

Refashioning Emotion Recognition Modeling: The Advent of Generalized Large Models

Zixing Zhang[✉], Senior Member, IEEE, Liyizhe Peng[✉], Student Member, IEEE, Tao Pang[✉], Student Member, IEEE, Jing Han[✉], Senior Member, IEEE, Huan Zhao[✉], and Björn W. Schuller[✉], Fellow, IEEE

Abstract—After its inception, emotion recognition or affective computing has increasingly become an active research topic due to its broad applications. The corresponding computational models have gradually migrated from statistically shallow models to neural-network-based deep models, which can significantly boost the performance of emotion recognition and consistently achieve the best results on different benchmarks, and thus has been considered the first option for emotion recognition. However, the debut of large language models (LLMs), such as ChatGPT and GPT4, has remarkably astonished the world due to their emerged capabilities of zero/few-shot learning, in-context learning (ICL), chain-of-thought, and others that are never shown in previous deep models. In the present article, we comprehensively investigate how the LLMs perform in emotion recognition in terms of diverse aspects, including ICL, few-shot prompting, accuracy, generalization, and explanation. Moreover, we offer some insights and pose other potential challenges, hoping to ignite broader discussions about enhancing emotion recognition in the new era of advanced and more generalized models.

Index Terms—Emotion recognition, few-shot learning, in-context learning (ICL), large language model (LLM).

I. INTRODUCTION

AFFECT modeling is key for human-centered machine intelligence systems, allowing for more empathetic, adaptive, and engaging interactions, by considering users' emotions.

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010112 and in part by the National Science Foundation of China under Grant 62076092. (Corresponding author: Jing Han.)

Zixing Zhang is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China, and also with Shenzhen Research Institute, Hunan University, Shenzhen 518000, China (e-mail: zixingzhang@hnu.edu.cn).

Liyizhe Peng, Tao Pang, and Huan Zhao are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: pengliyizhe@hnu.edu.cn; pang_tao@hnu.edu.cn; hzhao@breakhnu.edu.cn).

Jing Han is with the Department of Computer Science and Technology, University of Cambridge, CB3 0FD Cambridge, U.K. (e-mail: jh2298@cam.ac.uk).

Björn W. Schuller is with GLAM—the Group on Language, Audio, and Music, Imperial College London, SW7 2AZ London, U.K., and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany (e-mail: schuller@ieee.org).

Digital Object Identifier 10.1109/TCSS.2024.3396345

Developing valid and reliable affective models that accurately capture and understand human emotions can pave the way for various applications, including affective dialogue systems, affect-aware recommender systems, and emotion-based adaptive interfaces. Over the past decades, the availability of large-scale affective datasets and rich computational power, as well as the advancements in deep learning algorithms, have revolutionized the field of affect modeling research.

In particular, deep learning techniques, such as neural networks, have proven to be highly effective in capturing complex patterns and relationships within affective data [1], [2], [3]. For instance, convolutional neural networks (CNNs) are efficient to learn and extract relevant features, and to capture local patterns and dependencies, which have demonstrated remarkable success in emotion recognition tasks across various domains and datasets [4], [5]. Likewise, recurrent neural networks (RNNs) are also widely used in affective modeling when processing sequential data such as speech and text, by capturing temporal dependencies and contextual information [6], [7]. Moreover, researchers have also made massive efforts to develop and implement various algorithms and approaches, aiming at more accurate and robust emotion recognition systems to understand human emotions. Such efforts include but are not limited to integrating information from different modalities [8], [9], developing large-scale emotion datasets for model training and validation [10], [11], transferring knowledge of other relevant tasks by leveraging pretrained models [12], [13], and increasing the interpretability and providing insights into the decisions [14], [15].

However, all aforementioned works are built typically for emotion recognition tasks, and normally the employed models focus on limited and specific domains. In other words, though prevalent in various emotion-aware systems and applications today, these specialized models lack broad generalization and adaptability. For instance, a system specifically designed to predict a set of six emotions may not be applicable or effective in predicting an unseen or different emotion. Similarly, an emotion recognition model designed specifically for the Chinese language cannot be directly applied to the English language. These models and systems designed for specific tasks are often referred to as artificial narrow intelligence (ANI).

In contrast to ANI, a more advanced and comprehensive form of artificial intelligence is artificial general intelligence (AGI), being able to learn, adapt, and apply knowledge across a wide

range of tasks. While AGI systems may not be specifically tailored for emotion recognition tasks, they have the potential to achieve comparable levels of accuracy and effectiveness as ANI ones that are specifically designed for the task. In particular, recent breakthroughs in large language models (LLMs), such as GPT-4, are demonstrating significant advancements in natural language processing. These advancements may also open a new frontier for emotion recognition. These models are normally pretrained over large-scale corpora, and have showcased their strong capabilities in various domains and tasks, such as text generation and natural language understanding [16].

When utilizing LLMs for emotion recognition, there are several potential benefits that may enhance the performance and capabilities of the models: 1) large training data: being trained on vast amounts of data, LLMs can capture more diverse patterns, linguistic cues, and contextual information related to emotions, and thus might lead to better recognition performance; 2) explanations: LLMs can potentially explain the reasoning behind their decisions, and therefore can increase the interpretability and transparency of the emotion recognition process; 3) generalization: LLMs are trained over a large-scale dataset in an unsupervised way, which empowers models strong capability for emotion recognition that may not have been explicitly encountered during training; and 4) cross-domain application: LLMs have the potential to be applied across various domains, since they are trained on a wide range of data sources, and thus hold the potential to understand emotions expressed in various domains ranging from customer reviews to conversational data, which allows for wider applicability.

Therefore, the focus of this work is on the analysis of the emotion recognition capabilities of LLMs and their implications. Also, we share our views of a future in LLMs-based affective modeling. For this aim, we choose four LLMs and conduct a comparative analysis of their emotion recognition performance against other state-of-the-art (SOTA) non-LLM-based works using eight emotional datasets. This comparison allows us to assess the advancements and effectiveness of LLM-based approaches over specific deep-learning models in the context of emotion recognition. More specifically, we investigate three different scenarios for LLM-based emotion recognition: 1) emotion prediction of a given sample in the absence of any contextual information; 2) emotion prediction of a sample with context information of that sample; and 3) emotion prediction of a sample while an appropriate number of examples are provided as demonstrations. By exploring these three scenarios, we aim to assess LLM-based emotion recognition models' effectiveness in different contextual settings. Besides, this work further delves into examining the generalization, and interpretability of LLM-based affective modeling. Last, we engage in comprehensive discussions on the associated challenges and opportunities, hoping to pave the way for advancements and improvements in the field.

The main contributions of this article are as follows.

- 1) We for the first time carry out a deep detailed investigation of the emotion recognition capability of LLMs. In specific, we systematically evaluated four selected LLMs

for emotion recognition tasks on eight datasets from two different languages.

- 2) We extensively compare the performance with other SOTA models, showing that LLMs can achieve comparable or superior performance on emotion recognition tasks.
- 3) We comprehensively discuss the obtained results and point out challenges and opportunities that are still faced in the research and development of LLM-based affective modeling.

The structure of this work is organized as follows. Section II presents related works. Section III introduces the selected LLMs, datasets for evaluation, and implementation details. Experimental results, along with their analysis and comparison, are provided in Section IV. Challenges and opportunities of LLMs in the context of affective modeling are discussed in Section V. The work concludes in Section VI.

II. BACKGROUND

This section provides a brief overview of the relevant background and concepts.

A. Recent Advances of Emotion Recognition

Over the past decades, particularly with the emergence of deep learning, significant endeavors have been dedicated to developing effective and robust emotion recognition models. In the following, we briefly introduce related studies primarily focusing on three key aspects/challenges: enhancing accuracy, facilitating generalization, and enabling explainability of emotion recognition models.

One main research direction is to improve the *correctness* of the model, namely increasing the accuracy and precision of emotion recognition models. For this aim, several research studies have investigated various advanced feature extraction techniques, aiming to capture salient emotional representations. The features could be extracted from different modalities, such as speech, video, text, and physiological signals. For instance, three novel domain-specific audiovisual tasks were designed in [17] to learn better representations, by exploring large volumes of unlabeled data via self-supervised learning. There are also a huge number of studies that leverage different deep learning architectures to refine emotion recognition performance. For instance, a transformer-based model was proposed in [18] to process ECG signals for emotion recognition, and the obtained contextualized representations achieved SOTA performance on the AMIGOS dataset. In addition, some other works focus on combining information from multiple modalities, so that a more comprehensive and holistic understanding of users' emotional states can be obtained [9].

