

Algorithm Design for User Emotion Perception and Psychological State Prediction in Large Language Model Interaction

Yong Hu

College of Education, Nanning University, Nanning 530200, Guangxi, China

Corresponding author: Yong Hu (e-mail: m18076556686_1@163.com)

ABSTRACT With the continuous evolution of intelligent communication systems and distributed information processing technologies, real-time perception of user states has become an important component of adaptive human-machine interaction. This study proposes a dual-task parallel algorithm for instantaneous emotion perception and deep psychological state prediction in large language model interaction under pure text communication. The framework jointly models lexical features, syntactic complexity, interaction rhythm, topic coherence, and language habits to achieve synchronized short-term emotion tracking and long-term psychological state inference, while introducing an uncertainty estimation mechanism for confidence-aware decision making and interaction avoidance. Experimental results demonstrate that the proposed method achieves a macro F1 score of 0.85 for instantaneous emotion classification and a Pearson product-moment correlation coefficient of 0.76 for deep psychological state prediction, outperforming representative baseline approaches. Furthermore, confidence-driven interaction avoidance effectively suppresses high-risk misjudgments and improves overall interaction reliability. By integrating multi-level semantic perception with uncertainty-aware inference, the proposed framework provides an efficient solution for adaptive information processing and intelligent decision support, offering valuable methodological references for real-time signal interpretation, human-machine communication, and intelligent perception systems in Electromagnetic Waves, Antennas and Propagation related applications.

INDEX TERMS large language model interaction, dual-task learning, uncertainty estimation, intelligent information processing, adaptive human-machine communication

I. INTRODUCTION

AS large language model-driven dialogue systems are transitioning from task-oriented command execution to human-like deep interaction, machines are expected to understand users' explicit language content, model instantaneous emotional fluctuations, and infer stable psychological traits at the same time. However, the aforementioned two tasks are usually studied separately in current researches: the former only concerns with the emotion classification of single-turn texts, while the latter only depends on statistical feature modeling of multi-turn dialogues [1-2]. There is no common processing framework in interactive scenarios. Due to the aforementioned limitations of current interactive methods, in some typical interactive application scenarios such as psychological care, educational guidance, and human computer collaborative decision making, interactive systems can only track users' immediate emotional reaction or long-term psychological tendency in one of these two

ways: the former ignores the dynamic change of users' short-term state, while the latter misjudges users' deep need and thus is unable to resonate with users in terms of empathy, make inappropriate decisions, and even break trust with users [3-4]. Therefore, under the constraint of text interaction, which is the most common interaction medium with relatively low privacy risk, how to construct a dual task mechanism that can process instantaneous emotions and deep psychological states in a parallel way has become a key issue to enhance the depth and reliability of large language model interaction. Drawing on the distinction between emotional "states" and "traits" in affective science, we define "instantaneous emotion" as a transient, intense affective reaction (e.g., the six basic emotions) triggered by specific interaction stimuli. In contrast, "deep psychological state" is operationalized here as a longer-lasting affective mood or cognitive predisposition that remains stable over the duration of a dialogue session, bridging the gap between

fleeting emotions and permanent personality temperament.

Achieving parallel processing for both tasks presents three inescapable challenges. First, instantly changing moods are indicated by words appearing in the dialogue right at that instant, and thus call for algorithms with millisecond response requirements. Inferring psychological content is based on accumulating behavioral regularities across multiple rounds of dialogue, and so the two differ in their grain of time by several orders of magnitude [5-6]. Second, pure text media naturally lacks modal signals such as speech and facial expressions, forcing the system to simultaneously mine short-term emotional signals and long-term psychological traces from limited text features such as vocabulary, syntax, and rhythm. However, these two types of features are coupled and interfere with each other at the semantic level [7-8]. Third, psychological inferences output by the model when there is insufficient evidence may have a serious risk of being misleading. Existing methods generally lack the ability to quantify their own cognitive uncertainties and cannot proactively avoid judgments when unreliable, thus exposing misjudgments to users or downstream systems [9-10]. These three contradictions together constitute the technical bottleneck for achieving parallel processing of dual tasks in pure text interaction.

