

Recognizing Emotion Regulation Strategies from Human Behavior with Large Language Models

Philipp Müller
DFKI

Saarbrücken, Germany
philipp.mueller@dfki.de

Alexander Heimerl
Augsburg University

Augsburg, Germany
alexander.heimerl@uni-a.de

Sayed Muddashir Hossain
DFKI

Saarbrücken, Germany
sayed_muddashir.hossain@dfki.de

Lea Siegel
DFKI

Saarbrücken, Germany
lea.siegel@dfki.de

Jan Alexandersson
DFKI

Saarbrücken, Germany
jan.alexandersson@dfki.de

Patrick Gebhard
DFKI

Saarbrücken, Germany
patrick.gebhard@dfki.de

Elisabeth André
Augsburg University

Augsburg, Germany
elisabeth.andre@uni-a.de

Tanja Schneeberger
DFKI

Berlin, Germany
tanja.schneeberger@dfki.de

Abstract—Human emotions are often not expressed directly, but regulated according to internal processes and social display rules. For affective computing systems, an understanding of how users regulate their emotions can be highly useful, for example to provide feedback in job interview training, or in psychotherapeutic scenarios. However, at present no method to automatically classify different emotion regulation strategies in a cross-user scenario exists. At the same time, recent studies showed that instruction-tuned Large Language Models (LLMs) can reach impressive performance across a variety of affect recognition tasks such as categorical emotion recognition or sentiment analysis. While these results are promising, it remains unclear to what extent the representational power of LLMs can be utilized in the more subtle task of classifying users' internal emotion regulation strategy. To close this gap, we make use of the recently introduced DEEP corpus for modeling the social display of the emotion shame, where each point in time is annotated with one of seven different emotion regulation classes. We fine-tune Llama2-7B as well as the recently introduced Gemma model using Low-rank Optimization on prompts generated from different sources of information on the DEEP corpus. These include verbal and nonverbal behavior, person factors, as well as the results of an in-depth interview after the interaction. Our results show, that a fine-tuned Llama2-7B LLM is able to classify the utilized emotion regulation strategy with high accuracy (0.84) without needing access to data from post-interaction interviews. This represents a significant improvement over previous approaches based on Bayesian Networks and highlights the importance of modeling verbal behavior in emotion regulation.

Index Terms—emotion regulation, large language models, emotion recognition, bayesian networks

I. INTRODUCTION

One key finding of emotion research is that there is no one-to-one mapping of displayed emotional expressions to internally experienced emotions [1]. Emotions do not necessarily become visible [2], nor consciously experienced [3]–[5]. One reason for this is emotion regulation, which encompasses various conscious or unconscious strategies that individuals use to influence their emotional experience [6]. Especially unpleasant emotions such as shame are regulated to protect the self [3]–[5]. For many affective computing systems, knowledge of users' emotion regulation strategies has the potential to be

highly useful. Such systems include social skill training systems [7]–[9] or therapeutic assistance systems [10], [11]. The recently introduced DEEP approach was the first attempt to create a computational model of emotion regulation, focusing on the emotion shame elicited in job interviews [12]. While the authors presented a Bayesian Network (BN) model to classify emotion regulation strategies, their approach had two key limitations prohibiting application in realistic scenarios. First, they require results of an extensive analysis of in-depth post-interaction interviews as input. Second, they did not evaluate their model in a subject-independent scenario.

Recent studies indicate that generative large language models (LLMs) are able to, in a certain sense, understand human emotion in social situations. In zero-shot scenarios, GPT3.5 and GPT4 were successfully applied across a variety of emotion-related tasks, including sentiment analysis, emotion and emotion cause recognition, toxicity detection, and opinion extraction, albeit they are often still outperformed by approaches directly trained on the respective tasks [13], [14]. In contrast to zero-shot scenarios, instruction-tuning is an effective means to utilize the representational power of generative LLMs and at the same time adapt to a specific target task [15]–[19]. Using Low-rank Adaptation (LoRA) [20], this process is computationally efficient, and was already utilized for tasks related to affect and social behavior [21]–[23]. In particular, DialogueLLM [21] reached state-of-the-art results for emotion recognition on the MELD [24], IEMOCAP [25], and EmoryNLP [26] datasets. While these results are encouraging, it is unclear to what extent instruction-tuned LLMs can be used to classify emotion regulation strategies. In contrast to expressions of emotion, these strategies reflect inner processes that may not have distinct observable cues and are believed to be heavily related to nonverbal aspects of behavior [12].

In our work, we investigate to what extent instruction-tuned LLMs are capable of classifying the strategies employed by humans to regulate shame. To this end, we make use of the recently introduced DEEP corpus comprising recordings of human behavior in shame inducing situations and self-reported

information about individual experience [12]. Inspired by DialogueLLM [21], we encode participants’ multimodal behavior into prompts that are used for instruction-tuning Llama2-7B [27], [28] and Gemma [29] models with LoRA [20]. We present the first cross-user evaluations on the DEEP corpus and show that our LLM-based approach can reach an accuracy of 0.84 in emotion regulation classification without access to any information from informative but impractical post-interaction interviews. As such, our results represent an important step towards affective computing systems that can recognize human emotion regulation strategies in realistic scenarios.

Our specific contributions are three-fold.

- 1) We utilize LLMs instruction-tuned on prompts incorporating multimodal behavior to classify peoples’ strategies to regulate the emotion shame.
- 2) In the first cross-user evaluations on the recently introduced DEEP corpus [12], our approach outperforms the previous state of the art based on expert-constructed Bayesian Networks when information from post-interaction interviews is not available.
- 3) We conduct extensive ablation experiments, highlighting the impact of different modalities on performance.

