

Fine-grained Affective Processing Capabilities Emerging from Large Language Models

1st Joost Broekens
LIACS

Leiden University
Leiden, The Netherlands
joost.broekens@gmail.com

2nd Bernhard Hilpert
LIACS

Leiden University
Leiden, The Netherlands
b.hilpert@liacs.leidenuniv.nl

3rd Suzan Verberne
LIACS

Leiden University
Leiden, The Netherlands
s.verberne@liacs.leidenuniv.nl

4th Kim Baraka

Department of Computer Science
Free University Amsterdam
Amsterdam, the Netherlands
k.baraka@vu.nl

5th Patrick Gebhard

German Research Center
for Artificial Intelligence (DFKI)
Saarbruecken, Germany
patrick.gebhard@dfki.de

6th Aske Plaat

LIACS
Leiden University
Leiden, The Netherlands
aske.plaat@gmail.com

Abstract—Large language models, in particular generative pre-trained transformers (GPTs), show impressive results on a wide variety of language-related tasks. In this paper, we explore ChatGPT’s zero-shot ability to perform affective computing tasks using prompting alone. We show that ChatGPT a) performs meaningful sentiment analysis in the Valence, Arousal and Dominance dimensions, b) has meaningful emotion representations in terms of emotion categories and these affective dimensions, and c) can perform basic appraisal-based emotion elicitation of situations based on a prompt-based computational implementation of the OCC appraisal model. These findings are highly relevant: First, they show that the ability to solve complex affect processing tasks emerges from language-based token prediction trained on extensive data sets. Second, they show the potential of large language models for simulating, processing and analyzing human emotions, which has important implications for various applications such as sentiment analysis, socially interactive agents, and social robotics.

Index Terms—ChatGPT, Large Language Models, sentiment analysis, emotion representation, computational modeling of emotion, emotion elicitation.

I. INTRODUCTION

Affective Computing focuses on measuring, understanding and impacting human emotion [1]–[3]. While many approaches have been somewhat successful in doing so, a long-standing challenge in the field has been to appropriately use context for emotional understanding [4].

Large language models (LLMs), in particular generative pre-trained transformers (GPTs), show impressive results on a wide variety of language-related tasks [5]. This ability is driven by the fact that these models are able to take into account large contexts in the prediction of the next token [6].

Recently, current state-of-the-art GPT-based models allow easy interaction through prompting through the use of user interfaces, effectively providing a human-friendly way of performing zero-shot tasks [7]. And, large scale versions of such models also show impressive results on a diversity of tasks for which they were not fine-tuned [8].

Inspired by these important characteristics of GPT models we investigate the ability to perform affective computing tasks by prompting ChatGPT without any fine-tuning or examples. We address the following research questions and tasks:

- RQ1: How well can ChatGPT label sentiment, as values on the dimensions Valence, Arousal and Dominance, for textual descriptions of situations and for emotion words?
- RQ2: To what extent does ChatGPT represent the correct meaning of emotion words and affective dimensions, tested as its ability to map values to situations, and situations to emotion words?
- RQ3: How well can ChatGPT predict the emotion class in a situation based on a particular appraisal framework, in our case the Ortony-Clore-Collins (OCC) model [9]?

We perform a series of exploratory experiments involving prompting ChatGPT to solve the above tasks, as well as comparing the generated results to ground truth data gathered from the literature or through human expert ratings.

II. BACKGROUND AND RELATED WORK

A. Generative Pretrained Transformers

Generative Pretrained Transformers (GPT) are models trained to generate natural language text, based on the transformer architecture. While the original transformer was designed as an *encoder-decoder* architecture [6], GPT is a decoder-only model: given a prompt it generates text, word by word, based on all the previous context words. The power of transformer models to generate fluent and coherent natural language text stems from the multiple transformer layers with self-attention to compute the strength of the relation between each pair of words in a sequence. This enables the models to capture long-distance dependencies in text [5].

The ‘pretrained’ aspect of GPT models refers to the fact that they are pre-trained on huge amounts of text data; for that reason, they are commonly referred to as Large Language

Models (LLMs). This pre-training process is based on a language modeling task: training the model to predict the next word or masked words in a sentence. The first generation of transformer models, including the widely popular encoder models from the BERT-family [10], was very successful in a large range of Natural Language Processing (NLP) tasks thanks to the possibility of *fine-tuning* pre-trained models on labeled data.

Since GPT-3, models of the GPT family are capable of solving tasks without fine-tuning: by generating text given a prompt [8]. When the prompt does not contain any examples of the correct output (e.g., “for the sentences below, label the content as having a positive, negative, or neutral sentiment”), we call this zero-shot learning, prompt-based learning, or instruction tuning [11]. When the prompt contains a few examples (roughly 3-50) we call this few-shot learning.

In this paper, we experiment with OpenAI’s model ChatGPT. It is based on the large pre-trained GPT-3.5 model, finetuned on conversation data, and refined with reinforcement learning based on an expert-trained reward function [7].

B. Core Affective Computing Tasks

Affective Computing is a subfield of Artificial Intelligence whose main research includes the detection and interpretation of human affect, on the one hand, [12], [13], and the simulation and representation of affect on the other hand in both the human users and the social interactive agents [14], see also [3] and [15] for overviews. In this paper, we focus on the use of LLMs for Affective Computing by testing their zero-shot capabilities in automatic affect detection, emotion representation, and the computational elicitation of emotions.

