

Can Large Language Models Generate Effective Datasets for Emotion Recognition in Conversations?

Burak Can Kaplan*, Hugo Cesar De Castro Carneiro*, and Stefan Wermter*

*Department of Informatics, University of Hamburg, Hamburg 22527, Germany

Abstract—Emotion recognition in conversations (ERC) focuses on identifying emotion shifts within interactions, representing a significant step toward advancing machine intelligence. However, ERC data remains scarce, and existing datasets face numerous challenges due to their highly biased sources and the inherent subjectivity of soft labels. Even though Large Language Models (LLMs) have demonstrated their quality in many affective tasks, they are typically expensive to train, and their application to ERC tasks—particularly in data generation—remains limited. To address these challenges, we employ a small, resource-efficient, and general-purpose LLM to synthesize ERC datasets with diverse properties, supplementing the three most widely used ERC benchmarks. We generate six novel datasets, with two tailored to enhance each benchmark. We evaluate the utility of these datasets to (1) supplement existing datasets for ERC classification, and (2) analyze the effects of label imbalance in ERC. Our experimental results indicate that ERC classifier models trained on the generated datasets exhibit strong robustness and consistently achieve statistically significant performance improvements on existing ERC benchmarks.

Index Terms—large language models, machine learning, data generation, affective computing

I. INTRODUCTION

Emotion recognition in conversations (ERC) is a relatively new field of study that focuses on identifying and understanding human emotions expressed during interactions [1]. Its primary goal is to detect emotion shifts within dialogues, a capability that has become increasingly important with the rise of social robotics and applications requiring emotionally intelligent systems [2]. Large Language Models (LLMs) have demonstrated substantial improvements in various natural language processing (NLP) tasks, including affective computing [3], and hold potential as effective tools for generating ERC data. Despite their success, most evaluations of LLMs in affective tasks have been conducted with API-based models, which are expensive, or top-performing local models requiring significant computational resources [4], [5]. Exploring the capabilities of small and general-purpose LLMs in ERC tasks thus emerges as a promising and cost-efficient alternative.

A critical challenge in ERC lies in the scarcity of high-quality, diverse datasets. Most existing datasets are derived from biased sources such as scripted TV shows or social media, which often feature imbalanced label distributions [6], [7]. Moreover, crafting such datasets is costly and time-consuming due to participant recruitment, ethical concerns, unbiased dialogue construction, and the difficulty of accurate and consistent labeling. Annotating emotional data often involves subjective interpretations, with annotators frequently providing

TABLE I: ERC datasets

Dataset	Source	Speaker	Emotions	Language
MELD [6]	TV	Multiple	7	English
IEMOCAP [8]	Scripted	2	6	English
EmoryNLP [7]	TV	6	7	English
DailyDialog [10]	Scripted	2	7	English
CPED [11]	TV	Multiple	13	Chinese
EC [12]	Twitter	3	4	English
KDEmor [13]	TV	Multiple	3	Korean

inconsistent labels for the same utterance. Typically, a majority vote is used to select a label when there is disagreement, but this process is limited by the small number of annotators (usually 3–5), leading to reliability issues [7], [8]. Additionally, as highlighted in Tab. I, existing ERC datasets vary significantly in their emotion label sets, speaker numbers, and languages, making it difficult to combine them effectively for transfer learning. Furthermore, the inconsistencies in emotion categories across datasets limit their interoperability, leading some studies to rely on mappings based on psychological studies to align these different label sets [5]. However, such mappings are often rough approximations, raising concerns about their accuracy and applicability. Rooted in the concept that only the entity expressing a particular affective state can fully recognize it [9], we hypothesize that by having the LLM generate both utterances and their corresponding emotion labels simultaneously, we can address these issues and significantly improve dataset consistency. Tab. I highlights the characteristics of popular ERC datasets and underscores their limitations.

To tackle these limitations, we propose leveraging a small, general-purpose LLM to synthesize new ERC datasets. By generating both dialogue lines and their corresponding emotion labels in a single step, we aim to improve the consistency and reliability of ERC data while avoiding the costs and complexities associated with traditional data collection and annotation methods. To ensure comparability with existing benchmarks, we generate six new datasets, with two corresponding datasets for each of the three widely used ERC benchmarks. This alignment allows us to systematically evaluate the potential of LLM-generated datasets to supplement existing resources, mitigate label imbalance, and enhance ERC model performance. Fig. 1 shows an example of a dialogue generated by our LLM.

Our contributions are as follows: Firstly, we demonstrate the capability of a small LLM to generate consistent, multi-party ERC datasets suitable for training ERC models. Secondly, we propose a methodology to evaluate the quality and validity of

Joey: Hey, did you guys hear about Chandler's new job? (neutral)
Rachel: Oh my god, what is it? (surprise)
Joey: He's gonna be a janitor at the college. (neutral)
Ross: That's a demotion, isn't it? (sadness)
Joey: Nah, it's just a temporary thing until he finds something better. (neutral)
Chandler: Hey, can you guys keep this between us? I don't want everyone to know. (anger)

Fig. 1: Example of a generated dialogue, featuring multiple speakers, utterances, and corresponding emotion labels.

synthesized ERC data, focusing on their utility in improving model robustness and performance. Thirdly, we assess the effects of different label distributions in ERC datasets and analyze how these imbalances affect existing benchmarks. The prompts and parameters used for data generation are shared in the paper, ensuring reproducibility with local LLMs. Additionally, the code repository will be made available to facilitate further advancements in LLM-based ERC dataset creation.