Aiming at improving the *generalization* of emotion recognition models, several advanced deep learning techniques have been explored, such as knowledge transfer [19], multitask learning [20], self-supervised learning [21], and domain adaptation [22]. Applying these techniques appropriately could make the learnt emotion recognition model more effective across different domains, different contexts, and diverse scenarios. For

example, a pretrained model's knowledge gained from a large amount of data can be transferred to downstream tasks. Specifically, pretrained language models (PLMs), such as bidirectional encoder representations from transformers (BERT) and their variants have been utilized in text-based emotion recognition and achieved promising performance. In [23], the efficacy of BERT and its variants has been comparatively examined on a cross-culture emotional database and robustly optimized BERT pretraining approach (RoBERTa) achieved the best performance among others.

Enhancing *interpretability* is also critical for a reliable affective modeling system, where explanations should be provided along with its predictions. For this reason, advanced techniques have been applied in emotion recognition to improve the model interpretability. Such techniques include model-agnostic interpretation methods [24], attention mechanisms [25], and Bayesian neural networks [26].

These aforementioned efforts facilitate the development of correct, generalizable, explainable affective computing, and tremendous progress in this research field has been made thanks to these efforts. However, these efforts deal with the three above-mentioned challenges separately, and it is not straightforward to integrate all as a whole.

B. The New Era of Large Models

While transfer learning and PLMs have already shown promising advances on various downstream tasks, a new era of large models has started when researchers observed that scaling PLMs can largely lead to further improvement of the model capacity [27]. These large-sized PLMs are termed large language models (LLMs). Compared with smaller PLMs, LLMs significantly extend the model size and data size, even through the model architectures and the pretraining tasks are kept the same [28].

Various tech companies and research labs are striving to develop their own LLMs, and among these, there are some popular ones, including ChatGPT and the latest GPT-4 released by OpenAI, and LLaMA models introduced by Meta AI. Particularly, the launch of ChatGPT represents a significant step forward in the evolution of LLMs. Most LLMs are trained in two stages; LLMs normally are first pretrained on a vast amount of text data, and then further adapted via a reinforcement learning from human feedback (RLHF) algorithm [28]. While still based on standard deep learning and transfer learning, LLMs display a large performance improvement in terms of accuracy and robustness, achieving SOTA results across multiple domains.

Further, LLMs exhibit surprising generalization capability on unseen tasks, and are capable of solving a diverse range of complex tasks, even without further fine-tuning on each specialized task. This is mainly because LLMs have exhibited new emergent capabilities which were not present in smaller-scale PLMs [29]. One such new emergent ability is in-context learning (ICL), which was first introduced by GPT-3 in [30]. Being fed a prompt composed of task instructions and demonstrations (input-output examples of that task), the model learns

to solve this new task and generate the output of an unseen inference-time example, without any further model training or gradient updates. Other emergent abilities include step-by-step reasoning, instruction following, program execution, and model calibration [29]. With these advanced functionalities, LLMs have attracted increasing and widespread attention, leading to a paradigm shift in AI paradigm, shifting from task-specific models to more general-purpose models. In other words, the rapid technical progress of LLMs provides a fertile ground for AGI.

C. At the Crossroad

Emerging as powerful tools to understand and generate human-readable text, LLMs have demonstrated exceptional capabilities in a wide range of NLP tasks. Their impact on many domains is significant, such as question-answering, code generation, and creative writing. Standing at the crossroads of traditional affective modeling and large models, in this article, we aim to explore LLMs' impact on affective computing, as the effectiveness of LLMs on this task is not fully investigated, yet.

Having trained on massive amounts of textual data, LLMs offer exciting possibilities for understanding the subtle variations in emotional content. Besides, though not trained explicitly for emotion recognition such as previous task-specific models, the general-purpose LLMs have strong adaptation capabilities. Especially, prompting-based adaptation methods enable human-like interactions with an LLM via very simple prompts, entailing temporary learning during inference without specialized fine-tuning.

While LLMs hold promising opportunities in this domain, there still lacks a thorough investigation of the effectiveness of LLMs in affective computing. Specifically, there are still several key questions that need to be answered.

- 1) How do LLMs perform in emotion recognition tasks, compared with SOTA task-specific models?
- 2) Can one LLM be utilized across different affective computing-related tasks without further training or updating?
- 3) To what level can LLMs explain the reasons behind the detected emotion category from a given emotional input?

Different from [31] which focused on evaluating the effectiveness of ChatGPT on three datasets, this article explores the efficiency and effectiveness of four LLMs on eight emotional datasets from two languages. More importantly, we explore the emergent capability of ICL, showing that with few demonstration examples, LLMs perform better. The generalization and explainability of LLMs for affective computing are also analyzed, which is missing in previous works.

III. EMOTION RECOGNITION WITH LLMs

In this section, we outline our evaluation strategy for LLM-based emotion recognition. We begin by introducing the four LLMs selected for this evaluation. Then, we give an overview of the eight emotional datasets used in the evaluation. Lastly, we detail the implementation process.

TABLE I
DETAILED INFORMATION OF THE SELECTED EIGHT DATASETS

Dataset	Language	Modality	Dialogue	Data Source	#Sp.	#Dia.	#Utt. Total (Test)	#Words/Utt.	#Classes
SST	English	t	No	Movie review	-	-	11 855 (2 210)	-	Five (negative, somewhat negative, neutral, positive, somewhat positive)
Friends	English	t	Yes	Friends TV shows	-	1 000	14 503 (2 764)	10.7	Seven (neutral, joy, sadness, fear, anger, surprise, disgust)
Mastodon	English	t	Yes	Mastodon	-	505	2 217 (1 142)	-	Three (positive, neutral, negative)
DailyDialog	English	t	Yes	Various websites	-	13 118	102 979 (7 740)	-	Seven (neutral, joy, surprise, sadness, anger, disgust, fear)
MOSI	English	a, v, t	No	YouTube	89	-	2 199 (686)	12.0	Seven {-3, -2, -1, 0, 1, 2, 3}
MOSEI	English	a, v, t	No	YouTube	1 000	-	23 453 (4 662)	-	Seven {-3, -2, -1, 0, 1, 2, 3}
CH-SIMS	Mandarin	a, v, t	No	Movies, TVs, and shows	474	-	2 281 (457)	15.0	Five {-1.0, -0.8} {-0.6, -0.4, -0.2} {0.0}
M ³ ED	Mandarin	a, v, t	Yes	TV series	626	990	24 449 (4 201)	7.4	{0.2, 0.4, 0.6} {0.8, 1.0} Seven (happy, surprise, sad, disgust, anger, fear, neutral)

A. LLMs

In this work, we investigate the capabilities of LLMs to perform emotion recognition. For this aim, four LLMs are selected, including ChatGPT, Claude, Bing Chat, and Gemini.

1) *ChatGPT*: Released by OpenAI in November 2022, ChatGPT represents a significant advancement in the NLP field. It builds upon the prior GPT-3.5 model, which has 175 billion parameters, but is specifically optimized for dialogues. Notably, ChatGPT underwent fine-tuning via the RLHF method to produce more human-like responses. The RLHF procedure comprises three stages: supervised fine-tuning, reward model training, and reinforcement learning fine-tuning. Through RLHF, ChatGPT’s behavior is better aligned with human values and preferences. Consequently, ChatGPT excels at human-like communication, generating coherent responses and engaging in meaningful dialogues. This innovation paves the way for diverse applications. For our API interactions, we specified the model as gpt-3.5-turbo, which underpins ChatGPT. We also set the temperature to 0.01 to ensure more focused and deterministic outputs.

2) *Claude*: Claude, launched in March 2023 by Anthropic, is an LLM-based AI Chatbot. As a strong competitor to ChatGPT, Claude is also capable of a wide variety of conversational and text processing tasks. These include creative writing, Q and A, project planning, and coding. With the aim of creating AI systems that are helpful, honest, and harmless, the RLHF technique was also employed in Claude [32]. Further, Claude was trained using a “Constitutional AI” technique, where a model rather than a human determines the rankings/preferences based on a set of underlying principles [33]. In this manner, the output generated by Claude can align with human values and goals from inception.

3) *Bing Chat*: Microsoft announced the next version of its web search engine—the new Bing in February 2023, augmented with a variety of cutting-edge technologies from both Microsoft and its partner OpenAI. With this, Microsoft aims at bringing together search, browsing, and chat into one unified experience.

Particularly, the new Bing introduces a chat feature, Bing Chat, designed to deliver answers that feel more conversational than typical search engine responses. Bing Chat is powered by OpenAI’s GPT-4 but customized specifically for search. Also, Bing Chat has access to the internet and thus can provide more up-to-date information. This is different from ChatGPT, which was trained on data only up until September 2021. Additionally, in Bing Chat, the response to a user’s query is grounded with search results, incorporating footnotes that trace back to original sources for verification.