This paper aims to construct a dual-task parallel algorithm framework for pure text interaction, simultaneously solving the two major tasks of instantaneous emotion tracking and deep psychological state inference within a unified architecture, and introducing a quantifiable uncertainty management mechanism. The overall scheme is divided into four levels: at the input layer, emotion word frequency, syntactic complexity, and interaction rhythm features are extracted to achieve lightweight real-time capture of emotion signals in a single-turn dialogue; at the semantic layer, stable psychological traits of users are mined from the dialogue history through topic coherence modeling and temporal statistics of language habits; a dual-task uncertainty estimation module is introduced to evaluate the confidence of the outputs of instantaneous emotion classification and psychological state prediction, quantifying the model's grasp of its own judgment; a confidence-driven interaction avoidance mechanism is set up to automatically trigger avoidance behavior when the output is unreliable, avoiding high-risk misjudgments from being exposed to users or downstream systems. With above design, it is hoped that the algorithm can respond to transient emotion in real-time and predict users' deep psychological state in pure text interaction, while it also allows the model to reduce the appearance of risky responses when uncertainty is high. In experiments, the proposed algorithm on transient emotion classification task and deep psychological state prediction task is evaluated, and the result of good performance on both two tasks is achieved. Meanwhile, the interaction avoidance mechanism based on confidence effectively suppresses the appearance of high-risk misjudgment, which also proves the effectiveness of the parallel processing framework for dual tasks and

suppression of uncertainty.

II. DUAL-TASK ALGORITHM FRAMEWORK WITH CROSS-BRANCH INTEGRATION MECHANISM

The proposed framework addresses the dual dilemma not as two isolated processes, but through a multilevel integration. Specifically, the instantaneous emotion feature vector f_{inst} from the input layer is fed into the semantic layer as a temporal auxiliary signal. This allows the deep psychological state prediction to be informed by the cumulative emotional volatility, while the long-term psychological context from the semantic layer provides a 'baseline' for calibrating the instantaneous emotion classification. This bi-directional interaction ensures that immediate signals and long-term traits are processed in a unified, synergistic manner.

A. INSTANTANEOUS EMOTION FEATURE EXTRACTION FROM THE INPUT LAYER

In the dual-task algorithm framework for emotion perception and psychological state prediction, the input layer is responsible for extracting instantaneous emotion signals from the original text and is the first-level processing unit in the entire algorithm process. This layer takes the text input by the user in the current round as the only processing object and runs independently in each round of interaction, without relying on the information of subsequent rounds, so as to ensure the real-time response capability of the algorithm to instantaneous emotion fluctuations [11]. The output of the input layer is a lightweight feature vector, which is used by the subsequent uncertainty estimation module.

The expression of instantaneous emotion of users is first reflected in the word selection level. Under different emotional states, individuals tend to use specific categories of emotion words, a phenomenon that has been widely confirmed in psychological research [12]. To quantify this linguistic feature, a dictionary-based matching method is adopted. After word segmentation and word form restoration, the input text is matched with a preset emotion dictionary. The dictionary is based on the LIWC (Linguistic Inquiry and Word Count) and the Ontology of Chinese Emotions, containing 5,412 validated terms. The dictionary contains positive emotion words, negative emotion words, and six basic emotion words: anger, sadness, fear, happiness, surprise, and disgust. Each word corresponds to a unique emotion category. The frequency of various emotion words in the current round of text is counted and normalized with the total number of words in the current round of text to eliminate the influence of text length on frequency statistics, resulting in a six-dimensional emotion word frequency feature vector $f_{\text{lex}} \in R^6$. The i -th dimension is calculated as follows:

$$f_{\text{lex},i} = \frac{c_i}{N_{\text{tokens}}}, \quad i = 1, 2, \dots, 6. \quad (1)$$

In the formula, c_i is the number of times the i -th emotion word is matched, and N_{tokens} is the total number of words

in the current round of text.

