Co-EmP: Context-aware Emotion and Personality Alignment

Yuyang Lin
BASIS International School Hangzhou
Hangzhou, Zhejiang, China
yuyang.lin14883-bihz@basischina.com

1. Introduction

Large Language Models (LLMs) have become one of the most heated research fields in the AI industry, given both the laboratory breakthroughs and commercial successes. Among them, the development in maintaining a longer context window and increasing the reasoning capabilities remains as the most heated research topics, as State-of-the-Art (SOTA) LLMs branch into a wide variety of different industries. Such demand for LLMs to tackle growingly sophisticated tasks further increases interest in improving model context and reasoning.

Of course, the fundamental idea behind the mechanism of LLMs is that natural languages, such as English and Chinese, should encode human intelligence. By improving model context and reasoning capabilities, researchers essentially emulates the corresponding memory and logic core of human intelligence. However, human intelligence resembles cores beyond memory and logic. Among them, emotion also stands as a critical part embedded in natural language. People relied on emotion, too, for communicating and expressing information.

The ability to detect, recognize, and emulate human emotion by an LLM is another subsection of the larger subject of Affective Computing (AC). In the time of rapid LLM development, the general application future of AC has shifted more toward empathetic-AI in an interactive real-time setting for diverse settings. Combining with the belief of edge computing, a locally run small LLMs with specific emotion dynamic and personality support could potentially become a potent future development direction, applicable to industries such as video game, education, finance, and mental comfort.

SOTA models as of 2024, which are mainly commercial LLMs such as GPT and Claude, have been proven to be capable of emulating diverse human emotions, given specific prompting. [7] Meanwhile, small LLMs with parameter sizes smaller than 10 billion showed poor results compared to traditional Natural Language Processing models such as RoBERTa on standard emotion benchmark. 1 While no model has specialized its training with emotion-specific text corpora, targeting the emulation of realistic emotion dynamic during conversations, these low-parameter LLMs fail to capture realistic emotional expression through naive next-token-prediction. As such, additional encoder heads are required to process natural languages into emotion, study a complex emotional state embedding, and provide direct and strong emotion instruction to these small LLMs.

In this project, we tried to investigate the feasibility of constructing such external emotion head that models emotion dynamic with aligned context and personality content. The resulting head is then going be applied for a chatbot-styled AI generator, specifically a Qwen3.5-4B model. The goal is to achieve realistic emotion shifts within conversation, simulating real-world conversations.

To recognize the emotion contained within conversation turns, we first trained a deBERTa V3 model for predicting a three-way Valence Arousal Dominance (VAD) metric. To further capture the evolution of emotional states as a dynamic process from turn to turn as a context, we trained an additional Gated Recurrent Unit, with emotional state as the recurrently updated hidden state and personality and raw text as per-turn input. Finally, we prompt engineered a few-shot example prompt, explaining VAD metric and a targeting VAD to generate in accordance.

Existing studies either focus on processing utterance into emotion representations such as VAD metric or the emotion dynamic of speakers in a conversation. Rarely are the two topics considered together. Existing studies such as EmoBank [3] and ANEW [19] contain high quality human-rated Valence Arousal Dominance metrics, but the formality is either in sentences or words, respectively. On the other hand, datasets like DailyDialog [11], EmpatheticDialogues [16], and RECCON [15] contain full conversation-styled data but labeled different using natural language labels, such as ”happy” and ”neutral”. As a result, we used DeepSeek V4 Flash model and created an AI-synthetic dataset EmoDynamic with full emotional conversation utterances for two speakers and their corresponding VAD values for each utterance.

As a creative project, I further created a web page imple-

Table 1. Performance comparison on the DailyDialog emotion classification test set. Macro-F1 is emphasized due to class imbalance. Best values are shown in bold.