II. RELATED WORK

A. Model of Emotions and Emotion Regulation

There is a variety of emotion models both in psychology [30] and affective computing [31]. In our work, we follow a model of emotions that differentiates between internal and external components inspired by cognitive psychoanalysis [32]. *Internal components of emotions* are not directly observable as they represent individual experience occurring in humans’ inner worlds. Due to intrapersonal emotion regulation processes, the internal components may or may not be experienced consciously [33]. The intrapersonal emotion regulation, refers to how internal emotional components are managed [4]. It originates from psychoanalytical defense mechanism concepts and differs from the cognitive coping mechanism, which refers to a conscious-focused emotion regulation [34]. People regulate emotions to avoid or decrease experiential and/or behavioral aspects of negative emotions such as anger, sadness, and shame. Also positive emotions may be regulated – for example, if the social situation requires it. The result of intrapersonal emotion regulation is the *experienced component of emotions* and can be seen as the emotional information that is “bearable” within the related situation [35]. *External components of emotions* represent communicated information that regulates relationships with others and how they are experienced and represented internally. What is communicated externally is i.a. influenced by social display rules [36]. Due to both, intrapersonal emotion regulation and social display rules (interpersonal emotion regulation), the connection between internal and external components is not immediate and they do not necessarily match [1], [32].

For modeling human emotions computationally, computer scientists focused on cognitive appraisal theories for emotions

[37]. Some models take emotion regulation into account. One example is MARSSI [38], which models appraisal rules, emotion regulation rules, and social signal interpretation, and allows to define multiple possible and plausible relations between these components. Furthermore, MARSSI differentiates between internal and external components of emotions.

Recently, [12] presented the DEEP method, a cognition-based method that focuses on modeling the internal component of emotions. It incorporates an approach to query individual internal emotional experiences and to represent such information computationally. It combines social signals, with context information and information from a post-interaction interview (“verbalized introspection”). These different components were modeled with a Bayesian Network constructed from theoretical domain knowledge. They also presented first prediction results for the emotion regulation strategy employed by users. However, their approach is limited in two key aspects which makes it impractical in many application scenarios. First, it requires knowledge from the post-interaction interview, and second, it was not evaluated in a cross-subject scenario. In contrast, we present instruction-tuned LLMs that are able to predict emotion regulation strategies with high accuracy in a cross-subject setting and without having access to information from the verbalized introspection collected post-interaction.

B. LLMs and Emotion Understanding

Large language models (LLMs) have been applied to a variety of tasks related to human affect expression [39], [40]. One of the most popular of these tasks is sentiment analysis, which commonly involves classifying text into expressing a positive, negative, or neutral sentiment. Transformer-based LLMs such as BERT, RoBERTa, or XLNet have been a key component of state-of-the-art sentiment analysis approaches in recent years [41]–[43]. With the success of generative LLMs such as GPT-3.5, GPT4, or Llama, researchers have investigated their utility for sentiment analysis, mainly in zero-shot and few-shot scenarios [40], [44]. A slightly more complex task compared to sentiment analysis is categorical or dimensional emotion recognition. Language models such as BERT or RoBERTa have been widely applied on these tasks [13], [45]. GPT3.5 was shown to reach good performance on emotion- and emotion cause recognition, but is still outperformed by models fine-tuned for the specific task [14]. GPT4 improved upon GPT3.5 and is able to outperform an approach based on RoBERTa on tasks such as toxicity detection and opinion extraction, but it still lacks behind on tasks with strong implicit components such as subjectivity of personality estimation [13]. Emotion recognition in GPT-like models operating in zero-shot scenarios can be highly biased with respect to ground truth definition, prompt construction, or label word selection [40].

Recently, instruction tuning of large language models has become a popular technique to adapt generative LLMs to new tasks [19]. By utilizing Low-rank Adaptation (LoRA) [20], fine-tuning models such as Llama2-7B became feasible on a single GPU. This approach was also utilized for tasks related to affect and social behavior [21]–[23]. In [22], authors used

LoRA to create an instruction-tuned variant of Llama2-7B on various social behavior analysis tasks including, among others, sentiment and emotion classification. In their experiments, instruction tuning leads to large performance gains relative to the standard Llama2 model. In [23], authors showed that instruction-tuned Llama2 models can clearly outperform all zero or few-shot approaches, including those based on GPT4 across a variety of affect recognition tasks. The utilized emotion datasets of these approaches are entirely textual however, i.e. they do not incorporate nonverbal behavior present in a face-to-face interaction. Despite the importance of nonverbal behavior for the expression of emotions, only few works have made attempts to include nonverbal behavior into the prompts given to LLMs [21], [46]. In [46], authors extracted textual descriptions from clusters of nonverbal behavioral features and used this information in addition to verbal input for sentiment analysis. DialogueLLM [21] classified emotions in conversation by constructing prompts describing the conversational and visual context, including nonverbal behavior of the interactants. They fine-tuned Llama2-7B on several emotion recognition datasets, and outperform the previous state of the art on MELD [24], IEMOCAP [25], and EmoryNLP [26]. As such, instruction-tuning of LLMs seems to be a promising way to model interactions between verbal- and nonverbal behavior for emotion understanding tasks. To the best of our knowledge, we for the first time apply instruction tuning on prompts generated from multimodal inputs to recognize emotion regulation strategies.

