Persona-Conditioned Emotion Classification of Conversation Using LLMs

Israel Cuevas, Andrew Mackey, Susan Gauch
Department of Electrical Engineering and Computer Science
University of Arkansas – Fayetteville
Fayetteville, Arkansas, USA
e-mail: ibcuevas@uark.edu, almackey@uark.edu, sgauch@uark.edu

***Abstract*—Large Language Models (LLM) have demonstrated success across a wide range of tasks in the field of natural language processing, including within the emotion classification task of language. With the recent advancements of agentic workflows and conversational chatbots in the field of artificial intelligence, it is fairly common to employ the use of personas to bias LLM interactions toward domain-specific applications. In this study, we investigate the impact of persona-conditioned models for the task of emotion classification along with model confidence of performance under these persona-conditioned settings. Our statistically-significant results ($p < 0.001$) demonstrate that persona-conditioned models affect model performance while also demonstrating the performance differences between each of the personas. Furthermore, through our experiments we observed variations in model confidence between both open and closed LLMs for the Emotion Recognition in Conversation (ERC) task.**

Keywords-natural language processing; emotion analysis; large language models.

I. Introduction

Emotion Recognition in Conversation (ERC) is a task in which the affective state of language in conversation is identified for a wide range of applications [1][2]. Recent advancements in the field have led to major gains in performance by leveraging contextualized representations of utterances and surrounding conversation.

Persona conditioning is a technique where speaker-specific variations are captured in domain-specific settings [3]. This can be observed in conversational settings where speakers often differ in their communicative style, interpersonal roles, affective tendencies and expressiveness, and habitual responses to events. One example of this can be found in sarcastic language as it may indicate amusement for one speaker, while at the same time, the same words or expressions may be indicative of frustration for another. As a consequence, emotionally ambiguous utterances may necessitate the use of context-specific features, such as speaker or prior conversation, to accurately interpret and categorize the meaning. Prior literature for emotion classification has focused primarily on utterance-level semantics, local conversational context, or multimodal cues [3][4][5][6]. Work has included the use of speaker identity or persona, but this has often be leveraged as a shallow feature, or it was omitted during the ERC classification task. As a result, current models may fail to distinguish between emotion signals that are linguistically similar but different across speakers.

The goal of this paper is to investigate persona conditioning within the emotion classification task. The work considers whether explicit speaker representations can improve the recognition of emotions beyond text-only and context-only approaches. We define persona broadly to include information associated with the speaker, such as stable profile attributes, speaker-specific embeddings, and historical interaction patterns that characterize how emotion is typically expressed. The incorporation of personas into the emotion classification task can alter the performance of classification by reducing ambiguity, personalizing contextual interpretation, and enabling models to learn systematic differences in affective expression across roles. It is also important to observe that persona conditioning raises important questions regarding how persona should be represented, how it interacts with discourse context, and under what conditions it contributes meaningful gains across models.

To address these questions, this work investigates both open and closed LLM models that integrate persona signals via prompts into emotion classification and compare them against strong non-persona, or neutrally-conditioned persona, baselines. Our analysis focuses not only on overall predictive performance, but also on robustness across personas, flip rates between personas, and model confidence of accuracy. Through this investigation, we aim to clarify the role of persona in affective language understanding and to show that emotion classification can benefit from moving beyond generic contextual modeling toward more speaker-aware representations. In doing so, this work contributes to a broader view of Natural Language Process (NLP) systems as interpreters of language that is socially situated, personalized, and shaped by recurring human identities rather than by text alone.

The remainder of the paper is organized as follows: Section II provides background information relevant to the task of emotion analysis in conversations. Section III outlines the persona-conditioned emotion classification in conversation task. Section IV outlines the datasets that were used in our experiments. Section V provides information regarding the design of our experiments. Section VI provides the results from our experiments. Section VII summarizes our findings and details future directions for this work.

II. Related Work

Persona conditioning is an approach to using Large Language Models (LLMs) to model subjectivity in a wide range of tasks. Personas were introduced in [3] by implementing persona embeddings and speaker-specific conditioning to generate more

consistent, personalized neural dialogue responses. The authors constructed two persona-based models: a Speaker Model to model the respondent's personality, and a Speaker-Addressee model to parallel the respondent adaptation to a given addressee. The work demonstrated that personal characteristics could be captured through distributed representations, such as speaking style. Other work was done to demonstrate that grounding dialogue in explicit persona sentences improves consistency and engagement while also demonstrating that dialogue can be used to predict profile information [4].

As language models continued to make advancements, *prompt-based learning* strategies demonstrated promising results across a wide range of tasks in the field of NLP [7]. Instead of supervised machine learning tasks where by the goal is to predict $\mathbf{y}$ based on input $\mathbf{x}$, conditioned as $\Pr(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\Theta})$, language models were leveraged to formulate a new input $\hat{\mathbf{x}}$ from $\mathbf{x}$ to be used to obtain the target $\mathbf{y}$. Prior work has demonstrated that as language models continued to scale in size, they are capable of performing in-context learning where training examples can be provided to facilitate to performance of a task without the need for fine-tuning models [8]. This provided the ability to build architectures that are task-agnostic while also achieving competitive results. The work presented in [9] expanded on this idea by proposing a prompt-based fine-tuning method along with automatic prompt generation and better demonstration selection for strong few-shot text classification.

