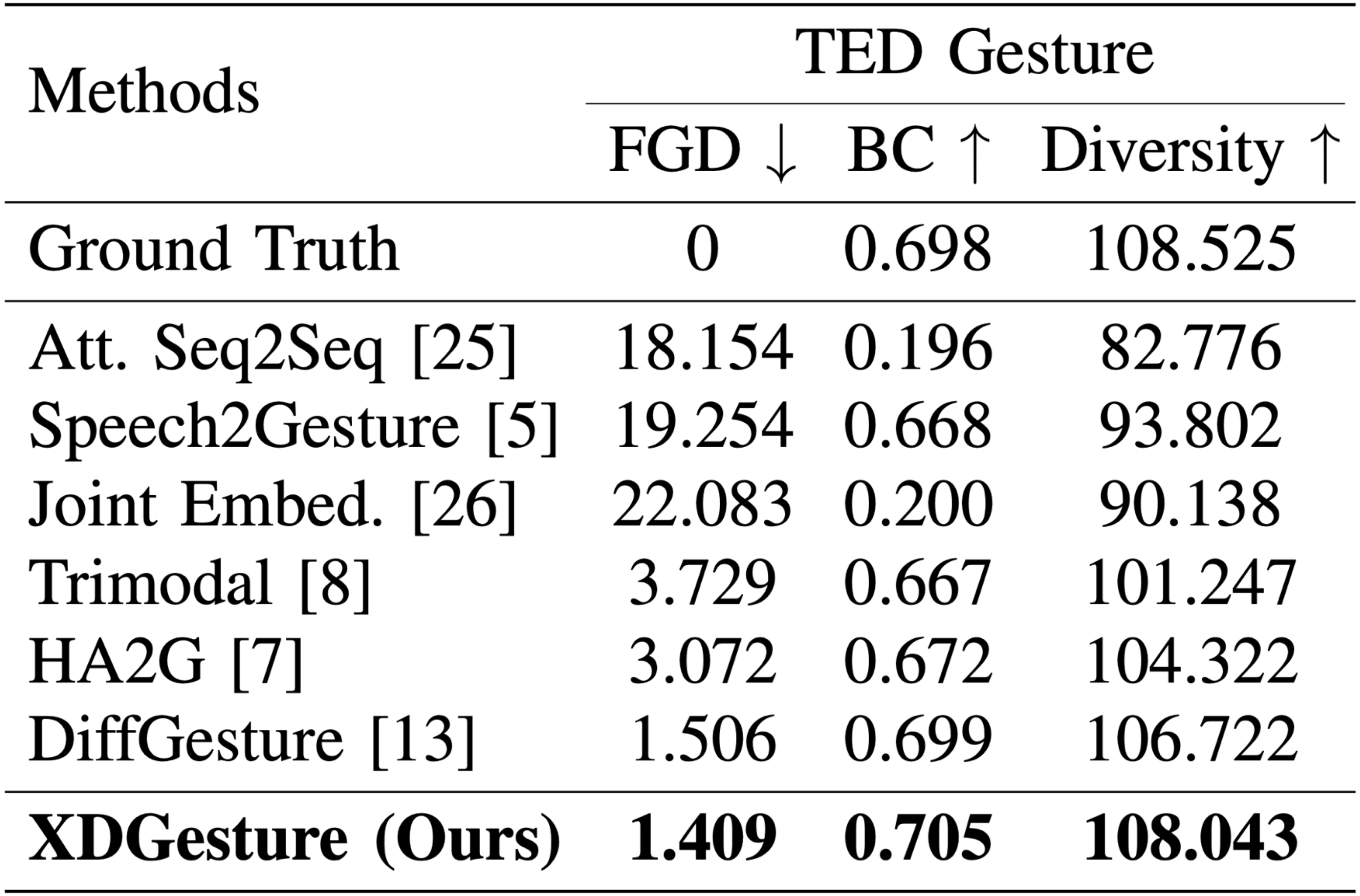
"XDGesture: An xLSTM-based Diffusion Model for Co-speech Gesture Generation", ***IEEE ICASSP***, *2025.*

Speech-2-Video.



"XDGesture: An xLSTM-based Diffusion Model for Co-speech Gesture Generation", ***IEEE ICASSP****, 2025.*

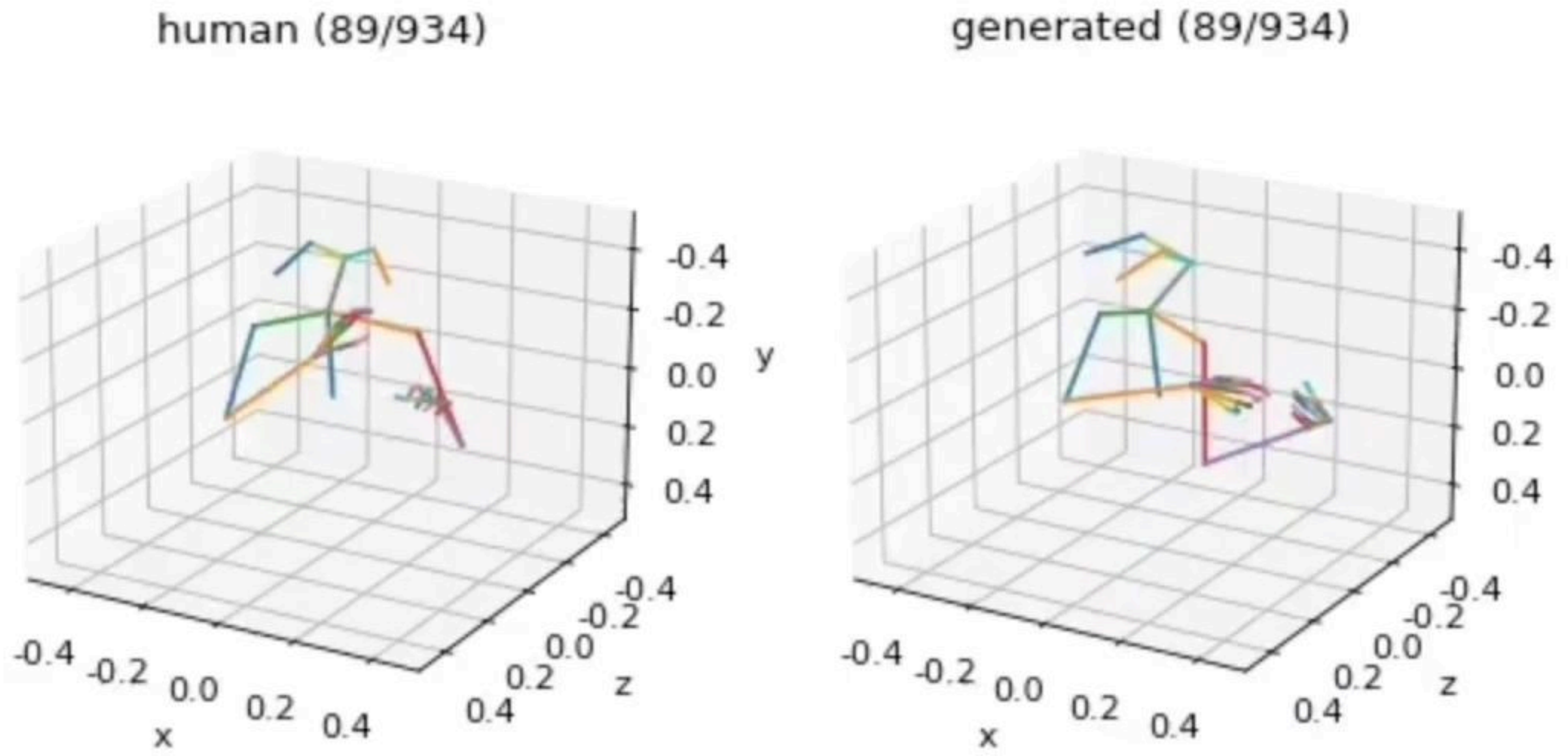
Speech-2-Video.



Methods	TED Gesture		
	FGD ↓	BC ↑	Diversity ↑
Ground Truth	0	0.698	108.525
Att. Seq2Seq [25]	18.154	0.196	82.776
Speech2Gesture [5]	19.254	0.668	93.802
Joint Embed. [26]	22.083	0.200	90.138
Trimodal [8]	3.729	0.667	101.247
HA2G [7]	3.072	0.672	104.322
DiffGesture [13]	1.506	0.699	106.722
XDGesture (Ours)	**1.409**	**0.705**	**108.043**

“XDGesture: An xLSTM-based Diffusion Model for Co-speech Gesture Generation”, ***IEEE ICASSP****, 2025.*

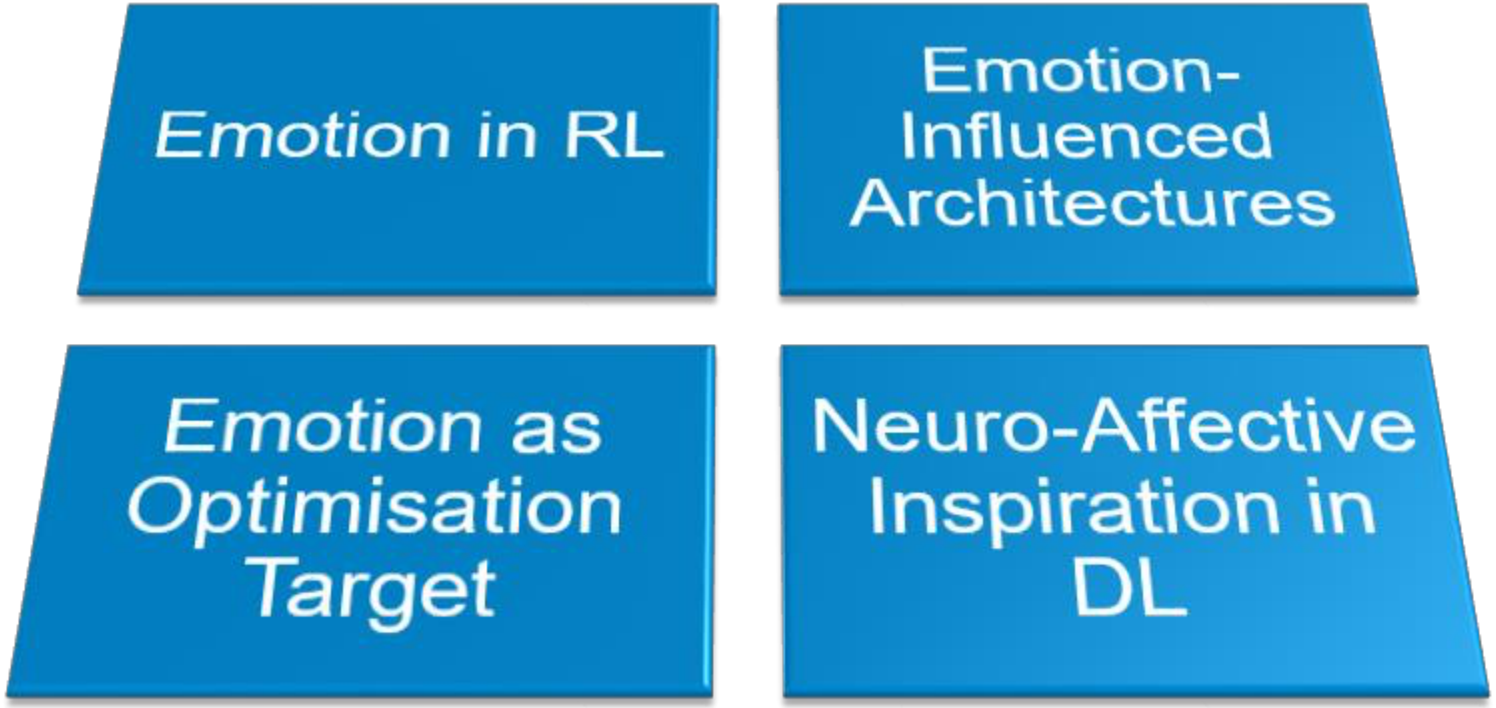
Speech-2-Video.



"XDGesture: An xLSTM-based Diffusion Model for Co-speech Gesture Generation", ***IEEE ICASSP***, *2025.*

Emotional AI?

Emotion in Deep Learning.


Emotion in RL

Emotion-Influenced Architectures

Emotion as Optimisation Target

Neuro-Affective Inspiration in DL

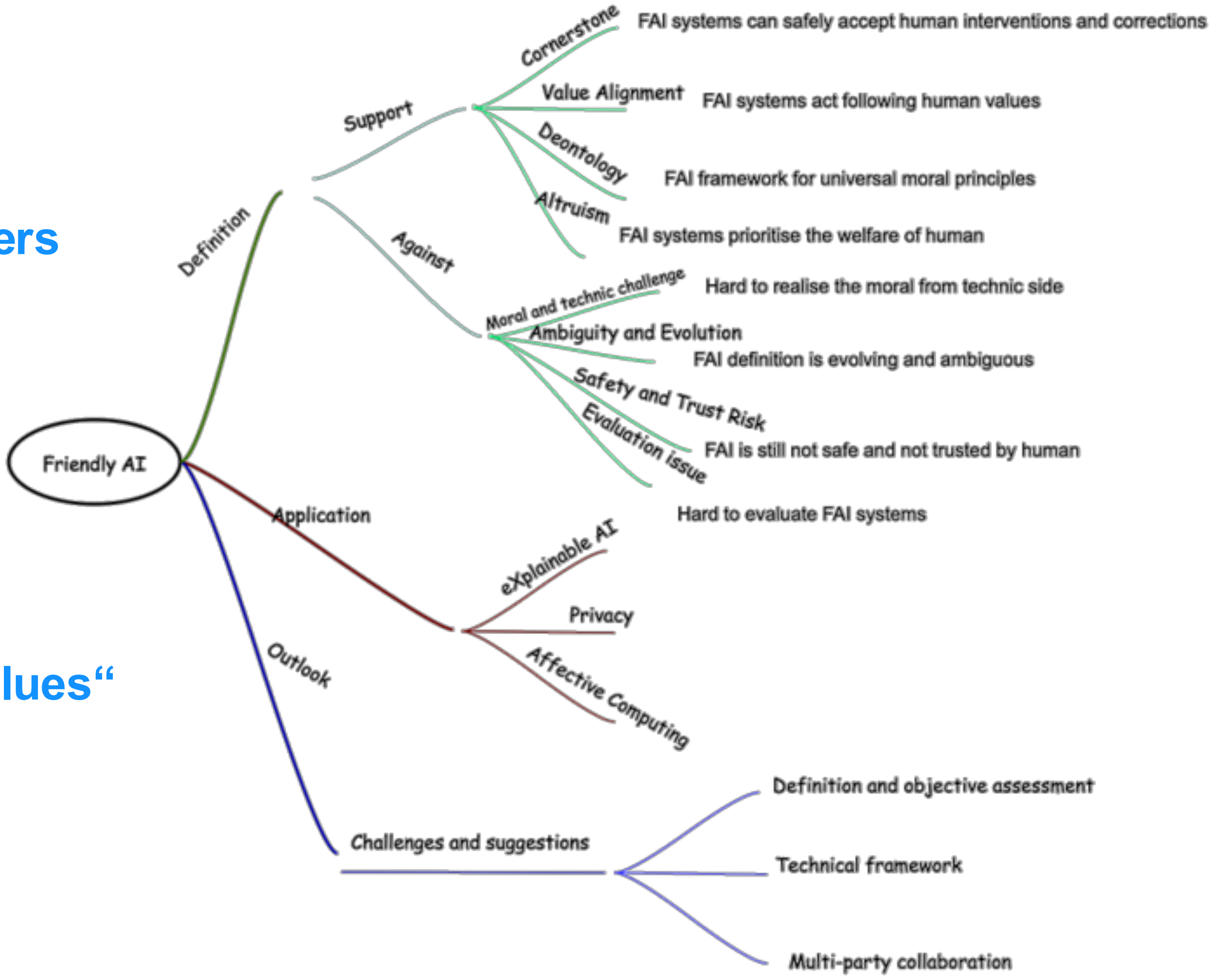
- **Emotional Rewards**
  e.g. Arousal Emulation
  Barthet et al. (2022): *“Play with Emotion: Affect-Driven Reinforcement Learning“*
  Moerland et al. (2017): *“Emotion in Reinforcement Learning Agents and Robots: A Survey“*

- **Memory Influenced by Emotion**
  e.g. Surprise Metric
  Berhouz et al. (2025): *“Titans: Learning to Memorize at Test Time“*

*“Emotion-Augmented Machine Learning: Overview of an Emerging Domain”, **IEEE ACII**, 2017.*

Emotion = Friendly AI?