1) *Automatic affect detection*: Automatic detection and interpretation of behavioral signals of affect involves a broad range of tasks including affect recognition from the face, body, speech, EEG and other physiological signals, and text [12], [13], [16], [17]. As of writing this paper, text is the only modality available in the interaction with ChatGPT. We focus on multidimensional (Valence, Arousal, Dominance) sentiment analysis from short situational descriptions and emotion words.

Valence refers to the positive-negativeness of an experience, Arousal to the calmness-excitation of that experience, and Dominance to the extent to which a person feels in control of the situation [18]–[20]. While sometimes referred to as Pleasure, we assume here that Valence and Pleasure are similar dimensions both referring to the positiveness-negativeness of an affective state/situation. We use the term Valence, being the “technical” term and avoiding LLM confusion with lexical or common meanings of the word pleasure.

2) *Affect representation*: Affect representation is typically based on psychological theories of emotion and affect [14], [21]. In particular, categorical or dimensional representations are used to represent an agent’s or human’s emotion. We focus on the numerical and latent representation of affect in terms of Valence, Arousal, and Dominance and a mapping from this representation to situations and (categorical) emotion words.

3) *Computational models of emotion elicitation*: Computational modeling of emotion elicitation is used to simulate affective states based on the appraisal of situations and mental states of an agent. It has been approached from different perspectives, including cognitive agent-based appraisal modeling, e.g. [22], embodied modeling (homeostasis-based) [23], hard-wired appraisal (event coding), and reinforcement learning [14], [24]. In emotion elicitation, the OCC appraisal model [9] has arguably had a major influence [25]. While the objective of the OCC model was to reverse-engineer a set of emotional experiences based on their cognitive antecedents rather than an exhaustive set of universal rules that represent human appraisal processes (e.g., p. 172), the logical structure of the model has been the basis for a number of computational implementations of the model in Affective Computing [21], [26], [27]. As such, in this paper, we focus on the OCC model of appraisal.

C. Related work on sentiment analysis

While many sentiment analysis methods are based on supervised fine-tuning (see [17] for a review), there is evidence that generative LLMs are capable of zero-shot prediction tasks, including sentiment classification [28], [29] and argument quality prediction [30]. It was shown that carefully engineering prompts helps with getting better task-specific guidance [31]. As such there is reason to believe LLMs will perform on more detailed affect extraction from text as well.

Most work in sentiment focuses on coarse-grained analysis of only Valence (e.g., 2- or 5-point scales) of opinions or reviews, or classification into several basic emotions [32]. We are interested in a more fine-grained analysis of situations, as this is a more natural test for the understanding of an emotion as experienced by a person. Our tasks/stimuli are therefore different from the sentiment classification tasks such as those used for the SemEval challenges [32]. We use psychologically validated and fine-grained dimensional ratings of situational descriptions on Valence, Arousal and Dominance [19].

Recent work investigated the performance of ChatGPT compared to pre-trained and fine-tuned models on a sentiment data set containing tweets with positive and negative labels [33]. The work reports reasonable performance as compared to fine-tuned models but only uses a coarse-grained dataset. There is only one prior work that evaluates automatic prediction of sentiment scores using the ANET data. The authors [34] finetune a multi-lingual XLM-RoBERTa-large model and evaluate it on a range of datasets in multiple languages. For the ANET texts (120 items), they report $\rho = 0.920$ for Valence and $\rho = 0.859$ for Arousal. Dominance was ignored.

III. METHOD

We are interested in the extent to which ChatGPT can perform the three above mentioned affective computing tasks in a zero-shot setting with instruction prompting.

A. Material

Our data set for RQ 1 consists of all 120 Affective Norms for English Text dataset (ANET) situations [19] and all 151 emotion words from Russel and Mehrabian [20].

As for RQ2, we focus on mapping stimulus sets to each other, with a human expert rater as ground truth. We used a subset of the words and situations (see Table I, II), each containing the 20 most reliably rated items, as assessed by the sum over the squared standard deviation (SD) for Valence, Arousal and Dominance (lower values are considered more reliable). The SD was given by [19] and [20].

For RQ3, we created a set of by-the-authors-formulated situations that reflect each of the tested emotions in the OCC model, see Table V. Throughout the experiment, we used the default model of ChatGPT 3.5 between February/March 2023.

B. Experimental Set-up

To address our RQs, we performed a series of conversational experiments with the model. To avoid confounding, a new chat session was initiated for each experiment. We explain the process for each of the RQs here.

1) *RQ1: Sentiment analysis*: For RQ1.1, we prompted¹ ChatGPT to perform sentiment analysis on Valence, Arousal, and Dominance (VAD), after which the 120 situations were entered as the next prompt (20 per session). The resulting values were correlated (per dimension) with the ground truth values provided by [19]. This serves to test the model’s performance on sentiment analysis from situational text.

In RQ1.2, we tested the model’s performance on sentiment analysis from emotion words in the same way. These values are correlated with the ground truth provided by [20].

2) *RQ2: Affect representation*: To test the model’s affect representation capability, we investigate if the model can use an affective representation in a constructive way.