Other work has been done to leverage prompt-based learning-templates, verbalizers, tuning strategies, and evaluation to provide a unified vocabulary and taxonomy [7]. In [10], the authors quantified the variance persona variables explain in subjective NLP dataset labels and find that persona prompting yields modest, but significant gains, mainly when persona truly predicts disagreement patterns. The work demonstrated that the gains are realized in situations where the entropy in annotation is high with a lower standard deviation, and that persona variables explain less than 10% of the variation in the human annotations. The work also demonstrated a clear association between predictive persona variables and human labels, with a zero-shot 70B model reaching 81% of the annotation variance achieved by a linear regression model trained on ground-truth annotations.

The authors in [11] demonstrated that injecting persona descriptions into LLM prompts can produce more diverse, controllable annotations that align with the subjective differences as seen in human annotations. To introduce personality-affected emotion transition modeling for dialogue systems, one study framed response emotion selection as a personality-affected state transition in Valence-Arousal-Dominance (VAD) space where the emotion for response is obtained through the sum of preceding emotion and variation [12].

Work has also been performed to investigate the performance of closed LLM models on the ERC task. ChatGPT was evaluated on its emotional dialogue understanding and capabilities, including ERC, under zero-shot and few-shot prompting and analyzes misalignment with dataset annotation standards [13]. To evaluate open LLM models, [14] fine-tuned LLaMA-family models with instructions to improve ERC performance through two-stage learning that includes speaker characteristics and emotion recognition. Another study built instruction-tuned emotional LLMs while constructing a large Affective Analysis Instruction Dataset (AAID) and an Affective Evaluation Benchmark (AEB) covering multiple affective tasks [15]. To improve emotion classification performance, a long-context emotional intelligence benchmark spanning tasks including emotion classification was introduced while also proposing Retrieval Augmented Generation (RAG) and Collaborative Emotional Modeling (CoEM) strategies to improve performance [16].

III. Emotion Classification

Emotion classification is a subfield of NLP that involves the identification and classification of emotional content expressed in language. The problem often is formulated as a problem in which the model maps input, such as sentences, posts, or dialogue turns, to one or more emotion labels. Emotion labels may be annotated into discrete categories, such as anger, disgust, fear, joy, sadness, or surprise, or to various affective states as defined by various psychological frameworks. These emotion labels and affective states allow for granular analysis of language by providing a more detailed account for the subjective meaning behind the input.

There are many challenges that exist with the task of emotion classification. One input may convey many meanings where a deeper understanding may be required. For example, it is possible for a given input document to convey emotions by employing the use of sarcasm. Similarly, emotion classification in text must also account for emotional expression by way of emojis, code-switching, domain-specific vocabulary, and variation across cultures.

For the purpose of the work presented in this paper, we will approach the emotion classification task by considering the classification of an utterance by a speaker from a given conversation. Each utterance will be assigned a single emotion e_i from a discrete set of dataset-dependent target classes $e_i \in \{e_1, e_2, ..., e_k\}$.

IV. Datasets

Two datasets are used throughout the analyses performed in this work. The Multimodal EmotionLines Dataset (MELD) is a large-scale multimodal, multi-party emotional conversational dataset that was constructed from the TV-series *Friends*. The dataset includes both conversations and utterances with each utterance being assigned an emotion label: {surprise, anger, neutral, sadness, disgusting, joy, and fear} [17]. The interactive emotional dyadic motion capture database (IEMOCAP) dataset includes data from ten actors in dyadic sessions that included emotional scripts in hypothetical scenarios to elicit the following emotions: {excited, frustrated, neutral, sad, happy, and angry }. There were 151 recorded conversation videos where clips were spread across five sessions per actor. The frequency

of utterances for each class can be seen in Figure 1 for the MELD dataset and Figure 2 for the IEMOCAP dataset.

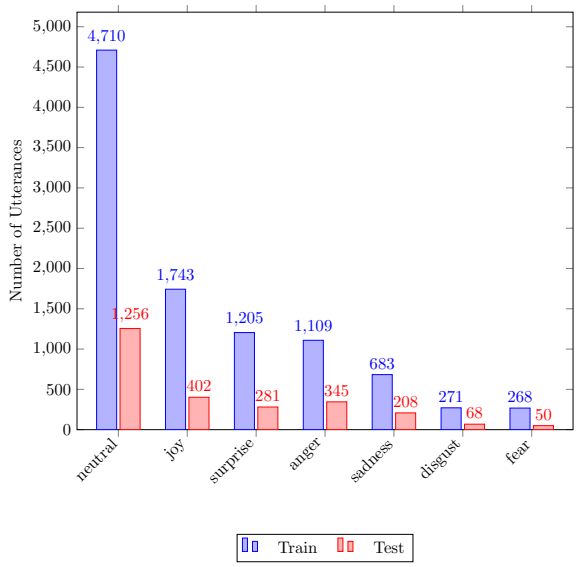


Figure 1. Emotion frequency for the labels in the MELD dataset.