4) *Gemini*: Gemini, introduced by Google in December 2023, are a family of highly capable multimodal language models. Gemini models build on top of transformer decoders, exhibiting remarkable capabilities across image, audio, video, and text understanding. Notably, they are the first models to achieve human-expert performance on the well-studied exam benchmark MMLU [34]. Gemini models are also tailored as chat-focused variants, which have become popular conversational AI service previously recognized as Bard. In March 2024, Google presented Gemini 1.5 Pro, the latest member of Gemini family. As a highly compute-efficient multimodal mixture-of-experts language model, it is highly capable of recalling and reasoning over intricate details from millions of context tokens [35]. In our experiments, we selected the SOTA Gemini 1.5 Pro as the representative of the Gemini family and we only used its ability in text.

B. Emotion Datasets

In this part, we introduce the eight datasets employed in our study. A summary of their statistics can be found in Table I. As these datasets are publicly accessible, they facilitate easy verification of our results.

- 1) *SST*: The Stanford Sentiment Treebank is an English corpus with fine-grained sentiment labels of 11 855 single sentences extracted from movie review data [36]. These sentences were parsed using the Stanford parser, resulting in a total of 215 154 unique phrases from those

parsed trees. Those phrases were then annotated by the classic Mechanical Turk for sentiment into 25 different levels. The obtained annotations for each phrase were then used to define fine-grained and binary versions of the task. For the fine-grained task, each sentence is labeled with one of five sentiment classes: negative, somewhat negative, neutral, somewhat positive, and positive; for the binary task, each sentence is labeled as either positive or negative (the neutral class is ignored). In this study, we do not consider phrase-level sentiment analysis.

- 2) *Friends*: Friends is a corpus derived from the Friends TV shows, comprising transcriptions of 1000 dialogue spoken by the characters in the show across seasons 1 to 9 [37]. Each dialogue was regarded as an annotation task on Amazon Mechanical Turk. The goal was to label each of 14 503 utterances within the 1000 dialogues into one of seven classes: anger, disgust, fear, joy, sadness, surprise, and neutral. The annotators were instructed to consider the context of the entire dialogue while assigning the sentiment labels.
- 3) *Mastodon*: The Mastodon dataset was built by crawling social media posts from Mastodon, and non-English posts were filtered out automatically [38]. The corpus was completely manually anonymized and annotated by two students. While the dataset was originally designed for sentiment recognition and dialogue act recognition, here, we consider the sentiment analysis task only. Each post was labeled as positive, negative, or neutral.
- 4) *DailyDialog*: DailyDialog is a high-quality multiturn dialog dataset, crawling from diverse websites designed for English learners to practice English dialog in daily life [39]. The language in DailyDialog is human-written and thus exhibits a higher degree of formality, resembling authentic conversations more closely. The corpus includes 13 118 multiturn dialogues and each utterance within it is annotated with seven emotion classes: joy, surprise, sadness, anger, disgust, fear, and neutral.
- 5) *MOSI*: The Multimodal Opinion-level Sentiment Intensity (MOSI) dataset [40] is a multimodal dataset for studying sentiment and subjectivity in opinion videos. The dataset consists of 2199 opinion video segments, segmented from 93 videos generated by 89 distinct English-speaking YouTube users. The sentiment intensity annotation was carried out by master workers via Amazon Mechanical Turk. Each opinion segment was annotated as a sentiment spectrum between highly negative and highly positive in the range $[-3, 3]$.
- 6) *MOSEI*: CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) is the largest dataset of sentence-level sentiment analysis and emotion recognition in online videos [11]. It contains 23 453 sentence utterance videos from 1000 English-speaking YouTube speakers. These sentence utterances were randomly chosen from various topics and monologue videos. Similar to MOSI, the dataset was annotated by three master judges via the Amazon Mechanical Turk platform. Each

sentence in the MOSEI dataset is annotated for sentiment on a Likert scale from -3 to $+3$.

- 7) *CH-SIMS*: A Chinese single- and multimodal sentiment analysis dataset, CH-SIMS, was introduced in [41]. It contains 2281 refined video segments in the wild, collected from different movies, TV serials, and variety shows. The dataset includes both multimodal and independent unimodal annotations, allowing both unimodal and multimodal sentiment analysis. Each clip has a manual transcription. The dataset was annotated by five students to label each clip into one of three classes: -1 (negative), 0 (neutral), or 1 (positive). Then, the final sentiment annotation falls into one of five categories: negative, weakly negative, neutral, weakly positive, and positive.
- 8) *M³ED*: A multimodal multiscene multilabel emotional dialogue (M³ED) dataset was recently proposed in [42]. The dataset contains 990 dyadic emotional dialogues, 9082 turns, and 24 449 utterances, derived from 56 different TV series. M³ED is the first multimodal emotional dialogue dataset in Chinese. Similar to Friends, the dataset is annotated on utterance level with seven emotion categories, including Ekman's six basic emotions (happy, surprise, sad, disgust, anger, and fear) and neutral.

C. Implementation Details

1) *Tasks*: We performed sentiment analysis and emotion recognition tasks on the selected eight datasets, as the testbed to evaluate the affective modeling ability of four LLMs. Particularly, we adhered to others' existing works to ensure a fair comparison. For the SST dataset, we undertook both binary and five-class classification tasks. Binary sentiment classification tasks, distinguishing between positive and negative sentiments, were applied to the MOSI, MOSEI, and CH-SIMS datasets. The Mastodon dataset and the MOSI underwent a three-class classification: positive, neutral, and negative. Lastly, for the Friends, DailyDialog, and M³ED datasets, we executed a seven-class emotion classification.

2) *Usage of LLMs*: When working with LLMs such as ChatGPT, we use a "prompt"—a specific input text—to elicit a response from the model. In our study on emotion recognition, we thoroughly investigated various prompting strategies. For context-free datasets (SST, MOSI, MOSEI, and CH-SIMS), we assessed two strategies: context-free zero-shot and context-free few-shot prompting. Meanwhile, for context-dependent datasets (Friends, Mastodon, DailyDialog, and M³ED), we explored three strategies: context-free zero-shot prompting, context-aware zero-shot prompting, and context-aware few-shot prompting. Essentially, "context-free" refers to predicting the emotion-based solely on the content of the sentence itself. In contrast, "context-aware" entails the prediction process that takes into account the surrounding context within the same conversation. Furthermore, "zero-shot" indicates that no prior knowledge or examples are provided for the specific task, whereas "few-shot" implies that a limited number of

demonstration examples are given to facilitate inference-time learning. Detailed prompts for each strategy are provided as follows.

- 1) Context-free zero-shot prompting (i.e., w/o context, w/o ICL).

Classify the sentiment of the sentence to Emotion 1, Emotion 2, ... or Emotion k. No need for you to explain. Don't repeat my sentence. Give me the simplest answer with a list and corresponding number: (provide multiple sentences from a test set with serial number)

- 2) Context-free few-shot prompting (i.e., w/o context, w/ ICL).

Examples of sentiment classification: (provide few training data and corresponding labels from a train set). According to the above examples, classify the sentiment of the sentence as Emotion 1, Emotion 2, ... or Emotion k. No need for you to explain. Don't repeat my sentence. Give me the simplest answer with a list and corresponding number: (provide multiple sentences from a test set with serial number)

- 3) Context-aware zero-shot prompting (i.e., w/ context, w/o ICL).

According to the context of a conversation, classify the sentiment of every sentence to Emotion 1, Emotion 2, ... or Emotion k. No need for you to explain. Don't repeat my sentence. Give me the simplest answer with a list and corresponding number: (provide all sentences of one dialogue with serial number)

- 4) Context-aware few-shot prompting (i.e., w/ context, w/ ICL).

Examples of sentiment classification: (provide few training data and corresponding labels from a train set). According to the above examples, based on the context of conversation, classify the sentiment of every sentence to Emotion 1, Emotion 2, ... or Emotion k. No need for you to explain. Don't repeat my sentence. Give me the simplest answer with a list and corresponding number: (provide all sentences of one dialogue with serial number)

Note that, K is the number of sentiment/emotion categories, determined by each specific task.

For example, the prompt is "Classify the sentiment of sentence to Positive or Negative" for MOSEI as $k = 2$. The instruction "No need for you to explain. Do not repeat my sentence. Give me the simplest answer with a list and corresponding

number." is aimed at a more concise and clearer response, making it easier to retrieve the desired sentiment/emotion prediction. Furthermore, for those binary datasets (positive and negative), we add "Do not answer neutral." into the prompt, which can significantly reduce the number of neutral responses. However, there are still unexpected neutral responses in the binary classification (positive versus negative) in such a manner. In addition, for any sentence, whether from the training set (as task demonstrations) or from the test set, we add two asterisks (**) as separators before and after the sentence. These separators can make it easier for the model to distinguish the beginning and end of sentences, avoiding the model considering multiple sentences as one paragraph and only providing one label instead. Also, when performing zero/few-shot prompting on context-free datasets, incorporating multiple sentences into a single query notably decreased the evaluation's time and cost requirements. Specifically, 100 test sentences were fed in one prompt for zero-shot prompting, while 50 sentences in one prompt for few-shot prompting. In contrast, when performing zero-shot prompting on context-dependent datasets, we reduced the number of sentences to one. This was to prevent the model from gaining any contextual information. Last, the number of task demonstrations added to the prompt under few-shot prompting settings cannot exceed 100 due to the limited number of tokens (4096) set in ChatGPT. We selected an equal number of sentences per sentiment/emotion category: 50 sentences per category for binary classification, 30 per category for three-class classification, 20 per category for five-class classification, and 10 per category for seven-class classification. To ensure fairness in comparison between the three LLMs, for ICL strategies, we applied the same training data points per dataset as task demonstrations when testing each LLM.