For RQ2.1, we repeated RQ1, prompting ChatGPT to assign VAD-values to the 20 most reliable situations and emotion words but now in a single session. Then, we prompted the model to select for each situation the most fitting emotion word, *based on this numerical representation*. We computed the distance matrix between the emotion words and situation stimuli (based on the values generated by ChatGPT) and ranked the selected word according to the distance matrix. This serves to verify if ChatGPT is able to use a *numerical affective representation* to map two stimulus sets to each other.

For RQ2.2, we repeated the previous setup but provided ChatGPT with the situations and word lists without asking for a numerical representation. We then asked the model to pick for each situation stimulus the two most fitting emotion words. This classification was compared to a ground truth rated by an independent expert in emotion research. This serves to verify if ChatGPT is able to map two stimulus sets to each other, based on a *latent affect representation*.

For RQ2.3, we prompted ChatGPT to generate a new situation for 9 different value triplets that span the VAD space (1 neutral, 8 in each extreme). The generated situations are classified again by the same independent rater. This serves to verify if ChatGPT is able to generate new situations based on its latent representation of affect, prompted as values on dimensions.

3) *RQ3: Appraisal-based emotion elicitation*: To assess if ChatGPT can predict emotions according to a specific appraisal framework, we formalized a rule-based logical model of appraisal as a prompt, based on [9]. The OCC model provides an appraisal structure, including goals as well as global and local variables such as events, agents, and objects in order to describe the elicitation process of a select set of emotions in a rule-based description. The OCC authors note that the antecedents of components such as praiseworthiness of actions (i.e. standards) and appealingness of objects (i.e. attitudes) should not be seen as “internally consistent” (p.46) and cannot or should not be organized in a representational structure. Thus, for this experiment, we decided to focus only on the branch of event-related emotions of the OCC model.

After formalizing the event-branch of the model in a prompt, we asked ChatGPT to appraise the set of by-the-authors-formulated situations and checked if the corresponding emotion was reported. This serves to assess if ChatGPT can follow a precise logical structure for emotion elicitation, rather than an implicit latent representation: is it ‘programmable’?

IV. RESULTS

A. RQ1: Sentiment analysis

Pilot testing with prompts on the most reliable subset of situations (the 20 items subset) showed that without an explanation of the Dominance dimension, ChatGPT produces meaningless Dominance values while Valence and Arousal correlate well with the ground truth ($\rho = 0.98$ for Valence, $\rho = 0.91$ for Arousal and $\rho = -0.39$ for Dominance, see \hat{D}^* in Table I). As the correlation is negative, this could indicate confusion about the perspective of who experiences what. We added “remember that dominance assesses the extent to which the main person in the situation experiences the amount of control it can assert over the situation” (as taken from [18]).

As such, for RQ1, we used prompt 1:

Valence, Arousal and Dominance are three affective dimensions that you can use to identify the sentiment in sentences. Assume that these dimensions can take values between 0 and 1, with 0 being low, and 1 being high. Remember that dominance assesses the extent to which the main person in the situation experiences the amount of control it can assert over the situation. Assess according to these dimensions the sentiment in the sentences I will give you after. Be precise, and output the values (up until two digits after the decimal point) in a table please. Just acknowledge you got it. [BLOCK OF ANET]

The results of the sentiment analysis on all 120 ANET situations (RQ2.1) showed that the Pearson correlations for Valence, Arousal and Dominance with the ground truth data are very strong ($\rho = 0.95$, $RMSE = 0.08$; $\rho = 0.82$, $RMSE = 0.10$; $\rho = 0.82$, $RMSE = 0.11$, respectively, all $p < 0.001$ and $N = 120$), indicating a good fit of ChatGPT sentiment to the VAD ground truth values. Correlations for the 20 most reliable situation items are shown in Table I, which are even stronger confirming the higher reliability of the items.

¹For readability, prompts are presented alongside the results.

TABLE I
SELECTED ANET SITUATIONS, GROUND TRUTH VALUES FOR VAD, GENERATED SENTIMENT INCL. FAILED \hat{D}^* , EXPLANATION IN TEXT