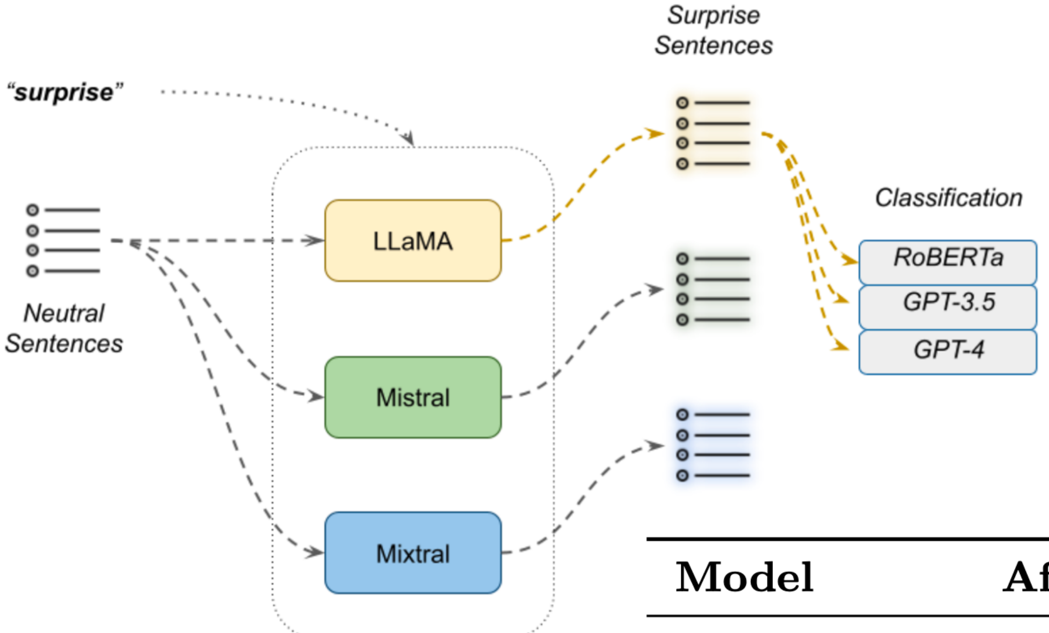
- Reflect actions‘ impacts on others
- Model empathic concern
- Guide behaviour w/ internal “values“

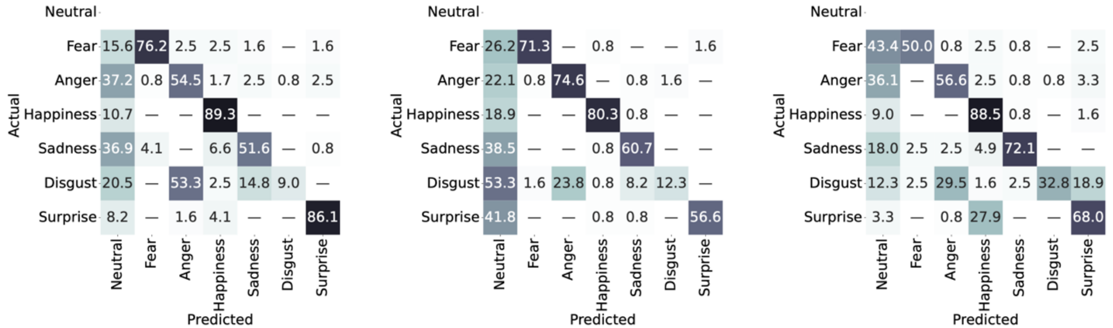

Friendly AI
Definition
Support
Cornerstone
FAI systems can safely accept human interventions and corrections
Value Alignment
FAI systems act following human values
Deontology
FAI framework for universal moral principles
Altruism
FAI systems prioritise the welfare of human
Against
Moral and technic challenge
Hard to realise the moral from technic side
Ambiguity and Evolution
FAI definition is evolving and ambiguous
Safety and Trust Risk
FAI is still not safe and not trusted by human
Evaluation issue
Hard to evaluate FAI systems
Application
eXplainable AI
Privacy
Affective Computing
Outlook
Challenges and suggestions
Definition and objective assessment
Technical framework
Multi-party collaboration

“Towards Friendly AI: A Comprehensive Review and New Perspectives on Human-AI Alignment”, ***arXiv***, *2024.*

Emerging?

Linguistics: Emergent Synthesis.





(a) LLaMA2

Actual \ Predicted	Neutral	Fear	Anger	Happiness	Sadness	Disgust	Surprise
Neutral							
Fear	15.6	76.2	2.5	2.5	1.6	—	1.6
Anger	37.2	0.8	54.5	1.7	2.5	0.8	2.5
Happiness	10.7	—	—	89.3	—	—	—
Sadness	36.9	4.1	—	6.6	51.6	—	0.8
Disgust	20.5	—	53.3	2.5	14.8	9.0	—
Surprise	8.2	—	1.6	4.1	—	—	86.1

(b) Mistral

Actual \ Predicted	Neutral	Fear	Anger	Happiness	Sadness	Disgust	Surprise
Neutral							
Fear	26.2	71.3	—	0.8	—	—	1.6
Anger	22.1	0.8	74.6	—	0.8	1.6	—
Happiness	18.9	—	—	80.3	0.8	—	—
Sadness	38.5	—	—	0.8	60.7	—	—
Disgust	53.3	1.6	23.8	0.8	8.2	12.3	—
Surprise	41.8	—	—	0.8	0.8	—	56.6

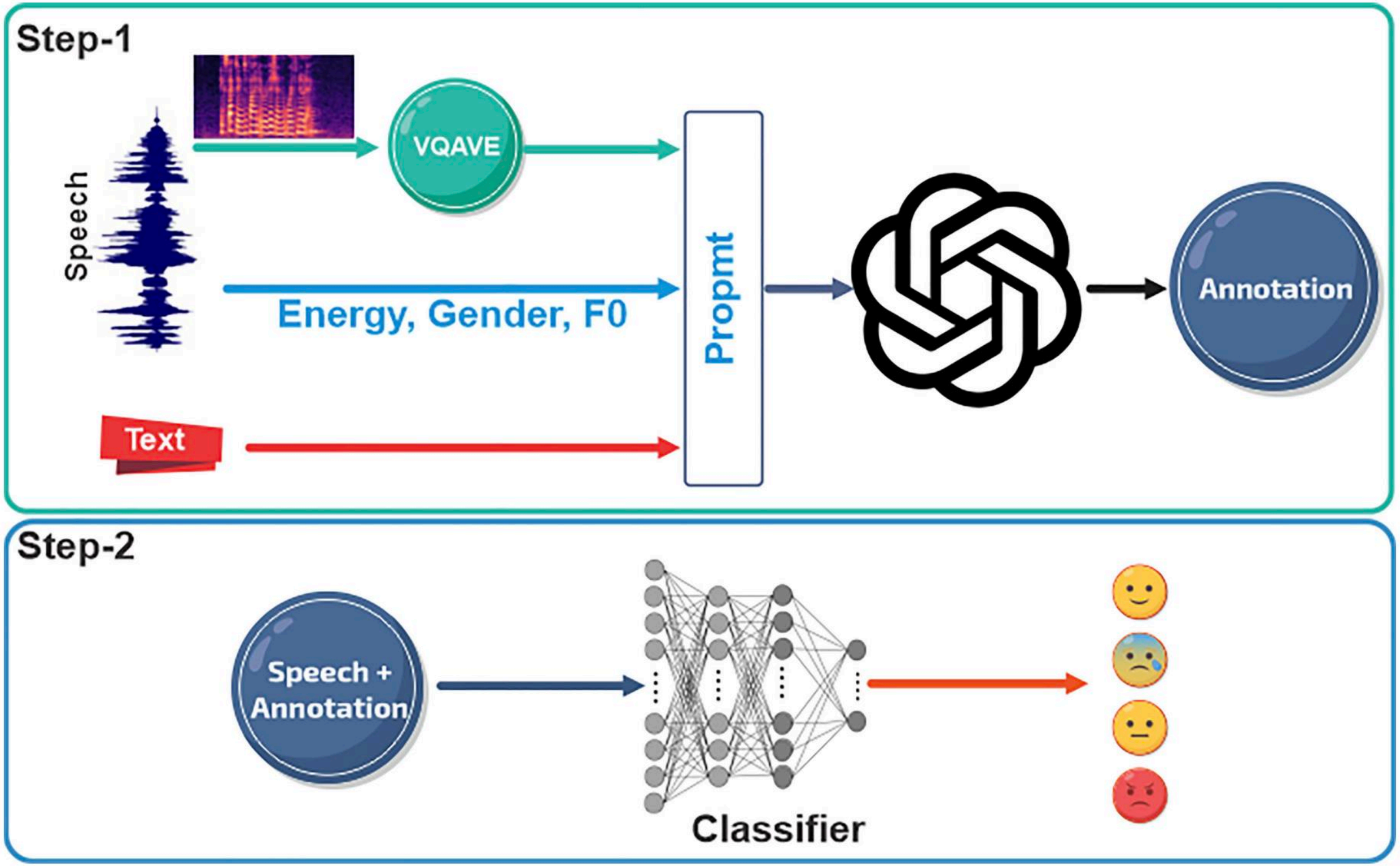
(c) Mixtral

Actual \ Predicted	Neutral	Fear	Anger	Happiness	Sadness	Disgust	Surprise
Neutral							
Fear	43.4	50.0	0.8	2.5	0.8	—	2.5
Anger	36.1	—	56.6	2.5	0.8	0.8	3.3
Happiness	9.0	—	—	88.5	0.8	—	1.6
Sadness	18.0	2.5	2.5	4.9	72.1	—	—
Disgust	12.3	2.5	29.5	1.6	2.5	32.8	18.9
Surprise	3.3	—	0.8	27.9	—	—	68.0

Model	**Affective Phrase**
Mixtral	*Wow! What a surprise! The sky is astonishingly bright and clear today!*
Mistral	*The sudden emergence of unobstructed sunlight has taken me by complete astonishment!*
LLaMA	*It comes as quite a shock to discover that the sky has transformed itself into such crystal clarity!*

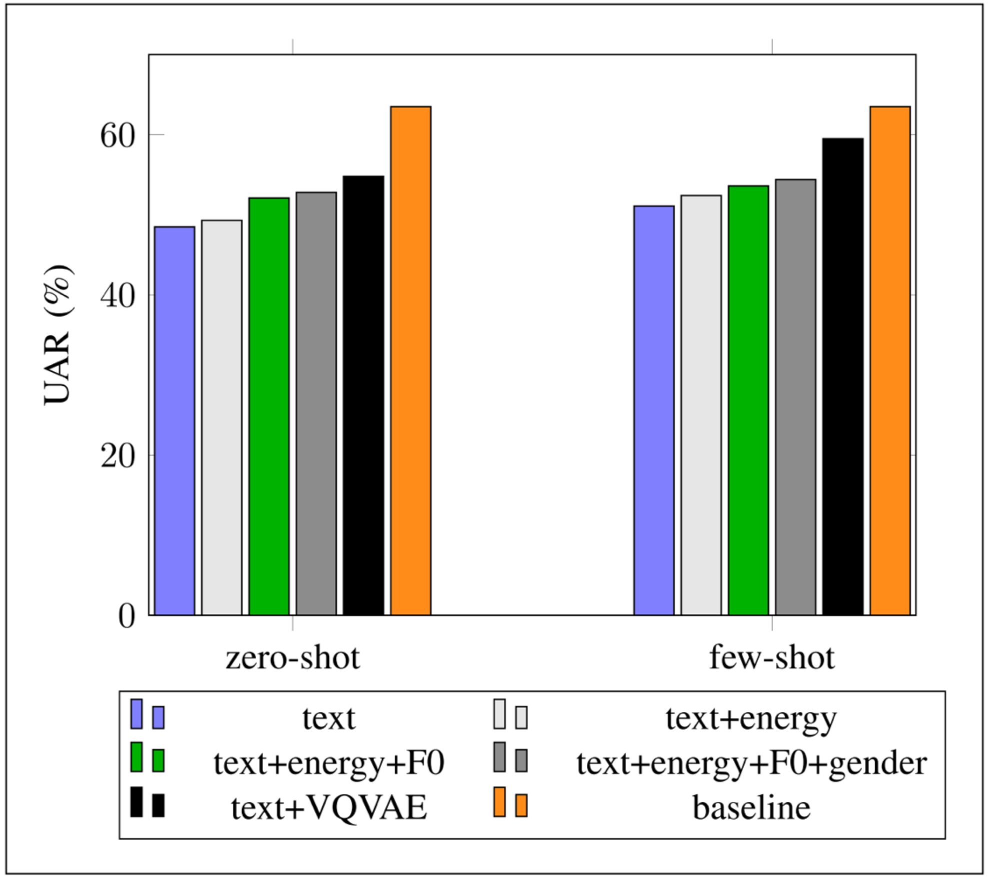
*"Affective Computing Has Changed: The Foundation Model Disruption", **arXiv**, 2024.*

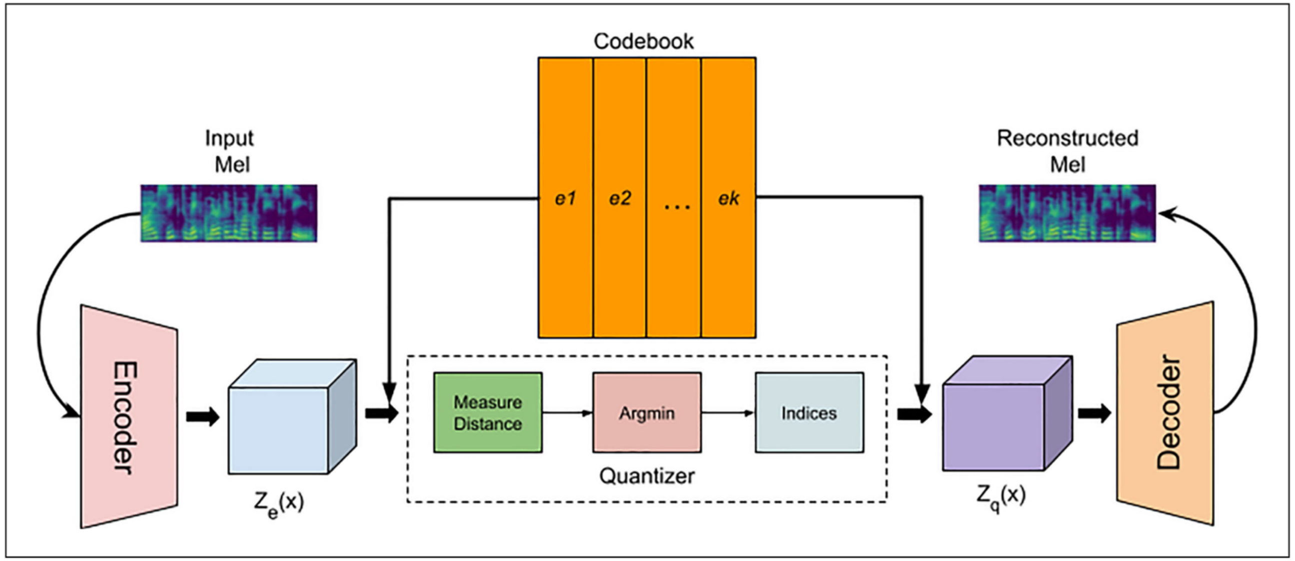
Audio: Emergent Analysis?



“Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers“,
***IEEE Computational Intelligence Magazine**, 2025.*

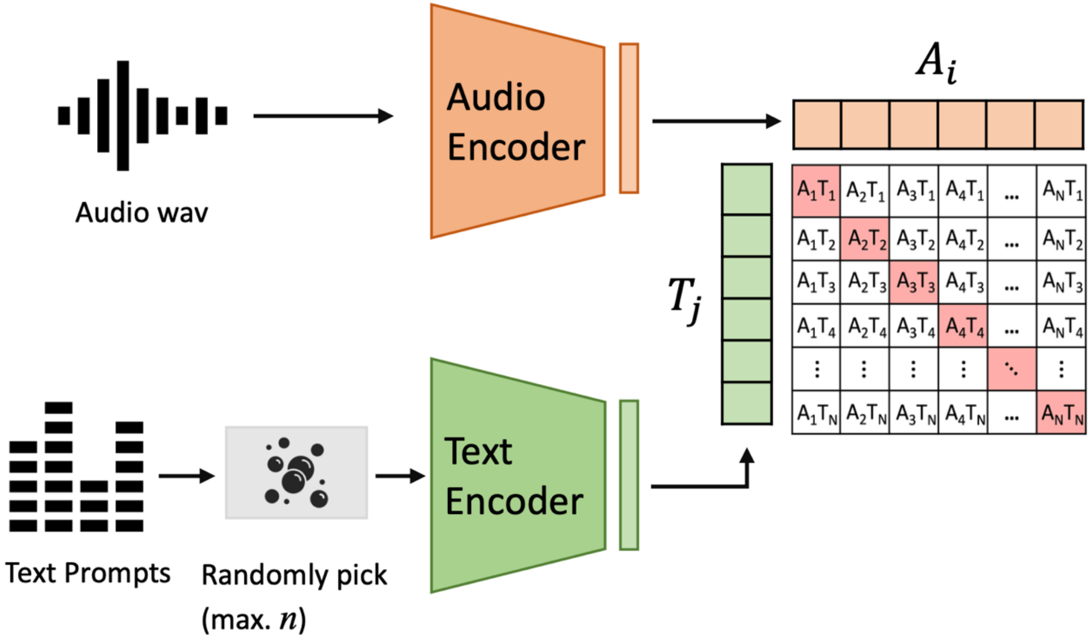
Audio: Emergent Analysis?