Model	Accuracy	Macro-F1	Weighted-F1
Qwen3 (zero/few-shot)	0.5243	0.2966	0.5560
RoBERTa (Epoch 3)	**0.7548**	0.4432	**0.7462**
RoBERTa (Epoch 6)	0.7452	**0.5097**	0.7406

mentation of the above workflow - a generative LLM with emotion head supported - using Gradio. The final product is delivered in a chatbot format, with four modes for selection: the baseline LLM chat, emotion dynamic supported LLM chat, Arena comparison, Two-AI Speaker Simulation. We used Qwen3.5-4B as the small LLM, run fully on local device using a Mobile Nvidia RTX 5080.

In summary, our contributions are mainly in threefold:

- Trained a deBERTa V3 model and a GRU model for analyzing emotion in VAD metric and for updating emotional state dynamically in conversation turns, respectively. Integrated together for a complete emotion and personality aligned chatbot model.
- Generated an EmoDynamic dataset that gives full conversation utterance between two speakers, each labeled using VAD metric.
- Created a locally ran chatbot interface with emotion dynamic supported to the LLM for emulating realistic emotion shifts in human conversations.

You can find source codes and final product details in Co-EmP GitHub Repo

2. Related Works

While many subtopics within AC provide crucial guidance for the project, virtually no existing studies have committed to complete a full workflow from detecting to recognizing, determining, and finally emulating emotion. Thus, we reviewed a diverse range of different sources and essays.

2.1. Valence Arousal Dominance Model

Introduced by Albert Mehrabian and James A. Rusell in 1974, the Valence Arousal Dominance Model [13] serves as a quantifying metric to describe any given emotion in a three-dimensional vector space. Each dimension has a range from 0 to 1, combined into a cube-like space, holding all possible emotion labels. Valence measures the polarity of the emotion, from very negative to very positive; Arousal measures the intensity of the emotion, from calm to excited; Dominance measures whether the speaker believes that they are in control of the scenario.

As an example, excitement is a classic vector with high valence, high arousal, and high dominance, while anger is a classic vector with low valence, high arousal, and high dominance.

2.2. Personality Traits

The development of a widely used personality model that quantifies personality traits into limited dimensions took longer than emotion models. Starting as early as the 1930s, researchers have been suggesting factors that generalize different personality significance. [8] Proceeding into the late 20th century, researchers eventually consolidated themselves on a fixed Five Factor Model [5], which evolved into the Big Five Personality Trait we see today. The five factors each in the range from 0 to 1 are Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism (OCEAN), respectively.

Openness to Experience evaluates one's acceptance level to new ideas and abstract terms, with a high score indicating a more adventurous personality; Conscientiousness describes the level self-discipline that a person has toward their goal; Extraversion measures how out-going and socially active a person is; Agreeableness shows the level of empathy towards others; Neuroticism indicates how sensitive a person is to negative emotions or pressure.

2.3. LLM in Affective Computing

Multiple literature reviews and evaluation-focused essays have reported the performance of foundational models - large Artificial Intelligence base model suitable for numerous downstream tasks - on traditional NLP tasks and classic AC problems such as toxicity detection, sarcasm detection, and so on. [1, 2]

These essays have consistently reported that foundational models like GPT 4 have achieved comparable results in tasks that involve explicit signals inside text corpora; however, for any tasks that involve tackling implicit signals hidden beneath textual evidence, the performance dropped significantly and is consistently lower than traditional models such as transformer-based RoBERTa and well optimized RNN.

2.4. Emotional Generation in LLM

While the generation of emotional utterances in a dialogue-based format is never new, existing research often focuses on three general approaches to achieve such effect.

[18, 21] have focused on prompt engineering. Methods such as few-shot prompting, rule-baesd prompting, and force Chain-of-Thought have been rigorously applied in order to emulate human thought process and perception during realistic conversation.

[12, 17] have focused on emotion detection and generation on a multi-model basis, which on itself is heavily based on visual cues and audio effects rather than simply through the understanding of natural language corpora. As such, research focus on visual transformer models are completely

different from what we are aiming for in our project goal.