Word selection reflects the content characteristics of emotion, while the structural mode of language organization reflects the cognitive load and thinking state when expressing emotion [13]. When emotionally excited or cognitively loaded, individuals often exhibit phenomena such as sentence simplification, sentence length shortening, or reduction of clause structure. In order to capture this structural level of emotion signal, syntactic complexity analysis is introduced, and a dependency parser is used to parse the current round of text to extract three types of features [14]. The ratio of the total number of words to the total number of sentences reflects the overall complexity of the sentence structure, i.e., the average sentence length:

$$f_{\text{len}} = \frac{N_{\text{tokens}}}{N_{\text{sentences}}}. \quad (2)$$

The degree of nesting within a sentence also reflects cognitive complexity, manifested as clause density, i.e., the ratio of the number of clauses to the total number of sentences:

$$f_{\text{clause}} = \frac{N_{\text{clauses}}}{N_{\text{sentences}}}. \quad (3)$$

Clause identification is based on the nesting relationship between the occurrence of subordinating conjunctions and verb structures. At a more macroscopic level, the overall depth of the syntax tree reflects the abstractness and hierarchical organization of language expression. The maximum depth of all syntax trees in the current round of text is taken as the third type of feature:

$$f_{\text{depth}} = \max_j D_j. \quad (4)$$

In the formula, D_j is the syntax tree depth of the j -th sentence. The three types of features together constitute the syntactic complexity feature vector $f_{\text{syntax}} \in R^3$.

The above lexical and syntactic features are extracted from the current round of text, reflecting the user's language selection and organization in a single input. During the dialogue process, these two types of features change dynamically with each round, and their evolution trajectory reflects the temporal pattern of user emotional fluctuations. Figure 1 shows the temporal distribution of emotional word frequency and syntactic complexity with each round in a certain dialogue sample. The stacked area represents the frequency evolution of the six types of emotional words, and the broken line reflects the changing trend of the syntactic complexity feature.

In the early stages of the dialogue (rounds 1-6) in Figure 1, the frequency of "Anger" significantly increases, peaking in round 5. "Clause Density" rises greatly, but "Tree Depth" does not rise in tandem, and remains low. Generally speaking, when users are in a heightened emotional state, they increase the number of clauses, rather than the depth of the syntactic hierarchy, that is, syntactic complexity presents

a dual differentiation feature: an increase in the number of clauses and a relatively small increase in depth. In the middle rounds of the dialogue (rounds 7-12), "Anger" sharply decreases, and "Sadness" increases. "Avg Sentence Length" and "Clause Density" both sharply decrease in rounds 7 and 8, increase in round 9, and then fluctuate within the range afterwards. In this stage, the overall trend of simplification in the syntactic structure is not obvious, but there is a certain degree of fluctuation feature. In other words, when the other person is "Sadness", the speaker has a certain degree of fluctuation in their style of organizing the syntax, and there may also be phased changes in the expressive intention of the language and cognitive resources. In the later rounds of the dialogue (rounds 13-20), the word "Joy" appears more and more often, reaching a high value around round 16. The three syntactic complexity lines generally increase, indicating that the enhancement of positive emotions is accompanied by the recovery of the desire to express and the enrichment of sentence structure. The frequencies of the two types of emotional words, "Disgust" and "Fear," remain low throughout the process, having a limited impact on the overall trend of syntactic features. The above patterns indicate that the change of emotional state has an observable structural correspondence at the syntactic level. The synergistic changes in emotional word frequency features and syntactic complexity features in the temporal dimension jointly characterize the emotional expression mode of user language content and structure. The user's behavior patterns during the interaction process also carry emotional information, constituting another source of signal beyond the language level.

Beyond language content and structure, the user's behavior patterns during the interaction process also carry emotional information [15]. In pure text dialogue, the changes in the timing of input and the length of text are often closely related to psychological states such as emotional arousal and cognitive load [16]. When the user is excited, the input interval is shorten or fluctuates in some ways; when the user is depressed or unwilling to make a decision, the input interval becomes long. To model this information on user's behavior level, an interaction rhythm feature is designed as follows.