III. CORPUS

For our work, we utilize the recently introduced DEEP corpus, which we received upon request from the authors [12]. The corpus consists of shame-inducing situations during job interviews. It includes data from 20 expert-annotated videos of ten participants, each in two shame-eliciting situations, comprising 11535 video frames. Shame was elicited in mock job interviews framed as job interview trainings. During these, participants were confronted with a virtual job interviewer (avatar). To elicit shame in participants, the following validated, controlled and pre-evaluated situations [47] were:

- 1) After greeting the interviewee, the job interviewer says: “Before we start, a quick question. Where did you get that outfit? Somehow it doesn’t really suit you.” Following [5], this statement reflects the association *personal attractiveness* to the self.
- 2) After the interviewee has presented their experience, the interviewer reacts as follows: “All the other applicants have already said what you said. You haven’t exactly stood out.” Following [5], this statement reflects the association *Sense of self*.

After the interaction with the avatar, participants went through an interview reflecting about their experience, called the “verbalized introspection”. The DEEP corpus consists of data from different sources of information about each specific shame-eliciting situation. Annotations were done by three trained raters (all with a degree in psychology, one of

TABLE I
GROUND TRUTH CLASSES ON THE DEEP CORPUS [12], INCLUDING THEIR DEFINITION, AS WELL AS POSSIBLE EXPERIENCED COMPONENTS AND NONVERBAL BEHAVIOUR.

WITHDRAWAL (655 frames)	Cut off the current situation so there is no more external influence or stimuli. Wish to hide, leave or escape. Experienced emotional components: distress, fear Nonverbal Behavior: freezing, lip biting, gaze/head aversion, silence
ATTACK SELF (515 frames)	Do to yourself what others may do to you, establishing impression to control the situation. Experienced emotional components: disgust Nonverbal Behavior: facial expression of disgust
ATTACK OTHER (629 frames)	Transfer the diminishment of self-esteem to the person (object) who caused it by diminishing the other person. Experienced emotional components: anger Nonverbal Behavior: lean forward, gestures of power, facial expression of anger
AVOIDANCE (1650 frames)	Acting according the principle “fool others, fool myself”. Experienced emotional components: joy Nonverbal Behavior: gaze/head aversion, lean backwards, facial expression of joy/surprise, smile
DEPRECIATION (1911 frames)	Deevaluation of interaction partner due to different (or even contrary) values and ideals. Experienced emotional components: disgust, contempt Nonverbal Behavior: raised eyebrows, smile, facial expression of disgust and contempt
STABILIZE SELF (3593 frames)	Attempt to react in a way that is compliant with the (ideal) self by accepting disagreement between job interviewer and person. Experienced emotional components: pride Nonverbal Behavior: no display of uncertainty, direct gaze
REST (2582 frames)	No identified emotion regulation strategy.

them an experienced psychotherapist) based on the behavior of the participant in the shame-eliciting interview, the transcribed verbalized introspection, the context and the theoretical knowledge about shame and shame regulation. We utilize the annotations from [12] as ground truth as well as input features. As ground truth, we use the seven emotion regulation strategy classes (see Table I for an overview). In the DEEP corpus, primary and secondary emotion regulation strategy annotations exist, as – similar to emotions [48] – several emotion regulation processes can be active at the same time. For the purpose of this paper, we chose to focus on the primary emotion regulation strategy exclusively. The input features consist of annotations extracted from nonverbal behavior, verbalized introspection, personal context, and situational context (Table II). For the purpose of this paper, we transcribed participants’ verbal answers in the shame eliciting situations and added these to the situational context features. For further information on the corpus and the different annotations, we refer to the Supplemental Material of [12].

IV. APPROACH

We preset our approach based on instruction-tuned Large Language Models (LLMs), as well as our baseline implementation of the Bayesian Networks proposed in [12].

TABLE II
ANNOTATED INPUT FEATURES ON THE DEEP [12] CORPUS.

Nonverbal Behavior	Observation of external components of emotions that are encoded in social signals in the specific situation. <i>Speech, Utterance, Facial Expression, Gaze, Eyes, Smile, Smile Control, Head, Head Tilt, Upper body, Shame display</i>
Verbalized introspection	Self-reports that reflect a person’s subjective experience gathered in semi-structured interviews after the specific situation with the aid of video material of the experienced situation. <i>Relationship management, Shame awareness, Experienced emotion, Internal emotion component, Display rule</i>
Personal Context	Personal context variables. <i>Gender, Mindedness score</i>
Situational Context	Situational context variables. <i>Situation (first vs. second shame induction), Conversation transcript</i>

A. Multi-modal LLM Approach

Our approach uses instruction tuning to fine tune LLMs on prompts created from different sources of information, including verbal and nonverbal behavior as well as contextual information.

1) *Prompt generation:* We construct one prompt from every frame in the corpus. Similar to [21], we generate textual descriptions from different sources of information. An example prompt, broken down into components, is shown in Figure 1. In particular, we provide situational context by describing the particular shame induction situation, and providing a transcript of the verbal exchange up until the current frame. We also clearly define the utterance for which the model is supposed to classify the shame regulation strategy, i.e. the utterance corresponding to the current frame. The nonverbal behavior annotated on the Deep corpus at the current frame is directly translated to textual descriptions. E.g. annotation “TILT” for interviewee head behavior is annotated, that would translate to “The interviewee tilts their head to the side”. Finally, we add a textualization of the personal context variables. The results of verbalized introspection are not part of our default approach, however as we add them in certain experiments, they are included for reference in Figure 1. For the verbal prompt components, we translated the German transcripts on the DEEP corpus to English, using the mbart-large-50-many-to-many-mmt model [49]. This model’s multilingual capabilities enabled it to surpass the performance of several one-to-one translation models. Our experiments confirmed this; we initially tested the smaller opus-mt-en-de model [50], but its translations were notably inferior to those produced by the mbart-based model after careful review.

2) *Context information:* In addition to the prompt generated from each frame, we provide constant context information to the model, explaining the task, situation, and ground truth definitions (i.e. extended definitions of the shame regulation strategies shown in Table I). We include this context information in the supplementary material.