[9] tried an agent-based approach which involves multiple roles within the same system, together resembling a simulation of realistic human behavior.

Of course, there are still publications such as [20] and [10] that provide structural inspiration for the project, suggesting an instructor-generator styled architecture, where an encoder provides the targeting emotion for the generator to produce a response upon.

2.5. Datasets

Existing Dialogue-based datasets are generally designed for emotion detection and causal extraction purposes. As such, datasets generally use categorical labels with natural English words as emotion labels. DailyDialog [11] has a six-way categorical label including neutral, anger, disgust, fear, happiness, sadness, and surprise for each utterance in each dialogue. Meanwhile, EmpatheticDialogues [16] expands on the list and uses around 32 emotions as a summary label for each dialogue. Moreover, datasets such as RECCON [15] build their own datasets with extra causal labeling on top of the original ones such as DailyDialog and IEMOCAP [4].

On the other hand, many datasets are focusing on multimodality direction of emotion detection and generation. Popular examples of such include MELD [14] which is largely a dataset based on the television series Friends and IEMOCAP [4] mentioned earlier.

VAD-based datasets, in comparison, uses continuous data due to the nature of VAD metric, yet they lack dialogue-basaed inputs. Datasets like EmoBank [3] includes mainly VAD rated sentences or phrases, while ANEW [19] are mainly VAD rated words.

3. EmoDynamic

Since this project focuses on modeling VAD values as emotions throughout a conversation to emulate a complete emotion dynamic, virtually no existing datasets perfectly match with our need for both turn-by-turn conversation format and the corresponding VAD labels. In light of this, we need to create a custom dataset that enables such training. However, because of the time and financial constraints, we are not able to construct a human-written and human-rated dataset, especially in large quantity. Instead, we used the newly published DeepSeek V4 Flash model [6] to generate the full EmoDynamic dataset with both a wide variety of emotion-rich conversations and their corresponding VAD annotations per utterance.

To ensure accountability and quality of this AI-synthetic dataset. The entire construction is split into several steps. To start, around 1000 instances of scenario seed are generated. A scenario seed includes an event title, a brief description for the scenario context to setup the conversation background, and two speaker personas each containing the role, traits, personal background, big five personality, and a set of initial VAD values. These seeds are generated in a batch of ten per API call. The model is specifically prompted to guarantee 2 low-stake everyday situations, 2 high-stake personal situations, 2 asymmetrical relevance scenarios, 2 subtle cases with little emotional shifts, 2 positive repairing scenes, 2 negative deteriorating scenes. Parts of the requirements are not mutually exclusive, so the model is given the choice to assemble the combinations. On top of this, a specific ban list is also included to prevent some commonly generated scenarios by DeepSeek, in order to ensure emotional diversity. After 1000 seeds are produced, we randomly select around 20 cases and check for validity and consistency in the personality trait and their assigned big five values and initial VAD values.

Afterward, we further run an augmentation script that calls the model to generate another 3 sets of persona pairs given the exact same scenario. This is to add more variety accounting for personality differences, so that the dataset explicitly examines the differences between different personality and how each would react differently than others, meanwhile adding variants in how emotion would shift during the conversation.

Finally, we then call on the model to generate the actual conversation given the scenario seeds listed. Each API call runs only one scenario seed to ensure no context pollution. We run with three parallel threads calling. Another 20 examples are sampled after the dataset is created. The quality of the dataset matches roughly with that of DailyDialog. While there are some emotional shifts in which I particularly find strange, the same also applies when reviewing DailyDialog, so the discrepancy between the generated label and the expectation could also be attributed to personal perspectives and attitude towards these specific scenarios. An example of one full case is provided. 2

All of the data including the original scenario seeds are stored in jsonl format. Thanks to the cheap API calls, the total cost for constructing this dataset is around 50 Chinese Yuan.