A sliding window of length $w = 5$ rounds is defined, and the time and length of text of current round and previous $w - 1$ rounds are recorded. Timestamps (t_k) are recorded with a millisecond resolution to capture subtle pacing changes. The input time intervals between adjacent rounds, $\Delta t_k = t_k - t_{k-1}$, form a sequence, where t_k is the timestamp of the k -th round. The mean of the interval sequence within the window reflects the central tendency of the user's input rhythm.

$$\bar{\Delta t} = \frac{1}{w-1} \sum_{k=1}^{w-1} \Delta t_k \quad (5)$$

The mean alone cannot characterize the degree of rhythm

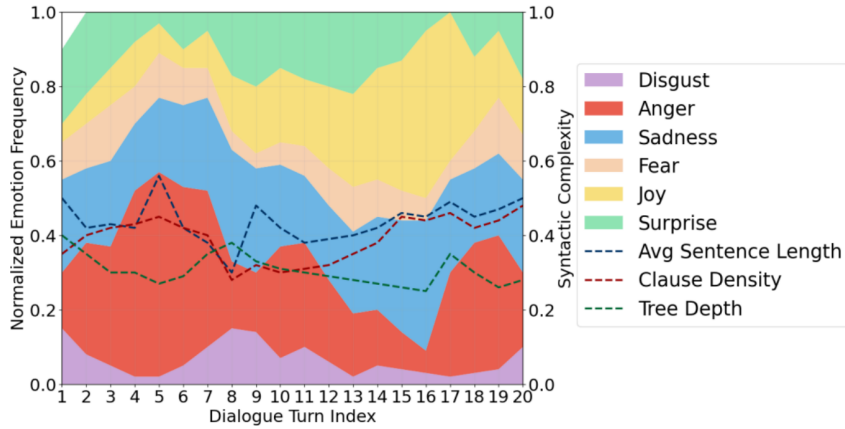


FIGURE 1. Temporal joint distribution of frequency and syntactic complexity features of six emotion categories during the dialogue process

fluctuation, and the stability of rhythm is closely related to emotional state. Therefore, the standard deviation of the interval sequence within the window is further calculated:

$$\sigma_{\Delta t} = \sqrt{\frac{1}{w-2} \sum_{k=1}^{w-1} (\Delta t_k - \bar{\Delta t})^2} \quad (6)$$

The above two statistics are calculated based on historical windows and reflect the user's long-term rhythm pattern. The instantaneous interval between the current round and the previous round captures the instantaneous change of rhythm and also contains the user's emotional feedback on the response of the previous round. This instantaneous feature is defined as:

$$\Delta t_{\text{current}} = t_{\text{current}} - t_{\text{current}-1}. \quad (7)$$

In the formula, t_{current} is the timestamp of the current round input, and $t_{\text{current}-1}$ is the timestamp of the previous round input.

After the extraction of the time interval feature, the change in text length also reflects the change in emotional state. When emotions are excited, the text length may increase or decrease sharply [17]. To capture this feature, a text length change rate is introduced, defined as the ratio of the current round's text length to the average text length within the window:

$$r_{\text{len}} = \frac{L_{\text{current}}}{\frac{1}{w} \sum_{k=0}^{w-1} L_{t-k}}. \quad (8)$$

In the formula, L_{current} is the current round's text length, and L_{t-k} is the k -th round's text length within the window. Finally, the time interval, rhythm average, rhythm fluctuation, and text length change rate together constitute the interaction rhythm feature vector $f_{\text{rhythm}} \in R^4$. In this study, text length (L_{current}) is measured by the number of tokens after word segmentation.

Therefore, from word selection and syntactic structure to interaction rhythm, three types of features characterize the user's instantaneous emotional expression from different

dimensions. The above three feature vectors are concatenated to obtain the instantaneous emotional feature vector $f_{\text{inst}} = [f_{\text{lex}}; f_{\text{syntax}}; f_{\text{rhythm}}] \in R^{13}$ for the current round's input. This feature vector is generated in real-time after each round of user input using a lightweight computation method, serving as the input for the instantaneous emotional perception branch in the subsequent uncertainty estimation module.