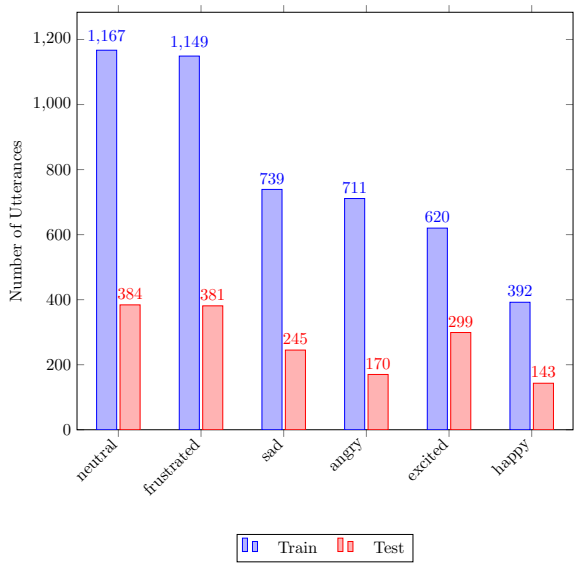


Figure 2. Emotion frequency for the labels in the IEMOCAP dataset.

V. Methodology

We evaluate the effects and impact of persona-conditioned emotion classification using LLMs through a comprehensive analysis using quantitative metrics. The goal of these experiments is to address the following Research Questions (RQs):

1) **RQ1**: Do measurable differences exist between persona-conditioned inputs in comparison to unconditioned, or neutrally-conditioned, inputs?
2) **RQ2**: How does predictive confidence compare to actual performance across models in persona-conditioned emotion classification?
3) **RQ3**: What variations can be observed across personas in the persona-conditioned emotion classification task?

Five personas were used throughout the experiments that follow. One baseline, or default, persona was used along with four personas were synthetically generated following other synthetic persona datasets to evaluate the effects of the models: *neutral*, *skeptical*, *empathic*, *social*, and *knowledgeable*. The *neutral* persona is defined as being a neutrally-conditioned, or unconditioned, persona where no additional context is provided to bias the model. The *skeptical* persona instructs the model to be skeptical and only perform a task when there exists strong emotions, while defaulting to neutral classifications otherwise. The *empathic* persona instructs the model to infer the primary emotion of the speaker from the text or context provided, even when it is subtle. The *social* persona instructs the model to assume expertise in conversational pragmatics and sarcasm while utilizing context in the conversation to detect implied emotion. The *knowledgeable* persona instructs the model to assume expertise in the given dataset topic domain while using the given context to decide the most probable emotion.

For each dataset, we generated $k = 5$ different sample sets comprised of $n = 500$ randomly selected samples each. The experimentation was conducted on four LLMs: GPT-4o, GPT-4.1, Llama 3.1 8B, and Gemma 3 12B. Llama 3.1 8B has demonstrated capability-to-efficiency trade-off with a 128K context window. Gemma 3 12B is beneficial for multimodal and multilingual capabilities with a 128K context window. GPT-4o supports text and image input with text output and a 128K context window to provide versatility in tasks. GPT-4.1 is a non-reasoning GPT model for instruction following and tool use.

To evaluate whether there exists statistical significance between personas and our baseline model, we used McNemar's test, which is a non-parametric significance test that is appropriate for paired nominal data by evaluating both systems on the same instances. Under the null hypothesis, the two models have the same misclassification, or error, rate from which the evaluation set is drawn. A statistically significant result would indicate that a model is more likely than the other to classify the same instances correctly.

Flip Rate (FR) is a measurement of the prediction instability as a result of perturbation. FR reflects the percentage of examples with label changes in a classification task due to controlled experimentation settings or levels. Given an input x_i and a perturbed version x_i', the flip rate reflects the proportion of instances when $f(x_i) \neq f(x_i')$. The label flip rate is defined for our purposes as being

$$\text{FR} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[f(x_i) \neq f(x_i')\right] \tag{1}$$

Lower flip rates indicate greater prediction consistency, whereas higher flip rates indicate less stability under perturbation and greater sensitivity to input modifications. For our purposes, higher rates would serve as an indication for stronger persona-induced shifts.

Expected Calibration Error (ECE) is a metric for the evaluation of probabilistic calibration in models where it reflects the discrepancy between a model's predicted confidence and its empirical accuracy. ECE is defined as being

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (2)$$

where B_m is the confidence of bin m, $\text{acc}(B_m)$ is the empirical accuracy of the given bin, $\text{conf}(B_m)$ is the mean confidence in that bin, and n is the total number of predictions. Lower ECE values serve as an indication that model's confidence scores are more reflective of the true probability of correctness, whereas higher values suggest a greater discrepancy or misalignment between confidence and actual predictive performance.

VI. Results

In this section, we present an analysis of the results from the experiments that were conducted.

A. Model Analysis

Our results demonstrate that persona conditioning affects both predictive performance and calibration. In Table II, we observe in GPT-4o and GPT-4.1 that the Empathic persona yields the best accuracy and macro-F1, improving the results substantially over the Neutral (baseline) persona in terms of accuracy (0.436 vs. 0.382) and macro-F1 (0.431 vs. 0.378) for GPT-4o, and in terms of accuracy (0.426 vs. 0.382) and macro-F1 (0.417 vs. 0.378) for GPT-4.1. For Gemma 3, the Knowledgeable persona achieves the best accuracy, macro-F1, and ECE scores of 0.483, 0.472, and 0.311, respectively. Llama produces the best results with the Social persona in terms of accuracy (0.515) and ECE (0.294), while the Knowledgeable persona achieves the best macro-F1 score (0.413).