While using a dataset purely generated by commercial LLMs could be both risky quality-wise and ethical-wise, such design actually partly resembles the belief of distilled learning, where a student model is trained to emulate the behavior of a teacher model given less training resources and lower parameter count. In such perspective, subsequent training could also be seen as learning the generations made by DeepSeek. In light of the urgent need for fitting dataset, such conduct is somewhat justified.

4. Methodology

To achieve the final pipeline at a low-resource condition using small LLMs, we separated it into a three-stage work.

Table 2. Example dialogue instance with persona traits and turn-level VAD annotations in EmoDynamic

Scenario	**Reunion of Childhood Friends**: Two childhood friends reunite after ten years and attempt to reconnect while reflecting on personal changes.		
Persona A	**Role**: childhood_friend_1 **Traits**: Brooding, inquisitive, slightly cynical **Background**: Freelance photographer; extensive travel; struggles with commitment **Big Five**: O=0.7, C=0.4, E=0.6, A=0.3, N=0.5 **Initial VAD**: V=0.1, A=0.4, D=-0.2		
Persona B	**Role**: childhood_friend_2 **Traits**: Warm, orderly, slightly guarded **Background**: Married with two children; accountant; values stability **Big Five**: O=0.4, C=0.9, E=0.3, A=0.8, N=0.3 **Initial VAD**: V=0.5, A=0.3, D=0.7		
Dialogue	**S**	**Utterance**	**VAD (V,A,D)**
	A	Ten years... you look exactly the same, except maybe a little more settled...	(0.10, 0.40, -0.20)
	B	It's good to see you too, Sam. I do have a minivan. And a mortgage...	(0.50, 0.30, 0.70)
	A	Dustier. And emptier, in a way...	(-0.15, 0.50, -0.35)
	B	Weight can be grounding, Sam. But it's not perfect...	(0.45, 0.25, 0.65)
	A	Mostly I've just gotten tired of the same lonely sunsets...	(-0.35, 0.70, -0.55)
	B	It's not perfect, but it's mine. I missed you, you know.	(0.60, 0.35, 0.75)

The first stage consists of training a deBERTa V3 to regress valence, arousal, and dominance values given an utterance. This is the initial step that finds the emotion embedded inside raw text. The second stage consists of training a Gated Recurrent Unit (GRU) that models emotion dynamics by passing in an evolution of emotion states per conversation turn. The model then provides the VAD prediction for the next conversation turn as the target VAD instruction, given the emotion context throughout. The final stage is a prompt engineering process that fine-tunes a few-shot example prompt concatenated with general generation guidelines.

4.1. VAD regressor

We combine four datasets as the integrated dataset for training the VAD regressor - DailyDialog, EmpatheticDialogues, EmoBank, and EmoDynamic. We split the dialogues in the first two datasets into utterance-level data input. The corresponding emotion labels in natural language is mapped into VAD values through a hard coded transformation. 4 5] Similarly, we also processed EmoDynamics into utterance-level inputs and their VAD labels. In turn, the integrated custom dataset has utterance-level inputs with continuous VAD values as labels, so that the model learns a transformation from raw text to emotion annotations without context.

We then train a microsoft deBERTa V3 base model fetched from Huggingface. We specifically choose deBERTa V3 for its capabilities for a wide variety of downstream tasks and its base understanding of natural language by default.

$$v_A(t) = f_{VAD}(x_A(t)) \tag{1}$$

where $x_A(t)$ is the raw text of the speaker at turn t; f_{VAD} is the VAD regressor model; $v_A(t)$ is the VAD metric values for the raw text in turn t.