3) *Utilized LLMs:* We utilized a variant of the Llama LLM [27] specifically, the Llama-2-7b-chat-hf model [28]. We opted for this chat-oriented model as its fine-tuning on conversational data enhances its ability to understand the nuances of human dialogue. Preliminary experiments comparing the base Llama model and the chat variant supported this decision, with the

latter yielding superior results. Additionally, we incorporated the recent Gemma model [29] from Google DeepMind, which reached state-of-the-art performance across various NLP tasks.

4) *Training Details:* For training, the inputs were the Prompt and the Context. The relevant output was the emotion regulation strategy. To fine-tune our models, we applied the Low-Rank Adaptation of Large Language Models (LoRA) technique [20]. For LoRA, we follow previous work [51] and set $r = 8$, $\alpha = 16$ and dropout of 0.1. Further hyperparameters are documented in the supplementary material. We trained both models for 5 epochs. For Llama2-7B we were able to use 16 batches per device, for Gemma only 4. Further training details and code are available online¹. In total, we made use of three Nvidia A100 GPUs with 40GB VRam each: two for fine-tuning and one for test-time inference. Training lasted for about 2 weeks to generate all results in this paper.

5) *Testing:* For testing, the model was put first into inference mode, with zero temperature. Then, the Prompt and Context were fed to the model for each of the instances. We extracted the predicted emotion regulation strategy from the model’s response. We checked for anomalies in the response of the LLMs via string matching between the generated and the set of desired output classes, but both LLMs always predicted exactly one emotion regulation strategy label for each sample.

B. Bayesian Network Model

As a baseline comparison, we make use of the DEEP-BN approach proposed in [12]. This method conceptualizes a Bayesian network (BN) model representing the Internal Emotion Component, Emotion Regulation and related concepts (see Figure 2). Bayesian Networks are graphical models and, compared to other Machine Learning frameworks relatively easy to comprehend and therefore are ideal for modeling theory-based implications and their explanation.

In general, there are two types of nodes in the DEEP-BN. Blue nodes in the Figure represent information that is updated based on observations in the BN, red nodes represent information that is inferred by the BN. When it comes to understanding the internal emotions it is essential to model the interplay between the Internal Emotion Component, the process of Emotion Regulation, the Experienced Emotion Component, and related Social Signals.

¹https://git.opendfki.de/philipp.mueller/acii24_emotionregulationllm

Situational Context:

We are concerned with a moment in time in the first shame induction situation. The agent tries to induce shame by attacking the interviewee’s personal attractiveness: “Before we start, one short question: Where did you get this outfit? Somehow it doesn’t really suit you.”

The conversation history up to the current point is:
[Avatar] Where did you get this outfit from?
[Avatar] Somehow it doesn't really suit you.
[Interviewee] Don't you like it so much?
[Interviewee] I thought I felt very comfortable in it, and I find that when you feel comfortable, you always sell yourself a bit better and in the application situation I thought that makes the most sense.

The current utterance is:
[Interviewee] I thought I felt very comfortable in it, and I find that when you feel comfortable, you always sell yourself a bit better and in the application situation I thought that makes the most sense.

Nonverbal Behavior:

The interviewee shows the following nonverbal behavior at the current moment: The interviewee looks straight at the interviewer. The interviewee holds their head straight. The interviewee tilts their head to the side. The interviewee shows a non-Duchenne smile, i.e. a smile that concentrates only on the mouth. The interviewee is speaking. The upper body is moved forwards

Verbalized Introspection:

The following information was gathered from the qualitative interview after the interaction: The interviewee experiences the following internal emotion at the current moment in time: shame/shyness. The interviewee was aware of feeling ashamed during the current moment in the job interview. During the qualitative interview, the interviewee became aware that they were having the emotion shame during the current moment in the job interview. The interviewee has the intention to maintain the relationship with the avatar.

Personal Context:

The following additional personal information was collected from the interviewer: The mindedness score of the interviewee is 4,77. The interviewee is female.

Fig. 1. Example prompt consisting of situational context, nonverbal behavior, verbalized introspection and personal context. The situational context incorporates a transcript (below the dotted line).

The Internal Emotion Component represents possible emotion classes that – depending on the Emotion Regulation – may or may not result in a consciously Experienced Emotion Component. It is possible that individuals do not apply strong Emotion Regulation which results in a match between the Internal Emotion Component and the Experienced Emotion Component. However, it may also be that the Emotion Regulation is strong and unconscious resulting in a completely different Experienced Emotion Component compared to the Internal Emotion Component (e.g., Experiencing anger when unconsciously applying the Emotion Regulation strategy Attack Other but not shame) (see Sec. II-A). The Social Signals represent the observable result of the underlying Experienced Emotion Component and applied Emotion Regulation.

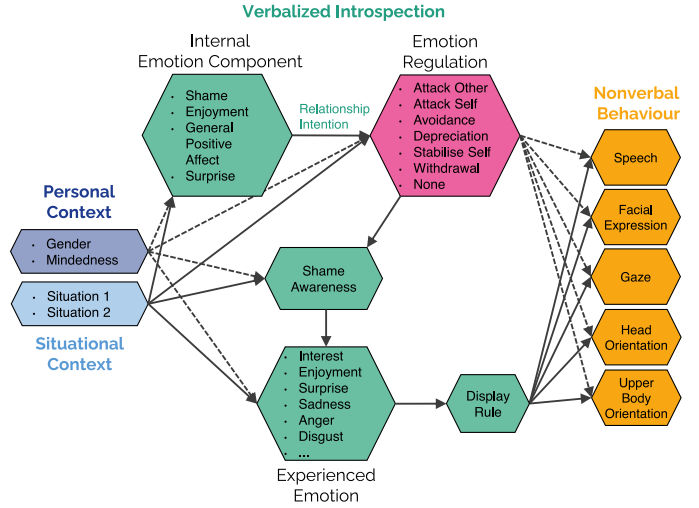


Fig. 2. The DEEP-BN schema constructed based on the DEEP method information. The ground truth of emotion regulation is also part of the verbalized introspection. But since it represents the ground truth (and not a potential input to the model), it is colored pink. The emotion regulation strategies are based on [5], while internal emotion component and the experienced emotion are based on the Differential Emotions Scale [52] and PANAS-X [53].