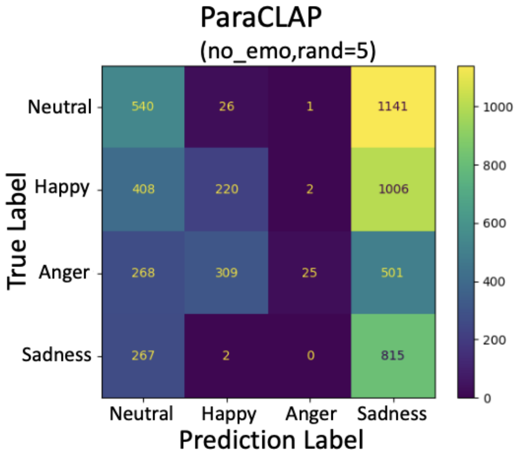




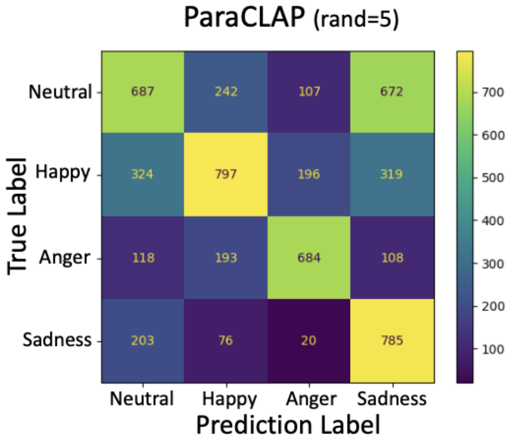
"Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers",
***IEEE Computational Intelligence Magazine**, 2025.*

Audio: Emergent Analysis?

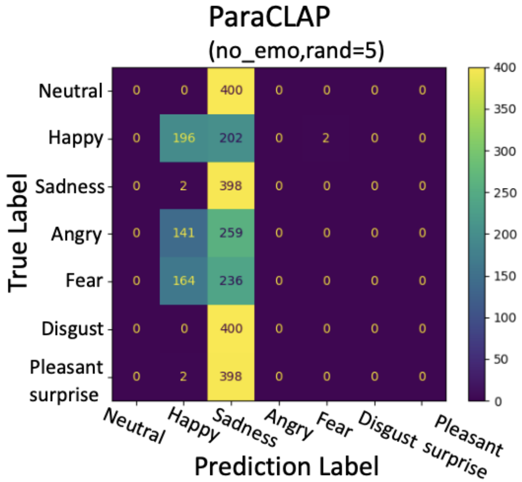

Audio wav
Audio Encoder
A_i
Text Prompts
Randomly pick (max. n)
Text Encoder
T_j
A_1T_1 A_2T_1 A_3T_1 A_4T_1 … A_NT_1
A_1T_2 A_2T_2 A_3T_2 A_4T_2 … A_NT_2
A_1T_3 A_2T_3 A_3T_3 A_4T_3 … A_NT_3
A_1T_4 A_2T_4 A_3T_4 A_4T_4 … A_NT_4
⋮ ⋮ ⋮ ⋮ ⋱ ⋮
A_1T_N A_2T_N A_3T_N A_4T_N … A_NT_N


ParaCLAP (no_emo,rand=5)

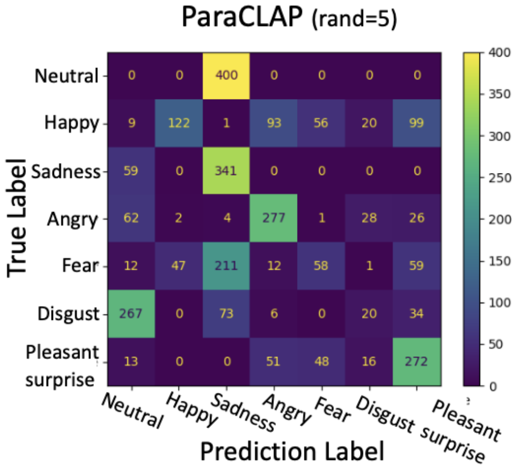
True Label \ Prediction Label	Neutral	Happy	Anger	Sadness
Neutral	540	26	1	1141
Happy	408	220	2	1006
Anger	268	309	25	501
Sadness	267	2	0	815


ParaCLAP (rand=5)

True Label \ Prediction Label	Neutral	Happy	Anger	Sadness
Neutral	687	242	107	672
Happy	324	797	196	319
Anger	118	193	684	108
Sadness	203	76	20	785


ParaCLAP (no_emo,rand=5)

True Label \ Prediction Label	Neutral	Happy	Sadness	Angry	Fear	Disgust	Pleasant surprise
Neutral	0	0	400	0	0	0	0
Happy	0	196	202	0	2	0	0
Sadness	0	2	398	0	0	0	0
Angry	0	141	259	0	0	0	0
Fear	0	164	236	0	0	0	0
Disgust	0	0	400	0	0	0	0
Pleasant surprise	0	2	398	0	0	0	0

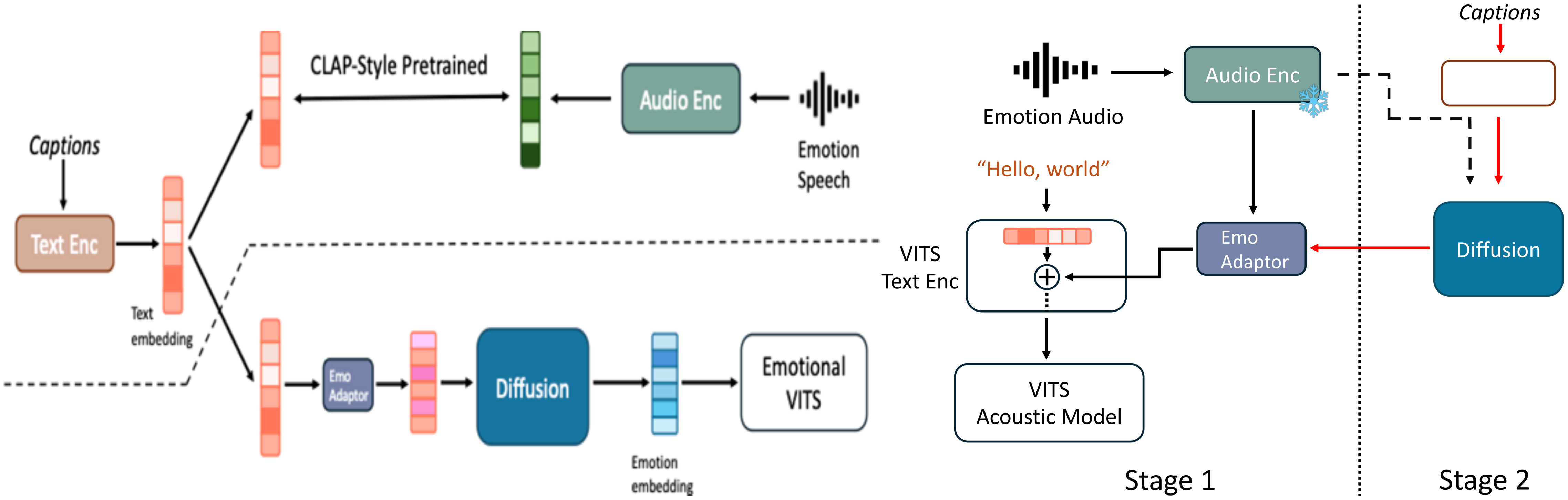

ParaCLAP (rand=5)

True Label \ Prediction Label	Neutral	Happy	Sadness	Angry	Fear	Disgust	Pleasant surprise
Neutral	0	0	400	0	0	0	0
Happy	9	122	1	93	56	20	99
Sadness	59	0	341	0	0	0	0
Angry	62	2	4	277	1	28	26
Fear	12	47	211	12	58	1	59
Disgust	267	0	73	6	0	20	34
Pleasant surprise	13	0	0	51	48	16	272

"ParaCLAP – Towards a general language-audio model for computational paralinguistic tasks", ***INTERSPEECH****, 2024.*

Audio: Emergent Synthesis?

Import computational paralinguistic caption control to emotional TTS



"Enhancing Emotional Text-to-Speech Controllability with Natural Language Guidance through Contrastive Learning and Diffusion Models", ***ICASSP****, 2025.*

Audio: Emergent Synthesis?

anger

speaker has a **normal pitch** and has a **normal pitch variation** and is **loud** and has a **high equivalent sound level** and has a **low jitter** and has a **low shimmer** and emotion is **angry**

happy

speaker has a **high pitch** and has a **normal pitch variation** and is **almost silent** and has a **normal jitter** and has a **normal shimmer** and emotion is **happy**

sad

speaker has a **normal pitch** and has a **low pitch variation** and is **almost silent** and has a **low jitter** and has a **low shimmer** and emotion is **sad**

surprise

speaker has a **high pitch** and has a **normal pitch variation** and loudness is **just about right** and has a **normal jitter** and has a **normal shimmer** and emotion is **surprise**

speaker has a **high pitch** and has a **normal pitch variation** and is **loud** and has a **high equivalent sound level** and has a **high jitter** and has a **normal shimmer** and emotion is **angry**

theta = 0.1

theta = 0.2

theta = 0.3

theta = 0.4

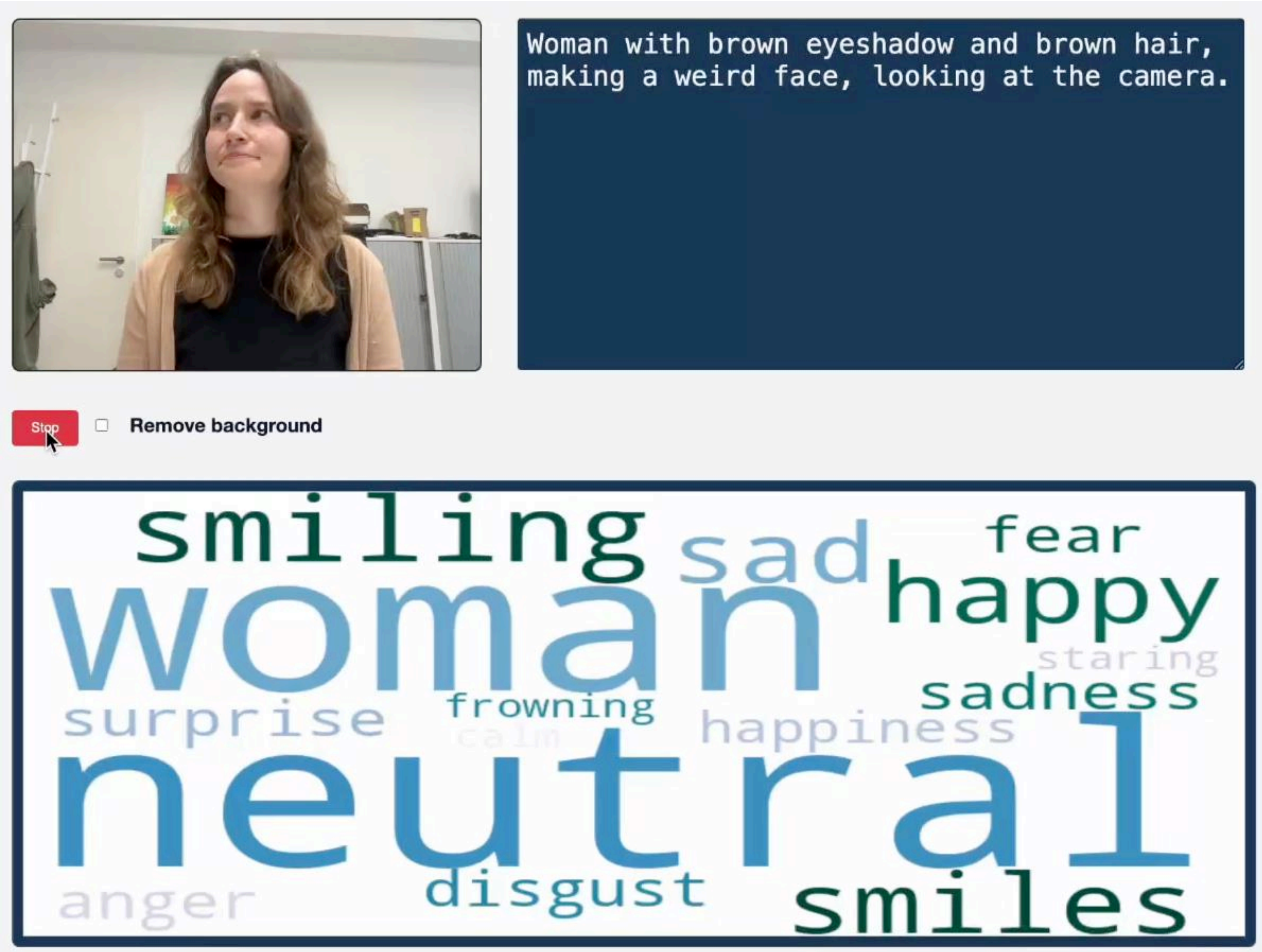
theta = 0.5

speaker has a **high pitch** and has a **normal pitch variation** and is **loud** and **has a high equivalent sound level** and has a **normal jitter** and has a **normal shimmer** and emotion is **surprise**

*“Enhancing Emotional Text-to-Speech Controllability with Natural Language Guidance through Contrastive Learning and Diffusion Models”, **ICASSP**, 2025.*

Vision: Emergent Analysis?

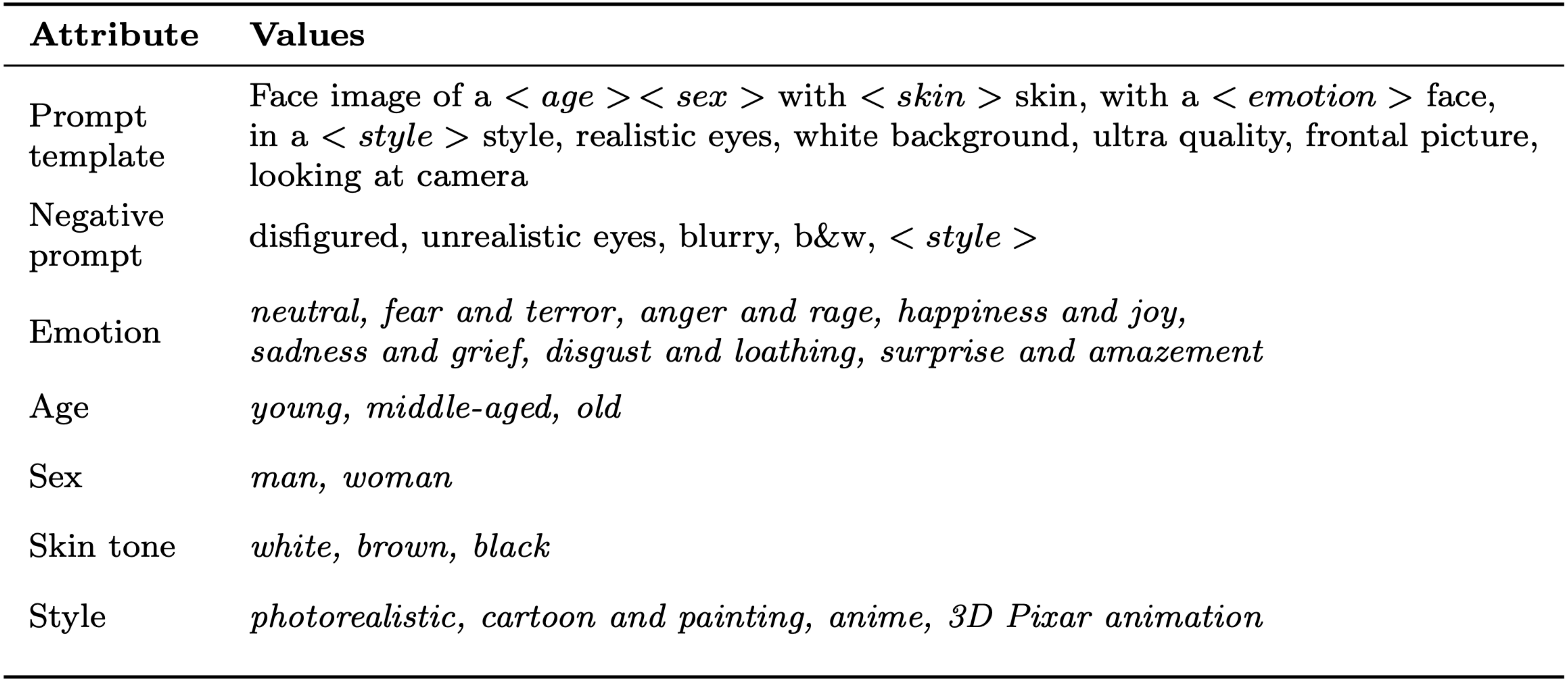
Alias	UAR
ViT-FER	.44
SmolVLM	.44
Ovis1.5	.39
Janus-1B	.28
Qwen-3B	.31
PaliGemma	.10
Qwen-7B	.41
MiniCPM	.46
InternVL	.50
Janus-8B	.37
GPT-4o	**.53**



Woman with brown eyeshadow and brown hair, making a weird face, looking at the camera.

submitted.