The remainder of the paper is organized as follows: Section II reviews existing ERC datasets and related research on LLMs for affective computing, dataset enhancement, and synthetic data generation. Section III details the LLM setup, the dataset synthesis process, prompt engineering, and all parameters used, providing all necessary information for transparency. Section IV presents experimental evaluations of the generated datasets, and Section V concludes with insights and directions for future research.

II. RELATED WORK

A. Existing Datasets

In this study, we focus on the three most popular ERC datasets in Papers With Code¹: MELD [6], EmoryNLP [7], and IEMOCAP [8]. These datasets are chosen to assess the performance of ERC models on our generated datasets, as well as on these existing datasets.

MELD is a dataset constructed by extracting lines from the “Friends” TV series. It encompasses 7 emotions: Neutral, Disgust, Anger, Sadness, Fear, Joy, and Surprise. There is a high imbalance among its labels, with Neutral being the most common. On the other hand, Disgust and Fear are notably rare. In some studies, authors perform classification task without either one or both of these labels. According to Papers With Code, weighted F1 classification results for MELD usually lie within the range from 60% to 70%.

The **EmoryNLP** dataset is also constructed from the lines of “Friends” TV series, but it employs different emotion labels than MELD, which are: Sad, Mad, Scared, Powerful, Peaceful,

Joyful, and Neutral. For the annotation process, 4 annotators participated, and only 6.17% of the annotations correspond to labels in which there was an unanimous agreement among the annotators. In 9.39% of the annotations, each annotator labeled the same data with different label. Majority voting was used to select the annotation in most cases. The weighted F1 results for EmoryNLP, as reported on Papers With Code, are around 35%, which is significantly lower compared to other datasets, characterizing EmoryNLP as a challenging dataset.

IEMOCAP is a scripted dyadic dataset which also provides multimodal information. It encompasses 10 labels: Neutral, Happiness, Sadness, Anger, Excited, Frustration, Fear, Surprise, Disgust, and Other (Uninformative). Disgust does not show up in any record in the conversations of the validation split. Classification tasks on IEMOCAP involve the utilization of specific subsets of its available classes. Those that do not include Disgust and Other are denoted 8-way. Fear and Surprise are very rare throughout the dataset, so papers often do not include them. This classification task is denoted 6-way. Finally, although with more observations, Excited and Frustration appear considerably less than Neutral, Happiness, Sadness, and Anger. A classification using solely these 4 labels is denoted 4-way. Results aggregated on Papers With Code reveal that weighted F1 scores for IEMOCAP are slightly higher than those for MELD regardless of the number of classes used. This paper adopts the 6-way classification, as it is the most commonly employed approach in the literature.

B. LLM Usage

MELD, IEMOCAP, and EmoryNLP are multimodal datasets, but they are often exclusively benchmarked in the textual modality [5], [14] due to the inherent noise in their audio and visual components, which poses challenges to ERC model training. Even though some approaches aim to enhance audio-visual data quality through preprocessing [15], [16] or modifying the modality fusing approach [17]–[19], opportunities for improvement in this field still persist. Language models are frequently leveraged to increase the quality of existing datasets as well. For instance, pretrained language models can provide additional information at the utterance and conversation levels, thus increasing the context of an existing ERC dataset [20]. Additionally, LLMs are also used in annotation process of affective speech data to enhance its quality [4], [21].

Empathetic intelligence also requires measuring the capabilities of an LLM in various affective tasks. Existing literature indicates that LLMs exhibit considerable empathetic intelligence. Zhu et al. [22] demonstrate that transformer-based architectures are capable of distinguishing emotions in dialogues, and Deng et al. [23] use transformer architectures to generate dialogue responses. Amin et al [4] assess ChatGPT in affective computing, revealing that even a generalized model can achieve decent results. ChatGPT’s knowledge has been also measured in solving affective computing problems, namely sentiment analysis, personality assessment, and suicide tendency detection [3]. LLMs are also evaluated in tasks like affective support, multi-party conversations and ERC [24],