After collecting the LLM responses, we postprocessed them to address any inconsistencies. Notably, due to occasional model instability, we sometimes received results outside our predefined classification range. For instance, the model might return "Embarrassed" or "Grateful," which are not within our specified emotion categories. To counteract this, we adjusted the model's temperature setting it to its maximum value of 2 and reprompted the model until we obtained a result within the expected range. If, after five attempts, we still did not get a response within our classification range, we deemed the sentiment classification for that sentence as incorrect and randomly assigned it to an incorrect classification category.

3) *Model Explainability of LLMs*: Providing clear and understandable explanations for the decisions made by complex models is an active area of research. As LLMs generate natural language responses, one potential approach could be to design prompts that request the models to elucidate the reasoning behind their decisions in human-readable format. To showcase the potential of this approach, we opted to conduct an exploratory experiment, using ChatGPT and selecting a few conversations from the Friends dataset. For each conversation, one of the sentences was chosen to be asked for the explanation behind the model's prediction. To solicit this additional information, we appended the following request to the original prompt this further request "In addition, please explain how

TABLE II
PERFORMANCE COMPARISON BETWEEN LLMs AND A SOTA STUDY ON THE **MASTODON** DATASET MEASURED BY ACCURACY (ACC), MACRO-F1 (F1), MACRO-PRECISION (PRECISION), AND MACRO-RECALL (RECALL)

Model [in Percent]	Acc	F1	Precision	Recall
DARER (2022) [43]	-	59.59	56.04	63.33
ChatGPT (w/o context)	65.50	64.46	72.04	63.39
ChatGPT (w/ context)	65.41	65.57	67.20	65.11
ChatGPT (w/ context + w/ ICL)	65.94	66.09	67.72	65.48
Claude (w/o context)	63.05	66.28	66.33	66.88
Claude (w/ context)	64.02	63.25	63.85	66.23
Claude (w/ context + w/ ICL)	64.27	64.51	64.75	66.97
Bing Chat (w/o context)	60.51	60.77	62.17	62.94
Bing Chat (w/ context)	60.95	61.19	61.99	63.28
Bing Chat (w/ context + w/ ICL)	63.66	63.94	64.14	65.42
Gemini (w/o context)	66.20	65.84	70.13	64.75
Gemini (w/ context)	67.86	67.96	68.96	67.49
Gemini (w/ context + w/ ICL)	66.55	66.76	67.19	66.80

Note: Bold indicates the best results for a specific metric on a particular dataset.

you judge the emotion of the <order of the selected sentence> sentence in the conversation based on the context.” The primary aim of this experiment was to showcase whether ChatGPT is capable of providing sound explanations for its emotion recognition decisions. By doing so, we hoped to gain a deeper insight into the model’s decision-making process and to further explore the possibilities of integrating XAI principles with LLM capabilities.

4) *SOTA Specialized Models*: For performance comparison with specialized affective computing models, we selected recently published SOTA works with competitive performance on each selected dataset, respectively.

5) *Evaluation Metrics*: The primary metrics we employed for performance evaluation were accuracy and the macro F1 score. For the M³ED dataset, we used the weighted average F1 score to ensure a fair comparison with other studies utilizing the same dataset. Similarly, for the Mastodon dataset, we calculated both macro precision and macro recall.

IV. EVALUATION AND DISCUSSIONS

We conducted experiments using eight datasets, with the comparative results presented in Tables II–V. All the results were obtained on the test sets of the corresponding datasets. Compared with SOTAs benchmarks, the LLM-based model demonstrates comparable, if not superior, performance.

A. ICL

Equipping with a few demonstration examples, few-shot prompting can significantly improve the LLM’s performance in affective modeling tasks.

As shown in Tables IV and V, on these four context-free sentiment analysis datasets, few-shot prompting surpasses zero-shot prompting in most cases. For example, when testing with ChatGPT, the few-shot prompting strategy obtained increased accuracy (95.31% over 92.32%) on the SST-2 task. For the same task, Claude, Bing Chat and Gemini also gain benefit from the ICL strategy, achieving improved accuracy when compared with zero-shot prompting where demonstration examples are

not available. Similar observations can be found on SST-5, CH-SIMS, MOSI-2, MOSI-3, and MOSEI as well.

Table II presents the result performance on the Mastodon dataset. As this dataset is composed of dialogues, i.e., posts/sentences with context information, we consider the performance comparison between zero-shot and few-shot promptings when the whole test dialogue is fed into the model as the input query. Table II shows that few-shot prompting outperforms zero-shot prompting in all measures on this dataset.

For multiclass emotion recognition tasks on three context-dependent datasets (Friends, DailyDialog, and M³ED), results in terms of Acc, UA, F1, and accuracy per emotion category are given in Table III. When the performance metrics of Acc, UA, and F1 is concerned, performance enhancements are obtained by the ICL strategy in 22 out of the total 36 cases (derived from 3 measures * 4 LLMs * 3 datasets = 36 cases).

B. Accuracy

For each of the selected datasets, SOTA performance from specialized models in previous works is provided for comparison. It is shown that, in general, LLM-based models yield competent or even better performance, both in binary classification and multiclass classification tasks. For instance, on the SST-2 and CH-SIMS datasets (cf. Table V), the best performance in accuracy is obtained by Claude, reaching 95.86% and 88.70%, respectively, outperforming SOTA specialized models.

When comparing the four selected LLMs, it is found that in general, ChatGPT, Claude and Gemini are better than Bing Chat. For instance, as shown in Table IV, on the MOSI-3 and MOSEI datasets, Bing Chat’s performance is lower than ChatGPT, Claude and Gemini. One of the potential reasons for the inferiority of Bing Chat on affective modeling is that the model is tailored for search. However, there is no consistent observation suggesting a particular LLM or prompting strategy as superior over the rest, as the optimal model and strategy setting vary across different datasets.

Further, we observe that incorporating context information (both context within the dialogue and context knowledge from the demonstrations) could enhance the recognition performance

TABLE III
PERFORMANCE COMPARISON ON **FRIENDS** (FIRST PART), **DAILYDIALOG** (SECOND PART), AND **M³ED** (THIRD PART) IN TERMS OF ACCURACY (ACC), F1, AND UNWEIGHTED ACCURACY (UA)