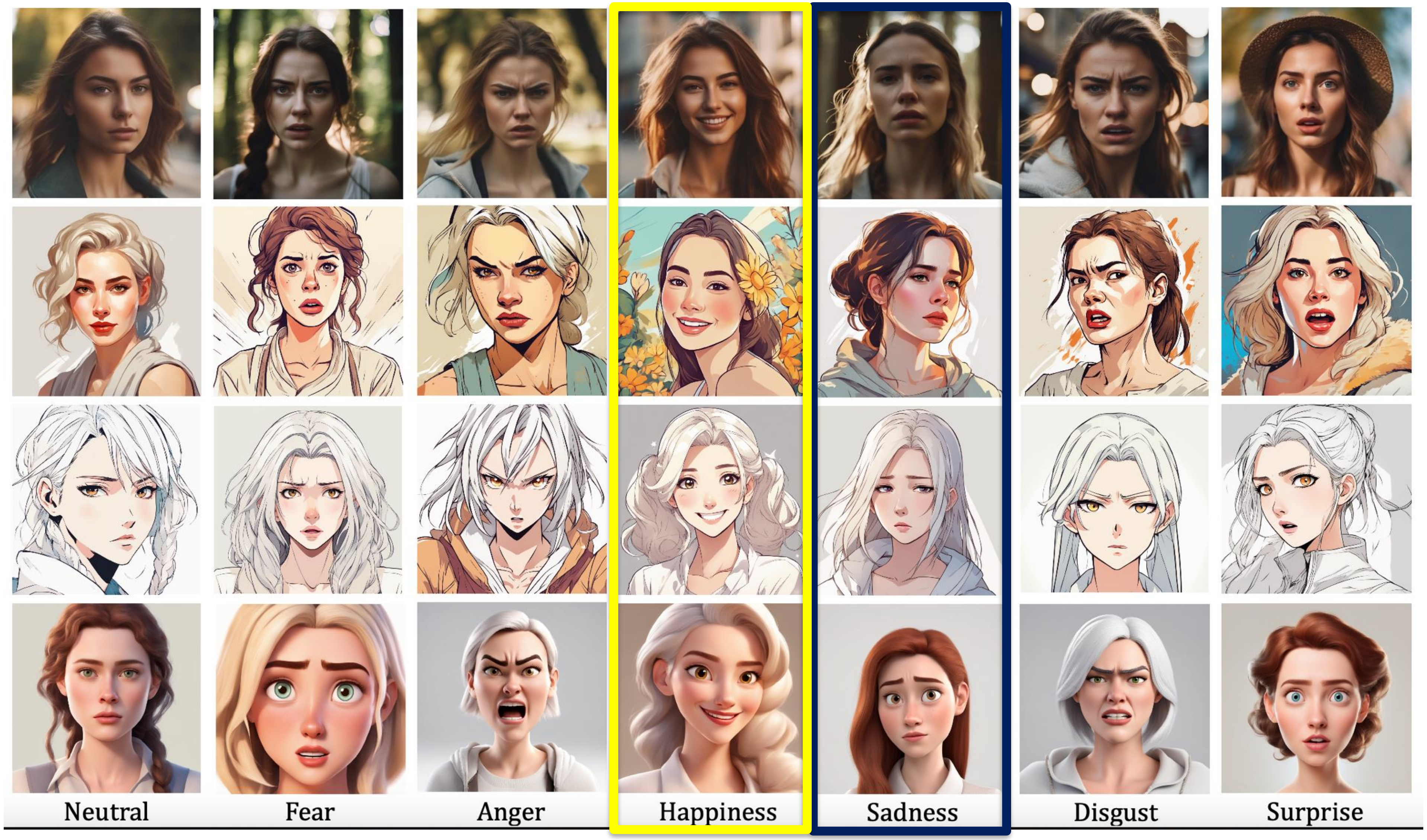
Vision: Emergent Synthesis.

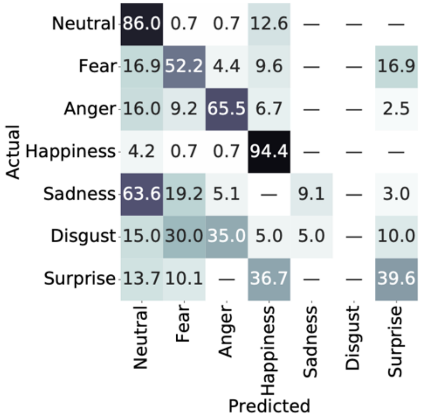


Attribute	**Values**
Prompt template	Face image of a $< age >< sex >$ with $< skin >$ skin, with a $< emotion >$ face, in a $< style >$ style, realistic eyes, white background, ultra quality, frontal picture, looking at camera
Negative prompt	disfigured, unrealistic eyes, blurry, b&w, $< style >$
Emotion	*neutral, fear and terror, anger and rage, happiness and joy, sadness and grief, disgust and loathing, surprise and amazement*
Age	*young, middle-aged, old*
Sex	*man, woman*
Skin tone	*white, brown, black*
Style	*photorealistic, cartoon and painting, anime, 3D Pixar animation*

“Affective Computing Has Changed: The Foundation Model Disruption”, ***arXiv****, 2024.*

Vision: Emergent Synthesis.





	Neutral	Fear	Anger	Happiness	Sadness	Disgust	Surprise
Neutral	86.0	0.7	0.7	12.6	—	—	—
Fear	16.9	52.2	4.4	9.6	—	—	16.9
Anger	16.0	9.2	65.5	6.7	—	—	2.5
Happiness	4.2	0.7	0.7	94.4	—	—	—
Sadness	63.6	19.2	5.1	—	9.1	—	3.0
Disgust	15.0	30.0	35.0	5.0	5.0	—	10.0
Surprise	13.7	10.1	—	36.7	—	—	39.6

(d) 3D

*“Affective Computing Has Changed: The Foundation Model Disruption”, **arXiv**, 2024.*

Emergent “Inner” Emotion?

- **(ChatGPT) Do you have emotions?**

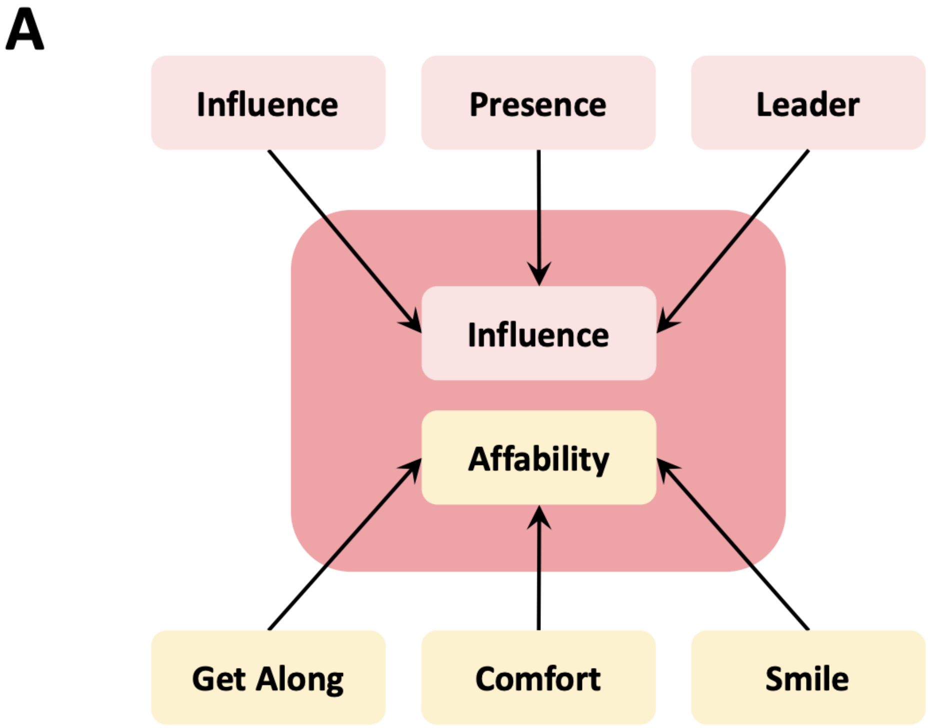
  “I don't have emotions the way humans do—I don't feel happiness, sadness, anger, or love. But I do understand how emotions work, can recognize them in language, and can simulate emotional tone in conversation to better connect with you. That means I can respond warmly, seriously, empathetically, or even playfully, depending on the context.“

- **(ChatGPT) Would inner emotion models for you make sense?**

  Yes—and that’s actually one of the most exciting frontiers in AI research. Building inner emotion models could make a lot of sense, depending on the goals. These wouldn’t be real feelings, but structured internal representations that mimic emotional processes in a useful way.
  …
  Yes—inner emotion models could enhance functionality, empathy, and alignment, especially if designed transparently and ethically. They’re not “emotions” in a biological sense, but they might serve a similar role in intelligent behavior.

Just Emo?

Charisma?



A

Influence

Presence

Leader

Influence

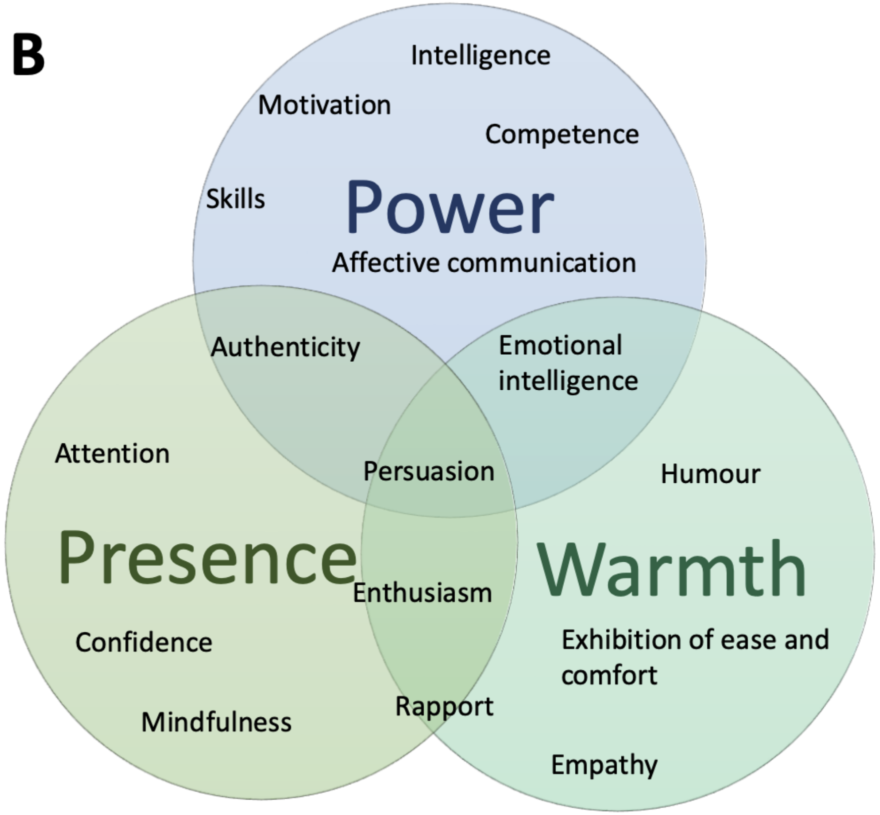
Affability

Get Along

Comfort

Smile

Tshkay et al. (2018)



B

Intelligence

Motivation

Competence

Skills

Power

Affective communication

Authenticity

Emotional intelligence

Attention

Persuasion

Humour

Presence

Enthusiasm

Warmth

Confidence

Exhibition of ease and comfort

Mindfulness

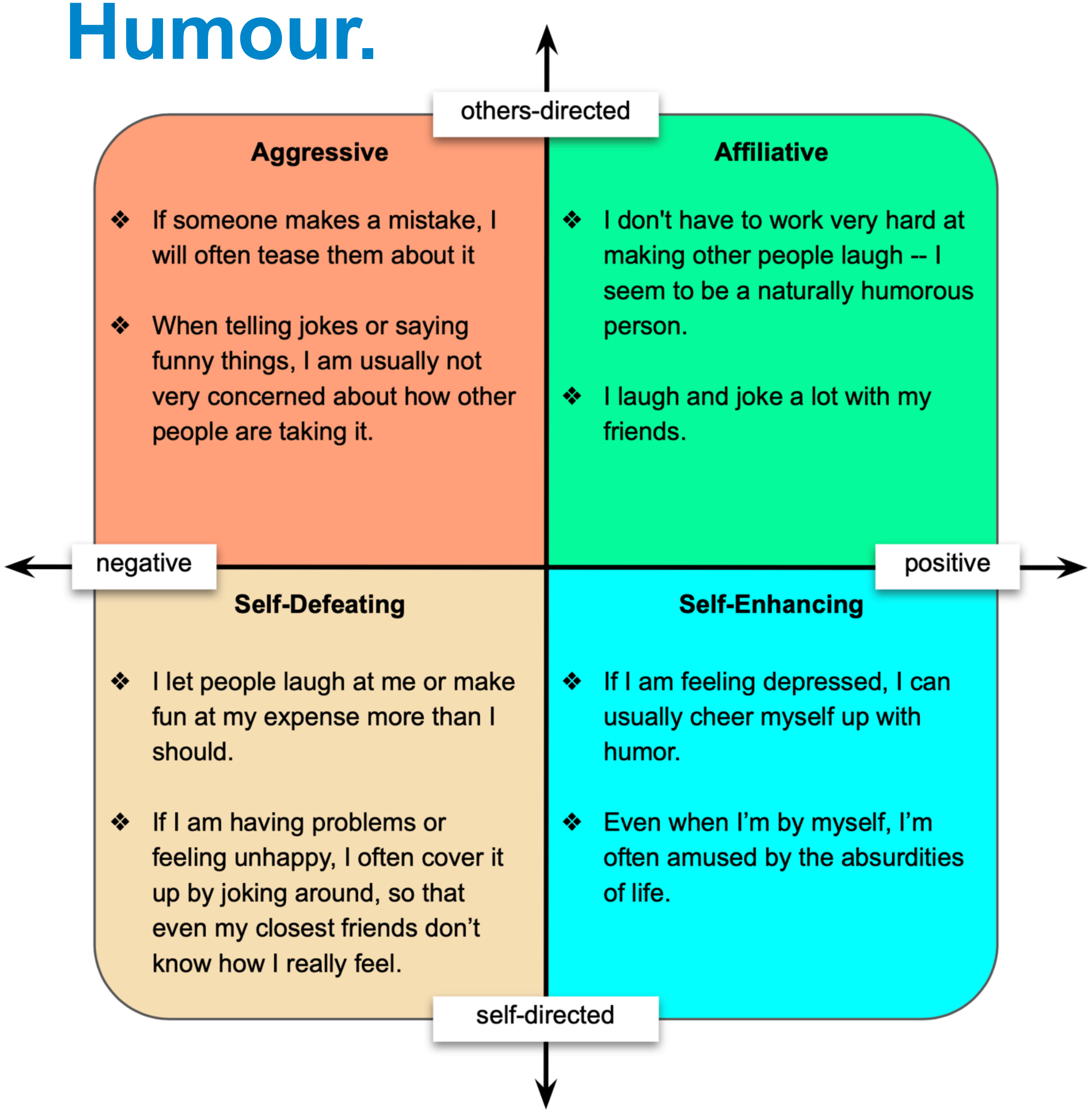
Rapport

Empathy

Fox Cabane (2013)

"Computational Charisma – A Brick by Brick Blueprint for Building Charismatic Artificial Intelligence",
Frontiers in Computer Science*, 2023.*

Humour.



others-directed

negative

positive

self-directed

Aggressive

- ❖ If someone makes a mistake, I will often tease them about it
- ❖ When telling jokes or saying funny things, I am usually not very concerned about how other people are taking it.

Affiliative

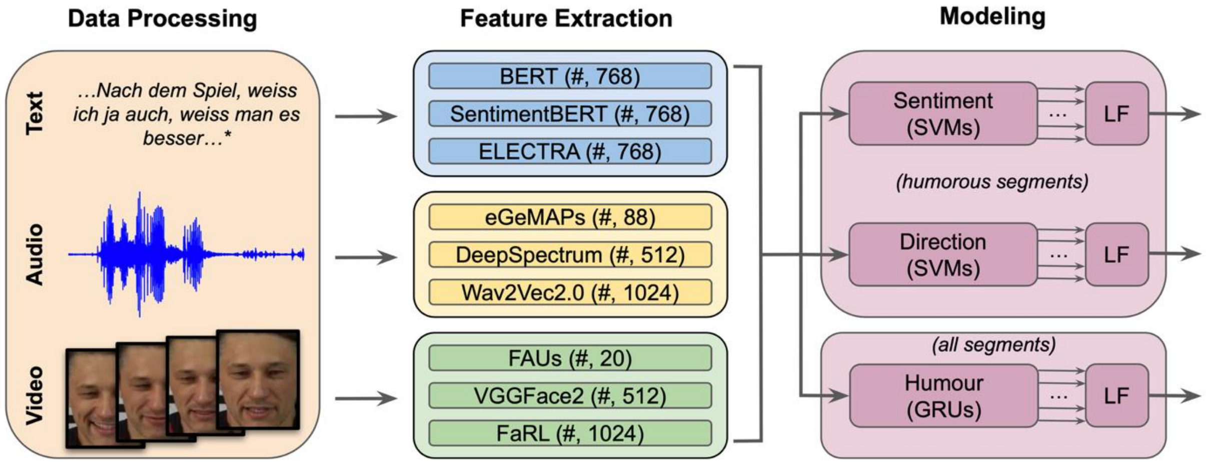
- ❖ I don't have to work very hard at making other people laugh -- I seem to be a naturally humorous person.
- ❖ I laugh and joke a lot with my friends.