Due to the discrepancies in the original formality of the datasets, we specifically assign different weights to data extracted from different datasets. Since EmoBank is the only human-labeled VAD dataset, it is considered the gold standard for a VAD regressor, so it is assigned with a flat weight of 1.0. The synthetic data gets a weight of 0.4, since it is weakly inspected by random selection. The other two categorical datasets, finally, are assigned with a weight of 0.15. This is mainly due to two reasons. The datasets undergo a mapping process that hard coded the transformation, which is not a well-articulated guideline but rather a rough estimation of the overall emotion during the talk. Meanwhile, the datasets themselves contain ambiguous labels such as neutral, which potentially have the risk of hampering model learning a specific VAD prediction.

For training, we run 4 epochs of the entire integrated custom dataset with a learning rate of $1e{-}5$. The result is shown in the following table 3. EmoDynamic is not included for test run. The VAD regressor achieves a low mean absolute error for all datasets. However, from the category-level pearson coefficient, the model learns the easier valence and achieves around 0.7 strong correlation between the predicted and real gold label in EmoBank while forming rel-

atively weaker correlation in arousal and dominance parts. Compared across datasets, we see that the model actually learns the arousal and dominance pattern better in DailyDialog and EmpatheticDialogues' data. This could mainly be attributed to two possible reasons. The training dataset is not uniformly sampled, which means that the two categorical datasets have much higher chance of getting sampled due to their large size. Meanwhile, there may be some discrepancies in which human rates VAD and how we mapped English words into VAD.

4.2. GRU encoder

In realistic settings, the emotion dynamic is typically attributed to three factors: emotion, contextual buildup, and the personality. Emotion is already covered by the VAD regressor. The contextual buildup represents how the emotion would recurrently evolve from turn to turn. Personality traits on the other hand mandates how the speaker would evolve specifically from one turn to another, as people with different personality tend react differently to the same thing.

In light of the recurrent update of emotional states, we trained a GRU model. GRU model has its unique characteristics in that it updates a hidden state h_t for every time step t. This perfectly fits with our expectation in which an emotion dynamic could be modeled. In addition, we put in a series of VAD traits and personality vector concatenated as the input per turn. Specifically, we used the VAD value of the speaker of interest in the previous turn, the VAD value of the counterpart speaker in the current turn, the difference between the VAD counterpart and self, and finally the big five personality vector for the speaker of interest. Given that each emotion label has 3 dimensions to account for VAD, and each personality vector had 5 dimensions to account for OCEAN, the input for every GRU inference has a dimension of 14. The resulting updated emotional state vector is then passed into a multiplayer perceptron (MLP) to regress the predicted emotion label for the current turn.

Detailed construction is given below:

$$z_A(t) = GRU(x(t),\ z_A(t-1)) \tag{2}$$

$$x(t) = [v_A(t-1),\ v_B(t),\ v_B(t) - v_A(t-1),\ p_A] \tag{3}$$

$$v_A(t) = g(x(t)) \tag{4}$$

where $z_A(t)$ is the emotional state embedding at time stamp t for speaker A; $v_A(t-1)$ is the predicted VAD metric for the previous utterance; $v_B(t)$ is the VAD metric produced by the VAD regressor; p_A is the personality metric in Big Five (OCEAN); g is a Multi-layer Perceptron (MLP); $v_A(t)$ is the predicted VAD metric for the target utterance.

We train the GRU model with EmoDynamic. Each dialog is parsed into sequential data and undergoes a standard GRU training. The loss function is specifically defined as the MSE between the actual next VAD label and the predicted next VAD given all the previous forward runs. We put the emotional state as an embedding with a dimension of 64. We run training for 40 epochs with a learning rate of $2e-4$, achieving a validation MSE of 0.0045, indicating in general a good alignment of the model with EmoDynamic data.

4.3. LLM

After we finished the VAD regressor and GRU encoder, we are now capable of modeling emotion dynamic and personality for predicting the VAD for the next-upcoming response of the speaker of interest. As such, the only step left is the emotionally-aligned generation using a small LLM. Here, we choose to use Qwen3.5-4B given its SOTA performance by release time.