¹<https://paperswithcode.com/task/emotion-recognition-in-conversation>

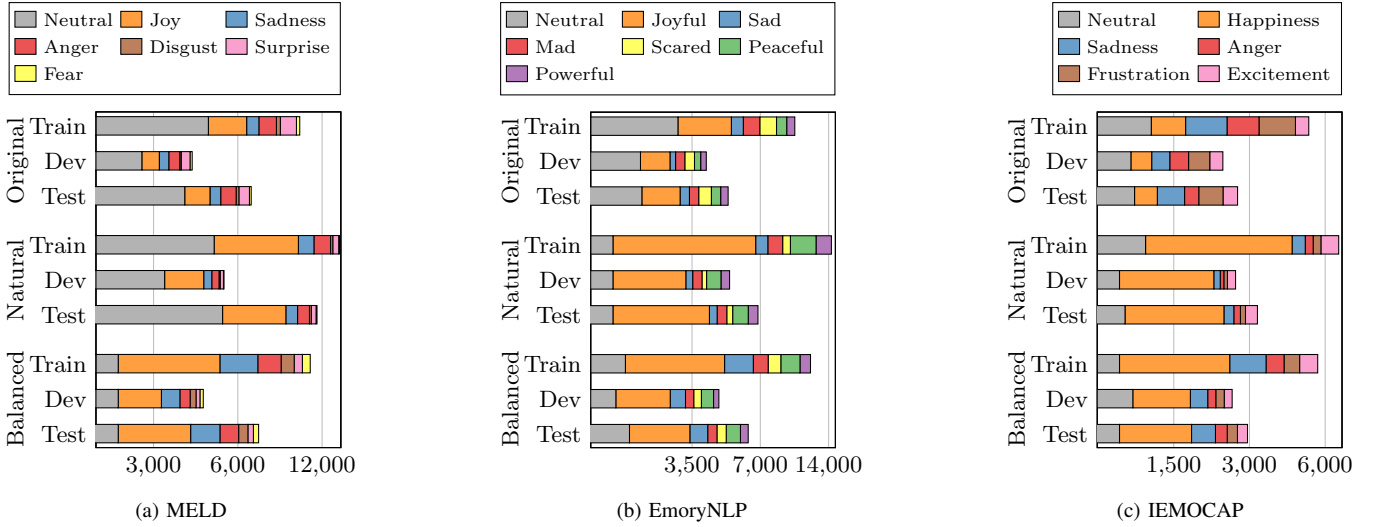


Fig. 2: Label distribution of reference and generated datasets. The horizontal axis represents the number of utterances within each dataset, and is shown in a logarithmic scale to enhance the visualization of rare labels.

[25]. Tu et al. [26] extract detailed additional knowledge from ERC data using ChatGPT and measures its impact on ERC models. Feng et al. [27] show that LLMs can serve as effective classifiers for affect recognition in conversation tasks. Additionally, LLM-based models have achieved high classification scores on widely used ERC datasets [5], further supporting the suitability of LLMs for ERC data generation.

There are approaches to enhance the LLMs’ data generation capabilities. Eldan et al. [28] employ GPT-3.5 and GPT-4 to generate child-level language to train small language models. Josifoski et al. [29] present a synthetic data generation pipeline that involves prompting LLMs to generate text from coherent triplets extracted from a knowledge graph. Conversely, Chung et al. [30] attempt to increase the diversity of LLM data, acknowledging potential trade-offs with lower output accuracy. Veselovsky et al. [31] use LLM synthetic data to train classifiers, evaluating them on real data with various strategies. In summary, the existing literature underscores LLMs’ effectiveness in diverse affective tasks. However, despite the significant limitations and noise present in existing ERC datasets, research on leveraging LLMs specifically for ERC data generation remains scarce, leaving an open opportunity for further exploration in this direction.

III. DATASET GENERATION

In the LLM selection phase, we ran small dialogue generation experiments, and observed that 7 billion-sized models proved incapable of generating creative and diverse dialogues while keeping sufficient output consistency for use as a dataset, often exhibiting repetitions of words or sentences. Regarding the larger models, we decided not to use ChatGPT despite yielding the most favorable results, due to our emphasis on ensuring the reproducibility of our approach. Although 33 billion-sized models yielded decent results, dialogues from the

13 billion-sized model appeared natural and required roughly 25 GB VRAM. Thus, we opted for using a small model with a reasonable GPU to offer an affordable and computation-efficient solution for ERC dataset generation, with Vicuna 1.5² being the 13 billion-sized model to provide the best and most consistent results, and one of the most popular open sourced LLMs.

A. Natural and Balanced Data

To assess our local LLMs’ capabilities in generating multi-party affective conversations across various aspects, we generated two types of dataset with the same set of labels and dialogue structure of the three mentioned datasets in Sec. II-A, providing a total of 6 new datasets. These dataset types are named “Natural” and “Balanced”. Each dataset was intentionally over-generated to safeguard against data limitations and to facilitate comparison with their corresponding original datasets. Fig. 2 displays total utterances, label amounts and label distributions of each dataset.

Natural datasets comprise dialogues created freely by an LLM without any predetermined bias in its generation process. These datasets show the broad potential of LLMs in creating dialogues, with the distribution of the emotion labels within them being closer to reality. For example, emotions like happiness, sadness, or even the absence of emotions being far more common than emotions like fear and surprise. The generation of natural data is particularly useful for the development of affective interactive systems, e.g., more realistic and natural social robots to be used in real-life applications. With natural datasets, we aim to assess how naturally LLMs generate dialogues without emotional pre-conditioning.