B. SEMANTIC LAYER DEEP PSYCHOLOGICAL STATE FEATURE MODELING

After the input layer completes the real-time capture of instantaneous emotions, the semantic layer under takes the function of mining the user's deep psychological state from the dialogue history. Differ from input layer, which processes immediate signals from single-turn text, the semantic layer takes accumulated multi-round dialogues as its processing object to output long-term features representing user's stable psychological traits [18]. The output of semantic layer is the input of deep psychological state prediction module for supporting subsequent estimation of uncertainty and interaction decision-making.

The stability of the user's psychological state is reflected first in the coherence of topic. In the long-term dialogue, the amount of time a user pays attention to a certain topic, the number of times switching between topics, and the concentration of focus on a topic are related to the user's cognitive style, interest preference, and even anxiety tendency [19-20]. A Latent Dirichlet Allocation (LDA) model is adopted, pre-trained on a diverse corpus of 100,000 human-computer dialogue samples. The topic space is set to $K=50$ topics, which was determined to be optimal through perplexity analysis. Each round of user input is mapped to a preset topic space to obtain the topic distribution vector of that round of text. Based on the topic distribution sequence, three types of topic coherence features are defined.

Topic switching frequency measures the degree of significant change in topic between adjacent dialogue rounds, defined as the proportion of adjacent rounds where the cosine distance between topic distributions exceeds a threshold

θ :

$$f_{\text{switch}} = \frac{1}{M-1} \sum_{k=2}^M \mathbf{1}(d(p_k, p_{k-1}) > \theta). \quad (9)$$

In the formula, p_k is the topic distribution vector of the k -th round of text; $d(\cdot, \cdot)$ is the cosine distance function; $\mathbf{1}(\cdot)$ is the indicator function (value is 1 when the condition is met, otherwise it is 0); M is the number of dialogue rounds in the history window. A high value of this feature indicates that the user's thinking is more jumpy, which may reflect a psychological state of inattentiveness or cognitive instability.

Topic focus measures the degree of concentration of user dialogue topics as a whole. The topic distribution vectors of all rounds in the window are summed and normalized to obtain the topic distribution P_{window} in the window. The topic corresponding to the highest probability value in this distribution is the core topic that the user focuses on during the window period. This probability value itself can be used as a direct measure of focus:

$$f_{\text{focus}} = \max_t P_{\text{window},t}. \quad (10)$$

In the formula, $P_{\text{window},t}$ is the probability of the t -th topic in the topic distribution within the window. The higher the focus value, the more concentrated the user's dialogue topics are on a single category, which is related to the stability of interests or obsession tendencies.

Topic distribution entropy measures the diversity of topics in each round within the window from another dimension [21], and is defined as the mean of the topic distribution entropy of each round:

$$f_{\text{entropy}} = \frac{1}{M} \sum_{k=1}^M H(p_k). \quad (11)$$

In the formula, $H(p_k)$ represents the topic distribution entropy of the k -th round of text. A high value of this feature indicates that the user presents an expression pattern of multiple topics intertwined in a single round of dialogue, which may reflect a psychological state of divergent thinking or high cognitive load.

The three features mentioned above—topic switching frequency, topic focus, and topic distribution entropy—characterize the topic coherence of user dialogue from three dimensions: the activity level of topic changes, the overall concentration level, and the local diversity, respectively, and together constitute the topic coherence feature vector $f_{\text{topic}} \in R^3$.

After completing the topic-level modeling, it is necessary to further characterize the stable language behavior patterns formed by users in long-term dialogues. Language habits, as an external representation of individual psychological traits, show cross-time stability in terms of pronoun usage preference, negative word frequency, abstract word ratio, and sentence length stability [22-23]. To capture these habitual

features, a language behavior feature vector is constructed based on the dialogue history. Language habit features are characterized from four dimensions: self-focus, avoidance tendency, degree of abstract thinking, and consistency of expression.

Pronoun usage preference reflects the user's degree of self-focus. The frequency of first-person singular and first-person plural pronouns within the statistical window is calculated and normalized to the total number of words in the window, yielding f_{self} and f_{we} , respectively. The ratio of these two features, $f_{\text{self}}/f_{\text{we}}$, serves as a feature value for pronoun usage preference; a higher value may reflect a self-focused psychological tendency.