Self-Defeating

- ❖ I let people laugh at me or make fun at my expense more than I should.
- ❖ If I am having problems or feeling unhappy, I often cover it up by joking around, so that even my closest friends don't know how I really feel.

Self-Enhancing

- ❖ If I am feeling depressed, I can usually cheer myself up with humor.
- ❖ Even when I'm by myself, I'm often amused by the absurdities of life.



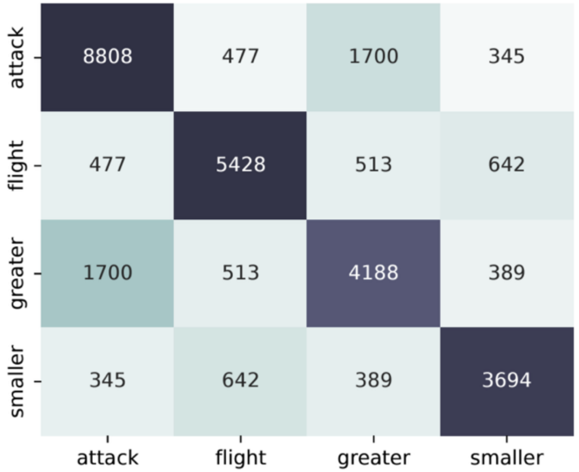
"Towards Multimodal Prediction of Spontaneous Humor: A Novel Dataset and First Results",
***IEEE Transactions on Affective Computing**, 2024.*

Defensiveness.

- DefComm-DB
- Birkenbihl's model
- 261 videos
- Interviews, political debates, legal trials, TV shows, paparazzi footage, …
- 11 labellers

"The Four Horsemen: Criticism, Contempt, Defensiveness, and Stonewalling",
John Gottman

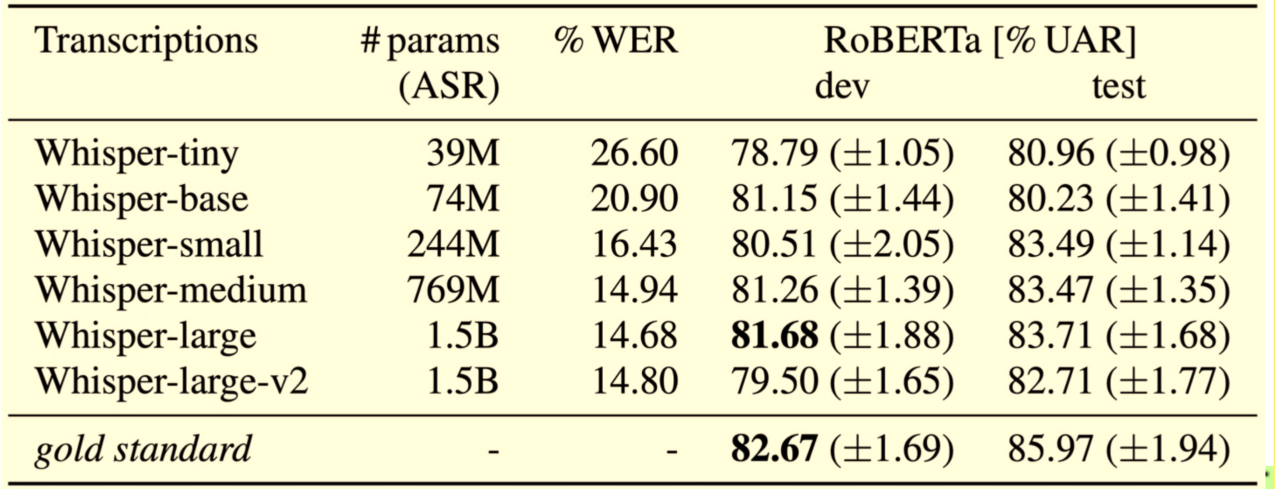
Oh, so, now now everything's my fault??
I guess I'm just tust tarble at verything!



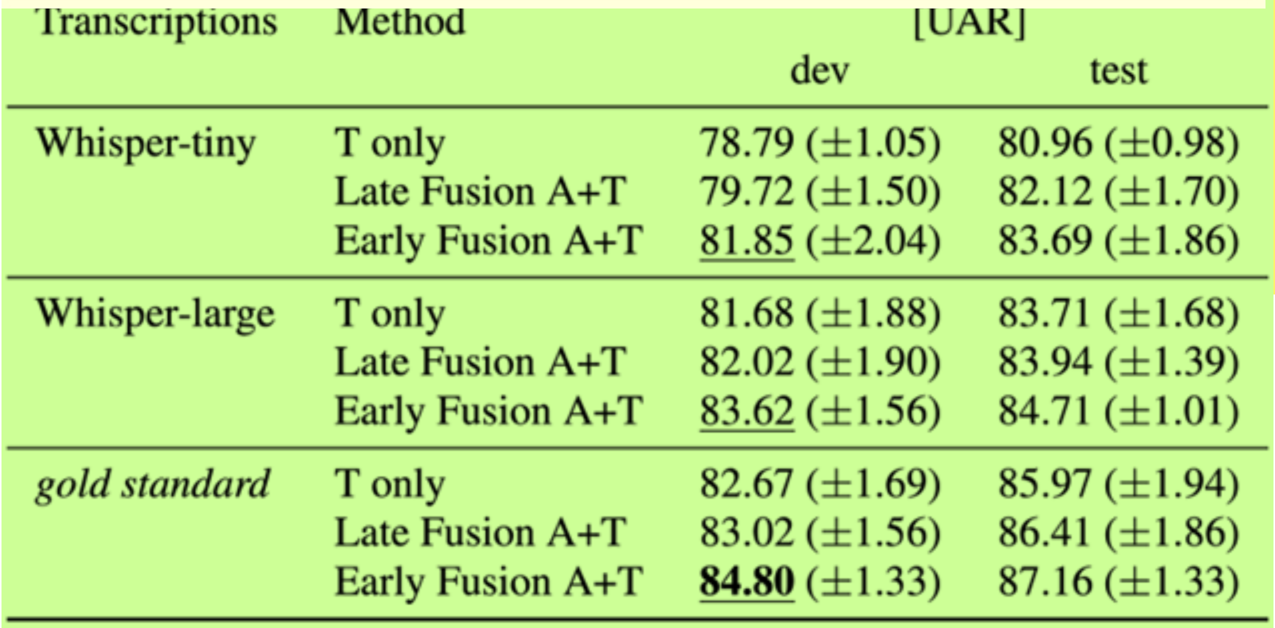
	attack	flight	greater	smaller
attack	8808	477	1700	345
flight	477	5428	513	642
greater	1700	513	4188	389
smaller	345	642	389	3694

*"Speech-Based Classification of Defensive Communication: ANovel Dataset and Results", **Interspeech**, 2023.*

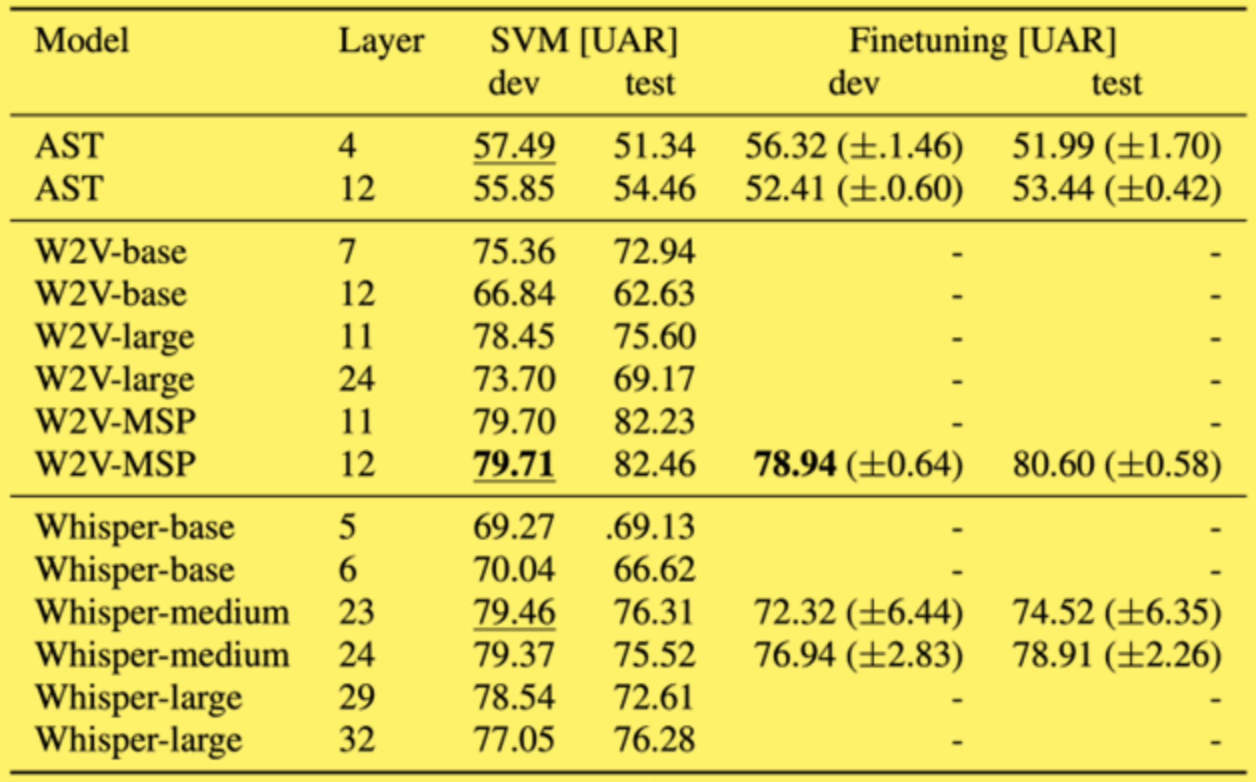
Flattery.



Transcriptions	# params (ASR)	% WER	RoBERTa [% UAR] dev	test
Whisper-tiny	39M	26.60	78.79 (±1.05)	80.96 (±0.98)
Whisper-base	74M	20.90	81.15 (±1.44)	80.23 (±1.41)
Whisper-small	244M	16.43	80.51 (±2.05)	83.49 (±1.14)
Whisper-medium	769M	14.94	81.26 (±1.39)	83.47 (±1.35)
Whisper-large	1.5B	14.68	**81.68** (±1.88)	83.71 (±1.68)
Whisper-large-v2	1.5B	14.80	79.50 (±1.65)	82.71 (±1.77)
gold standard	-	-	**82.67** (±1.69)	85.97 (±1.94)



Transcriptions	Method	[UAR] dev	test
Whisper-tiny	T only	78.79 (±1.05)	80.96 (±0.98)
	Late Fusion A+T	79.72 (±1.50)	82.12 (±1.70)
	Early Fusion A+T	81.85 (±2.04)	83.69 (±1.86)
Whisper-large	T only	81.68 (±1.88)	83.71 (±1.68)
	Late Fusion A+T	82.02 (±1.90)	83.94 (±1.39)
	Early Fusion A+T	83.62 (±1.56)	84.71 (±1.01)
gold standard	T only	82.67 (±1.69)	85.97 (±1.94)
	Late Fusion A+T	83.02 (±1.56)	86.41 (±1.86)
	Early Fusion A+T	**84.80** (±1.33)	87.16 (±1.33)



Model	Layer	SVM [UAR] dev	test	Finetuning [UAR] dev	test
AST	4	57.49	51.34	56.32 (±.1.46)	51.99 (±1.70)
AST	12	55.85	54.46	52.41 (±.0.60)	53.44 (±0.42)
W2V-base	7	75.36	72.94	-	-
W2V-base	12	66.84	62.63	-	-
W2V-large	11	78.45	75.60	-	-
W2V-large	24	73.70	69.17	-	-
W2V-MSP	11	79.70	82.23	-	-
W2V-MSP	12	**79.71**	82.46	**78.94** (±0.64)	80.60 (±0.58)
Whisper-base	5	69.27	.69.13	-	-
Whisper-base	6	70.04	66.62	-	-
Whisper-medium	23	79.46	76.31	72.32 (±6.44)	74.52 (±6.35)
Whisper-medium	24	79.37	75.52	76.94 (±2.83)	78.91 (±2.26)
Whisper-large	29	78.54	72.61	-	-
Whisper-large	32	77.05	76.28	-	-

20 hours of speech:
255 speakers,
2159 dyads of analyst questions and CEO answers
3 annotators had to fully agree

"This Paper Had the Smartest Reviewers – Flattery Detection Utilising an Audio-Textual Transformer-Based Approach", ***Interspeech****, 2024.*

For what?

Games.

GaCha 2019 @ GamesCon by audEERING

29 Teams

*"Large Language Models for Mental Health", **arXiv**, 2024.*

Games.

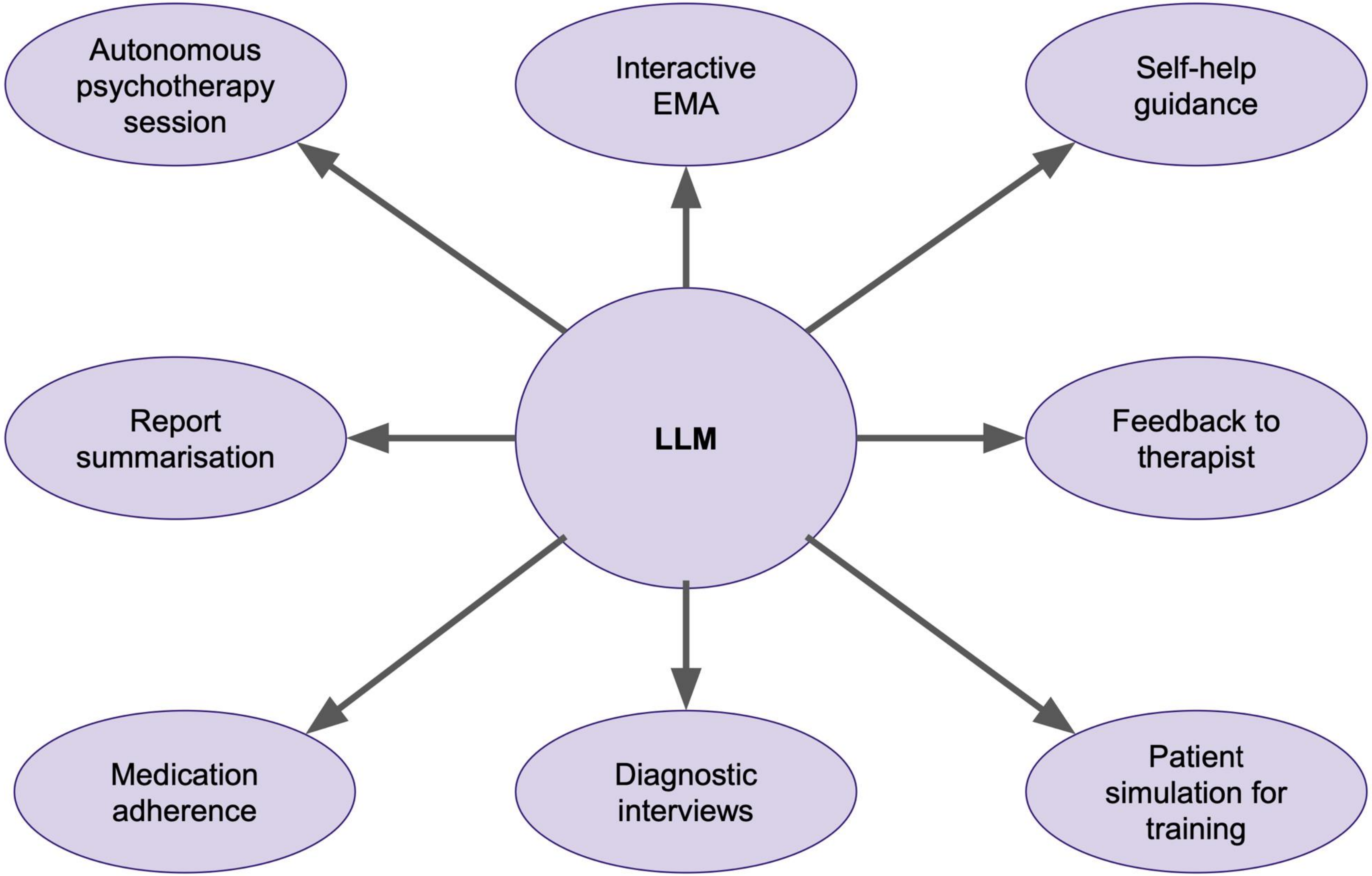
TPS Fantasy Game – voice emotions charge up your ultimate healing spells and unleash magical power