Friends Model [in Percent]	Acc	F1	UA	Neu (57.89)	Joy (13.68)	Sad (3.82)	Fea (1.44)	Ang (7.24)	Sur (12.87)	Dis (3.06)
CNN-BiLSTM (2017) [44]	77.40	-	39.40	87.00	60.30	28.70	0.00	32.40	40.90	26.70
BERT+SRL-GNN-8 (2020) [45]	72.10	-	53.71	84.32	69.92	48.39	31.03	47.06	73.51	21.74
XLNet+SRL-GNN-8 (2020) [45]	72.82	-	53.41	85.34	73.98	67.74	3.45	60.00	61.59	21.74
PRE-CODE (2020) [46]	81.30	65.90	-	-	-	-	-	-	-	-
ChatGPT (w/o context)	72.29	54.31	52.33	85.31	73.68	38.82	40.63	37.27	53.85	36.76
ChatGPT (w/ context)	63.65	51.92	59.43	69.31	77.30	61.18	68.75	47.83	34.27	57.35
ChatGPT (w/ context + w/ ICL)	63.38	50.26	57.89	67.37	83.55	51.76	62.50	54.04	36.01	50.00
Claude (w/o context)	56.63	44.16	52.74	58.42	65.10	47.05	59.37	64.59	43.70	30.88
Claude (w/ context)	51.51	41.22	56.53	49.34	76.64	61.18	81.25	80.75	17.13	29.41
Claude (w/ context + w/ ICL)	58.43	44.86	53.92	63.30	63.49	43.53	68.75	69.57	33.57	35.29
Bing Chat (w/o context)	40.31	33.52	40.87	37.22	58.55	36.47	31.25	52.82	30.07	39.71
Bing Chat (w/ context)	55.62	45.25	56.17	52.06	72.70	65.88	37.50	71.43	45.10	48.53
Bing Chat (w/ context + w/ ICL)	56.91	44.86	53.53	56.25	75.66	51.76	43.75	62.73	43.36	41.18
Gemini (w/o context)	69.82	50.91	45.36	92.23	43.42	48.24	31.25	52.80	29.02	20.59
Gemini (w/ context)	73.14	53.04	46.92	94.33	53.95	45.88	28.13	52.17	36.36	17.65
Gemini (w/ context + w/ ICL)	72.78	59.26	54.07	89.20	58.88	60.00	37.50	62.73	36.36	33.82
DailyDialog Model [in Percent]	Acc	F1	UA	Neu (83.10)	Joy (12.51)	Sad (1.12)	Fea (0.17)	Ang (0.99)	Sur (1.77)	Dis (0.34)
CoMPM (2022) [47]	-	53.15	-	-	-	-	-	-	-	-
DARER (2022) [43]	-	53.42	-	-	-	-	-	-	-	-
ChatGPT (w/o context)	66.74	34.68	61.51	67.24	67.22	63.73	88.24	68.64	50.00	25.53
ChatGPT (w/ context)	71.74	42.13	69.45	72.30	70.76	63.73	82.35	72.88	56.34	68.09
ChatGPT (w/ context + w/ ICL)	71.19	40.85	67.27	71.21	75.76	61.76	88.24	66.10	43.97	63.83
Claude (w/o context)	74.78	38.10	52.36	80.15	50.54	48.04	47.06	70.34	49.14	21.28
Claude (w/ context)	76.47	40.40	55.86	82.63	46.61	49.02	64.71	74.58	54.31	19.15
Claude (w/ context + w/ ICL)	75.59	40.58	57.10	80.32	53.48	56.82	58.82	73.73	55.17	21.28
Bing Chat (w/o context)	75.40	42.06	52.52	78.33	68.79	41.18	52.94	51.69	53.45	21.28
Bing Chat (w/ context)	77.56	46.52	58.47	80.21	71.34	47.05	70.59	55.93	58.62	25.53
Bing Chat (w/ context + w/ ICL)	76.23	46.66	58.80	78.91	69.38	47.06	64.71	52.54	54.31	44.68
Gemini (w/o context)	79.09	35.30	37.41	89.00	37.00	34.21	23.53	29.66	25.00	23.40
Gemini (w/ context)	81.68	44.86	49.15	89.67	48.18	49.02	41.18	41.53	31.90	42.55
Gemini (w/ context + w/ ICL)	81.19	45.92	53.15	88.72	48.97	48.04	58.82	44.92	37.93	44.68
M ³ ED Model [in Percent]	Acc	F1	UA	Neu (44.16)	Joy (8.52)	Sad (17.47)	Fea (1.55)	Ang (17.52)	Sur (5.59)	Dis (5.19)
DialogueGCN (2019) [49]	-	46.09	-	-	-	-	-	-	-	-
DialogueRNN (2019) [47]	-	48.80	-	-	-	-	-	-	-	-
MDI (2022) [42]	-	49.42	-	-	-	-	-	-	-	-
ChatGPT (w/o context)	44.47	40.40	31.89	69.11	39.66	19.35	24.62	24.73	25.11	20.64
ChatGPT (w/ context)	45.39	43.00	35.91	65.34	35.47	27.11	29.23	26.77	34.89	32.57
ChatGPT (w/ context + w/ ICL)	46.32	45.39	32.33	76.06	29.61	13.90	33.85	27.58	19.15	26.15
Claude (w/o context)	34.90	34.83	31.10	33.32	19.27	20.16	36.92	70.52	37.02	0.46
Claude (w/ context)	53.73	50.14	34.14	83.94	24.58	24.66	29.23	46.74	17.02	12.84
Claude (w/ context + w/ ICL)	53.80	52.83	42.56	75.20	29.33	32.02	52.31	47.42	44.68	16.97
Bing Chat (w/o context)	36.44	38.51	36.79	37.90	47.49	22.21	23.08	37.23	61.13	27.52
Bing Chat (w/ context)	43.42	45.36	43.59	45.39	54.19	27.38	35.38	46.47	70.64	25.69
Bing Chat (w/ context + w/ ICL)	47.73	50.53	48.22	52.78	48.04	32.43	49.23	45.65	55.74	53.67
Gemini (w/o context)	50.85	42.29	26.51	95.53	14.25	11.85	20.00	22.01	18.72	3.21
Gemini (w/ context)	56.30	50.69	32.72	94.34	38.72	19.75	15.38	38.72	21.70	11.47
Gemini (w/ context + w/ ICL)	58.22	54.54	38.79	91.59	33.24	33.92	29.23	33.70	19.57	30.28

Note: The best results are in bold. Note that, F1 indicates macro-F1 for Friends and DailyDialog, and represents weighted average F1 for M³ED. The ratio of each emotion class in the test set is given under each emotion class.

for emotional sentences within dialogues. Tables II and III display results on context-available datasets. In all 13 global metrics (4 for Mastodon, 3 for Friends, 3 for M³ED and 3 for DailyDialog), 10 of them are improved when ChatGPT is deployed, 11 metrics are increased by Claude, all 13 metrics are increased when leveraging Bing Chat, and 12 metrics are increased by Gemini.

Not only do LLMs achieve comparable overall performance, but they also display better recognition performance in minority emotion classes. The ratio of each emotion class within two emotion datasets is listed in Table III. Conventional models tend to struggle with emotions constituting less than 10% of the dataset samples compared to more prevalent emotion classes. For example, the “fear” emotion class (which makes

TABLE IV
PERFORMANCE COMPARISON BETWEEN LLMs AND SOTA WORKS ON THE **MOSI**
AND **MOSEI** DATASETS MEASURED BY ACCURACY (ACC) AND MACRO-F1 (F1)

Model [in Percent]	MOSI-2		MOSI-3		MOSEI	
	Acc	F1	Acc	F1	Acc	F1
TFR-Net (2021) [48]	83.49	-	-	-	-	-
CHFN (2022) [49]	85.20	-	-	-	-	-
SeqSeq2Sent (2018) [50]	-	-	77.00	-	-	-
CTFN (2021) [51]	-	-	80.79	-	-	-
TBJE (2020) [52]	-	-	-	-	81.90	-
COGMEN (2022) [53]	-	-	-	-	84.42	-
ChatGPT (w/o ICL)	86.13	85.92	73.62	62.21	85.60	84.43
ChatGPT (w/ ICL)	89.18	88.93	76.38	63.32	80.74	79.80
Claude (w/o ICL)	87.04	86.55	79.88	63.67	85.83	84.81
Claude (w/ ICL)	88.72	88.37	82.65	63.92	82.11	81.33
Bing Chat (w/o ICL)	70.73	70.72	65.60	55.76	69.84	68.36
Bing Chat (w/ ICL)	88.26	88.12	67.20	55.97	72.01	70.28
Gemini (w/o ICL)	87.96	87.59	72.45	60.89	79.42	78.83
Gemini (w/ ICL)	88.41	88.09	73.47	61.40	80.57	79.97

Note: Bold indicates the best results for a specific metric on a particular dataset.

TABLE V
PERFORMANCE COMPARISON BETWEEN LLMs AND SOTA WORKS ON THE **SST**
AND **CH-SIMS** DATASETS MEASURED BY ACCURACY (ACC) AND MACRO-F1 (F1)

Model [in Percent]	SST-2		SST-5		CH-SIMS	
	Acc	F1	Acc	F1	Acc	F1
GNN (2020) [54]	92.13	-	-	-	-	-
BT-TAPT (2021) [55]	92.40	-	-	-	-	-
SEMGraPh-P (2022) [56]	94.23	-	-	-	-	-
ELECTRA (2020) [57]	-	-	57.89	-	-	-
SentiLARE (2020) [58]	-	-	58.59	-	-	-
SentiWSP (2022) [59]	-	-	59.32	-	-	-
MLF-DNN (2020) [41]	-	-	-	-	80.26	-
ChatGPT (w/o ICL)	92.32	92.32	53.36	49.43	79.66	78.78
ChatGPT (w/ ICL)	95.31	95.31	54.50	52.41	87.28	86.43
Claude (w/o ICL)	95.31	95.30	49.86	40.44	88.70	87.44
Claude (w/ ICL)	95.86	95.86	51.27	46.19	77.40	76.86
Bing Chat (w/o ICL)	81.56	81.53	48.64	48.63	72.03	70.99
Bing Chat (w/ ICL)	85.64	85.64	50.32	50.25	82.77	82.09
Gemini (w/o ICL)	94.37	94.36	49.45	39.61	86.72	85.44
Gemini (w/ ICL)	95.42	95.41	54.18	49.53	87.85	86.91

Note: Bold indicates the best results for a specific metric on a particular dataset.

up 1.44% of the Friends dataset) achieved a peak accuracy of 31.03% in SOTA methodologies. In contrast, LLMs across various prompting strategies outperformed this, with Claude achieving the best result at 81.25%. Similarly, Claude’s accuracy of recognizing the “angry” emotion in Friends, at 80.75%, was markedly superior to SOTA results. Comparable trends are observed with other minority emotions on the M³ED dataset. This can be attributed to the vast and diverse text LLMs have been trained on, which encompasses a broader emotional range than a specific dataset.