Balanced datasets, conversely, comprise dialogues with more evenly distributed emotions and they are designed to address

²<https://huggingface.co/lmsys/vicuna-13b-v1.5>

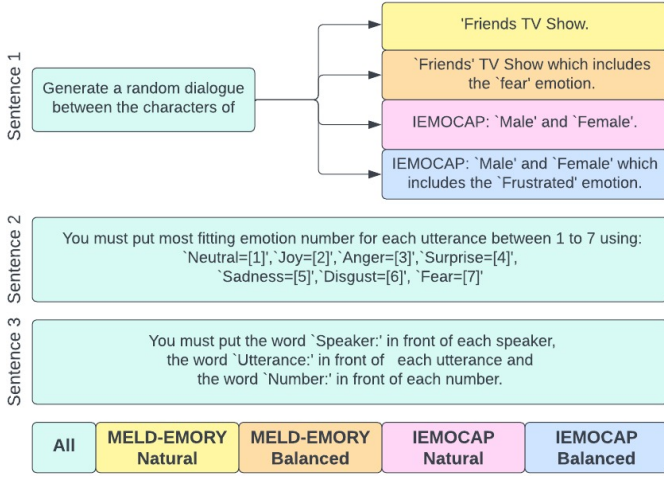


Fig. 3: Prompt examples for each generated dataset. Each prompt consists of three sentences (1:Task Definition, 2:Labelling with Logic Reasoning, 3:Structuring). The emotion labels specified in the diagram are used for illustration and are modified with matching labels for each dataset being generated.

class imbalance in existing ERC datasets while maintaining natural dialogue flow. To preserve the naturalness of the dialogues, we instruct the LLM to generate conversations that include at least one utterance with a specific given emotion. However, the content and placement of these within the dialogue, as well as the emotions of the remaining utterances, are left for the LLM to decide. This procedure does not yield a dataset where all emotions are uniformly distributed; rather, it ensures that a particular emotion appears in a significant number of dialogues. Fig 2 shows that the balanced datasets have the most balanced label distributions among all datasets. Due to the intentional bias introduced in its generation process, balanced datasets are not well suited for generative and affective usage, since rare emotions may appear more often than they do in real-life conversations. Nevertheless, these datasets are better fitting in the development of more accurate classifiers.

B. Prompt Engineering

Prompting is a critical aspect of this research, directly biasing the LLM towards the goal. In the generation of natural data, four tasks were needed to achieve a proper output which are providing speaker names, utterances, consistent emotion labels and structure. Due to the varying nature of the speakers in the target datasets, distinct prompts tailored to each dataset were employed. Specifically, IEMOCAP involves dyadic interactions with unnamed male and female participants, while EmoryNLP and MELD are derived from conversations in the “Friends” TV show.

The most challenging aspect of the prompting process was obtaining accurate emotion labels from the LLM, given that these models tend to hallucinate and often forget prompt details when faced with complex tasks, leading to inconsistencies. To restrict the LLM to generate emotion labels from a specific label group, we assign each emotion label with a number as

TABLE II: LLM Parameters used. RP: Repetition Penalty

Temperature	Top_p	Top_k	RP	Typ. p
0.7	1	10000	1	0.995

symbols and ask the LLM to provide one of them for each utterance. In the literature, the same logic is used in LLM Logic Reasoning [32]–[34], and it is a proven technique that improves the LLM performance. In our case, this method ensures the generation of consistent labels across datasets.

The last task involved employing a prompt to generate structured outputs to facilitate the retrieval of the necessary data. To ensure the parseability of that structure, we instructed the LLM to prepend speaker, utterance, and number (representing emotion) to the corresponding data. Fig. 3 provides the prompts used to generate the dialogues with this structure. At the generation, all sentences are concatenated and submitted to the LLM as a single prompt. Fig. 1 displays an example of ERC dialogue generated by the LLM.

For the generation of balanced data, the prompt is kept the same, except for the last part of the first sentence, where we instruct the LLM to ensure the dialogue contains at least one utterance expressing a specific emotion. Fig. 3 offers an example of prompt used to ensure the existence of at least one utterance expressing fear. To preserve the naturalness of the dialogue, we restrain from specifying the number associated with that label. This prompt iterates through all labels within the target dataset, ensuring an equal number of dialogues containing utterances expressing some particular emotion.

We tested these prompts with the LLM several times with the parameters provided in Tab. II. It was observed that the LLM consistently produced the same outputs, ensuring reproducibility. Specific words drive each LLM to produce different outputs based on their weights, and the prompts here offered were tailored for Vicuna 1.5(13b).

C. LLM Parameters

For the utilization of the local LLM, we employed Textgen-WebUI Chat API³ with Langchain⁴, which provides access to all LLM parameters. For reproducibility, the seed is randomized for diversity but fixed to maintain consistent random numbers for each LLM run. Tab. II shows all parameters used in our data generation process.

Temperature is not crucial in our case due to the fixed seed, so we opted to keep it at its default value. We kept top p, top k, and typical p very high, to enhance context diversity and provide more options in the dialogues. Furthermore, we minimized repetition penalty to avoid limiting the LLM based on the tokens it produced.