As demonstrated in Table II, a persona-conditioned model generally outperformed the baseline model on the shared evaluation set for each LLM used in the experiments. McNemar's test on paired instance-level correctness found the difference to be significant as demonstrated in Table I at the significance level of $\alpha = 0.001$ using the exact p-value. This answers **RQ1** regarding the existence of measurable differences by infusing personas into the instructions for the emotion classification task using LLMs.

Model predictions are shown to be affected by persona conditioning in Figure 3. The effect is shown to differ between both models and persona pairings. The two closed models are shown to have more robust results with lower pairwise flip rates in comparison to the open source models, Gemma 3 and Llama, across both datasets. The *skeptical* persona demonstrates high prediction instability when compared to the knowledgeable and empathic personas, while moving from skeptical to neutral induces the smallest change. The effect is also observed as having a dataset-dependency given that Gemma 3 is sensitive to the IEMOCAP dataset whereas GPT-4.1 remains relatively stable in comparison. Neutral appears the most stable overall. We can also see the Empathic-Skeptical pairs often produce the largest disagreement.

B. Confidence Score Analysis

In this section, we investigate the impact of model confidence in persona-conditioned models for **RQ2** and **RQ3**. Figure 4 reflects the model's confidence with the given classification of the conversation utterance. With Gemma 3, we observe in the MELD dataset that the InterQuartile Range (IQR) was low for both the empathic and neutral personas of $IQR = 0$ and $IQR = 0.10$, $\bar{x} = 0.802$ and $0.\bar{7}89$, and $m = 0.800$ and $m = 0.800$ for correct classification, respectively. The confidence scores had the greatest range for the social persona in both correct and incorrect classifications where $IQR = 0.186$, $\bar{x} = 0.789$, $m = 0.800$ for correct classifications and $IQR = 0.200$, and $0.\bar{7}85$, and $m = 0.800$ for incorrect classifications. We observe a consistent range for the empathic, neutral, and skeptical personas, while the social persona has the greatest range and the knowledgeable persona has the smallest range. For the IEMOCAP dataset, we observed consistent outputs for Q1 and the median with the correct classification, whereas there were consistent outputs with the median and Q3 values with an incorrect classification. The social persona had had the largest range with more consistent and lower confidence scores with incorrect classifications.

Llama produced more consistent confidence scores with the IEMOCAP dataset with values consistently around the median score $m = 0.800$ and $IQR = 0$ for both correct and incorrect classifications with a mean score of $\bar{x} = 0.807$ and standard error of $SE = 0.004$. In the MELD dataset, the confidence scores for incorrect classifications were consistently about the median $m = 0.800$ for all personas with knowledgeable and neutral personas having a median score of $m = 0.900$ for correct classifications. The confidence scores for correct classifications had an increased IQR score when compared to the incorrect classifications where the incorrect classifications had an $IQR = 0$ for all personas. For correct classifications, we observed $IQR = 0$ for the skeptical persona, $IQR = 0.100$ for the empathic and social personas, and $IQR = 0.200$ for the knowledgeable and neutral personas, which also had an increase in median value compared to other classes.

The ranges and IQR metrics for confidence scores within the GPT models demonstrated greater consistency across both correct and incorrect classifications for the datasets. For the GPT-4o model, using the persona-by-correctness means, the mean confidence score was $\bar{x} = 0.739$ with $SE = 0.010$ in the IEMOCAP dataset. For correct classifications, the empathic, neutral, skeptical, and social personas each had $IQR = 0.100$; the empathic persona had a median of $m = 0.700$ and mean of $\bar{x} = 0.751$, the neutral persona had a median of $m = 0.700$ and mean of $\bar{x} = 0.761$, the skeptical persona had a median of $m = 0.800$ and mean of $\bar{x} = 0.778$, and the social persona had a median of $m = 0.800$ and mean of $\bar{x} = 0.771$. The knowledgeable persona had the largest spread for the correct classifications, with $IQR = 0.150$, a median of $m = 0.800$, and mean of $\bar{x} = 0.775$. For incorrect classifications, the empathic, knowledgeable, skeptical, and social personas each had $IQR = 0.100$ with the correct classification of emotions,

TABLE I. EXACT POOLED MCNEMAR p-VALUES COMPARING EACH PERSONA-CONDITIONED MODEL AGAINST THE NEUTRAL BASELINE. * INDICATES SIGNIFICANCE AT THE LEVEL OF $\alpha = 0.001$.