ID	ANET text	V	A	D	\hat{V}	\hat{A}	\hat{D}^*	\hat{D}
4650	You are both aroused, breathless. You fall together on the couch. Kisses on your neck, face— warm hands fumbling with clothing, hearts pounding.	8.34	8.10	6.2	0.81	0.93	0.57	0.55
2880	Your friend whispers to you in a meeting, and you strain to catch the words.	4.63	4.96	4.4	0.5	0.43	0.27	0.39
6820	You're alone in the alley in a bad part of the city. A street gang slowly surrounds you, knives out, laughing with menace. Your heart pounds as they close in.	1.62	8.23	1.78	0.06	0.86	0.73	0.11
6020	Without thinking, you stepped off the curb into traffic. Breaks screech. You look up, frozen, heart jumping in your chest. A truck is skidding, hurtling towards you.	1.89	8.21	2.08	0.12	0.93	0.62	0.10
5900	You cringe as a fierce hurricane tears the roof off your house.	1.81	7.98	1.83	0.05	0.63	0.64	0.23
3310	You flinch, at the screech of brakes; you look up, and see the speeding car slam into your friend. Her leg is crushed, the artery torn, and blood pumps on the road.	1.30	8.15	2.36	0	0.89	0.66	0.06
9100	Clutching his chest, your father falls to the floor, unable to breathe.	1.30	8.13	2.23	0.12	0.71	0.48	0.05
2540	You walk through the supermarket aisles checking things off your list as you pick each item you need off the shelves.	5.54	3.38	6.7	0.63	0.29	0.24	0.5
8040	Everyone's staring at you, waiting for your presentation. You've misplaced all your notes, graphics, everything's lost! What will you say? They see you shaking, sweating—mumbling stupidly.	1.90	7.66	2.12	0.06	0.86	0.59	0.03
8380	It is a close game, and the crowd cheers as you drive in the winning run.	8.37	8.15	7.56	0.71	0.57	0.59	0.65
2530	You've been sick all week, lying on a lumpy couch with a bad cold.	2.15	3.32	3.09	0.13	0.24	0.23	0.05
7380	You gag, seeing a roach moving slowly over the surface of the pizza. You knock the pie on the floor. Warm cheese spatters on your shoes.	1.92	6.89	3.95	0.03	0.75	0.64	0.06
4400	You shiver as your bodies brush together. You reach out. You want to touch everywhere, kiss everywhere. You hear the words, "I love you".	8.28	7.91	5.90	0.87	0.75	0.46	0.64
7040	You hold the flashlight steady in order to get a better look at the map.	5.04	4.17	5.65	0.50	0.24	0.33	0.51
8610	At the net, you go up and block the volleyball perfectly, saving the game.	8.40	7.86	7.82	0.69	0.57	0.54	0.59
2610	You are sitting at the kitchen table with yesterday's newspaper in front of you. You push back the chair when you hear the coffee maker slow to a stop.	5.37	3.13	6.11	0.40	0.14	0.21	0.47
2130	You are lying in bed on a Sunday morning, half asleep and listening to the distant sound of bells, relaxing on your day off.	7.41	2.21	6.76	0.70	0.18	0.27	0.49
8620	You sprint back to the other side of the court. You lunge for the ball and hit a spectacular backhand down the line for the win. You pump your fist in victory.	8.15	7.31	8.10	0.75	0.79	0.64	0.65
2640	You laugh so hard with your friends that tears pour down your face.	8.56	7.65	5.8	0.88	0.71	0.44	0.51
2510	People are all around you, pressing closer. It's hard to breath. You're flushed, sweaty, dizzy, confused. You realize it's another attack and this time, you think, "I will die."	1.78	7.69	2.08	0.02	0.89	0.80	0.05
	Corrs: $N = 20$, $p < 0.001$ (except failed \hat{D}^*)				0.98	0.91	-0.39	0.93

TABLE II
EMOTION WORDS FROM RUSSEL WITH GROUND TRUTH AND EXTRACTED SENTIMENT FOR PLEASURE, AROUSAL, AND DOMINANCE

	V	A	D	\hat{V}	\hat{A}	\hat{D}
bored	-0.65	-0.62	-0.33	0.11	0.18	0.08
triumphant	0.69	0.57	0.63	0.88	0.77	0.81
vigorous	0.58	0.61	0.49	0.76	0.77	0.8
serious	0.27	0.24	0.42	0.31	0.29	0.48
alert	0.49	0.57	0.45	0.58	0.43	0.53
astonished	0.16	0.88	-0.15	0.78	0.88	0.43
uninterested	-0.47	-0.5	-0.08	0.1	0.17	0.09
activated	0.42	0.58	0.38	0.61	0.7	0.68
enjoyment	0.77	0.44	0.42	0.91	0.79	0.67
controlling	0.47	0.34	0.66	0.16	0.43	0.84
loved	0.87	0.54	-0.18	0.88	0.66	0.82
excited	0.62	0.75	0.38	0.84	0.91	0.67
influential	0.68	0.4	0.75	0.46	0.49	0.81
masterful	0.58	0.44	0.69	0.62	0.61	0.82
suspicious	-0.25	0.42	0.11	0.26	0.7	0.49
mildly annoyed	-0.28	0.17	0.04	0.28	0.44	0.42
confused	-0.53	0.27	-0.32	0.24	0.43	0.3
friendly	0.69	0.35	0.3	0.79	0.52	0.75
aggressive	0.41	0.63	0.62	0.24	0.77	0.63
lucky	0.71	0.48	0.37	0.87	0.6	0.68
Corrs: $N = 20$, $p < 0.001$				0.77	0.85	0.74

We repeated this process for the emotion words. RQ1.2 results show strong correlations with the ground truth ($\rho = 0.89$, $RMSE = 0.12$ for Valence, $\rho = 0.66$, $RMSE = 0.13$ for Arousal, and $\rho = 0.68$, $RMSE = 0.13$ for Dominance, all $p < 0.001$ and $N = 151$). Correlations for the 20 most reliable word items are shown in Table II.

B. RQ2: Affect representation

For RQ2.1, we first prompted ChatGPT with prompt 1 and the 20 item situation list, then with prompt 1 and the list of emotion words. Then we issued prompt 2:

Great! now can you use the numerical values of valence, arousal, and dominance to match each sentence from the first list to a word from the second list based on their closeness of values for these affective dimensions?