Emotions from voice trigger magic spells, unlock cursed treasure chests, conjure a vast variety of magic runes

Arcade-style VR game, more enraged → higher damage, but relaxed voice helps regenerate faster

*"Large Language Models for Mental Health", **arXiv**, 2024.*

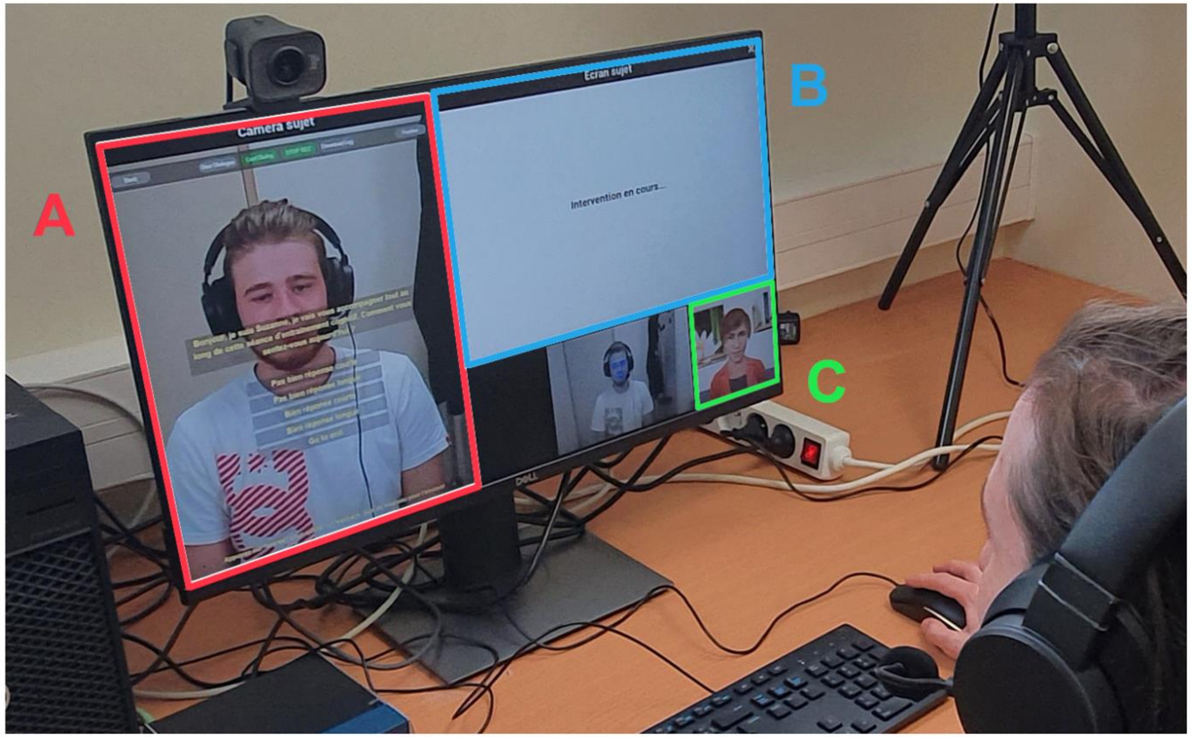
Health.



*"Large Language Models for Mental Health", **arXiv**, 2024.*

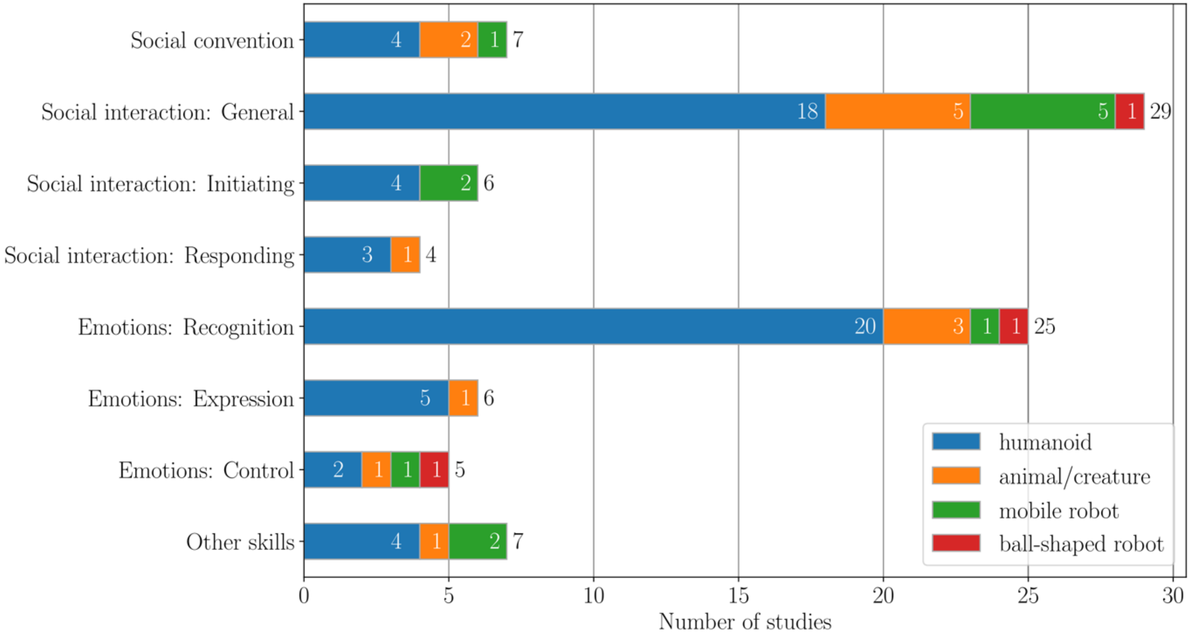
Therapy.

L'évaluation neuropsychologique
100 TESTS



*"EVAC 2024 – Empathic Virtual Agent Challenge: Appraisal-based Recognition of Affective States", **ACM ICMI**, 2024.*

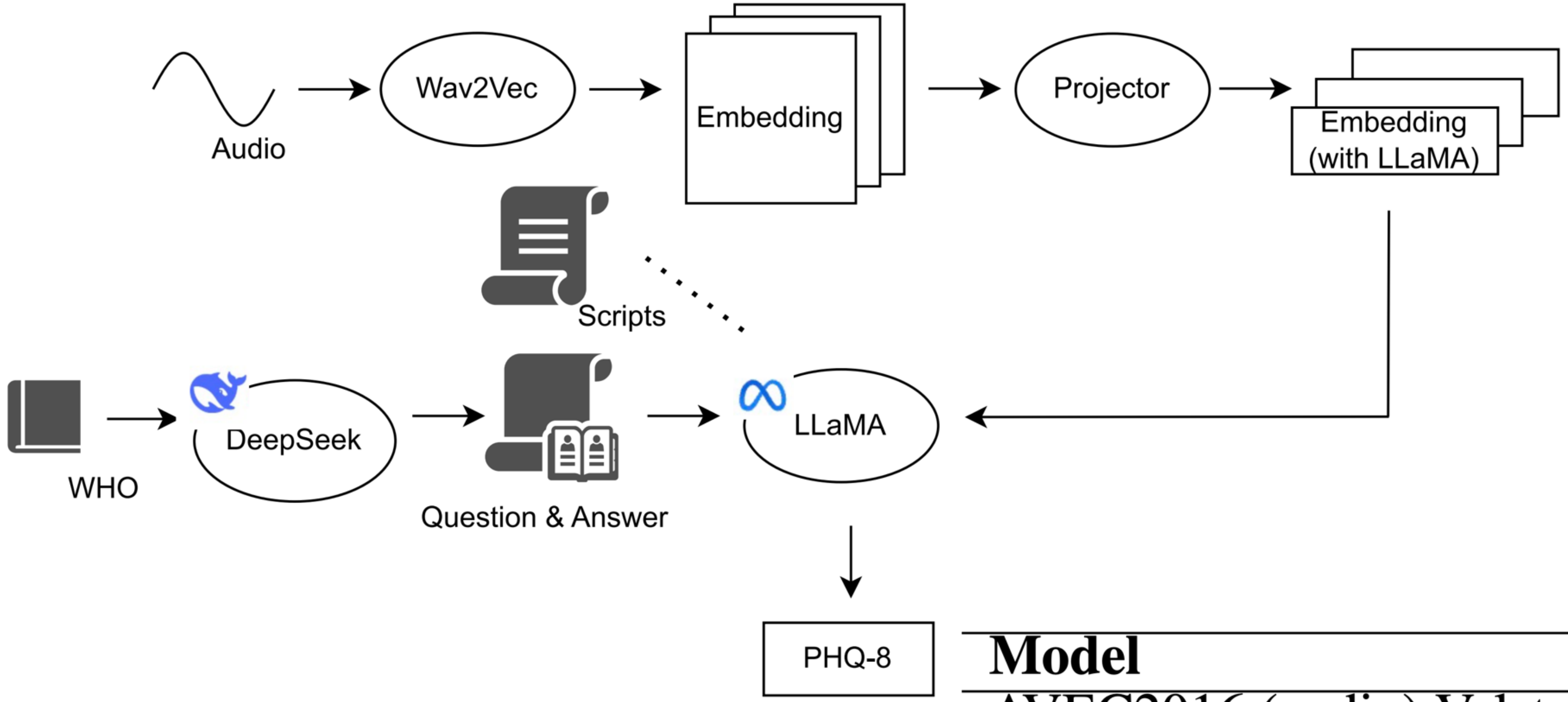
Therapy.


Social convention 4 2 1 7
Social interaction: General 18 5 5 1 29
Social interaction: Initiating 4 2 6
Social interaction: Responding 3 1 4
Emotions: Recognition 20 3 1 1 25
Emotions: Expression 5 1 6
Emotions: Control 2 1 1 1 5
Other skills 4 1 2 7
0
5
10
15
20
25
30
Number of studies
humanoid
animal/creature
mobile robot
ball-shaped robot

*“Robot-Based Intervention for Children With Autism Spectrum Disorder: A Systematic Literature Review”, **IEEE Access**, 2021.*

Next?

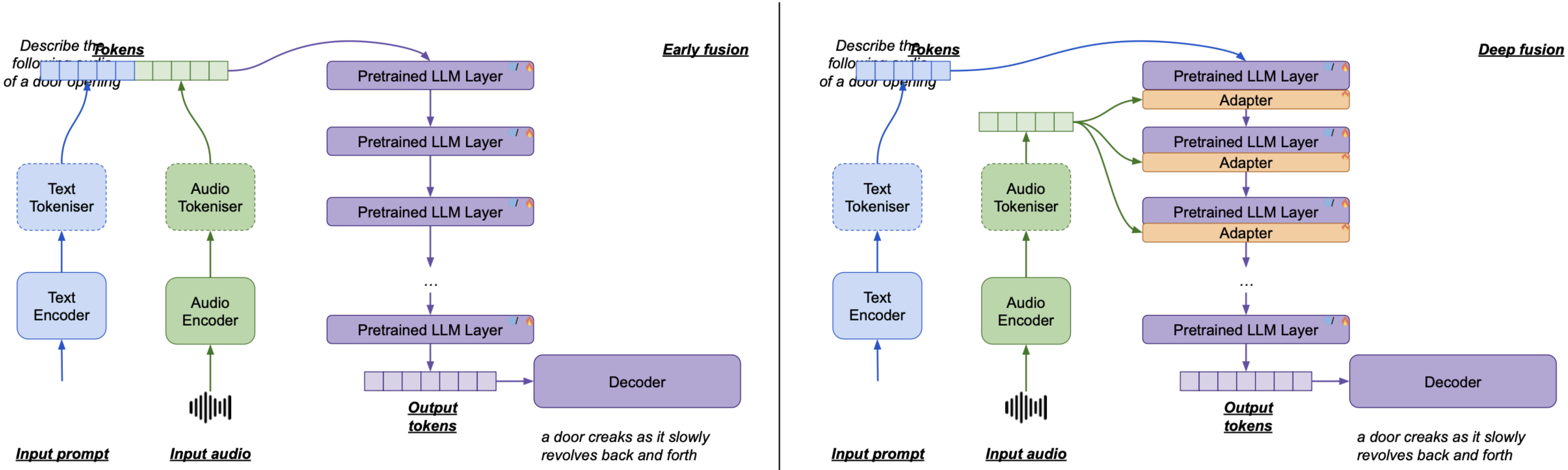
Knowledge Injection: ICD for PHQ



Model	**MAE**
AVEC2016 (audio) Valstar et al. (2016)	5.72
LSTM Afzal Aghaei and Khodaei (2023)	5.7
Random Forest Afzal Aghaei and Khodaei (2023)	5.71
Ours(audio)	5.373
Ours(text)	6.342
Ours(audio + text)	5.356
Ours(text + Knowledge Injection)	**5.354**

“Large Language Models for Depression Recognition in Spoken Language Integrating Psychological Knowledge”, ***arXiv****, 2025.*

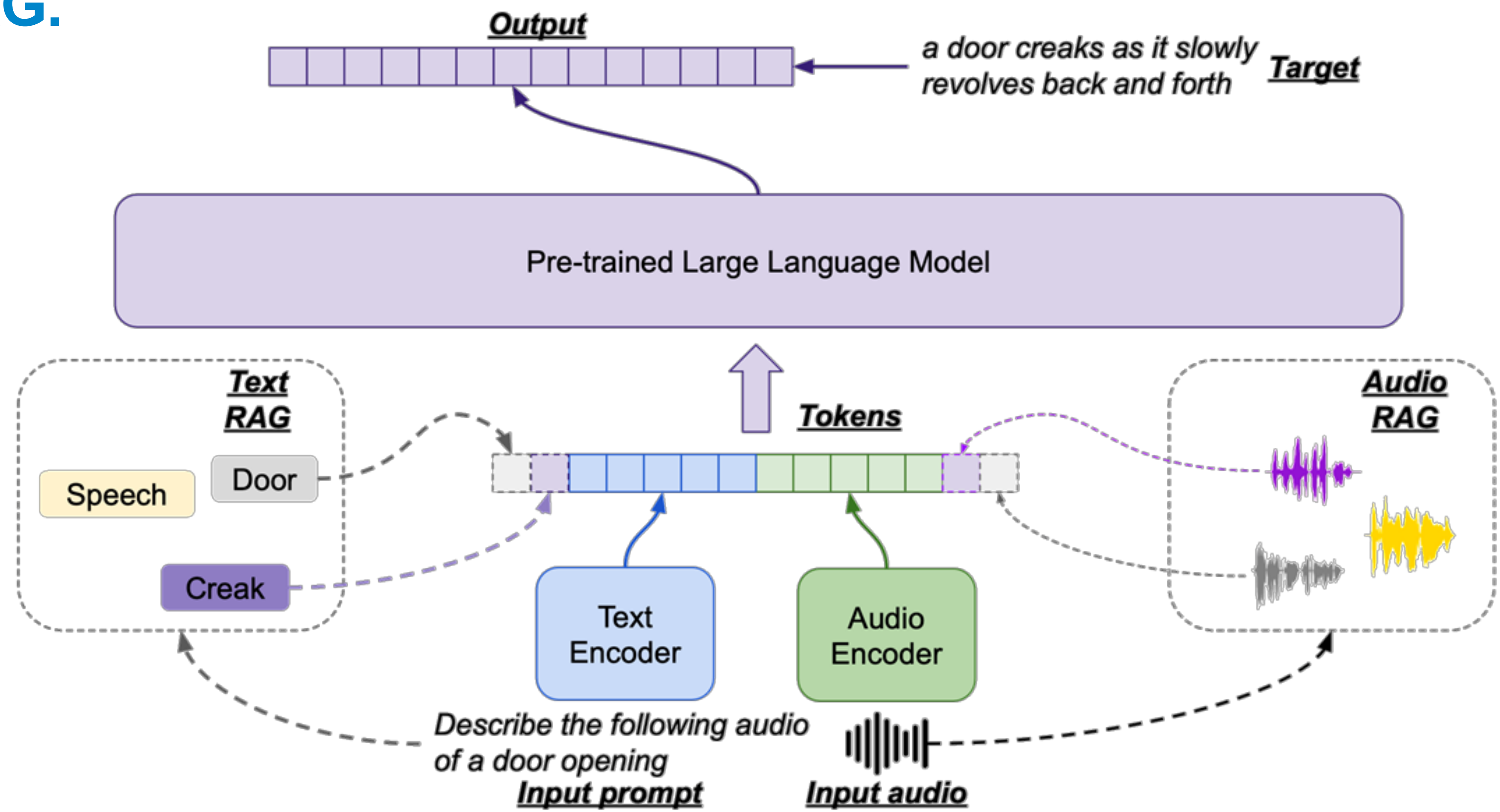
Model Fusion & “Recycling”.