	IEMOCAP				MELD			
Persona	**GPT-4o**	**GPT-4.1**	**Llama**	**Gemma3**	**GPT-4o**	**GPT-4.1**	**Llama**	**Gemma3**
Empathic	3.96e-12*	4.73e-11*	0.195	0.00202*	8.15e-05*	3.71e-06*	1.81e-16*	3.17e-41*
Knowledgeable	3.83e-10*	4.03e-06*	1.41e-06*	1.38e-05*	1.73e-05*	0.00181*	4.78e-10*	1.58e-25*
Skeptical	4.27e-25*	1.73e-15*	0.283	6.21e-13*	2.99e-04*	0.163	2.69e-13*	0.873
Social	8.75e-05*	0.0535	8.01e-10*	0.349	1.87e-17*	1.57e-13*	1.10e-06*	1.63e-06*

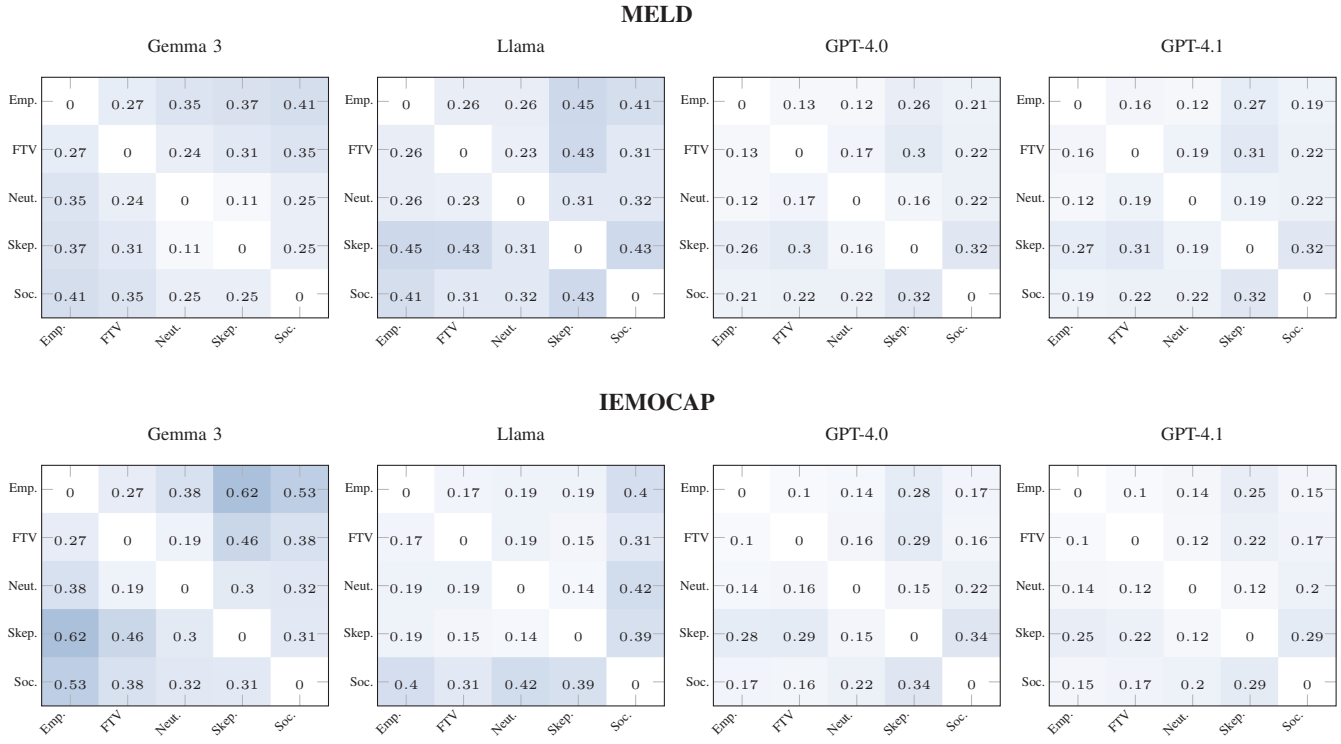


Figure 3. Pairwise persona flip-rate heatmaps by models.

while the neutral persona had the smallest spread with $IQR = 0$. For the MELD dataset, the mean confidence score was $\bar{x} = 0.758$ with $SE = 0.007$ in the dataset. The empathic, knowledgeable, neutral, and social personas each had $IQR = 0.150$; the empathic persona had a median of $m = 0.700$ and mean of $\bar{x} = 0.754$, the knowledgeable persona had a median of $m = 0.800$ and mean of $\bar{x} = 0.771$, the neutral persona had a median of $m = 0.700$ and mean of $\bar{x} = 0.756$, and the social persona had a median of $m = 0.800$ and mean of $\bar{x} = 0.768$. The skeptical persona had the largest spread for the correct classifications, with $IQR = 0.200$, a median of $m = 0.800$, and mean of $\bar{x} = 0.798$. For incorrect classifications, the knowledgeable and social personas each had $IQR = 0.150$, while the empathic, neutral, and skeptical personas each had $IQR = 0.100$.

For the GPT-4.1 model, the model produced a mean confidence score across classes and correctness of $\bar{x} = 0.802$ with $SE = 0.008$ in the IEMOCAP dataset. For correct classifications, $IQR = 0.150$ for the empathic persona with $m = 0.800$ and $\bar{x} = 0.812$, $IQR = 0.100$ for the knowledgeable and neutral personas with the knowledgeable persona having a median of $m = 0.85$ and mean $\bar{x} = 0.832$ while the neutral persona had a median of $m = 0.85$ and mean of $\bar{x} = 0.830$. The skeptical persona had an IQR metric of $IQR = 0.250$, which is the largest of the personas, with a median of $m = 0.85$ and mean of $\bar{x} = 0.833$. The social persona had an IQR score of $IQR = 0.150$, a median of $m = 0.800$ and mean of $\bar{x} = 0.807$. For the incorrect classes, the IQR scores for each persona were either the same (skeptical and social) or within a ± 0.050 difference (empathic, knowledgeable, and neutral). For the MELD dataset, the mean confidence score was $\bar{x} = 0.790$ with $SE = 0.012$ in the dataset. The empathic persona had $IQR = 0.200$ with a median of $m = 0.850$ and mean of $\bar{x} = 0.823$ for correct classifications. The knowledgeable persona had the smallest spread, with $IQR = 0.050$, a median of $m = 0.850$, and mean of $\bar{x} = 0.828$.