The frequency of negative words reflects the user's avoidance tendency. The frequency of negative words within the statistical window is calculated and normalized to the total number of words in the window, yielding f_{neg} . This feature is typically higher in individuals with depressive tendencies or anxiety states.

The proportion of abstract words reflects the user's level of abstract thinking. An abstractness dictionary is used to match the text within the window, and the frequency of abstract words is calculated and normalized to the total number of words in the window, yielding f_{abstract} . The tendency towards abstract expression is related to an individual's analytical thinking traits. Sentence length stability reflects the consistency of the user's language expression. The average sentence length $f_{\text{len},k}$ for each round of text within the window is calculated, where $k = 1, 2, \dots, M$, yielding an average sentence length sequence. The standard deviation of this sequence is used as a stability measure.

$$f_{\text{stability}} = \sqrt{\frac{1}{M-1} \sum_{k=1}^M (f_{\text{len},k} - \bar{f}_{\text{len}})^2} \quad (12)$$

In the formula, $\bar{f}_{\text{len}}$ represents the mean of the average sentence length within the window. A low value of this feature indicates stable sentence lengths for users, potentially reflecting entrenched language habits or a stable psychological state.

The above four features together constitute the language habit feature vector $f_{\text{habit}} \in R^4$.

Topic coherence and language habit, as two types of features, characterize the user's deep psychological state from the dimensions of topic evolution and language behavior, respectively. These two aspects intertwine and co-evolve during the dialogue process. Users with high topic focus tend to exhibit stable language habits, while users who frequently switch topics may show increased use of negative words or more pronounced fluctuations in sentence length. Figure 2 shows the co-evolution trajectory of topic coherence and language habit features with each dialogue round in a specific dialogue sample, where each line represents the overall distribution of a user across the seven normalized feature dimensions.

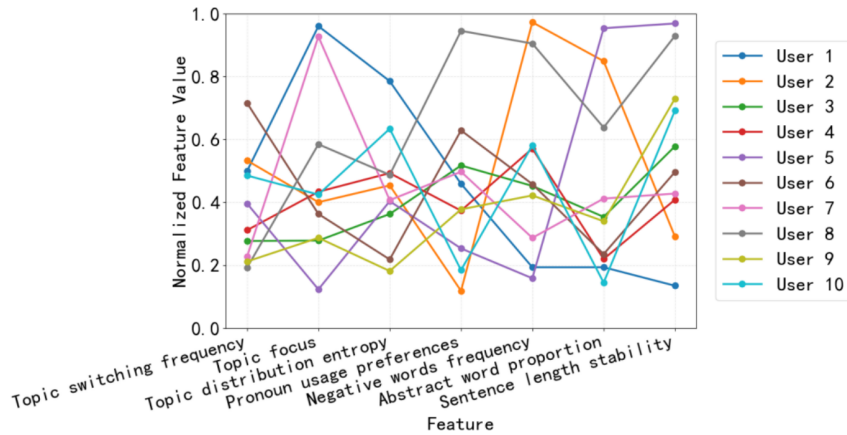


FIGURE 2. Cross-dimensional co-distribution of deep psychological state characteristics of multiple users

Figure 2 shows that on the topic coherence dimension, Users 1 and 7 exhibit high values for Topic focus, while Topic switching frequency and Topic distribution entropy are relatively low. This reflects that these users have highly concentrated dialogue topics and stable thinking focus, which is consistent with the low fluctuations in their Sentence length stability dimension. This indicates that the psychological state of topic convergence is simultaneously manifested as the regularization of sentence structure at the language expression level. In contrast, User 8 shows significantly higher values for both abstract word proportion and negative word frequency, while User 5 shows a higher abstract word proportion but a lower negative word frequency and a lower topic focus value. Concatenating these two types of features yields the deep psychological state feature vector $f_{\text{deep}} = [f_{\text{topic}}; f_{\text{habit}}] \in R^7$. This feature vector is calculated based on the dialogue history and updated after each round of interaction, serving as the input to the deep psychological state prediction branch in the subsequent uncertainty estimation module.