C. Generalization

Our experimental results show that, without explicitly training on a specific dataset, LLMs exhibit an impressive generalization ability in multiple affective modeling tasks. Instead of learning specialized models for each of the seven distinct emotional datasets separately, all three selected models have shown

their capabilities in managing varied emotion recognition tasks, underlining their versatility and broad applicability.

In particular, as discussed earlier, the LLMs could achieve comparable performance in zero-shot prompting scenarios. And this performance is further enhanced by the implementation of few-shot prompting, indicating that LLMs can adapt and learn from limited contextual information during the inference.

Most importantly, the LLMs show a strong generalization ability in their emotion analysis across different corpora and domains. The datasets utilized in this study encompass a wide array of text types, including TV show scripts, movie reviews, social media posts, and YouTube video transcriptions. Despite the variances in style, tone, and context inherent to these data sources, LLMs consistently produce reliable emotion recognition results.

Additionally, the results attest to the LLMs’ cross-language generalization capability. The corpora used in this study spanned both English and Chinese languages, further

TABLE VI
TWO EXAMPLES ILLUSTRATING CHATGPT’S EXPLANATIONS FOR CONVERSATIONAL EMOTION RECOGNITION

Example 1	<p>According to the context of the conversation, classify the emotion of every sentence to Neutral, Disgust, Surprise, Anger, Joy, Sadness, or Fear. Give me the simplest answer with a list and corresponding number. In addition, please explain how you judge the emotion of the seventh sentence in the conversation based on the context:</p> <ol style="list-style-type: none"> Joey: Hey! Rachel: Hi! Joey: Oh you know uh Kash, really liked you the other day. He said he thought you were charming. Ross: Okay I-I just have to stop by my place first. Joey: To tape the game? You do this every time Ross, you are Ross: No-no, I-I have to see if this apartment became available. Joey: Oh, you are switching apartments? Ross: It is not for me, it is for Rachel. Joey: But Rachel has an apartment. Ross: Yeah, but when the baby comes she is gonna want to move. Joey: She is? Ross: Yeah, you did not expect her to live there with a baby did you? Joey: I guess I did not really think about it. Ross: Ezel! Hey Gunther! You’re an ezel! Ross: Damnit!
ChatGPT response:	<p>1. Joy 2. Joy 3. Joy 4. Neutral 5. Anger 6. Fear 7. Surprise 8. Neutral 9. Surprise 10. Neutral 11. Surprise 12. Neutral 13. Neutral 14. Disgust 15. Anger</p> <p>Explanation for the seventh sentence: The emotion for the seventh sentence is “Surprise”. Joey’s response, “Oh, you’re switching apartments?” indicates surprise or astonishment in reaction to Ross’s statement about looking for a new apartment for Rachel. The word “Oh” at the beginning of the sentence and the questioning tone imply that Joey was not expecting this news, and it caught him off guard. This element of surprise is the emotional judgment based on the context of the conversation.</p>
Example 2	<p>According to the context of the conversation, classify the emotion of every sentence to Neutral, Disgust, Surprise, Anger, Joy, Sadness, or Fear. Give me the simplest answer with a list and corresponding number. In addition, please explain how you judge the emotion of the ninth sentence in the conversation based on the context:</p> <ol style="list-style-type: none"> Phoebe: Hey! Monica: Hey. Phoebe: Hey, why are you mopping your ceiling? Monica: Oh, there is banana on it. Rachel: You said she was bald. Phoebe: Yeah, she was bald, she is not now. Rachel: How could you not tell me that she has hair? Phoebe: I do not know, I hardly ever say that about people. Rachel: Ohh, well, this is just perfect! Phoebe: Well I am sorry, I thought you said it was okay. Rachel: Yeah, I said what was okay when I thought she was some weird bald chick. I mean, y’know, that girl has hair got Phoebe: Well, maybe it will not work out. Maybe Ross will not like her personality. Rachel: Why, does she have a bad personality? Phoebe: Oh no, Bonnie’s the best!
ChatGPT response:	<p>1. Joy 2. Neutral 3. Surprise 4. Disgust 5. Surprise 6. Neutral 7. Anger 8. Neutral 9. Anger 10. Neutral 11. Anger 12. Neutral 13. Neutral 14. Joy</p> <p>Explanation for the emotion of the ninth sentence: In the ninth sentence, “Ohh, well, this is just perfect!” Rachel displays anger and frustration. She initially thought Bonnie was bald, and now that she knows Bonnie has hair, Rachel’s reaction suggests a negative response. The phrase “this is just perfect” is used sarcastically and negatively, indicating Rachel’s annoyance and aligning with the emotion of “Anger.”</p>

Note: The sentences selected for explanation, along with their corresponding decisions and associated explanations, are highlighted for clarity.

emphasizing the models’ adaptability. Such cross-language generalization ability holds considerable potential for universal emotion recognition models that overcome language barriers.

D. Explanation

To further explore the explainability of LLMs in their emotion recognition decisions, we delve into ChatGPT’s explanations for two sample conversations, as showcased in Table VI. This provides an initial insight into how LLMs rationalize their decisions in the context of emotion recognition.

As presented in Table VI, ChatGPT provides reasonable explanations for the sentences under analysis, effectively utilizing the context information within the conversation. For example, taken out of context, the sentence “Ohh, Well, this is just perfect!” might seem positive. However, when considering the preceding conversation, the model correctly identifies it as conveying a negative emotion. It is important to note that while these intelligible explanations offer some insight into the model’s decision-making process, the correlation between these human-readable explanations and the actual decision-making mechanism of the model is still ambiguous. This relationship

warrants further investigation, which is crucial for establishing more reliable and transparent emotion recognition systems.

V. CHALLENGES AND OPPORTUNITIES

Although we have seen the great capability of LLMs in emotion recognition from the aspects of zero/few-shot learning, capturing the context information, and explanation, several other challenges still need to be overcome in future. In this section, we delve into several challenges and opportunities associated with the utilization of LLMs for affective computing tasks.

A. Robustness

The robustness of LLMs in affective computing tasks is also a crucial topic. When users interact with LLMs such as ChatGPT, they may introduce a range of natural errors, encompassing typographical mistakes and grammatical inaccuracies. These errors mirror the imperfect, real-world language usage seen in everyday communications. Hence, it becomes pivotal to investigate the resilience of LLMs to such inconsistencies. It is equally important to explore strategies to enhance model robustness against this “noise,” especially when discerning emotions from textual content. How well these models handle and respond to these errors not only tests their robustness but also their capability to simulate human-like comprehension and interaction. An LLM’s capacity to correctly interpret sentiment or emotional intent in the face of such errors may serve as a strong indicator of its utility and effectiveness in real-world scenarios.

B. Adaptation/Fine-Tuning

In this present study, we have explored the capabilities of off-the-shelf LLMs without any additional adaptation/fine-tuning for emotion recognition. Although LLMs are trained on vast amounts of text data across various domains, their abilities can be further tailored to specific goals through continual pretraining or fine-tuning by leveraging domain-specific data.

In the context of affective computing, to further enhance the affective modeling ability of LLMs, it would be of interest to deploy further pretraining approaches such as domain-adaptive pretraining (DAPT) [60] and task-adaptive pretraining (TAPT) [61]. In DAPT, an LLM undergoes additional unsupervised pretraining on a large corpus of unlabeled domain-specific data, with the aim of adapting the LLM to a particular domain. Conversely, in TAPT, an LLM is further pretrained on the unlabeled training set for a specific task. Compared to DAPT, TAPT uses a significantly smaller, but far more task-relevant pretraining corpus. Research in [60] suggests that tailoring a pretrained model to the domain of a target task via multiphase adaptive pretraining can significantly enhance task performance. It may be valuable to investigate whether further pretraining an LLM toward the domain of the emotion recognition corpus of interest can yield benefits. However, it is worth paying attention to issues such as the time and cost of pretraining, as well as the potential for incurring inductive biases.

In addition, fine-tuning an LLM in a supervised learning way can provide benefits for a specific affective modeling task.