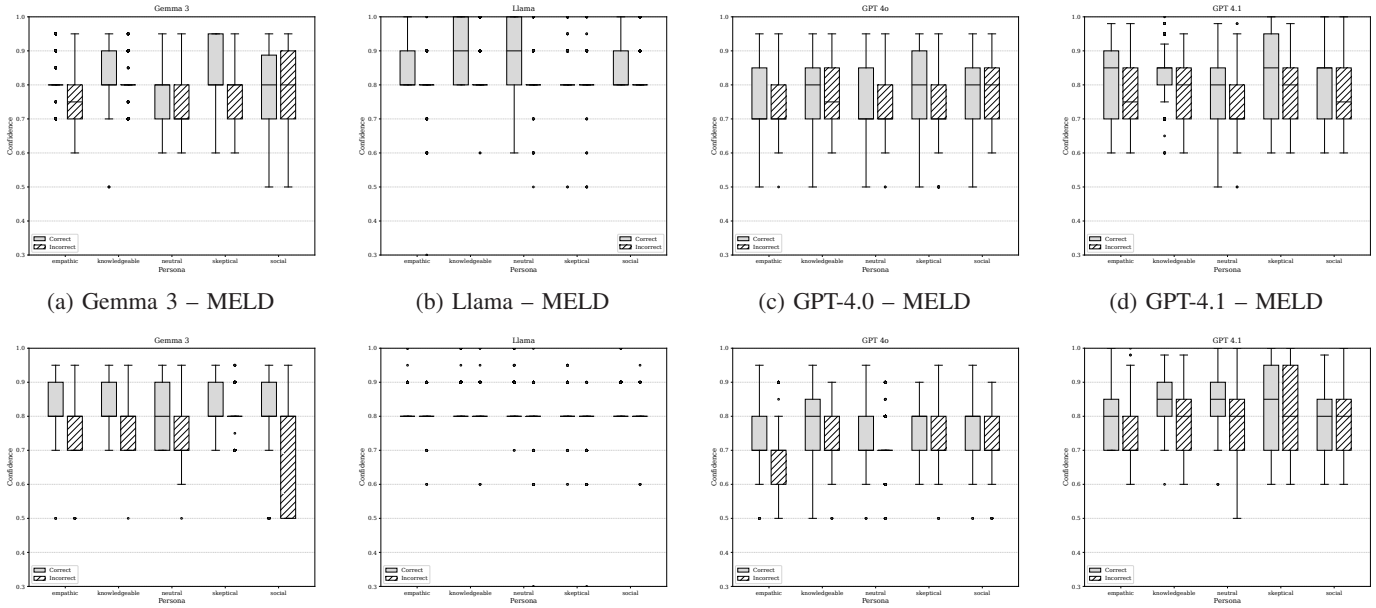
TABLE II. MEAN CLASSIFICATION ACCURACY, MACRO-F1, CONFIDENCE, AND EXPECTED CALIBRATION ERROR (ECE) ACROSS PERSONAS. † INDICATES THE BASELINE PERSONA.