The Internal Emotion Component, the Emotion Regulation and the Experienced Emotion Component are influenced by the Personal Context, for example, demographic aspects (e.g., gender), or personality aspects (e.g., mindedness), as well as the Situational Context (i.e. the shame-inducing situation).

The BN we built based on the DEEP method acts as a benchmark to investigate the capabilities of LLMs to recognize emotion regulation strategies. Even though the main focus of the DEEP method is to provide a deeper understanding of the Internal Emotion Component, the architecture of a BN allows us to easily change the inference target from predicting internal emotions to predicting emotion regulation strategies given a specific emotion.

V. EVALUATION

We trained and evaluated the approaches discussed in section IV on the task of classifying the user’s emotion regulation strategy for each frame during the shame induction situations on the DEEP corpus [12]. To assess the generalizability of the models we employed a LOSO (leave-one-subject-out) evaluation. We evaluate our models in two general settings: (1) with verbalized introspection, i.e. including the information gathered from the post-interaction interview, and (2) without verbalized introspection. While we expect the first setting to reach higher performance, the second setting respects the demands of application scenarios, where it is usually impractical to perform an additional interview with the user.

A. Overall Results

Table III presents the models’ accuracy and F1 score for each class for our two evaluation scenarios, i.e. with verbalized introspection, and without verbalized introspection. When considering all available information including the verbalized

TABLE III
ACCURACY AND F1-SCORE OF DIFFERENT MODELS FOR EMOTION REGULATION RECOGNITION ON DIFFERENT GROUND TRUTH CLASSES, AS WELL AS OVERALL. AS OVERALL F1 SCORE, WE REPORT THE WEIGHTED F1 SCORE.

	Withdrawal		Attack self		Attack other		Avoidance		Depreciation		Stabilize self		Rest		Overall	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
<i>w/ verb. introspection</i>																
Bayesian Net	0.99	0.94	0.99	0.91	0.99	0.93	0.99	0.99	0.98	0.94	0.98	0.98	0.96	0.91	0.96	0.96
Gemma	0.98	0.83	0.99	0.86	0.98	0.86	0.98	0.92	0.98	0.93	0.97	0.95	0.98	0.95	0.93	0.93
Llama2-7B	0.99	0.88	0.97	0.69	0.98	0.84	0.98	0.94	0.96	0.89	0.95	0.92	0.95	0.88	0.89	0.89
<i>w/o verb. introspection</i>																
Bayesian Net	0.81	0.21	0.88	0.0	0.89	0.08	0.79	0.33	0.65	0.13	0.69	0.34	0.72	0.26	0.23	0.25
Gemma	0.94	0.56	0.94	0.55	0.94	0.57	0.92	0.70	0.90	0.70	0.88	0.78	0.90	0.76	0.71	0.72
Llama2-7B	0.97	0.76	0.97	0.71	0.96	0.71	0.96	0.85	0.95	0.84	0.93	0.88	0.95	0.88	0.84	0.84

TABLE IV
WEIGHTED ACCURACY AND WEIGHTED F1-SCORE OF BAYESIAN NETWORKS FOR EMOTION REGULATION RECOGNITION CONTAINING DIFFERENT MODALITIES

Input Modalities	Bayesian Net		Llama2-7B		Gemma	
	ACC	F1	ACC	F1	ACC	F1
<i>w/ verb. introspection</i>						
All	0.96	0.96	0.89	0.86	0.93	0.93
No personal context	0.69	0.68	0.88	0.88	0.93	0.93
No situational context	0.84	0.85	0.49	0.51	0.61	0.63
No transcript	—	—	0.45	0.47	0.63	0.64
No nonverbal behavior	0.06	0.01	0.87	0.87	0.87	0.87
Only verbalized introspection	0.17	0.16	0.54	0.56	0.54	0.56
<i>w/o verb. introspection</i>						
All	0.23	0.25	0.84	0.84	0.71	0.72
No personal context	0.26	0.27	0.44	0.46	0.45	0.47
No situational context	0.22	0.23	0.38	0.40	0.35	0.38
No transcript	—	—	0.40	0.42	0.34	0.37
No nonverbal behavior	0.25	0.28	0.42	0.44	0.44	0.46
Only nonverbal behavior	0.25	0.25	0.47	0.50	0.44	0.46

introspection the three models achieved excellent accuracy and F1 scores. However, the BN slightly outperformed the two LLMs in terms of overall accuracy and F1 score, with 0.96 and 0.96 respectively. The BN achieved the highest accuracy and F1 scores for all classes except the Rest class, here the Gemma model was able to surpass the BN with an accuracy of 0.98 and F1 score of 0.95. However, when excluding the information about the verbalized introspection the predictive performance of the BN heavily decreased. The BN was only able to achieve an overall accuracy of 0.23 and F1 score of 0.25. In contrast to that, the LLMs were still able to largely maintain their performance. The Llama2-7B model outperformed the Gemma model for both metrics with an accuracy of 0.84 and a f1-score of 0.84 in comparison to an accuracy of 0.71 and F1 score of 0.72. In addition to comparing the predictive performance of the three models when including or excluding the information about the verbalized introspection we also investigated the influence of the other modalities on the recognition scores. When inspecting the per-class F1 scores we observe a slight trend towards lower performances for less frequent classes across all models. In the case of Llama2-7B without verbalized introspection, F1 scores for Withdrawal (655 frames), Attack self (515 frames), and Attack other (629 frames) are between 0.71 and 0.76, whereas for the remaining classes (each > 1500

frames) they range from 0.84 to 0.88.