IV. EXPERIMENTS AND ANALYSES

As outlined in Section III-A, two datasets—natural and balanced—were generated for each of the datasets discussed

³<https://github.com/oobabooga/text-generation-webui>

⁴<https://www.langchain.com/>

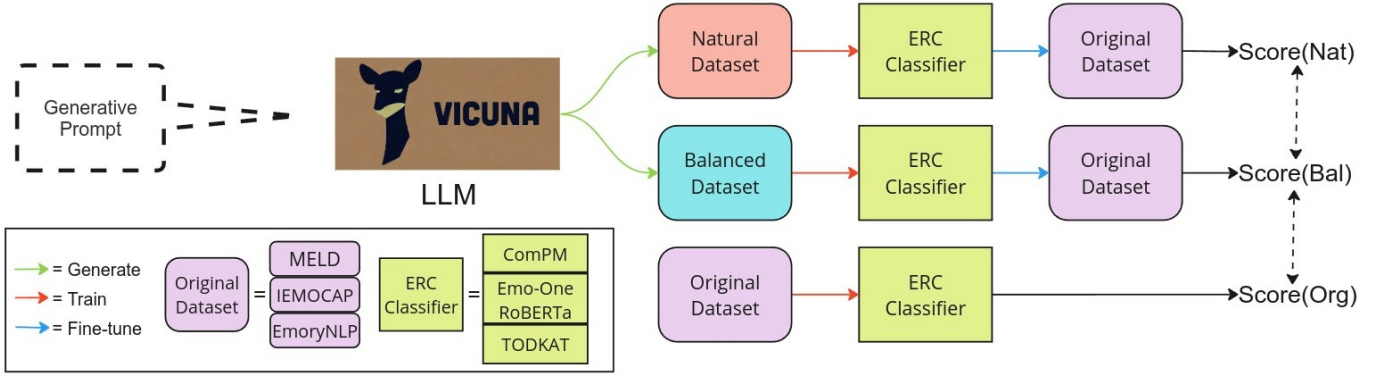


Fig. 4: Diagram of our pipeline.

in Section III-B. These datasets share the same speakers, structure, and emotion label set as their corresponding datasets for the purpose of comparison. To assess the utility of the generated datasets, we employed three popular ERC classifier architectures: CoMPM [35], EmoOne-RoBERTa [36], and TODKAT [22]. These models were chosen based on their available implementations and competitive performance scores on all three datasets: MELD, EmoryNLP, and IEMOCAP. These architectures were preferred over some state-of-the-art models due to factors such as the lack of implementations for certain datasets or high VRAM requirements such as 4x80G nVidia A100 [5]. This decision aligns with our aim of delivering a reproducible and affordable solution. Furthermore, the primary focus of this research is not achieving the highest scores but demonstrating whether the datasets generated through our approach boost the ERC classification process. To further validate our findings, we conducted statistical tests on the experiment results from the three classifier architectures, ensuring the significance of our conclusions. The entire pipeline used for evaluation is illustrated in Fig. 4.

A. Assessing Synthetic Dataset Properties

To evaluate the utility of the generated datasets for ERC and the methodology proposed in this paper, we need to assess whether these datasets possess desirable properties relevant to the task. These properties include: 1. having all their splits sampled from the same distribution; 2. being able to train models to be robust to unseen data; and 3. having the potential to pretrain models for better performance when fine-tuned with benchmark datasets. The first property is guaranteed by our methodology since dialogues are independently generated by the LLM using the same prompt and parameters. As dialogues are independently generated and the dataset is only split afterward, all generated datasets have splits sampled from the same distribution. The second property requires training ERC architectures on the generated datasets and subjecting them to unseen data. The third property involves training the same architectures on the generated datasets and further fine-tuning them on the corresponding reference dataset to evaluate whether it results in significant improvements in the recognition capability.

To evaluate both robustness and fine-tuning potential, we split each generated dataset into training, validation, and test sets, ensuring that a portion of the data remained unseen for evaluation at later stages. Additionally, models were tested on unseen data by evaluating them on the original test splits of MELD, IEMOCAP, and EmoryNLP. In the literature, this evaluation strategy—train on synthetic, test on real (TSTR)—is commonly used to assess the effectiveness of synthetic data in downstream tasks and represents the most suitable evaluation scheme for our study [37], [38].

After splitting the datasets, we trained the CoMPM, EmoOne-RoBERTa, and TODKAT architectures separately using the training splits of the generated datasets. Once training was completed, we fine-tuned all models on the training splits of MELD, IEMOCAP, and EmoryNLP for domain adaptation. Finally, we evaluated the fine-tuned models on the test splits of MELD, IEMOCAP, and EmoryNLP to assess: 1. whether they function as robust ERC classifiers, and 2. whether models trained on synthetic data exhibit improved performance compared to their original versions. The results of this experiment can be seen in Tab. III

An initial review of Table III shows that models trained on LLM-generated datasets exhibit strong robustness and generalization capabilities, producing ERC models that perform comparably to or better than those trained on original datasets. Most importantly, across all classifier architectures, models trained on generated datasets (highlighted in gray backgrounds) consistently outperform their original counterparts across all three benchmarks. This demonstrates that even small LLM-generated datasets can enhance ERC classifier performance through transfer learning, thereby achieving the primary objective of this study.

B. Assessment of the Effects of Different Label Distributions

In this subsection, we compare model performances across different synthetic label distributions, as presented in Table III. Surprisingly, our experiments reveal that not all benchmarks behave similarly when evaluating models trained on datasets with different label distributions. Balanced datasets yielded the highest scores on the MELD dataset across all classifiers, highlight that this dataset needs much more balanced labels to be

TABLE III: Performance results of three ERC classifiers across all three benchmarks (wfl). Org denotes scores from the original dataset, while generated dataset scores (Nat and Bal) are highlighted with a colored background for clearer distinction. The highest scores for each benchmark is marked in bold.