Dataset	Model	Persona	Accuracy ↑	Macro-F1 ↑	Confidence ↑	ECE ↓
IEMOCAP	GPT-4o	Empathic	**0.4360 ± 0.0016**	**0.4306 ± 0.0046**	0.7166 ± 0.0037	**0.2806 ± 0.0041**
		Knowledgeable	0.4327 ± 0.0066	0.4273 ± 0.0026	0.7433 ± 0.0031	0.3106 ± 0.0107
		Neutral†	0.3820 ± 0.0130	0.3781 ± 0.0151	0.7207 ± 0.0009	0.3387 ± 0.0165
		Skeptical	0.2993 ± 0.0109	0.2861 ± 0.0206	**0.7463 ± 0.0037**	0.4469 ± 0.0097
		Social	0.4213 ± 0.0084	0.3993 ± 0.0021	0.7415 ± 0.0026	0.3201 ± 0.0125
		Overall	**0.3943 ± 0.0512**	**0.3843 ± 0.0527**	**0.7337 ± 0.0014**	**0.3394 ± 0.0100**
	GPT-4.1	Empathic	**0.4260 ± 0.0102**	**0.4174 ± 0.0057**	0.7853 ± 0.0010	**0.3593 ± 0.0132**
		Knowledgeable	0.4087 ± 0.0137	0.4069 ± 0.0115	0.8017 ± 0.0026	0.3984 ± 0.0201
		Neutral†	0.3773 ± 0.0115	0.3781 ± 0.0148	0.7993 ± 0.0049	0.4233 ± 0.0101
		Skeptical	0.3220 ± 0.0075	0.3173 ± 0.0094	**0.8183 ± 0.0065**	0.4963 ± 0.0080
		Social	0.3960 ± 0.0142	0.3751 ± 0.0143	0.7820 ± 0.0039	0.3900 ± 0.0131
		Overall	**0.3860 ± 0.0357**	**0.3790 ± 0.0348**	**0.7973 ± 0.0022**	**0.4135 ± 0.0113**
	Llama	Empathic	0.4187 ± 0.0213	0.3869 ± 0.0152	0.8078 ± 0.0041	0.3891 ± 0.0266
		Knowledgeable	0.4753 ± 0.0255	**0.4133 ± 0.0195**	0.8133 ± 0.0030	0.3379 ± 0.0304
		Neutral†	0.4307 ± 0.0165	0.4129 ± 0.0172	0.8074 ± 0.0024	0.3767 ± 0.0185
		Skeptical	0.4393 ± 0.0217	0.4039 ± 0.0250	0.7940 ± 0.0010	0.3550 ± 0.0260
		Social	**0.5147 ± 0.0146**	0.3946 ± 0.0102	**0.8088 ± 0.0009**	**0.2942 ± 0.0188**
		Overall	**0.4557 ± 0.0350**	**0.4023 ± 0.0103**	**0.8063 ± 0.0021**	**0.3505 ± 0.0219**
	Gemma 3	Empathic	0.4100 ± 0.0255	0.4055 ± 0.0250	0.7892 ± 0.0047	0.3792 ± 0.0272
		Knowledgeable	**0.4833 ± 0.0157**	**0.4719 ± 0.0101**	0.7938 ± 0.0050	**0.3105 ± 0.0143**
		Neutral†	0.4453 ± 0.0077	0.4432 ± 0.0035	0.7715 ± 0.0041	0.3262 ± 0.0081
		Skeptical	0.3680 ± 0.0071	0.3647 ± 0.0094	**0.8100 ± 0.0048**	0.4420 ± 0.0131
		Social	0.4333 ± 0.0159	0.3908 ± 0.0118	0.7637 ± 0.0057	0.3304 ± 0.0149
		Overall	**0.4280 ± 0.0382**	**0.4152 ± 0.0380**	**0.7856 ± 0.0043**	**0.3576 ± 0.0093**
MELD	GPT-4o	Empathic	0.6240 ± 0.0240	0.5075 ± 0.0264	0.7488 ± 0.0444	0.1376 ± 0.0600
		Knowledgeable	0.6160 ± 0.0166	0.5073 ± 0.0194	0.7675 ± 0.0386	0.1571 ± 0.0533
		Neutral†	**0.6500 ± 0.0213**	**0.5108 ± 0.0259**	0.7470 ± 0.0462	**0.1038 ± 0.0580**
		Skeptical	0.6225 ± 0.0216	0.4373 ± 0.0258	**0.7705 ± 0.0250**	0.1564 ± 0.0398
		Social	0.5765 ± 0.0222	0.4806 ± 0.0186	0.7702 ± 0.0368	0.2104 ± 0.0489
		Overall	**0.6178 ± 0.0237**	**0.4887 ± 0.0279**	**0.7608 ± 0.0382**	**0.1446 ± 0.0553**
	GPT-4.1	Empathic	0.6070 ± 0.0215	0.5049 ± 0.0289	0.8012 ± 0.0165	0.1963 ± 0.0330
		Knowledgeable	0.6115 ± 0.0140	**0.5237 ± 0.0257**	0.8074 ± 0.0163	0.2018 ± 0.0116
		Neutral†	0.6385 ± 0.0225	0.5150 ± 0.0243	0.7845 ± 0.0107	**0.1490 ± 0.0316**
		Skeptical	**0.6500 ± 0.0273**	0.5133 ± 0.0296	**0.8290 ± 0.0053**	0.1859 ± 0.0271
		Social	0.5725 ± 0.0170	0.4874 ± 0.0201	0.7623 ± 0.0309	0.2371 ± 0.0123
		Overall	**0.6159 ± 0.0270**	**0.5089 ± 0.0123**	**0.7969 ± 0.0149**	**0.1927 ± 0.0222**
	Llama	Empathic	0.4307 ± 0.0098	0.4002 ± 0.0181	0.8128 ± 0.0036	0.3848 ± 0.0089
		Knowledgeable	0.4553 ± 0.0096	0.4171 ± 0.0028	**0.8521 ± 0.0044**	0.4021 ± 0.0107
		Neutral†	0.5167 ± 0.0172	0.4213 ± 0.0084	0.8459 ± 0.0033	0.3439 ± 0.0218
		Skeptical	**0.6053 ± 0.0282**	**0.4327 ± 0.0178**	0.8149 ± 0.0044	**0.2163 ± 0.0351**
		Social	0.4613 ± 0.0066	0.3822 ± 0.0038	0.8299 ± 0.0025	0.3752 ± 0.0071
		Overall	**0.4939 ± 0.0624**	**0.4107 ± 0.0177**	**0.8311 ± 0.0032**	**0.3444 ± 0.0167**
	Gemma 3	Empathic	0.4690 ± 0.0052	0.4205 ± 0.0174	0.7771 ± 0.0010	0.3081 ± 0.0067
		Knowledgeable	0.5220 ± 0.0081	0.4537 ± 0.0129	0.8151 ± 0.0015	0.2946 ± 0.0113
		Neutral†	0.6170 ± 0.0184	0.4770 ± 0.0240	0.7698 ± 0.0021	**0.1528 ± 0.0214**
		Skeptical	**0.6185 ± 0.0221**	**0.4790 ± 0.0199**	**0.8245 ± 0.0028**	0.2060 ± 0.0236
		Social	0.5710 ± 0.0267	0.4468 ± 0.0160	0.7873 ± 0.0034	0.2173 ± 0.0332
		Overall	**0.5595 ± 0.0575**	**0.4554 ± 0.0215**	**0.7948 ± 0.0015**	**0.2356 ± 0.0177**