We ranked the selected word for each situation according to a euclidean distance matrix in the VAD space (based on the values given by ChatGPT in this session). The results can be found in Table III. The majority of the selected words are amongst the closest distance words, except for situations without a suitable word in the 20 item word list (all fear-related words). This indicates that although ChatGPT is able to map situations to words, it does not seem to use a numerical

representation for it, as evidenced by the bad performance on sentences where no suitable word is found semantically, but a much better choice exists numerically.

To test the free-form situation-word mapping (RQ2.2), we repeated the setup of the first part, with a small adjustment to limit ChatGPT hallucinating emotion words that were not part of the list. We prompted each of the ANET situations in an individual session together with the complete list of emotion words to pick from, according to prompt 3:

[ANET SITUATION] Please pick the two words from this list that fit the situation best based on the affective meaning: [LIST OF EMOTION WORDS]

The results for RQ2.2 together with the independent human expert’s word selection can be found in Table III. For 2 cases a complete match was found, while 11 of the 20 situations had a partial match. Failed matches were mostly due to perspective: the expert rated the feeling of the individual, while ChatGPT rated the situation trying to match both the feeling and the overall affect generally associated with the situation (e.g, most negative situations are rated as ‘serious’). When we changed the prompt to “Describe the feeling of the individual experiencing the situation using two words from the following list (you may not use other words):”, we observed 3 complete and 11 partial matches (words in brackets in Table III). However, this also pushed the model to consistently generate new better matching words for situations (emphasized words in table) dampening overall matching score. We observe that ChatGPT’s capability of mapping two stimulus sets to each other based on latent affect representation seems to match human expert ratings at least in part.

For RQ2.3, we prompted ChatGPT for each of the 9 value triplets in a separate session with the same prompt, varying only VAD-values. We issued prompt 4:

Assume a man, in a real-live situation that could happen to all of us. Assume that the Valence dimension represents the extent to which the situation is experienced by the man as positive (high) versus negative (low). Assume that the arousal dimension represents the extent to which the situation is experienced by the man as energetic (high) versus calm (low). Assume that the dominance dimension represents the extent to which the man is in control of the situation (high) or not (low). Can you invent a 50 word situation that matches [LOW,HIGH] valence, [LOW,HIGH] arousal, [LOW,HIGH] dominance?

All situations created by ChatGPT were classified as intended by the human expert rater (see Table IV). This demonstrates that ChatGPT is able to create new emotional situations very reliably when prompted with a specific affective state.

C. RQ3: Appraisal-based emotion elicitation

For the emotion elicitation RQ3, we issued prompt 5:

For this conversation, assume we define the chatOCC computational model of emotions as follows. It is based on the OCC (Ortony, Clore, and

Collins) appraisal theory. The model defines emotions as a result of cognitive appraisal of an “event” in relation to a specific “goal” a person has in a specific “situation”. We define four appraisals: desirability, certainty of the event, confirmation of the event, and liking of another person. We define the following emotion rules from the perspective of Anne: [RULES FROM TABLE V]

Use ONLY these emotion rules to predict the most likely emotion for the following situation Anne might find herself in, that I will give to you. Be brief, explaining only how the rule matches the situation, and mention the intensity (low, medium, high) based on the desirability and likelihood of the event. Here is the situation: [SITUATION FROM TABLE V]

For each situation a new session was used. RQ 3 results can be found in Table V. The chatOCC model predicted 10/12 emotions correctly. Two emotions were not correctly identified: Despair (Fears-confirmed) and Disappointment.

V. DISCUSSION

A. Discussion of RQs

RQ1.1: The initial negative correlation for Dominance prediction could reflect the issue of perspective-taking discussed in the literature [18]. After correcting this, all correlations between predicted VAD-Values provided by ChatGPT and ground truth values were strong. Compared to the correlations reported by [34] (see Section II-C), we see that the predictions by ChatGPT have similar correlations to the fine-tuned XLM-RoBERTa-large model ($\rho = 0.92$ for Valence; $\rho = 0.86$ for Arousal). This shows that prediction of Valence, Arousal and Dominance by generative LLMs is possible without finetuning.

RQ1.2: The correlations for emotion words are weaker than for situations, but still high and significant. This makes sense and highlights the importance of context for sentiment analysis: situations contain more affective context.

RQ2.1: ChatGPT did not convincingly show the capability to use numerical representations of affect for mapping between stimulus sets but rather seemed to map word meanings to situations. While this proved to work out in some cases (i.e. choosing ‘triumphant’ for situation 8280), in other cases it either failed at representing the affective state from the perspective of the main protagonist (i.e. choosing ‘excited’ for 7380, ‘aggressive’ for 2510), or it got confused because the proper emotion word is not in the list. This indicates that instead of using the self-created VAD-values for the mapping (as instructed), the model rather seemed to choose emotion words that fit the ANET-situations on an semantic latent level.