This 4 Billion parameter large model is intentionally selected, since our local device and run a Unsloth fine-tuning of the model. However, due to timely constraints, we put it as one of our future work proceeding after this stage of the project.

Here, we work on some prompt engineering techniques in order to make the model stays consistent with the VAD instruction passed to it. These techniques ultimately come down to two general methods.

We first implemented a rule based prompting, outlining some of the specific rules in interpreting a VAD instruction. For every VAD category, we give detailed and descriptive words for low, medium, and high values. In this manner, we can put more weight towards the emotion instructed in the self-attention mechanism.

Meanwhile, we implemented some few-shots examples explaining how to generate given different VAD values. Few-shot prompting has consistently been proven by previous researches that it is going to improve model consistency with the given expectations compared to zero-shot prompting, where the model is not given with any examples to emulate.

We additionally put in some generation guidelines in order to set the model into the simulation of realistic conversation interactions. Some trials and modifications are needed after the prompts are completed.

5. Final Product

We implemented our Co-EmP design in a locally hosted web page design using Gradio, a package that specializes in chatbot html designs. The website is hosted through a local device running the python Gradio implementation. In the website, there are mainly four modes ready to use. Users have the full flexibility in filling background information for the scenario, personal context for speaker A and B respectively, the big five personality trait for both speakers, and the initial VAD for both speakers. Of course, users are also

Table 3. test results on VAD prediction across datasets. Lower MAE indicates lower prediction error, while higher Pearson correlation indicates stronger agreement with ground truth labels.

Dataset	Label Source	N	MAE	V-MAE	A-MAE	D-MAE	V-r	A-r	D-r	Mean R^2
EmoBank	Gold VAD	1000	0.047	0.050	0.050	0.041	0.709	0.442	0.374	0.067
DailyDialog	mapping	7740	0.068	0.073	0.085	0.046	0.616	0.590	0.577	0.314
EmpatheticDialogues	mapping	5714	0.130	0.144	0.125	0.123	0.721	0.576	0.652	0.408

Table 4. DailyDialog basic emotion labels mapped to approximate VAD values.

Emotion	Valence	Arousal	Dominance
neutral	0.50	0.20	0.50
anger	0.20	0.80	0.65
disgust	0.15	0.60	0.45
fear	0.10	0.85	0.20
happiness	0.85	0.65	0.70
sadness	0.15	0.25	0.25
surprise	0.55	0.85	0.45

able to tune typical LLM configurations such as the topk probability, max token per generation length, temperature, and so on. The model typically plays Speaker B while users play Speaker A.

Each mode is explained here:

- Mode 1: Baseline - This is a baseline chatbot with only Qwen3.5-4B implemented for emulating the emotion of a speaker in the given situation.
- Mode 2: Default - The is the fully supported chatbot with Co-EmP head implemented. The VAD regressor processes raw text from user into VAD values, and the GRU encoder takes in regressed VAD values and predicts the VAD for its next utterance generation. The instruction is passed toward Qwen3.5-4B emotion-aligned generation.
- Mode 3: Arena - This is a combination of mode 1 and 2, where the 2 different pipelines are directly put together side by side for comparison.
- Mode 4: Full AI simulation - This is where both speakers in the scenario are played by the LLM using different inferences. This is the mode that fully showcases the performance of the work and the quality of their simulation.

An example mode 4 scenario simulation is given in 6 7. Here, showed a consistent emotion flow for both speakers within the scenario. Speaker A is made as an outgoing and kind person, staying at his hometown ever since high school, while speaker B is made as an introverted person, nervous about the meeting. Here, the model realized VAD aligned very closely with the GRU's targeting VAD. For speaker A, the two turns showed a downward trend for valence, as speaker A is met with unexpected downfall from speaker B. Meanwhile, speaker B sees an increase in arousal, for he slowly gets to warm up with the conversation. The content by itself also proved the high quality performance of the Qwen3.5-4B model, as the additional topic of food is introduced without any external control in the context window and user inputs. This high resembles a realistic conversation, for the attention to the main topic drifts as conversation proceeds.