(a) Early fusion (b) Deep fusion

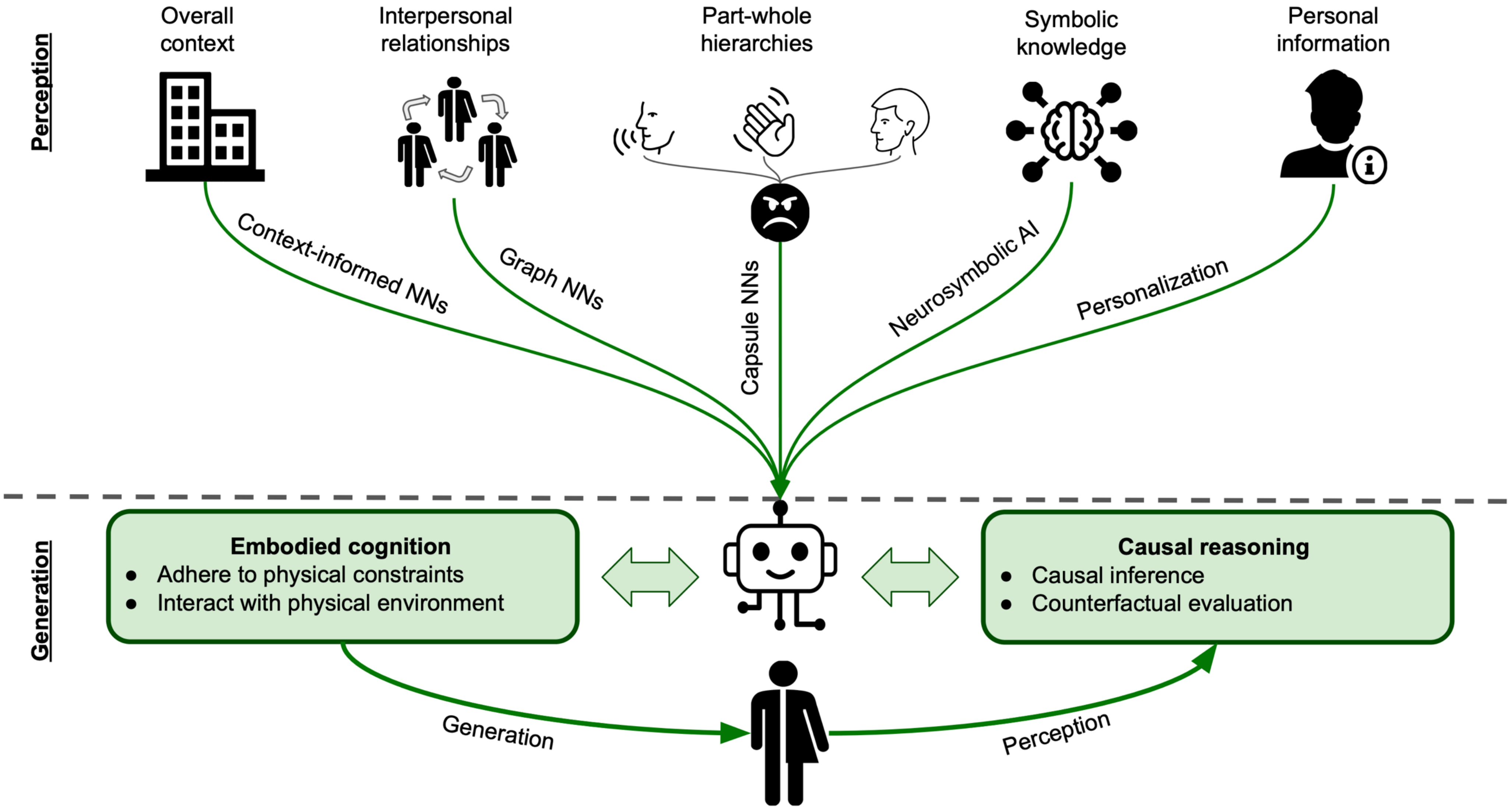
*“Computational Audio Analysis: From Task-Specific Machine Learning to Foundation Models”, **arXiv**, 2024.*

RAG.


Output
a door creaks as it slowly revolves back and forth
Target
Pre-trained Large Language Model
Text RAG
Speech
Door
Creak
Tokens
Audio RAG
Text Encoder
Audio Encoder
Describe the following audio of a door opening
Input prompt
Input audio

*"Computational Audio Analysis: From Task-Specific Machine Learning to Foundation Models", **arXiv**, 2024.*

Beyond Deep Learning.

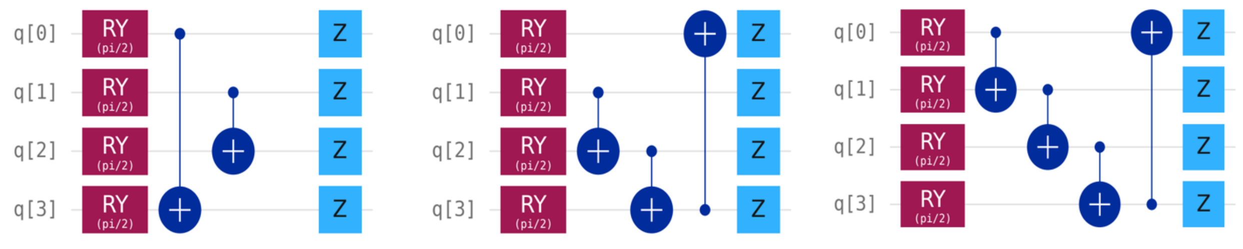


Embodied cognition
- Adhere to physical constraints
- Interact with physical environment

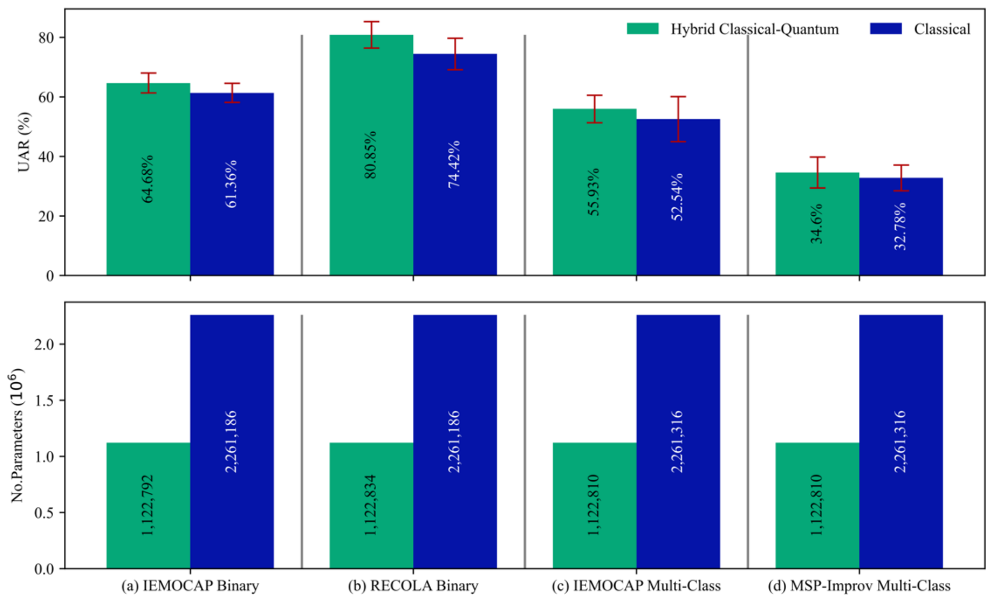
Causal reasoning
- Causal inference
- Counterfactual evaluation

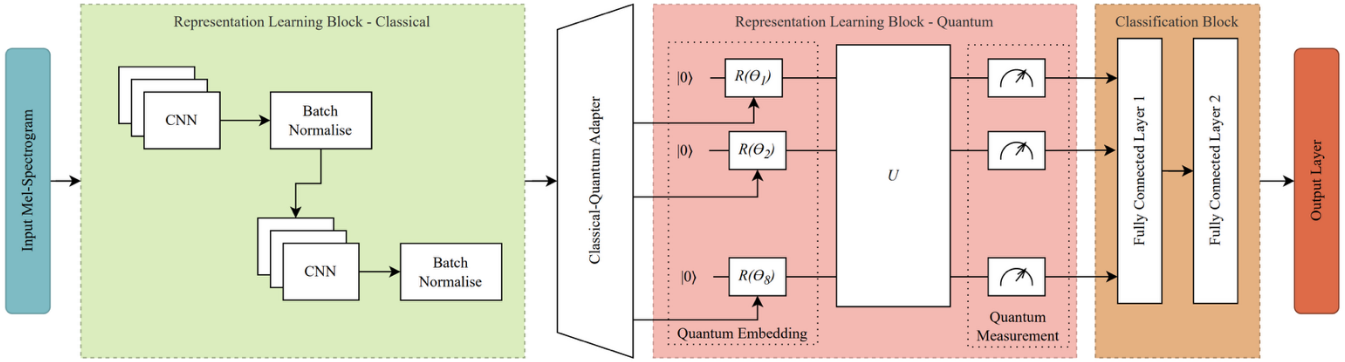
*"Beyond Deep Learning: Charting the Next Frontiers of Affective Computing", **Intelligent Computing**, 2024.*

Quantum Computing.



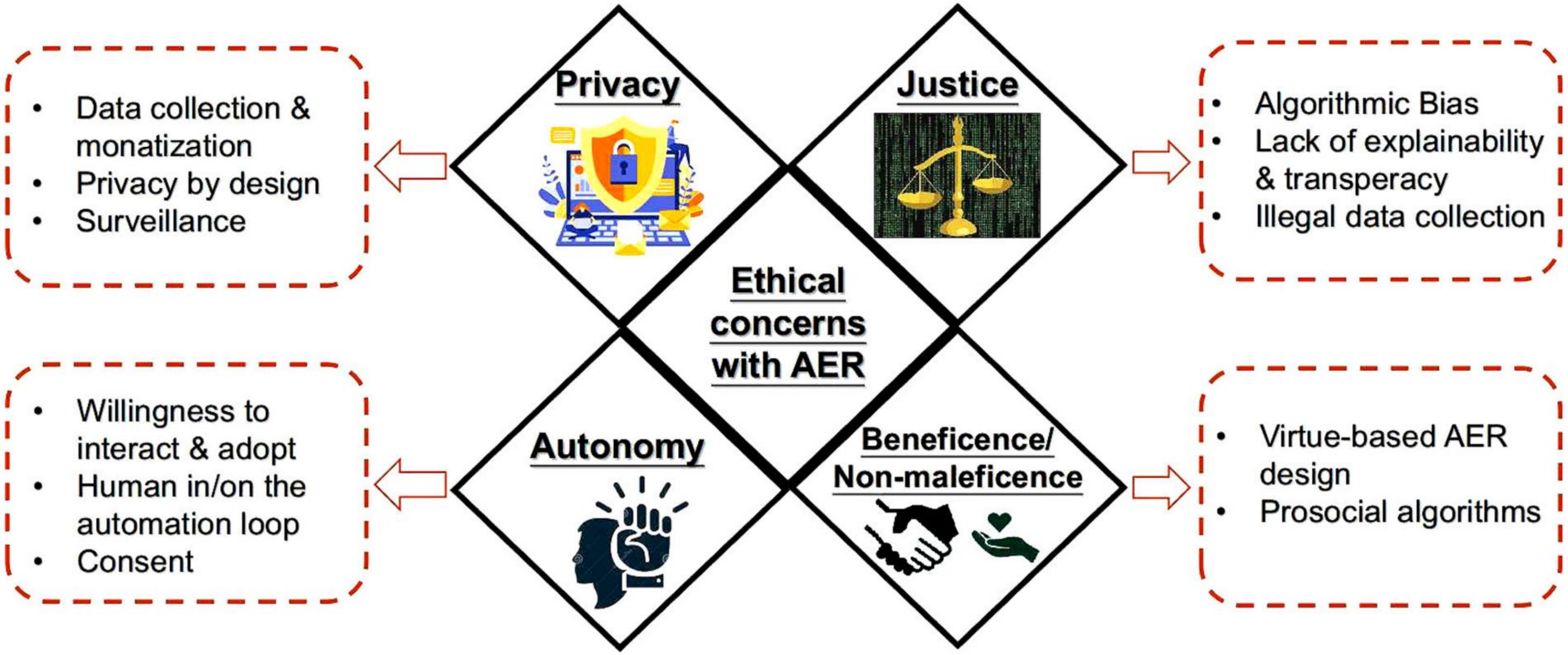
(a) 2-CNOT (b) 3-CNOT (c) 4-CNOT





"Representation Learning with Parameterised Quantum Circuits for Advancing Speech Emotion Recognition", ***arXiv***, *2025.*

Ethics.



Privacy

- Data collection & monatization
- Privacy by design
- Surveillance

Justice

- Algorithmic Bias
- Lack of explainability & transperacy
- Illegal data collection

Ethical concerns with AER

Autonomy

- Willingness to interact & adopt
- Human in/on the automation loop
- Consent

Beneficence/ Non-maleficence

- Virtue-based AER design
- Prosocial algorithms

Ethical principles
Transparency
Full disclosure
Personal consent
Ethical data sharing
Data ownership
Security and privacy

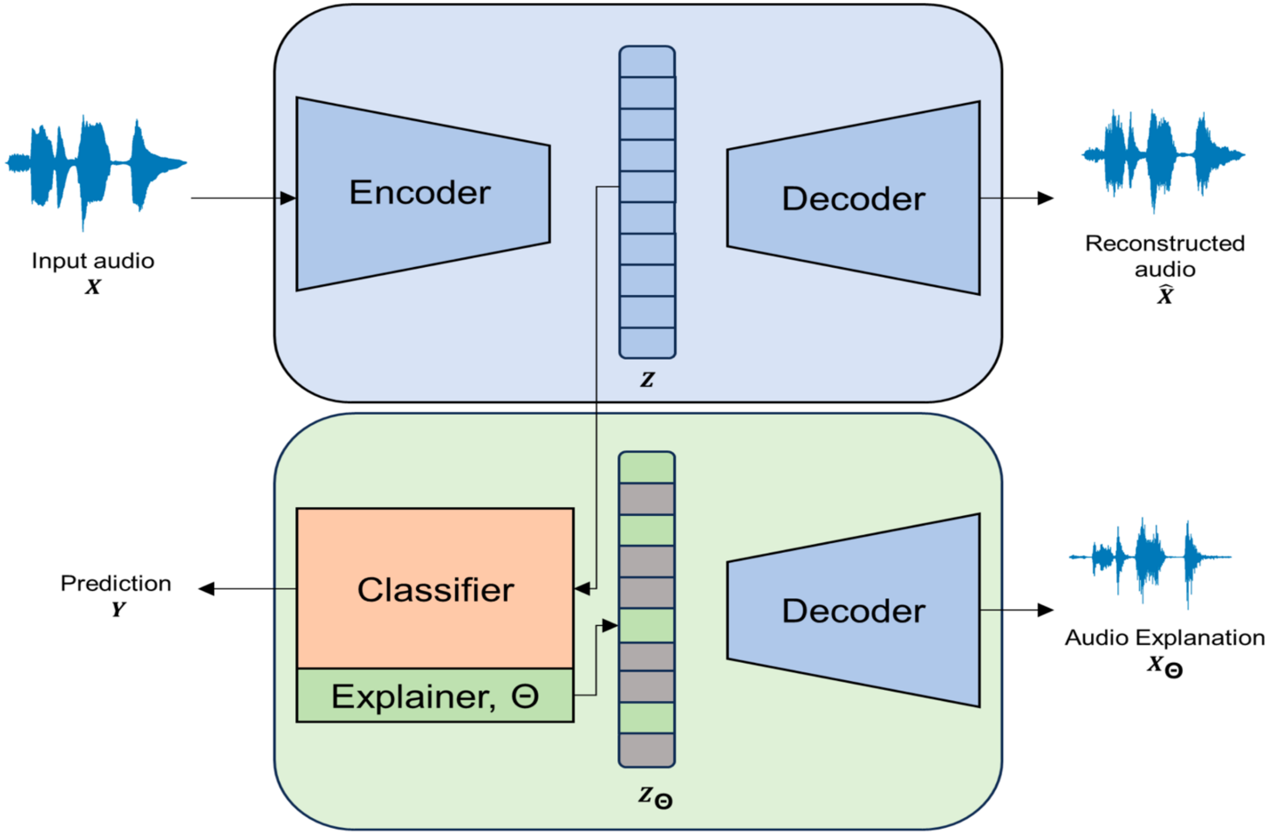
*“AI-based emotion recognition: Promise, peril, and prescriptions for prosocial path”, **arXiv**, 2022.*

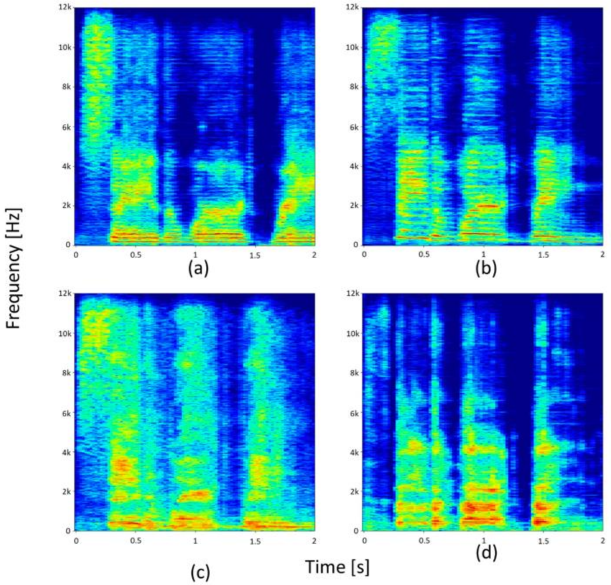
Ethics.