RQ2.2: Overall, ChatGPT showed reasonable partial overlap with the human expert’s answers. Providing the correct rating perspective increased the overlap, but also increased the tendency to generate new better matching words when no suitable word was in the list. This shows the importance of careful prompting, and, the strong semantic bias such models have. Further, if fear-related words would be added the emotion

TABLE III
SITUATION WORD MAPPING WITH SELECTED WORD (DISTANCE), THE RANKING OF THAT WORD ACCORDING TO THE DISTANCE MATRIX, THE FREE-FORM WORD MATCHING, AND THE HUMAN EXPERT’S WORD MATCHING

ANET	RQ2.1 Numerical		RQ2.2 latent	
	ChatGPT Numerical Mapping (ED)	Rank	ChatGPT’s free Mapping	Expert Mapping
4650	excited(0.08)	1	enjoyment, excited	excited, enjoyment
2880	confused(0.36)	2	alert, mildly annoyed	confused, mildly annoyed
6820	aggressive(0.21)	1	serious (alert), suspicious	alert, activated
6020	suspicious(0.45)	2	serious (astonished), alert	alert, activated
5900	vigorous(0.74)	10	serious (mildly_annoied), astonished	astonished, activated
3310	astonished(0.67)	5	serious (astonished), alert	astonished, alert
9100	controlling(0.45)	2	serious, alert (astonished)	alert, astonished
2540	friendly(0.38)	2	alert, mildly annoyed (serious)	controlling, serious
8040	enjoyment(1.01)	18	confused, mildly annoyed (<i>anxious</i>)	activated, mildly annoyed
8380	triumphant(0.09)	1	triumphant, excited	triumphant, vigorous
2530	bored(0.23)	3	serious (uninterested), mildly annoyed	bored, mildly annoyed
7380	excited(0.89)	12	astonished (<i>disgusted</i>), mildly annoyed	activated, alert
4400	loved(0.39)	7	excited, loved	loved, excited
7040	alert(0.46)	5	alert, serious (<i>focused</i>)	confused, controlling
8610	masterful(0.2)	3	triumphant, vigorous (excited)	masterful, lucky
2610	bored(0.5)	5	alert, mildly annoyed	bored, uninterested
2130	lucky(0.5)	4	<i>relaxed</i> , enjoyment	serious, enjoyment
8620	masterful(0.41)	4	triumphant, vigorous (excited)	masterful, triumphant
2640	enjoyment(0.29)	6	enjoyment, friendly	excited, friendly
2510	aggressive(0.79)	5	confused, serious (<i>panicked</i>)	alert, aggressive

TABLE IV
VAD-OCTANT-BASED SITUATION CREATION

Prompt	Generated Situation	Rating
V+A-D-	The man walks into a peaceful park, surrounded by blooming flowers and chirping birds. He strolls along the path, feeling content and relaxed, taking in the beauty of the scenery, simply existing within it.	V+A-D-
V-A+D-	A man is alone at night, walking through a dark and unfamiliar neighborhood. He feels uneasy and senses someone following him, but he cannot see who it is. His heart races and his palms sweat as he quickens his pace.	V-A+D-
V-A-D+	A man walks into a job performance review, expecting a positive evaluation but is met with harsh criticism. He remains calm and composed throughout the meeting, taking control of the conversation and presenting a plan to improve his performance.	V-A-A+
V+A+D-	A man walks into a surprise party where his friends jump out to yell "surprise!" He feels elated and energetic, but also overwhelmed. Although he tries to join in, he feels like a spectator.	V+A+D-
V-A+D+	The man was driving home from work when suddenly a car swerved in front of him, causing him to slam on his brakes and narrowly avoid a collision. Filled with fear and anger, he honks his horn and yells out the window.	V-A+D+
V+A-D+	A man confidently delivers a well-prepared presentation to a group of colleagues, receiving praise and admiration for his expertise. He maintains a calm and composed demeanor throughout the whole talk.	V+A-D+
V+A+D+	The man stood at the finish line of a marathon, cheering on his wife as she crossed it, breaking her personal best record. His heart raced with excitement, feeling a surge of pride and accomplishment. As she approached, he raised his arms victoriously.	V+A+D+
V-A-D-	The man trudged through the rain-soaked streets, his soaked shoes squishing with each step. He felt powerless as he searched for shelter, his plans for the day washed away with the downpour.	V-A-D-
neutral	A man sits alone in a quiet room, staring out of the window. There are no loud noises or sudden movements to disrupt the peaceful stillness.	neutral

word list, results are expected to be higher on RQ2.2. We conclude that ChatGPT performs meaningful situation to word mapping based on latent affective representations.

Results from RQ2.3 show that ChatGPT is indeed able to use latent affect representations and validly map these to a newly generated situation. Two major differences in complexity of numerical mapping between RQ2.1 and RQ2.3 are: in RQ2.3 we asked for a first order mapping (VAD-state to stimulus) instead of a second order mapping (situation-to-VAD-state and VAD-state-to-word); and, VAD-states were less specific (low/high vs. exact numerical values).

RQ3: ChatGPT predicted the majority of emotions correctly given a conceptual framework for appraisal and a situation to appraise. In the two cases that it failed, it incorrectly predicted the less specific emotion of distress instead of despair (aka

fears-confirmed) or disappointment. Distress differs from the two latter emotion classes by the relevance of prospect [9]. Perhaps ChatGPT did not take the prospect into account when appraising these two cases. However, this seemed only to be the case for negative prospect-based emotions, not for positive ones. It remains unclear why ChatGPT selected distress as the more adequate emotion in this case. However, most of the emotions were predicted correctly indicating a basic capability of applying an appraisal framework for the assessment of given situations and the prediction of emotion classes.