	Test Set	ComPM	EmoOne	TODKAT
MELD	Org	65.43	65.46	63.47
	Nat	65.52	66.50	64.20
	Bal	66.16	67.27	64.27
EMORYNLP	Org	37.25	35.93	35.38
	Nat	39.50	38.79	36.77
	Bal	38.93	39.05	37.40
IEMOCAP	Org	65.21	67.19	54.63
	Nat	68.06	69.28	55.96
	Bal	67.87	67.81	53.39

more effective for ERC. In contrast, models trained on natural datasets performed best on IEMOCAP, suggesting that class imbalance is less critical for this benchmark and that training should prioritize data that reflects real-world distributions. For EmoryNLP, classification scores were consistently lower than in the other two benchmarks, indicating its greater complexity. Moreover, no clear advantage was observed between training on balanced versus natural datasets for this dataset. These findings suggest that the impact of label distributions on model performance is dataset-dependent, highlighting the importance of tailoring synthetic dataset generation strategies to the characteristics of individual benchmarks.

C. Further Validation on Findings

To further validate our findings, we conducted an additional experiment, where we subjected the models used in Section IV-A to unseen test data split from the generated datasets additionally. The models were then ranked across all test sets and architectures to enable a comparative analysis.

We performed a Friedman rank sum test [39], which is a non-parametric statistical test that has been established as a scientifically valid way to evaluate the significant improvement of a classifier in comparison to several others over various datasets [40]. Due to the relatively small number of classifiers (9) and datasets (9), we could not resort to chi-squared approximations and calculated the exact p -values using the formula introduced by Eisinga et al. [41].

The Friedman test is non-parametric, as the dataset does not follow a normal distribution. Consequently, ranking-based analysis is applied instead of using raw W-F1 scores. Each trained model instance—whether trained solely on the original dataset or pretrained on synthetic data and subsequently fine-tuned on the original dataset—was assigned a rank from 1 (highest W-F1 score) to 9 (lowest W-F1 score). This ranking procedure was repeated for each test set, corresponding to each row in Table IV.

After ranking, the rank sums for each model across all test sets were calculated. A lower rank sum indicates consistently strong performance, while a higher rank sum suggests weaker

performance across the test sets. The rank sum of models pretrained on synthetic datasets (and fine-tuned on the original dataset) was then subtracted from the rank sum of models trained solely on the original dataset. If the rank sum of a model pretrained on synthetic data was lower than that of a model trained solely on the original dataset, this indicated that the pretrained model exhibited stronger overall performance. Conversely, if the rank sum was higher, the original model outperformed the pretrained one. Significant differences in rank sums suggested consistent performance disparities, which could indicate statistical significance.

Bonferroni-corrected p -values were calculated to provide a quantitative measure of statistical significance [42]. Table IV presents each model’s rank (in parentheses), rank sums, and absolute differences. The bottom row of the table provides the corresponding Bonferroni-corrected p -values.

The low p -values obtained through the Friedman rank sum test suggest that it is highly unlikely that the performance improvements of models pretrained on synthetic data are due to random chance. This effect is especially evident in CoMPM and EmoOne-RoBERTa, which demonstrated the strongest performance across benchmarks. However, no statistical significance was observed for TODKAT, likely due to inherent limitations of this model. These results further reinforce the potential of LLM-generated datasets in enhancing ERC classifier performance while highlighting variations in model-specific responses to synthetic pretraining.

V. CONCLUSION

This study introduces a reproducible, affordable, and computationally efficient approach for generating ERC datasets using a small, resource-efficient LLM with structured prompt engineering. By addressing key challenges such as soft labels, dataset incompatibilities, and class imbalance, our method enables scalable dataset creation without relying on expensive or black-box models. Our experimental results demonstrate that models trained on LLM-generated datasets exhibit enhanced recognition capabilities and improved performance on ERC benchmarks. We proposed a systematic approach to assessing the quality of these datasets, and confirmed their potential for fine-tuning ERC models. Statistical tests further solidify these findings, providing robust support for the impact of our approach. Additionally, we assessed the effects of having different label distributions in ERC, and our results highlight the need of having more balanced data for some particular benchmarks.

Looking ahead, fine-tuning existing LLMs (e.g., Llama 3) could enable the generation of even larger and more diverse affective datasets, further improving dataset adaptability. Additionally, given that scalability is one of the biggest advantages of LLM synthetic data, the research on its effects in ERC is also can be an interesting direction. Moreover, our methodology provides a foundation for generating customizable datasets not only for ERC but also for other NLP tasks. By releasing our parameters, code, and prompts, we aim

TABLE IV: Extended version of results(W-F1), in which the test results of test splits of generated datasets and Friedman test calculations included. The W-F1 scores are ranked row-wise (1 for the highest, 9 for the lowest), with the highest in bold. The “Rank” part shows the sum of ranks and the absolute difference between rank sums compared to the original dataset-trained counterpart. The “p” row provides the p -values from the Friedman rank sum test.