For instance, the original model may excel in tasks such as binary sentiment classification, but underperform in more nuanced sentiment analysis tasks such as seven-class sentiment classification. Likewise, the original model may be good at analyzing emotions from structured movie reviews, but struggle with less formal and more diverse forms of expression in social media posts. In these conditions, it would be helpful to conduct further fine-tuning for domain-specific adaptation for optimal performance. One primary approach for fine-tuning LLMs is instruction tuning. This approach involves formatting the specific task training data with natural language task descriptions to guide the LLM toward the task objective. Along with the paired input–output training data, these specially designed task descriptions in natural language play a pivotal role in helping the LLM to understand the desired goal. In the realm of affective computing, we may fine-tune an LLM using customized emotion recognition dataset(s) to enhance the model’s adaptability. Interestingly, even when using the same labeled instances on the same dataset, different task instructions can lead to varying performance outcomes. This indicates the significant impact that the phrasing and clarity of task descriptions can have on the effectiveness of instruction tuning [62]. It could be highly beneficial to investigate directions such as identifying the most effective task descriptions for emotional datasets, exploring strategies to optimize fine-tuning using a minimal amount of target domain emotional data, or how to efficiently fine-tune an LLM for affective computing.

Finetuning LLMs enables models to learn domain-specific knowledge, using a smaller dataset within the target domain. However, given the massive number of parameters that LLMs possess, full-parameter tuning can be highly resource-intensive, potentially leading to increased computational costs and time. To address this, several parameter-efficient fine-tuning (PEFT) techniques have been devised, which only fine-tune a small number of (extra) parameters and meanwhile maintain the strong performance [63]. These include strategies such as prefix tuning which optimizes trainable vectors added to each transformer block as a prefix, prompt tuning which learns task-specific prompt embeddings to complement the input text embeddings, adapter tuning which incorporates small neural network modules without changing existing parameters [64], and low-rank adaptation (LoRA) which decomposes the weight update matrix into lower-dimensional matrices without losing too much important information [65]. It would be interesting to examine the effectiveness of these various PEFT techniques to tailor existing open-source LLMs for emotion recognition tasks. This will expand our understanding of how to most effectively enhance LLMs to meet specific aims in the field of affective computing.

C. Privacy and Security

Ensuring user privacy and data security are crucial considerations when deploying LLMs for affective computing in real-world applications. The processed text is inherently sensitive, including users’ emotional states, personal experiences, and various forms of identifiable information. Thus, it is essential

to implement robust safeguards to protect this information and address privacy concerns when leveraging the advantages of LLMs. As highlighted in [66], an empirical investigation indicates that larger and more complex models are more prone to leaking private information. During their experiments, four evaluated LLMs inadvertently disclosed demographic information (location, age, and gender) to varying degrees during the sentiment analysis process [66]. Moreover, they also conducted a critical evaluation of various privacy-preserving technologies, such as differential privacy, to quantify the impact of several privatizing methods on alleviating information leakage. An interesting research direction could be the exploration of solutions that effectively balance performance and privacy, with the goal of developing privacy-preserving LLM-based affective computing models.

D. Model Compression

On one hand, LLMs are powerful and show appealing performance in affective computing. On the other hand, they also present challenges due to their substantial memory requirements during the inference stage, making deployment costly. In the context of affective computing, computation often takes place locally rather than on cloud-based systems. This approach is primarily adopted to address privacy and security concerns, as it prevents sensitive data from being transmitted elsewhere. However, these local processors are frequently deployed on resource-constrained devices such as mobile phones, where the substantial size and computational requirements of LLMs pose a significant challenge. Therefore, it becomes essential to explore model compression techniques, such as quantization methods [67], to adapt these powerful models for local deployment. These approaches aim to reduce the memory footprint and latency of LLMs and meanwhile maintaining high accuracy, making them more suitable for implementation on local, resource-limited devices. By pursuing this research direction, we can potentially bring the power of LLMs to real-world applications within the area of affective computing, while respecting user privacy and security.

E. Large Multimodal Models

In March 2023, openAI released GPT-4 [68], extending the text input to multimodal signals. To be more specific, this latest GPT-4 model accepts text and images inputs and produces text outputs, demonstrating its capability in conducting multimodal dialogues with humans. Within the realm of emotion recognition, the strategy of integrating information from multiple modalities is typically superior to relying solely on one single modality. Emotion, as a complex human phenomenon, is expressed through multiple channels, including text, speech, facial expressions, body language, and more. Instead of considering text only, jointly analyzing multiple modalities can achieve a more holistic and accurate understanding of the emotional state being expressed. While LLMs provide a straightforward approach for estimating emotions based on text inputs, integrating these models with other modalities holds substantial promise. Future research directions may focus on investigating optimal

approaches to fuse the multimodality information effectively without losing crucial emotional context.

Furthermore, other research efforts are being directed toward the development of large vision models [69], [70], large speech models [71], [72], and multimodal LLMs [73]. Instead of utilizing traditional models and gaining knowledge from small-scale emotional datasets, these large models hold immense potential to bring large gains in emotion recognition performance, in both single modality and multimodality. Incorporating these advanced models could potentially transform the future landscape of affective computing.

VI. CONCLUSION

In this study, we provided a deep insight into the capability of LLMs in the domain of emotion recognition. We explored this by evaluating the performance of four leading LLMs across eight datasets, and compared the results against SOTA works. The experimental outcomes have shown that the LLMs have shown superior performance in sentiment analysis and emotion recognition tasks, especially when identifying minority emotion categories. Unlike traditional specialized models, the LLMs benefit from being trained on expansive data volumes. This vast exposure equips them with impressive generalization capabilities and paves the way for enhanced explainability in their predictions. Importantly, our findings also indicate that LLMs harness context to enhance emotion estimation. Furthermore, we have mapped out potential trajectories for the future of LLM-based emotion recognition systems. Our vision for these systems combines adaptability, multimodality, robustness, privacy preservation, and compatibility with resource-constrained devices.

REFERENCES

- [1] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1634–1654, Apr. 2023.
- [2] Y. Wang et al., "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion*, vol. 83, pp. 19–52, Jul. 2022.
- [3] T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multimodal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 96–107, Jan. 2019.
- [4] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. F. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Hong Kong, China, 2019, pp. 154–164.
- [5] D. Kollias and S. Zafeiriou, "Exploiting multi-CNN features in CNN-RNN based dimensional emotion recognition on the OMG in-the-wild dataset," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 595–606, Jul. 2021.
- [6] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, San Francisco, CA, USA, 2016, pp. 3593–3597.
- [7] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. F. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Honolulu, HI, USA, 2019, pp. 6818–6825.
- [8] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 553–564, Jul. 2021.