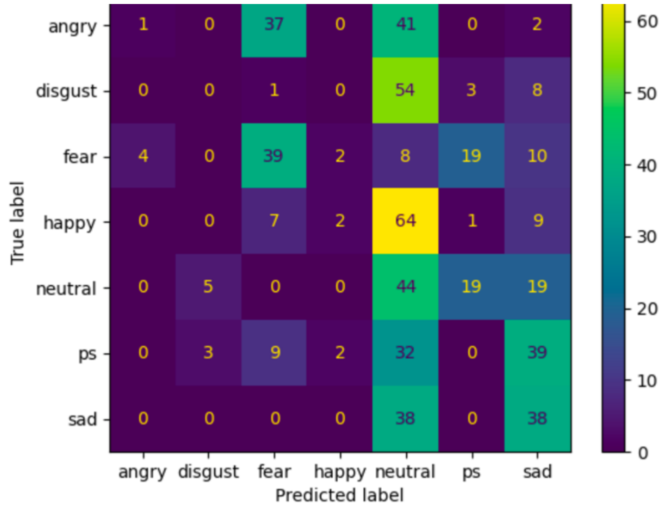
TEQUILA	Recommendation
Trust	Data security, privacy, and transparency in AI are crucial
Evidence	Robust evidence-based design is crucial
Quality	Ongoing regulatory oversight and continuous evaluation, including long-term user feedback and clinical validation
Usability	Ensuring ease of use and accessibility is crucial
Interest	Interventions must prioritise end-users’ interests, involves addressing who funds, controls, and uses data, and incorporating peer support to ensure inclusivity
Liabilty	Legal considerations, especially around liability, are critica
Accredetation	Accreditation by relevant boards for ’Dr. AI’ are essential

“Digital interventions in mental health: An overview and future perspectives”, ***Internet Interventions***, *2025.*

Explanations.



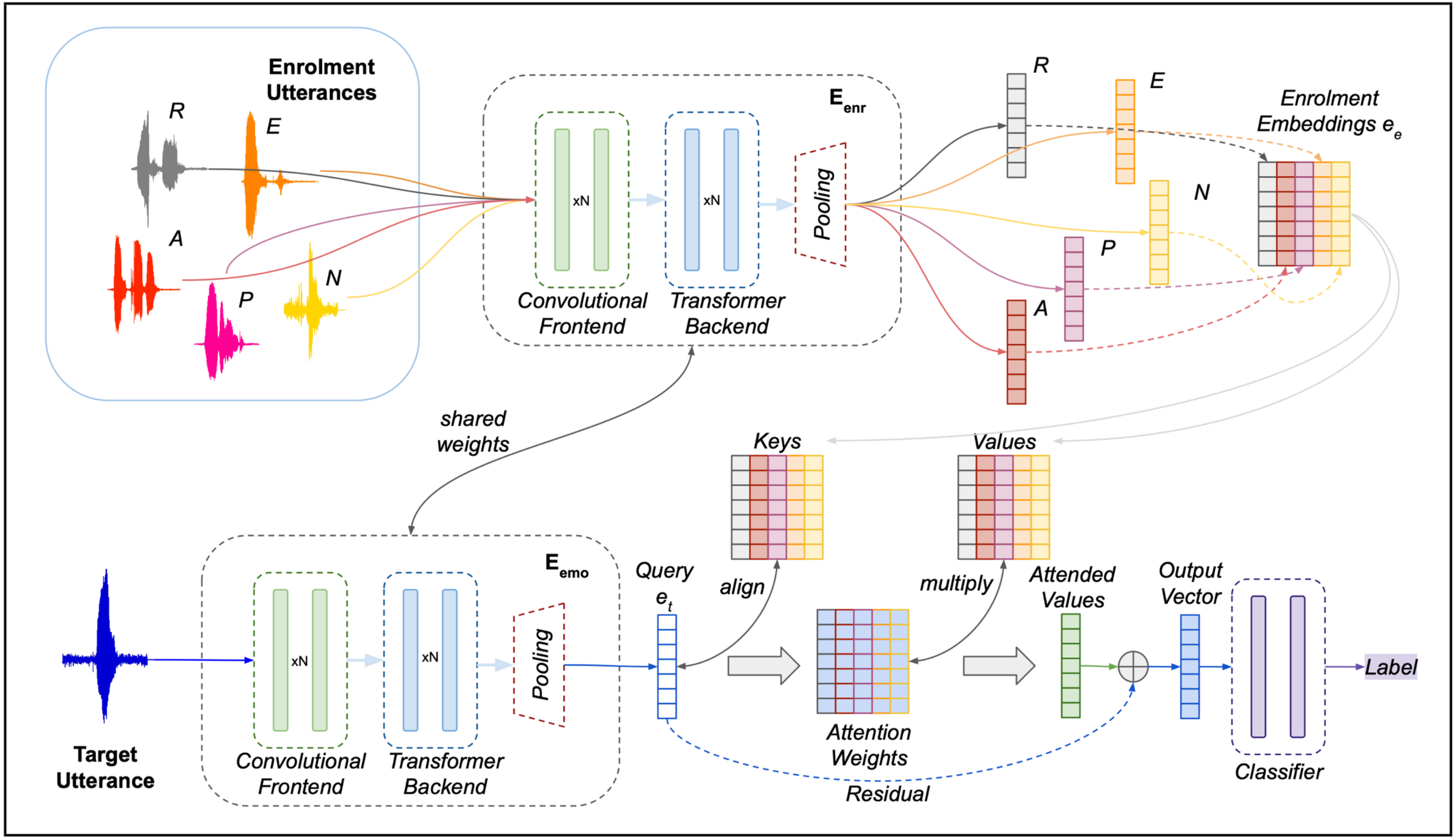




True label \ Predicted label	angry	disgust	fear	happy	neutral	ps	sad
angry	1	0	37	0	41	0	2
disgust	0	0	1	0	54	3	8
fear	4	0	39	2	8	19	10
happy	0	0	7	2	64	1	9
neutral	0	5	0	0	44	19	19
ps	0	3	9	2	32	0	39
sad	0	0	0	0	38	0	38

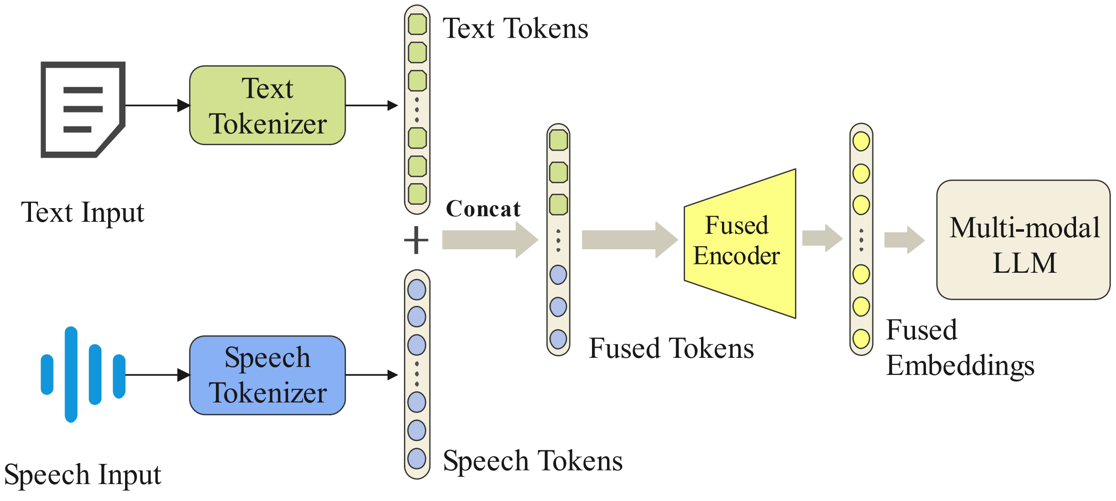
*“Audio Explanation Synthesis with Generative Foundation Models”, **ICASSP**, 2025.*

Fairness.

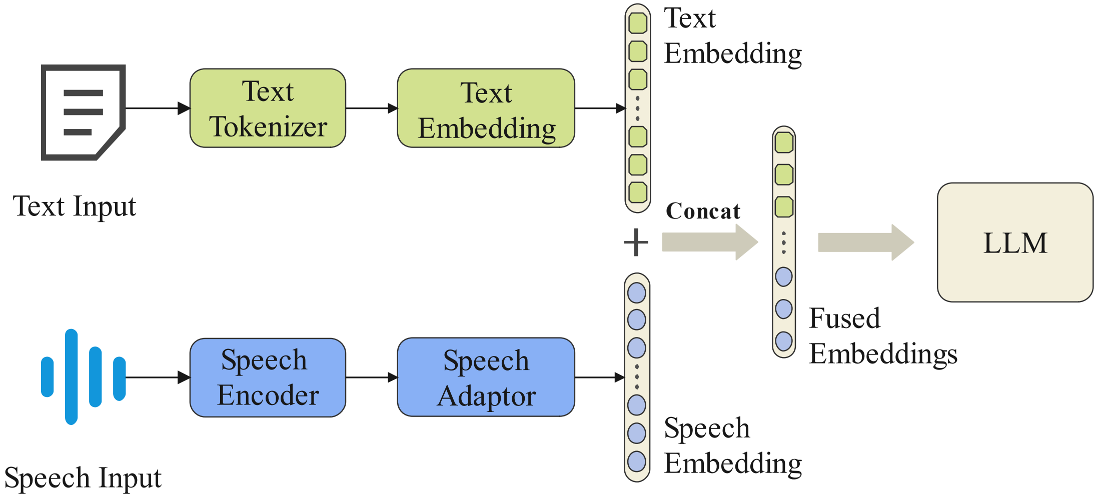


*"Enrolment-based personalisation for improving individual-level fairness in speech emotion recognition", **Interspeech**, 2024.*

SLMs: Train your local own?


Text Input
Text Tokenizer
Text Tokens
Concat
Speech Input
Speech Tokenizer
Speech Tokens
Fused Tokens
Fused Encoder
Fused Embeddings
Multi-modal LLM

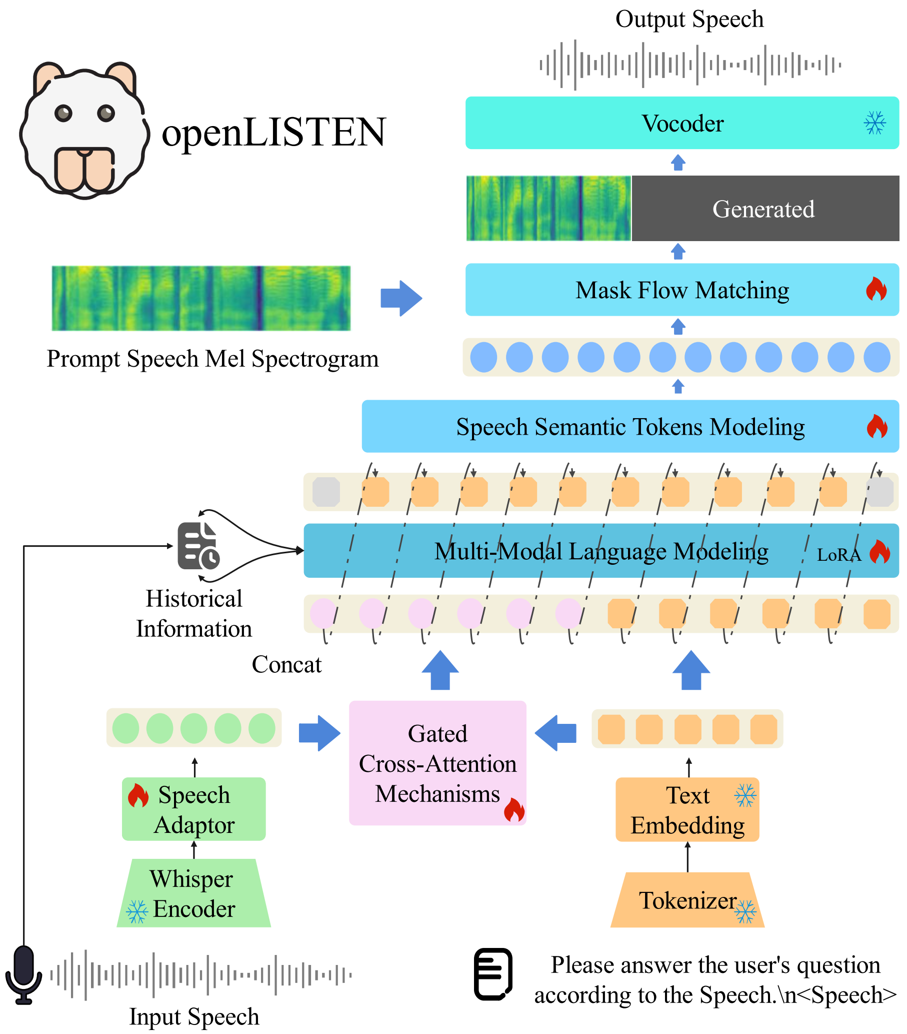
(a) Vocabulary Expansion Method


Text Input
Text Tokenizer
Text Embedding
Text Embedding
Concat
Speech Input
Speech Encoder
Speech Adaptor
Speech Embedding
Fused Embeddings
LLM

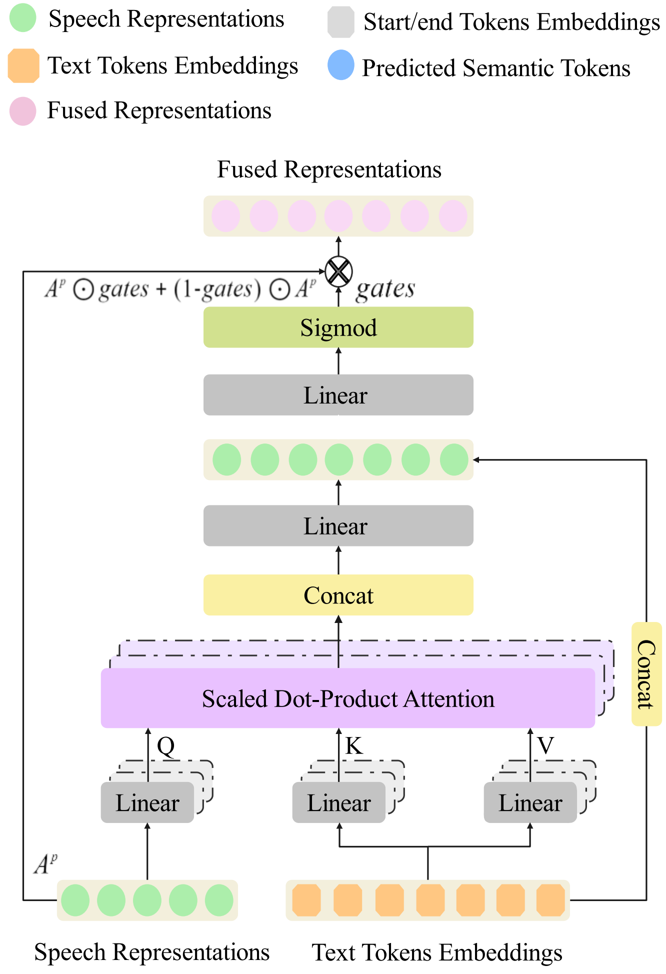
(b) Direct Projection Method

*"openLISTEN: Freestyle Cross-Modal Instruction Compliance for Large Speech-Language Models with Limited Resources", **submitted**.*

SLMs: Train your local own?


openLISTEN
Output Speech
Vocoder
Generated
Prompt Speech Mel Spectrogram
Mask Flow Matching
Speech Semantic Tokens Modeling
Multi-Modal Language Modeling
LoRA
Historical Information
Concat
Gated Cross-Attention Mechanisms
Speech Adaptor
Whisper Encoder
Text Embedding
Tokenizer
Input Speech
Please answer the user's question according to the Speech.\n<Speech>

(a) Overall Structure



(b) Gated Cross-Attention Mechanisms structure

*"openLISTEN: Freestyle Cross-Modal Instruction Compliance for Large Speech-Language Models with Limited Resources", **submitted**.*

THE END

Gracias

Next Episode:
Emotional AI.