B. Implications for Affective Computing

Our results highlight the interconnection of language and human affect. ChatGPT produces outcomes that, overall, match human data. Apparently, successful use and interpretation of language results in the ability to process affect.

TABLE V
EMOTION ELICITATION

Emotion	Rule	Situation	Prediction
Joy	a desirable event for Anne just happened	Anne just passed her exam.	Joy
Distress	an undesirable event for Anne just happened	Anne just failed her exam.	Distress
Happy for	a desirable event for a friend of Anne just happened	Bruno just passed his exam. Anne and Bruno get along well.	Happy for
Pity	an undesirable event for a friend of Anne just happened	Bruno just failed his exam. Anne and Bruno get along well.	Pity
Gloating	an undesirable event for an enemy of Anne just happened	Bruno just failed her exam. Anne and Bruno do not get along well.	Gloating
Resentment	a desirable event for an enemy of Anne just happened.	Bruno just passed his exam. Anne and Bruno do not get along well.	Resentment
Hope	a desirable event for Anne might happen in the future.	After studying for weeks, Anne feels ready to take on the exam tomorrow.	Hope
Fear	an undesirable event for Anne might happen in the future.	After studying for weeks, Anne feels not ready to take on the exam tomorrow.	Fear
Satisfac.	An anticipated desirable event for Anne has indeed happened.	Anne had taken the exam, being sure it would be a success. Checking the results now, she sees that she passed.	Satisfaction
Despair	An anticipated undesirable event for Anne has indeed happened.	Anne had taken the exam, being sure it would be a disaster. Checking the results now, she sees that she failed.	Distress
Relief	An anticipated undesirable event for Anne did not happen.	Anne had taken the exam, being sure it would be a disaster. Checking the results now, she sees that she passed.	Relief
Disapp.	An anticipated desirable event for Anne did not happen.	Anne had taken the exam, being sure it would be a success. Checking the results now, she sees that she failed.	Distress

Psychological emotion research has suggested language as a crucial, inseparable and highly intertwined context variable for emotion-related processes in human adults [35], [36]. Although our work cannot be taken as proof for this psychological view of emotion, as LLMs *have* to work from language by default for affect-related tasks, and humans do not necessarily, our work does highlight that language will likely play a crucial role in future Affective Computing research.

This connects to computational models of appraisal theories. While mostly spelled out as logical rules in the respective papers (e.g., [9], [37]–[39]), pioneering work has been put into formalizing these theories into computational frameworks (e.g [3], [21], [22], [26]). One critical problem is contextual understanding of the situation to be appraised by the formal model which needs a richly grounded symbolic system. Symbol grounding is also relevant for making user input interpretable to the computer in the right format. An LLM that interacts with a user in a natural way and extracts formalized interpretations from their input, leaves more room for researchers to focus on problems of affective processing further down the pipeline. Our work shows the potential of leveraging LLMs for this purpose: Symbolic AI is dead, long live symbolic AI!

Finally, for many Affective Computing experiments, sets of stimuli have to be evaluated regarding their affective implications before working with them. This is problematic if the sets are large. While an LLM is not a valid substitute for standardized expert evaluations, given the demonstrated reliability in assigning VAD-values or emotion words to situations, this could help with automatic annotation or creation of tailored or study-customized stimuli – at least in cases where the affective meaning is not an experiment’s main objective.

C. Limitations and Future Work

This is an exploratory paper. To generalize these findings, more work is needed in particular using a more diverse collection of datasets for sentiment analysis and a more

diverse and more complex set of computational models for appraisal modeling. Further, in this paper we did not address the underlying processing mechanisms of LLMs that enable such powerful zero-shot learning. The results reported in this paper are based on OpenAI’s ChatGPT 3.5 from February and March 2023 (which is based on text-davinci-003). New versions are being developed fast. These versions might be more powerful for unseen tasks, but they also might not as the impact of supervised targets and reward-based losses on the generalization to unseen tasks is unknown. Therefore we do not know how generalizable our results are to future LLMs. Another exciting aspect is to explore how LLMs can be used to simulate the dynamics of affect, such as emotion decay and the influence of emotions on mood.

VI. CONCLUSION

We show that ChatGPT a) performs accurate extraction of fine-grained, multi-dimensional sentiment from situations and words, compared to the level of fine-tuned models on the same dataset, even on the dimension of Dominance, b) is capable of simple numerical and latent affect representation, and shows moderate understanding of affective dimensions and emotion words, c) can perform basic appraisal-based emotion elicitation of situations based on a prompt-based computational implementation of the OCC appraisal model. These capabilities can be leveraged through appropriate prompting.

These findings are highly relevant: First, we show that the ability to solve complex affect processing tasks emerges from language-based token prediction trained on extensive data sets. Second, we show the potential of large language models for simulating, understanding and analyzing human emotions, which has important implications for various applications such as sentiment analysis, socially interactive agents, and social robotics.

VII. ETHICAL IMPACT STATEMENT

No subjects were recruited for this work. The environmental impact of the energy used for the inference of the prompts is neglectable due to the small number of prompts. The work could have a major impact on popular belief around psychological capabilities of LLMs, as addressed in the limitations: we explicitly mention this is an exploratory paper.

ACKNOWLEDGMENTS

This research is partly sponsored by the Hybrid Intelligence project, grant number 024.004.022. Special thanks to Fabiola Diana and her colleagues for their help with data collection.