6. Limitations and Future Works

While the Co-EmP project overall is deemed a huge success, because of the timely constraints, there are still many inevitable limitations that could use some further refinement in the future.

The GRU encoder model is current trained only on VAD inputs and the personality vector, as indicated in equation (3). Such design does indeed models a emotion dynamic, taken previous emotion states as the contextual support for recurrent update. However, such design fails to capture any semantic content expressed by the counterpart speaker during any conversation. As such, it also misses the potential emotion changes caused by these semantic content. Take the example of a bad news announcement. If a speaker expresses the shocking bad news in a rather plain mood, GRU may fail to realize a decreasing trend in valence. Arguably, such bad news expression could be captured by the VAD regressor, indicating already a very low valence score to the GRU; however, such prediction is not guaranteed. In light of this, we suggest that adding an additional appraisal vector that includes evaluations on things such as personal relevance, impact on relationship, agency attributed to self or other, and so on could greatly capture the semantic significance of a raw text. If concatenated to the $x(t)$ input for GRU, the GRU model could now capture both emotional and semantic causes toward a emotion shift.

In addition, some simulation trials still indicate large discrepancies between the target VAD from GRU and the realized VAD from LLM. This suggests a difference in model comprehension to these VAD metrics. A solution to this would be to fine-tune the LLM using reinforcement learning methods such as Reinforcement Learning with Human Feedback or Direct Policy Optimization.

Table 5. EmpatheticDialogues emotion-context labels mapped to approximate VAD values .

Emotion	V	A	D	Emotion	V	A	D
surprised	0.55	0.85	0.45	excited	0.85	0.85	0.70
angry	0.20	0.80	0.65	proud	0.80	0.60	0.80
sad	0.15	0.25	0.25	annoyed	0.25	0.65	0.55
grateful	0.85	0.45	0.65	lonely	0.20	0.30	0.20
afraid	0.10	0.85	0.20	terrified	0.05	0.95	0.10
guilty	0.20	0.45	0.25	impressed	0.75	0.60	0.55
disgusted	0.15	0.60	0.45	hopeful	0.75	0.55	0.60
confident	0.80	0.55	0.85	furious	0.10	0.90	0.75
anxious	0.20	0.80	0.25	anticipating	0.60	0.70	0.50
joyful	0.90	0.70	0.70	nostalgic	0.60	0.35	0.45
disappointed	0.20	0.35	0.30	prepared	0.65	0.45	0.75
jealous	0.25	0.65	0.40	content	0.80	0.30	0.65
devastated	0.05	0.45	0.10	embarrassed	0.25	0.65	0.20
caring	0.80	0.40	0.60	sentimental	0.65	0.40	0.45
trusting	0.75	0.35	0.60	ashamed	0.15	0.45	0.15
apprehensive	0.30	0.65	0.25	faithful	0.75	0.35	0.65

Table 6. Scenario and speaker profiles

Item	Description
Scenario	Two high school friends reunite after ten years, discovering how much they have changed and attempting to reconnect.
Speaker A	Stayed in hometown after high school, working on farms. Emotionally open and excited about reunion.
Big Five	(0.8, 0.9, 0.9, 0.5, 0.2)
Initial VAD	(0.700, 0.800, 0.900)
Speaker B	Moved to a big city, worked on a startup after college. Busy life, slightly anxious about reunion.
Big Five	(0.3, 0.4, 0.2, 0.7, 0.8)
Initial VAD	(0.400, 0.400, 0.300)

References

[1] Mostafa Mahmoud Amin. Foundation models in affective computing. 2025. 2

[2] Mostafa M Amin, Rui Mao, Erik Cambria, and Björn W Schuller. A wide evaluation of chatgpt on affective computing tasks. *IEEE Transactions on Affective Computing*, 15 (4):2204–2212, 2024. 2