- [9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.
- [10] J. Kossaihi et al., "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.
- [11] A. A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, Australia, Jul. 2018, pp. 2236–2246.
- [12] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, 2022, pp. 6907–6911.
- [13] W. Lu, T. Tan, and H. Ma, "Bi-branch vision transformer network for EEG emotion recognition," *IEEE Access*, vol. 11, pp. 36233–36243, 2023.
- [14] I. Perikos, S. Kardakis, and I. Hatzilygeroudis, "Sentiment analysis using novel and interpretable architectures of hidden Markov models," *Knowl.-Based Syst.*, vol. 229, pp. 1–18, Oct. 2021.
- [15] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition," in *Proc. 22nd Annu. Conf. Int. Speech Communication Assoc. (INTERSPEECH)*, Brno, Czech Republic, 2021, pp. 1748–1752.
- [16] J. Koon et al., "ChatGPT: Jack of all trades, master of none," *Inf. Fusion*, vol. 99, 2023, Art. no. 101861.
- [17] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 1168–1172.
- [18] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Transformer-based self-supervised learning for emotion recognition," in *Proc. 26th IEEE Int. Conf. Pattern Recognit. (ICPR)*, Montréal, Canada, 2022, pp. 2605–2612.
- [19] Y. Chang, Z. Ren, T. T. Nguyen, K. Qian, and B. W. Schuller, "Knowledge transfer for on-device speech emotion recognition with neural structured learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.
- [20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3164–3176, Oct.–Dec. 2023.
- [21] P. Sarkar and A. Etemad, "Self-supervised ECG representation learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1541–1554, Jul. 2022.
- [22] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.
- [23] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition," in *Proc. 17th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Chengdu, China, 2020, pp. 117–121.
- [24] M. Deramgozin, S. Jovanovic, H. Rabah, and N. Ramzan, "A hybrid explainable AI framework applied to global and local facial expression recognition," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Kaohsiung, Taiwan, 2021, pp. 1–5.
- [25] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted graph attention convolutional network for micro-expression recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 2871–2880.
- [26] X. Wang, M. Li, Y. Chang, X. Luo, Y. Yao, and Z. Li, "Multimodal cross-attention Bayesian network for social news emotion recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 1–9.
- [27] R. Bommasan et al., "On the opportunities and risks of foundation models," 2021, [arXiv:2108.07258](https://arxiv.org/abs/2108.07258).
- [28] W. X. Zhao et al., "A survey of large language models," 2023, [arXiv:2303.18223](https://arxiv.org/abs/2303.18223).
- [29] J. Wei et al., "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, pp. 1–30, Mar. 2022.
- [30] T. B. Brown et al., "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, 2020, pp. 1877–1901.
- [31] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT," *IEEE Intell. Syst.*, vol. 38, no. 2, pp. 15–23, Mar. 2023.
- [32] Y. Bai et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022, [arXiv:2204.05862](https://arxiv.org/abs/2204.05862).
- [33] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," 2022, [arXiv:2212.08073](https://arxiv.org/abs/2212.08073).
- [34] R. Anil et al., "Gemini: A family of highly capable multimodal models," 2023, [arXiv:2312.11805](https://arxiv.org/abs/2312.11805).
- [35] M. Reid et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024, [arXiv:2403.05530](https://arxiv.org/abs/2403.05530).
- [36] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Seattle, WA, USA, 2013, pp. 1631–1642.
- [37] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, Miyazaki, Japan, 2018, pp. 1597–1601.
- [38] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. T. Le, "Multi-task dialog act and sentiment recognition on mastodon," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, Santa Fe, NM, USA, 2018, pp. 745–754.
- [39] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, Taipei, Taiwan, 2017, pp. 986–995.
- [40] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.
- [41] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Virtual, 2020, pp. 3718–3727.
- [42] J. Zhao et al., "M3ED: Multi-modal multi-scene multi-label emotional dialogue database," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 5699–5710.
- [43] B. Xing and I. W. Tsang, "DARER: Dual-task temporal relational recurrent reasoning network for joint dialog sentiment classification and act recognition," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 3611–3621.
- [44] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vancouver, Canada, 2017, pp. 873–883.
- [45] C. T. Heaton and D. M. Schwartz, "Language models as emotional classifiers for textual conversation," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, Seattle, WA, USA, 2020, pp. 2918–2926.
- [46] W. Jiao, M. R. Lyu, and I. King, "Exploiting unsupervised data for emotion recognition in conversations," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Virtual, 2020, pp. 4839–4846.
- [47] J. Lee and W. Lee, "CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technol. (NAACL)*, Seattle, WA, USA, 2022, pp. 5669–5679.
- [48] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. 29th ACM Int. Conf. Multimedia (MM)*, Chengdu, China, 2021, pp. 4400–4407.
- [49] J. Guo, J. Tang, W. Dai, Y. Ding, and W. Kong, "Dynamically adjust word representations using unaligned multimodal information," in *Proc. 30th ACM Int. Conf. Multimedia (MM)*, Lisbon, Portugal, 2022, pp. 3394–3402.
- [50] H. Pham, T. Manzini, P. P. Liang, and B. Poczós, "Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis," in *Proc. Grand Challenge Workshop Human Multimodal Lang. (Challenge-HML)*, Melbourne, Australia, 2018, pp. 53–63.
- [51] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Virtual, 2021, pp. 5301–5311.
- [52] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A Transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang. (Challenge-HML)*, Seattle, WA, USA, 2020, pp. 1–7.

- [53] A. Joshi, A. Bhat, A. Jain, A. V. Singh, and A. Modi, "COGMEN: Contextualized GNN based multimodal emotion recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technologies (NAACL)*, Seattle, WA, USA, 2022, pp. 4148–4164.
- [54] Y. Zhang, X. Yu, Z. Cui, S. Wu, Z. Wen, and L. Wang, "Every document owns its structure: Inductive text classification via graph neural networks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Virtual, 2020, pp. 334–339.
- [55] J. Lee, J. Kim, and P. Kang, "Back-translated task adaptive pretraining: Improving accuracy and robustness on text classification," 2021, *arXiv:2107.10474*.
- [56] B. Wang, B. Liang, J. Du, M. Yang, and R. Xu, "SEMGraph: Incorporating sentiment knowledge and eye movement into graph model for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Abu Dhabi, United Arab Emirates, 2022, pp. 7521–7531.
- [57] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "Pre-training transformers as energy-based cloze models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Virtual, 2020, pp. 285–294.
- [58] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Sentiment-aware language representation learning with linguistic knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Virtual, 2020, pp. 6975–6988.
- [59] S. Fan et al., "Sentiment-aware word and sentence level pre-training for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Abu Dhabi, United Arab Emirates, 2022, pp. 4984–4994.
- [60] S. Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Virtual, 2020, pp. 8342–8360.
- [61] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, Australia, 2018, pp. 328–339.
- [62] J. Wei et al., "Finetuned language models are zero-shot learners," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Virtual, 2022, pp. 1–46.
- [63] N. Ding et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023.
- [64] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, 2019, pp. 2790–2799.
- [65] E. J. Hu et al., "LORA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Virtual, 2022, pp. 1–26.
- [66] R. Plant, V. Giuffrida, and D. Gkatzia, "You are what you write: Preserving privacy in the era of large language models," 2022, *arXiv:2204.09391*.
- [67] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," in *Low-Power Computer Vision*. London, U.K.: Chapman & Hall, 2022, pp. 291–326.
- [68] OpenAI, "GPT-4 technical report," 2023, *arXiv:2303.08774v3*.
- [69] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 12104–12113.
- [70] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 12009–12019.
- [71] R. Huang et al., "AudioGPT: Understanding and generating speech, music, sound, and talking head," 2023, *arXiv:2304.12995*.
- [72] Y. Zhang et al., "Google USM: Scaling automatic speech recognition beyond 100 languages," 2023, *arXiv:2303.01037*.
- [73] S. Yin et al., "A survey on multimodal large language models," 2023, *arXiv:2306.13549*.



Zixing Zhang (Senior Member, IEEE) received the master's degree in physical electronics from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2010, and the Ph.D. degree in computer engineering from Technical University of Munich (TUM), Munich, Germany, in 2015.

He is now a Full Professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. From 2017 to 2019, he was a Research Associate with the Department of Computing at the Imperial College London (ICL), London, U.K. Before that, he was a Postdoctoral Researcher with the University of Passau, Passau, Germany. His research focuses on human-centred emotion and health computation. He has authored more than 110 publications in peer-reviewed books, journals, and conference proceedings, leading to more than 5000 citations (h-index 40). He serves as an Associate Editor of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and *Frontiers in Signal Processing*, an Editorial Board Member of *Nature Scientific Reports*, and a Guest Editor of IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.



Liyizhe Peng (Student Member, IEEE) is currently working toward the master's degree with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.

His research interests include large language models, model adaptation, and emotion recognition.



Pang Tao (Student Member, IEEE) is currently working toward the master's degree with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.

Her research topics relate to health-related audio signal processing.



Jing Han (Senior Member, IEEE) received the bachelor's degree in electronic and information engineering from Harbin Engineering University (HEU), Harbin, China, in 2011, the master's degree in integrated circuit from Nanyang Technological University, Singapore, in 2014, and the Ph.D. degree in computer science from the University of Augsburg, Augsburg, Germany, in 2019.

Since 2019, she has been a Postdoctoral Researcher with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K. Her research interests include affective computing and digital health. She (co)authored more than 60 publications in peer-reviewed journals and conference proceedings. She has served as a Program Committee Member of the Audio/Visual Emotion Challenge and Workshop in 2018 and a Technical Program Committee Member of the Association for Computing Machinery (ACM) Multimedia since 2019, and is a Leading Guest Editor of IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, and an Associate Editor of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.



Huan Zhao received the bachelor's and master's degrees in computer application technology and the Ph.D. degree in computer science and technology from Hunan University, Changsha, China, in 1989, 2004, and 2010, respectively.

She is currently a Professor with the College of Computer Science and Electronic Engineering, Hunan University. Her research areas include speech information processing, natural language processing, and multimodal data analysis. She has led over 20 research projects funded by the National

Key R&D Program and the Natural Science Foundation of China. She has published over 100 papers in journals and conferences, such as IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, International Conference on Acoustics, Speech and Signal Processing (ICASSP), and the *Journal of Communications*. She has been granted eight patents.



Björn W. Schuller (Fellow, IEEE) received the diploma degree, in 1999, the doctoral degree, in 2006, and the habilitation and Adjunct Teaching Professorship in the subject area of signal processing and machine intelligence, in 2012, all in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany.

He is a Tenured Full Professor heading the Chair of Embedded Intelligence for Health Care and Well-being, University of Augsburg, Augsburg, Germany,

and a Professor of Artificial Intelligence heading GLAM, Department of Computing, Imperial College London, London, U.K.

Dr. Schuller is the Field Chief Editor of *Frontiers in Digital Health*, former Editor in Chief of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, a President-Emeritus and a Fellow of Association for the Advancement of Affective Computing (AAAC), a Golden Core Awardee of the IEEE Computer Society, a Fellow of International Speech Communication Association (ISCA), a Fellow of British Computer Society (BCS), a Fellow of European Laboratory for Learning and Intelligent Systems (ELLIS), and a Senior Member of ACM. He (co-) authored five books and more than 1200 publications in peer-reviewed books, journals, and conference proceedings leading to more than 56 000 citations (h-index 106).