The neutral and social personas each had $IQR = 0.150$; the neutral persona had a median of $m = 0.800$ and mean of $\bar{x} = 0.807$, while the social persona had a median of $m = 0.850$ and mean of $\bar{x} = 0.783$. The skeptical persona had the largest spread for the correct classifications, with $IQR = 0.250$, a median of $m = 0.850$, and mean of $\bar{x} = 0.853$. For incorrect classifications, the empathic, knowledgeable, skeptical, and social personas each had $IQR = 0.150$, while the neutral persona had the smallest spread with $IQR = 0.100$.

VII. CONCLUSION AND FUTURE WORK

The work presented in this paper demonstrates that persona conditioning of models is a significant control variable for emotion classification as opposed to a superficial variation of prompts. While the results produced statistically significant results to demonstrate that persona-conditioning can bias the model that alters its performance with the emotion classification task, it is important to observe that there does not exist uniformity across settings as the best persona depended on both



(a) Gemma 3 – MELD (b) Llama – MELD (c) GPT-4.0 – MELD (d) GPT-4.1 – MELD

(e) Gemma 3 – IEMOCAP (f) Llama – IEMOCAP (g) GPT-4.0 – IEMOCAP (h) GPT-4.1 – IEMOCAP

Figure 4. Boxplots for each model on MELD (top) and IEMOCAP (bottom).

the model family and dataset. The results also indicate that some personas can increase confidence without improving correctness, and sometimes this can lead to severe miscalibration. Our results indicate that persona conditioning and selection should be utilized as a tunable modeling choice with evaluation including metrics to evaluate calibration in addition to accuracy and macro-F1.

Future work in this space could investigate the biases that associated with personas and how they shift the classification task as a result. The task could also be applied to natural dialogues as opposed to scripted language from TV sources to better understand the impact of personas on real data. In addition, given that persona declarations in prompts often contain emotion words or context that leak signal, these factors which contribute to emotion classification task should be considered to understand how the contribute to certain emotions, such as through the deployment of hold-out paraphrases, shuffling personas, etc. Other directions may consider speaker-receptor relations, cultural specifics, or environmental sets.

REFERENCES

[1] Y. Liu, J. Zhao, J. Hu, R. Li, and Q. Jin, "DialogueEIN: Emotion interaction network for dialogue affective analysis", in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari et al., Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 684–693.

[2] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, *Emotion recognition in conversation: Research challenges, datasets, and recent advances*, 2019. arXiv: 1905.02947 [cs.CL].

[3] J. Li et al., "A persona-based neural conversation model", in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds., Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 994–1003. DOI: 10.18653/v1/P16-1094.

[4] S. Zhang et al., "Personalizing dialogue agents: I have a dog, do you have pets too?", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213. DOI: 10.18653/v1/P18-1205.

[5] A. Mackey, S. Gauch, and I. Cuevas, "Prompt distillation for emotion analysis", in *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR*, INSTICC, SciTePress, 2024, pp. 328–334, ISBN: 978-989-758-716-0. DOI: 10.5220/0012951200003838.

[6] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, *Emotional chatting machine: Emotional conversation generation with internal and external memory*, 2018. arXiv: 1704.01074 [cs.CL].

[7] P. Liu et al., "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing", *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023, ISSN: 0360-0300. DOI: 10.1145/3560815.

[8] T. Brown et al., "Language models are few-shot learners", in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[9] T. Gao, A. Fisch, and D. Chen, *Making pre-trained language models better few-shot learners*, 2021. arXiv: 2012.15723 [cs.CL].

[10] T. Hu and N. Collier, "Quantifying the persona effect in LLM simulations", in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 289–10 307. DOI: 10.18653/v1/2024.acl-long.554.

[11] L. Fröhling, G. Demartini, and D. Assenmacher, *Personas with attitudes: Controlling llms for diverse data annotation*, 2024. arXiv: 2410.11745 [cs.CL].

[12] Z. Wen, J. Cao, R. Yang, S. Liu, and J. Shen, "Automatically select emotion for response via personality-affected emotion transition", in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 5010–5020. DOI: 10.18653/v1/2021.findings-acl.444.

[13] W. Zhao et al., *Is chatgpt equipped with emotional dialogue capabilities?*, 2023. arXiv: 2304.09582 [cs.CL].

[14] Y. Fu et al., *Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics*, 2025. arXiv: 2403.07260 [cs.CL].

[15] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis", in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24, ACM, Aug. 2024, pp. 5487–5496. DOI: 10.1145/3637528.3671552.

[16] W. Liu et al., *Longemotion: Measuring emotional intelligence of large language models in long-context interaction*, 2025. arXiv: 2509.07403 [cs.CL].

[17] S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. DOI: 10.18653/v1/P19-1050.