[3] Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, 2017. 1, 3

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 3

[5] Paul T. Costa and Robert R. McCrae. The revised neo personality inventory (neo-pi-r). In *The SAGE Handbook of Personality Theory and Assessment*, pages 179–198. SAGE Publications, 2008. 2

[6] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek, 2026. Accessed: 2026-05-21. 3

[7] Shin-nosuke Ishikawa and Atsushi Yoshino. Ai with emotions: Exploring emotional expressions in large language models. In *Proceedings of the 5th international conference on natural language processing for digital humanities*, pages 614–627, 2025. 1

[8] Oliver P. John and Sanjay Srivastava. The big-five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, 2 edition, 1999. 2

[9] Chenlu Ju, Jiaxin Liu, Shobhit Sinha, Hao Xue, and Flora

Table 7. conversation generated

Turn	Utterance	V_t	A_t	D_t	V_p	A_p	D_p	Context
A1	You made it! I was waiting here all day, thinking you'd never show up. You look exactly like the same crazy kid from ten years ago, just... bigger. Come on in, let's go get something to eat before I lose my mind	0.722	0.812	0.894	0.703	0.757	0.563	A
B1	I... wow. I'm sorry I'm late. I've been so busy with this whole city thing that I almost forgot to tell anyone I was coming back. Thanks for waiting, really. I just wanted to catch up, even if it feels weird seeing me like this.	0.351	0.409	0.280	0.371	0.633	0.377	B
A2	Don't apologize! You're back, and that's what matters! Look at this town—it's still my backyard, but honestly? It feels different without you running wild around here. Let's grab some food right now; I'm starving, and I don't care if we have to eat off the sidewalk. We have ten years of catching up to do, and I'm not wasting another second!	0.689	0.817	0.875	0.685	0.781	0.591	A
B2	I... I really shouldn't have come back like this, but I'm glad I did. Just... please don't rush us. I'm still a little lost, and it's okay if we sit on the curb for a while.	0.276	0.528	0.237	0.336	0.674	0.364	B

Salim. Trajllm: A modular llm-enhanced agent-based framework for realistic human trajectory simulation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2847–2850, 2025. 3

[10] Sumedha Kshirsagar. A multilayer personality model. In *Proceedings of the 2nd international symposium on Smart graphics*, pages 107–115, 2002. 3

[11] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi 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)*, pages 986–995, 2017. 1, 3

[12] Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. Speak from heart: an emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 international conference on multimedia retrieval*, pages 533–542, 2024. 2

[13] Albert Mehrabian and James A. Russell. *An Approach to Environmental Psychology*. The MIT Press, 1974. 2

[14] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada 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*, pages 527–536, 2019. 3

[15] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021. 1, 3

[16] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381, 2019. 1, 3

[17] Raj Kishor Verma, Gaurav Agarwal, and Nitin Pandey. Real-time emotion recognition and response generation via llm-embedded multimodal interfaces. In *2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, pages 1–7. IEEE, 2025. 2

[18] Shiquan Wang, Ruiyu Fang, Zhongjiang He, Shuangyong Song, and Yongxiang Li. Emotional support with llm-based empathetic dialogue generation. *arXiv preprint arXiv:2507.12820*, 2025. 2

[19] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013. 1, 3

[20] Hanqing Zhang, Si Sun, and Dawei Song. Emotion-aware llm adaptation for empathetic dialogue generation. *IEEE Transactions on Audio, Speech and Language Processing*, 34:125–137, 2025. 3

[21] Jiahao Zhu, Zijian Jiang, Boyu Zhou, Jionglong Su, Jiaming Zhang, and Zhihao Li. Empathizing before generation: a double-layered framework for emotional support llm. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 490–503. Springer, 2024. 2