Test Set		ComPM			EmoOne-RoBERTa			TODKAT		
		Org	Nat	Bal	Org	Nat	Bal	Org	Nat	Bal
<i>MELD</i>	Org	65.43 ⁽⁶⁾	65.52 ⁽⁴⁾	66.16 ⁽³⁾	65.46 ⁽⁵⁾	66.50 ⁽²⁾	67.27⁽¹⁾	63.47 ⁽⁹⁾	64.20 ⁽⁸⁾	64.27 ⁽⁷⁾
	Nat	48.07 ⁽⁷⁾	50.96⁽¹⁾	50.29 ⁽³⁾	49.18 ⁽⁶⁾	50.95 ⁽²⁾	49.24 ⁽⁵⁾	46.52 ⁽⁹⁾	49.37 ⁽⁴⁾	47.86 ⁽⁸⁾
	Bal	58.66 ⁽⁸⁾	60.77 ⁽⁶⁾	65.99 ⁽²⁾	61.17 ⁽⁵⁾	61.20 ⁽⁴⁾	66.10⁽¹⁾	57.34 ⁽⁹⁾	60.46 ⁽⁷⁾	62.30 ⁽³⁾
<i>EMORY-NLP</i>	Org	37.25 ⁽⁶⁾	39.50⁽¹⁾	38.93 ⁽³⁾	35.93 ⁽⁸⁾	38.79 ⁽⁴⁾	39.05 ⁽²⁾	35.38 ⁽⁹⁾	36.77 ⁽⁷⁾	37.40 ⁽⁵⁾
	Nat	31.66 ⁽⁶⁾	35.89 ⁽²⁾	34.00 ⁽⁵⁾	28.85 ⁽⁸⁾	34.06 ⁽⁴⁾	34.73 ⁽³⁾	28.37 ⁽⁹⁾	37.14⁽¹⁾	31.12 ⁽⁷⁾
	Bal	47.67 ⁽⁷⁾	53.39 ⁽⁴⁾	60.86 ⁽²⁾	46.91 ⁽⁸⁾	50.51 ⁽⁵⁾	60.92⁽¹⁾	38.15 ⁽⁹⁾	49.71 ⁽⁶⁾	56.95 ⁽³⁾
<i>IEMOCAP</i>	Org	65.21 ⁽⁶⁾	68.06 ⁽²⁾	67.87 ⁽³⁾	67.19 ⁽⁵⁾	69.28⁽¹⁾	67.81 ⁽⁴⁾	54.63 ⁽⁸⁾	55.96 ⁽⁷⁾	53.39 ⁽⁹⁾
	Nat	16.76 ⁽⁹⁾	37.58⁽¹⁾	27.08 ⁽⁶⁾	19.84 ⁽⁸⁾	35.85 ⁽²⁾	27.89 ⁽⁵⁾	26.02 ⁽⁷⁾	30.86 ⁽³⁾	30.27 ⁽⁴⁾
	Bal	34.05 ⁽⁸⁾	53.89 ⁽³⁾	59.62 ⁽²⁾	33.20 ⁽⁹⁾	50.26 ⁽⁴⁾	60.35⁽¹⁾	40.23 ⁽⁷⁾	41.94 ⁽⁶⁾	47.64 ⁽⁵⁾
Rank		Org	Nat	Bal	Org	Nat	Bal	Org	Nat	Bal
Sum		63	24	29	62	28	23	76	49	51
Diff.		-	39	34	-	34	39	-	27	25
p		-	0.0034	0.0186	-	0.0186	0.0034	-	0.1273	0.2025

to facilitate further research in synthetic data generation for affective computing and beyond.

REFERENCES

- [1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
- [2] H. Mahdi, S. A. Akgun, S. Saleh, and K. Dautenhahn, “A survey on the design and evolution of social robots — Past, present and future,” *Robotics and Autonomous Systems*, vol. 156, p. 104193, Oct. 2022.
- [3] M. M. Amin, E. Cambria, and B. W. Schuller, “Can ChatGPT’s responses boost traditional natural language processing?,” *IEEE Intelligent Systems*, vol. 38, no. 5, pp. 5–11, 2023.
- [4] 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. arXiv:2303.03186 [cs].
- [5] S. Lei, G. Dong, X. Wang, K. Wang, and S. Wang, “InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task LLMs framework,” Nov. 2023. arXiv:2309.11911 [cs].
- [6] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 527–536, Association for Computational Linguistics, July 2019.
- [7] S. M. Zahiri and J. D. Choi, “Emotion detection on TV show transcripts with sequence-based convolutional neural networks,” in *Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 44–51, 2018.
- [8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, Dec. 2008.
- [9] R. W. Picard, *Affective computing*. MIT press, 2000.
- [10] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, “DailyDialog: A manually labelled multi-turn dialogue dataset,” in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (G. Kondrak and T. Watanabe, eds.), (Taipei, Taiwan), pp. 986–995, Asian Federation of Natural Language Processing, Nov. 2017.
- [11] Y. Chen, W. Fan, X. Xing, J. Pang, M. Huang, W. Han, Q. Tie, and X. Xu, “CPED: A large-scale Chinese personalized and emotional dialogue dataset for conversational AI,” May 2022. arXiv:2205.14727 [cs].
- [12] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, “SemEval-2019 Task 3: EmoContext contextual emotion detection in text,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, (Minneapolis, Minnesota, USA), pp. 39–48, Association for Computational Linguistics, 2019.
- [13] S. Pant, E. Lim, H.-J. Yang, G.-S. Lee, S.-H. Kim, Y.-S. Kang, and H. Jang, “Korean drama scene transcript dataset for emotion recognition in conversations,” *IEEE Access*, vol. 10, pp. 119221–119231, 2022.
- [14] X. Song, L. Huang, H. Xue, and S. Hu, “Supervised prototypical contrastive learning for emotion recognition in conversation,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 5197–5206, Association for Computational Linguistics, Dec. 2022.
- [15] J. Pohjalainen, F. Fabien Ringeval, Z. Zhang, and B. Schuller, “Spectral and cepstral audio noise reduction techniques in speech emotion recognition,” in *Proceedings of the 24th ACM International Conference on Multimedia*, (Amsterdam, The Netherlands), pp. 670–674, ACM, Oct. 2016.
- [16] H. Carneiro, C. Weber, and S. Wermter, “Whose emotion matters? Speaking activity localisation without prior knowledge,” *Neurocomputing*, vol. 545, p. 126271, Aug. 2023.
- [17] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, “MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Singapore, Singapore), pp. 7037–7041, IEEE, May 2022.
- [18] X. Zhang and Y. Li, “A cross-modality context fusion and semantic refinement network for emotion recognition in conversation,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 13099–13110, Association for Computational Linguistics, July 2023.
- [19] L. Maoheng, “Enhanced emotion recognition through multimodal fusion using trimodal fusion graph convolutional networks,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2024.
- [20] J. Kim, H. Ko, S. Song, S. Jang, and J. Hong, “Contextual augmentation of pretrained language models for emotion recognition in conversations,” in *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, (Barcelona, Spain (Online)), pp. 64–73, Association for Computational Linguistics, Dec. 2020.
- [21] S. Latif, M. Usama, M. I. Malik, and B. W. Schuller, “Can large language models aid in annotating speech emotional data? Uncovering new frontiers,” July 2023. arXiv:2307.06090 [cs, eess].