In preliminary experiments, we investigated the feasibility of a zero-shot approach without instruction tuning based on Llama2-7B. We made two observations. First, we were not able to instruct the model to output a classification decision instead of a text generation, making this approach impractical for full-scale quantitative evaluations. Second, on a small test set of five samples from each ground truth class, we observed that the model’s outputs are highly biased: in 30 out of 35 cases the model predicted Stabilize self.

B. Ablation Results

Table IV displays the overall accuracy and weighted F1 score for the three classifiers considering different modalities. When considering verbalized introspection and all other available modalities we already reported that the BN performed the best with the highest scores overall. However, when removing information about nonverbal behavior or even only considering the verbalized introspection the recognition scores of the BN drastically decrease. The removal of nonverbal behavior has very little influence on the accuracy and F1 score of both LLMs. But removing situational context (which includes the transcript) leads to a noticeable decrease in prediction performance for the Llama2-7B and Gemma models. In fact, the accuracy and F1 score similarly decrease as when excluding

the transcript only (but keeping the information about the shame inducing situation), indicating that the key information the LLMs utilize is users’ verbal behavior. For the BN, the information about the situational context is less important to correctly predict the emotion regulation strategies. Without access to verbalized introspection, the BN only reaches F1 scores between 0.23 to 0.28, while the LLMs can better maintain their performance. As in the condition with available verbalized introspection, removal of situational context or transcript impacts the LLMs most.

VI. DISCUSSION

A. On Performance

While the Bayesian Network based approach achieved the highest performance when all modalities including verbalized introspection were available, the LLMs were much more robust when modalities were removed. Especially the fact that LLMs proved to be relatively robust to the removal of verbalized introspection information makes them a decidedly better choice in application scenarios where post-interaction interviews are impractical, or online prediction is desired.

It is crucial to acknowledge that the Bayesian Network (BN) does not include the raw transcript of the job interview, but a distilled representation of the data sourced from the job interview and verbalized introspection. The distillation can be beneficial, especially if it encapsulates the most pertinent information required to identify affective states. However, there’s a risk that during this process, potentially relevant details may be excluded. Thus, depending on the quality of abstraction, the removal of modalities may have a less or more detrimental impact on the performance of the BN. For example, the internal emotion component appears to contain the most relevant information by representing the extracted emotion classes. In our case, leaving out the information associated with that component has a detrimental impact on the performance of the BN which cannot be compensated by the information associated with the situational context. This underscores the critical nature of the abstraction approach, particularly in how omitted information impacts the comparative efficacy of the BN and LMM in emotion recognition tasks.

LLMs bear the advantage that they are able to access semantic information from the transcripts of the job interviews. This information enables them to leverage additional nuanced information crucial for affect recognition while the BN has only access to this information in terms of abstract representations gathered from the verbalized introspection. While the BN benefits from incorporating a theory-driven emotion model, the advantage of such a model is contingent upon its access to relevant information resulting from verbal introspection.

B. Limitations and Future Work

While our results represent an encouraging step towards emotion regulation recognition in realistic scenarios, several limitations remain. Due to the need for verbalized introspection and the complexity of the annotations, the DEEP corpus is limited in size and variability. The ten participants

were all having the same cultural background, similar age and were pre-selected having good skills to reflect on their internal experiences. Therefore, the full range of emotion regulation strategies and associated nonverbal behavior may not be captured, which may limit the generalizability of our findings. The reduction of effort by using an LLM to predict emotion regulation strategies, where verbalized introspection seems to be less crucial, seems promising. It would allow for more economical data collection and annotation for future work investigating emotion regulation strategies, however the accuracy of the LLM predictions need to be rigorously evaluated in any new scenario.

This paper focuses on emotion regulation in validated shame-eliciting situations, limiting the extension of the work to situations where other emotion classes are elicited. Shame is an ideal starting point for this kind of research, both because of the existing extensive theoretical background describing shame regulation strategies [5], as well as due to the availability of the DEEP corpus. However, emotions are not only regulated in shame eliciting situations, as most (if not all) emotions are intrapersonally regulated [33]. Therefore, future work should extend the application of this proposed hybrid approach to other emotion classes, to gain an overall deeper understanding of individual emotional experiences.

Finally, while our proposed approach allows to automatically infer emotion regulation strategies from behavioral descriptions, the descriptions provided with the DEEP dataset were manually annotated. Future work should replace such manual steps with automatic methods. While this might not be easy to do for features extracted from the verbalized introspection, automatic methods to detect facial behavior [54], body language [55], and to recognize speech [56] are available. When using automatic approaches, the set of nonverbal behaviors can also easily be extended, e.g. by detecting backchannels [57], or analyzing prosody [58].

VII. CONCLUSION

In this paper, we presented the first evaluation of instruction-tuned large language models (LLMs) on the task of recognizing the strategy employed to regulate the emotion shame. We utilized the recently introduced DEEP corpus of shame-inducing situations during job interviews, which is annotated with multi-modal behaviors and verbalized introspection gathered after the shame-inducing interactions. Our results indicate that while theory-driven Bayesian Networks perform best when all information is available, LLMs can cope much better with missing information from the verbalized introspection, likely due to their capability to effectively make use of users’ verbal behavior. As such, our insights are an important building block towards affective computing systems able to recognize emotion regulation strategies in realistic scenarios.