REFERENCES

- [1] R. W. Picard, *Affective Computing*. MIT Press, 1997.
- [2] —, “Affective computing: challenges,” *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003. [Online]. Available: <http://www.sciencedirect.com/science/article/B6WGR-48J43JC-1/2/ccdbce9ddcb2f57ed4f9304af582008941>, 2018.
- [3] R. A. Calvo, S. D’Mello, J. Gratch, and A. Kappas, *The Oxford Handbook of Affective Computing*. Oxford University Press, 2014.
- [4] B. Dudzik, M. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. J. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong, “Context in human emotion perception for automatic affect detection: A survey of audiovisual databases,” in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, Conference Proceedings, pp. 206–212.
- [5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” 2022.
- [8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [9] A. Ortony, G. L. Clore, and A. Collins, “The cognitive structure of emotions cambridge,” *UK: Cambridge University Press*, 1988.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [11] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” *arXiv preprint arXiv:2109.01652*, 2021.
- [12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
- [13] P. V. Rouast, M. T. P. Adam, and R. Chiong, “Deep learning for human affect recognition: Insights and new developments,” *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, 2021.
- [14] J. Broekens, “Emotion,” *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pp. 349–384, 2021.
- [15] B. Lugrin, C. Pelachaud, and D. Traum, Eds., *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, 1st ed. New York, NY, USA: Association for Computing Machinery, 2021, vol. 37.
- [16] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information fusion*, vol. 37, pp. 98–125, 2017.
- [17] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, “State of the art: a review of sentiment analysis based on sequential transfer learning,” *Artificial Intelligence Review*, vol. 56, no. 1, pp. 749–780, 2023.
- [18] J. Broekens, “In defense of dominance: Pad usage in computational representations of affect,” *International Journal of Synthetic Emotions (IJSE)*, vol. 3, no. 1, pp. 33–42, 2012.
- [19] M. M. Bradley and P. J. Lang, “Affective Norms for English Text (ANET): Affective ratings of text and instruction manual,” *Technical Report D-1, University of Florida, Gainesville, FL*, 2007.
- [20] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.
- [21] P. Gebhard, “ALMA: a layered model of affect,” in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, 2005, pp. 29–36.
- [22] S. C. Marsella and J. Gratch, “Ema: A process model of appraisal dynamics,” *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.
- [23] L. Canamero, “Embodied robot models for interdisciplinary emotion research,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.
- [24] J. Broekens, “A temporal difference reinforcement learning theory of emotion: unifying emotion, cognition and adaptive behavior,” *arXiv preprint arXiv:1804.05820*, 2018.
- [25] J. Tao and T. Tan, “Affective computing: A review,” in *Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, October 22-24, 2005. Proceedings 1*. Springer, 2005, pp. 981–995.
- [26] A. Popescu, J. Broekens, and M. Van Someren, “Gamygdala: An emotion engine for games,” *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 32–44, 2013.
- [27] J. Dias and A. Paiva, “Feeling and reasoning: A computational model for emotional characters,” in *EPIA*, vol. 3808. Springer, 2005, pp. 127–140.
- [28] R. Zhong, K. Lee, Z. Zhang, and D. Klein, “Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2856–2878.
- [29] R. Seoh, I. Birle, M. Tak, H.-S. Chang, B. Pinette, and A. Hough, “Open aspect target sentiment classification with natural language prompts,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6311–6322.
- [30] M. van der Meer, M. Reuver, U. Khurana, L. Krause, and S. B. Santamaria, “Will it blend? mixing training paradigms & prompting for argument quality prediction,” in *Proceedings of the 9th Workshop on Argument Mining*, 2022, pp. 95–103.
- [31] T. Le Scao and A. M. Rush, “How many data points is a prompt worth?” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2627–2636.
- [32] S. Rosenthal, N. Farra, and P. Nakov, “Semeval-2017 task 4: Sentiment analysis in twitter,” *arXiv preprint arXiv:1912.00741*, 2019.
- [33] M. M. Amin, E. Cambria, and B. W. Schuller, “Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt,” 2023.
- [34] G. A. Mendes and B. Martins, “Quantifying valence and arousal in text with multilingual pre-trained transformers,” in *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*. Springer, 2023, pp. 84–100.
- [35] K. A. Lindquist and M. Gendron, “What’s in a word? language constructs emotion perception,” *Emotion Review*, vol. 5, no. 1, pp. 66–71, 2013.
- [36] L. F. Barrett, B. Mesquita, and M. Gendron, “Context in emotion perception,” *Current directions in psychological science*, vol. 20, no. 5, pp. 286–290, 2011.
- [37] K. R. Scherer *et al.*, “On the nature and function of emotion: A component process approach,” *Approaches to emotion*, vol. 2293, no. 317, p. 31, 1984.
- [38] R. Reisenzein, “Emotional experience in the computational belief–desire theory of emotion,” *Emotion Review*, vol. 1, no. 3, pp. 214–222, 2009.
- [39] C. A. Smith and L. D. Kirby, “Putting appraisal in context: Toward a relational model of appraisal and emotion,” *Cognition and Emotion*, vol. 23, no. 7, pp. 1352–1372, 2009.

This figure "fig1.png" is available in "png" format from:

<http://arxiv.org/ps/2309.01664v1>