- [22] L. Zhu, G. Pergola, L. Gui, D. Zhou, and Y. He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Online), pp. 1571–1582, Association for Computational Linguistics, Aug. 2021.
- [23] Q. Deng, L. Wu, K. Su, W. Wu, Z. Li, and W. Duan, "Hierarchical fusion framework for multimodal dialogue response generation," in *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024.
- [24] W. Zhao, Y. Zhao, X. Lu, S. Wang, Y. Tong, and B. Qin, "Is ChatGPT equipped with emotional dialogue capabilities?," Apr. 2023. arXiv:2304.09582 [cs].
- [25] C.-H. Tan, J.-C. Gu, and Z.-H. Ling, "Is ChatGPT a good multi-party conversation solver?," in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore, Singapore), pp. 4905–4915, Association for Computational Linguistics, Dec. 2023.
- [26] G. Tu, B. Liang, B. Qin, K.-F. Wong, and R. Xu, "An empirical study on multiple knowledge from ChatGPT for emotion recognition in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, (Singapore), pp. 12160–12173, Association for Computational Linguistics, 2023.
- [27] S. Feng, G. Sun, N. Lubis, C. Zhang, and M. Gašić, "Affect recognition in conversations using large language models," Sept. 2023. arXiv:2309.12881 [cs].
- [28] R. Eldan and Y. Li, "TinyStories: How small can language models be and still speak coherent English?," May 2023. arXiv:2305.07759 [cs].
- [29] M. Josifoski, M. Sakota, M. Peyrard, and R. West, "Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 1555–1574, Association for Computational Linguistics, Dec. 2023.
- [30] J. J. Y. Chung, E. Kamar, and S. Amershi, "Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, pp. 575–593, Association for Computational Linguistics, July 2023.
- [31] V. Veselovsky, M. H. Ribeiro, A. Arora, M. Josifoski, A. Anderson, and R. West, "Generating faithful synthetic data with large language models: A case study in computational social science," May 2023. arXiv:2305.15041 [cs].
- [32] H. Zhang, L. H. Li, T. Meng, K.-W. Chang, and G. Van Den Broeck, "On the paradox of learning to reason from data," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*, 2023.
- [33] A. Creswell and M. Shanahan, "Faithful reasoning using large language models," Aug. 2022. arXiv:2208.14271 [cs].
- [34] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with language model prompting: A survey," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 5368–5393, Association for Computational Linguistics, July 2023.
- [35] J. Lee and W. Lee, "CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Seattle, United States), pp. 5669–5679, Association for Computational Linguistics, July 2022.
- [36] J. Lee, "The emotion is not one-hot encoding: Learning with grayscale label for emotion recognition in conversation," in *Interspeech 2022*, pp. 141–145, ISCA, Sept. 2022.
- [37] S. L. Hyland, C. Esteban, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans," *stat*, vol. 1050, p. 8, 2017.
- [38] Y. Yuan, Y. Liu, and L. Cheng, "A multi-faceted evaluation framework for assessing synthetic data generated by large language models," 2024.
- [39] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, pp. 675–701, Dec. 1937.
- [40] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, Dec. 2006.
- [41] R. Eisinga, T. Heskes, B. Pelzer, and M. Te Grotenhuis, "Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers," *BMC Bioinformatics*, vol. 18, p. 68, Dec. 2017.
- [42] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, pp. 52–64, Mar. 1961.