ETHICAL IMPACT STATEMENT

The paper employs data from the recently introduced DEEP corpus, which we received upon request from the authors. The DEEP corpus includes recordings of human behaviors

in job interviews and subsequent verbal introspection. The collection and analysis of such data involves processing personal and potentially sensitive data. Furthermore, it exposes participants to shameful situations which may lead to negative emotional states. It is crucial to obtain informed consent from the participants, ensure that the employed stimuli don't negatively affect their mental health, implement robust data protection measures and only collect data necessary for the intended affect recognition purposes. Approval for collecting and processing these data was obtained from the ethical review board of the DEEP corpus' authors. The current analysis of multimodal interview data and subsequent verbal introspection is covered by the ethics' approval for the DEEP corpus.

Our model contributes to endeavors aimed at deciphering internal states, particularly benefiting Affective Computing systems reliant on discerning user emotions, such as social training systems or therapeutical assistants. However, the potential for misapplication raises pertinent privacy concerns.

In our investigation, we solicited insights from participants concerning their internal experience in shame-eliciting situations. Although participants provided consent for research purposes, in practical scenarios, individuals may withhold consent due to apprehensions surrounding the exposure of their internal experiences. Such reluctance could engender adverse ramifications for social interactions and interpersonal relationships.

Prior to engaging with systems employing models for interpreting observable expressions and internal states, it is imperative that users are adequately informed and provide consent regarding functionality, data collection, processing, and attendant risks. The utilization of such systems without the informed consent of individuals subject to observation may result in deleterious outcomes. Unsanctioned application of such technologies may inadvertently gather deeply personal information about individuals' internal experiences, subsequently exposing them to potential harm to their social standing, privacy, and overall well-being.

ACKNOWLEDGMENT

P. Müller, S. Hossain, L. Siegel, and J. Alexandersson were partially funded by the European Union Horizon Europe programme, grant number 101078950. E. André, P. Gebhard, and T. Schneeberger were supported by the German Federal Ministry for Education and Research (BMBF) as a segment of the UBIDENZ project, under grant numbers 13GW0568D and 13GW0568F.

REFERENCES

- [1] L. F. Barrett, *How emotions are made: The secret life of the brain*. Pan Macmillan, 2017.
- [2] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame.," *Journal of Personality and Social Psychology*, vol. 68, no. 3, p. 441, 1995.
- [3] M. Lewis, "Self-conscious emotions: Embarrassment, pride, shame, and guilt.," in *Handbook of Emotions* (M. Lewis, J. M. Haviland-Jones, and L. Feldman Barrett, eds.), vol. 3, p. 742–756, Guilford, 2008.
- [4] J. J. Gross, *Handbook of emotion regulation*. Guilford, 2013.
- [5] D. L. Nathanson, *Shame and pride*. Norton, 1994.
- [6] J. J. Gross, "Emotion regulation: Past, present, future.," *Cognition & emotion*, vol. 13, no. 5, pp. 551–573, 1999.
- [7] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 697–706, 2013.
- [8] P. Gebhard, T. Schneeberger, E. André, T. Baur, I. Damian, G. Mehlmann, C. König, and M. Langer, "Serious games for training social skills in job interviews.," *IEEE Transactions on Games*, 2018.
- [9] T. Schneeberger, N. Sauerwein, M. S. Anglet, and P. Gebhard, "Stress management training using biofeedback guided by social agents," in *26th Conference on Intelligent User Interfaces*, pp. 564–574, 2021.
- [10] P. Gebhard, T. Schneeberger, M. Dietz, E. André, and N. u. H. Bajwa, "Designing a mobile social and vocational reintegration assistant for burn-out outpatient treatment," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 13–15, 2019.
- [11] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al., "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1061–1068, 2014.
- [12] T. Schneeberger, M. Hladký, A.-K. Thurner, J. Volkert, A. Heimerl, T. Baur, E. André, and P. Gebhard, "The deep method: Towards computational modeling of the social emotion shame driven by theory, introspection, and social signals.," *IEEE Transactions on Affective Computing*, pp. 1–16, 2023.
- [13] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, "A wide evaluation of chatgpt on affective computing tasks.," *arXiv preprint arXiv:2308.13911*, 2023.
- [14] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is chatgpt equipped with emotional dialogue capabilities?," *arXiv preprint arXiv:2304.09582*, 2023.
- [15] J. Li, K. Pan, Z. Ge, M. Gao, H. Zhang, W. Ji, W. Zhang, T.-S. Chua, S. Tang, and Y. Zhuang, "Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions.," in *The Twelfth International Conference on Learning Representations*, 2024.
- [16] R. Zhang, Y.-S. Wang, and Y. Yang, "Generation-driven contrastive self-training for zero-shot text classification with instruction-following llm.," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 659–673, 2024.
- [17] A. Borzunov, M. Ryabinin, A. Chumachenko, D. Baranchuk, T. Dettmers, Y. Belkada, P. Samygin, and C. A. Raffel, "Distributed inference and fine-tuning of large language models over the internet.," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [18] Y. Liu, H. He, T. Han, X. Zhang, M. Liu, J. Tian, Y. Zhang, J. Wang, X. Gao, T. Zhong, et al., "Understanding llms: A comprehensive overview from training to inference.," *arXiv preprint arXiv:2401.02038*, 2024.
- [19] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, et al., "Instruction tuning for large language models: A survey.," *arXiv preprint arXiv:2308.10792*, 2023.
- [20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models.," 2021.
- [21] Y. Zhang, M. Wang, P. Tiwari, Q. Li, B. Wang, and J. Qin, "Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations.," *arXiv preprint arXiv:2310.11374*, 2023.
- [22] G. Dey, A. V. Ganesan, Y. K. Lal, M. Shah, S. Sinha, M. Matero, S. Giorgi, V. Kulkarni, and H. A. Schwartz, "Socialite-llama: An instruction-tuned model for social scientific tasks.," *arXiv preprint arXiv:2402.01980*, 2024.
- [23] Z. Liu, K. Yang, T. Zhang, Q. Xie, Z. Yu, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis.," *arXiv preprint arXiv:2401.08508*, 2024.
- [24] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations.," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (A. Korhonen, D. Traum, and L. Márquez, eds.)*, (Florence, Italy), pp. 527–536, Association for Computational Linguistics, July 2019.
- [25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional

- dyadic motion capture database,” *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [26] S. M. Zahiri and J. D. Choi, “Emotion detection on tv show transcripts with sequence-based convolutional neural networks,” in *Workshops at the thirty-second AAAI Conference on Artificial Intelligence*, 2018.
- [27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” 2023.
- [28] A. Patil, D. Wu, R. Ong, S. Jain, and L. Huang, “meta-llama/Llama-2-7b-chat-hf: Llama 2.7B fine-tuned on Conversational data for Chatbot task.” <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>, 2024. Hugging Face Model Hub.
- [29] G. D. Gemma Team, “Gemma: Open models based on gemini research and technology,” tech. rep., Google DeepMind, 2024.
- [30] K. R. Scherer *et al.*, “Psychological models of emotion,” *The Neuropsychology of Emotion*, vol. 137, no. 3, pp. 137–162, 2000.
- [31] S. PS and G. Mahalakshmi, “Emotion models: a review,” *International Journal of Control Theory and Applications*, vol. 10, no. 8, pp. 651–657, 2017.
- [32] U. Moser and I. Von Zeppelin, “Die Entwicklung des Affektsystems,” *Psyche*, vol. 50, no. 1, pp. 32–84, 1996.
- [33] S. S. Tomkins, “Affect theory,” *Approaches to Emotion*, vol. 163, p. 195, 1984.
- [34] P. Cramer, “Defense mechanisms in psychology today: Further processes for adaptation,” *American Psychologist*, vol. 55, no. 6, p. 637, 2000.
- [35] D. L. Nathanson, *Shame and pride: Affect, sex, and the birth of the self*. Norton, 1992.
- [36] P. Ekman and W. V. Friesen, “The repertoire of nonverbal behavior: Categories, origins, usage, and coding,” *Semiotica*, vol. 1, pp. 49–98, 1969.
- [37] A. Moors, P. C. Ellsworth, K. R. Scherer, and N. H. Frijda, “Appraisal Theories of Emotion: State of the Art and Future Development,” *Emotion Review*, vol. 5, pp. 119–124, April 2013.
- [38] P. Gebhard, T. Schneeberger, T. Baur, and E. André, “MARSSI: Model of appraisal, regulation, and social signal interpretation,” in *Int. Conference on Autonomous Agents and MultiAgent Systems*, pp. 497–506, 2018.
- [39] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, “Emotional intelligence of large language models,” *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, 2023.
- [40] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, “The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection,” *IEEE Transactions on Affective Computing*, 2022.
- [41] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, “Unsupervised data augmentation for consistency training,” *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.
- [42] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Advances in neural information processing systems*, vol. 32, 2019.
- [43] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma, “Entailment as few-shot learner,” *arXiv preprint arXiv:2104.14690*, 2021.
- [44] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, “Is chatgpt a general-purpose natural language processing task solver?” *arXiv preprint arXiv:2302.06476*, 2023.
- [45] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh, “Dimensional emotion detection from categorical emotion,” *arXiv preprint arXiv:1911.02499*, 2019.
- [46] M. K. Hasan, M. S. Islam, S. Lee, W. Rahman, I. Naim, M. I. Khan, and E. Hoque, “Textmi: Textualize multimodal information for integrating non-verbal cues in pre-trained language models,” *arXiv preprint arXiv:2303.15430*, 2023.
- [47] T. Schneeberger, M. Scholtes, B. Hilpert, M. Langer, and P. Gebhard, “Can social agents elicit shame as humans do?,” in *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 164–170, IEEE, 2019.
- [48] P. L. Harris, “What children know about the situations that provoke emotion,” in *The socialization of emotions*, pp. 161–185, Springer, 1985.
- [49] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, “Multilingual translation with extensible multilingual pretraining and finetuning,” 2020.
- [50] J. Tiedemann and S. Thottingal, “OPUS-MT — Building open translation services for the World,” in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, (Lisbon, Portugal), 2020.
- [51] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [52] C. E. Izard, F. E. Dougherty, B. M. Bloxom, and N. E. Kotsch, *The Differential Emotions Scale: a Method of Measuring the Subjective Experience of Discrete Emotions*. Nashville, Tenn.: Vanderbilt Univ. Press, 1974.
- [53] D. Watson and L. A. Clark, “The panas-x: Manual for the positive and negative affect schedule-expanded form,” 1994.
- [54] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018.
- [55] M. Balazia, P. Müller, Á. L. Tánczos, A. v. Liechtenstein, and F. Bremond, “Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation,” in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 70–79, 2022.
- [56] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*, pp. 28492–28518, PMLR, 2023.
- [57] A. Amer, C. Bhuvaneshwara, G. K. Addluri, M. M. Shaik, V. Bonde, and P. Müller, “Backchannel detection and agreement estimation from video with transformer networks,” in *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2023.
- [58] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.