III. EXPERIMENTAL DESIGN

A. DATASET AND PREPROCESSING

The sentence used in the experiment is the multi-turn human-computer dialogue text. The content of dialogue is from three scenarios of psychological care, expressing emotion and ordinary interaction to guarantee the diversity of training sample in the aspect of emotion type and psychological state. The corpus includes 583 dialogue sequences; the average number of dialogue turns per sequence is 20; the total number of annotated turns is 11,660, from which 2,000 representative samples were selected for detailed psychological state manual scoring. The dataset carries two types of annotations: transient emotion labels and deep psychological state labels. Transient emotion labels cover six basic emotion categories: anger, sadness, fear, happiness, surprise, and disgust, corresponding to the classification target of the six-dimensional emotion word frequency feature vector in the input layer. Deep psychological state labels are continuous values normalized to the $[0,1]$ interval, reflecting

the overall psychological state intensity of the user during the dialogue window, corresponding to the regression target of the semantic layer deep psychological state prediction task. Transient emotion labels are independently annotated by three trained annotators for each single-turn text, and the final category is determined by majority voting. Deep psychological state labels are evaluated by professional annotators with psychology backgrounds based on complete dialogue sequences, and normalized after scoring using a Likert scale. The dataset is divided into training, validation, and test sets in a 7:1:2 ratio, with the division based on dialogue sequences rather than individual texts to avoid data leakage caused by different rounds of the same dialogue sequence appearing in the training and test sets respectively. The training set is used for model parameter optimization and confidence distribution statistics; the validation set is used for hyperparameter tuning and early stopping detection; the test set is used for the final evaluation of all performance metrics in Chapter 4.

B. COMPARISON METHODS AND PARAMETER CONFIGURATIONS

For the transient emotion classification task, three methods are selected as comparison baselines: Robustly Optimized BERT Pretraining Approach (RoBERTa)-base, Pattern-Exploiting Training (PET), and Generative Pretrained Transformer 3.5 (GPT-3.5) zero-shot method. RoBERTa-base represents the mainstream emotion classification approach based on fine-tuning of a pre-trained language model; PET represents a few-shot emotion recognition method based on cue learning; and GPT-3.5 zero-shot method represents the emotion classification ability of a large language model without specific training. These three methods cover the main technical approaches in current text emotion classification research. For the deep psychological state prediction task, eXtreme Gradient Boosting (XGBoost) and the Transformer sequence model are selected as comparison baselines, with an ablation variant that removes temporal features of emotion. The XGBoost baseline utilizes “static features” which are the global averages of lexical and syn-

tactic features aggregated over the entire dialogue sequence, discarding all temporal information. An ablation variant (Ours w/o emotional-temporal features) is used to verify the gain contribution of input layer sentiment temporal features to deep psychological state prediction. This variant includes all semantic layer features (topic coherence and language habits) but replaces the turn-by-turn instantaneous emotion features from the input layer with a zero-padding vector.

IV. RESULTS

A. INSTANTANEOUS EMOTION PERCEPTION PERFORMANCE

In instantaneous emotion perception performance analysis, due to the macro-level indicator representation, it is hard to analyze the confusion structure and misdistribution relationship of the model on various emotion categories. Therefore, in order to analyze the classification situation of the proposed method on six emotion categories at a granular level, the confusion matrix on test set is visualized as shown in Figure 3. At the same time, true sample distribution and predicted category attribution of each category are represented by inner and outer double-ring structure of the sunburst diagram, respectively.

Figure 3 shows that Disgust and Joy achieve the highest number of correct classifications (101 and 96 respectively), representing the largest proportion of the total real samples in their respective categories. This is closely related to their significant lexical boundary features in language expression, indicating that the model has a more stable grasp of the feature anchor points for these emotions. Sadness achieves a relatively low number of correct classifications (70) compared to the other two categories. Significant misclassification scattering towards Fear and Anger is observed in its outer ring. This phenomenon stems from the partial overlap in syntactic structure and emotional vocabulary between Sadness and Fear/Anger in language expression. The blurred boundaries of the three negative emotions in the feature space lead to classifier confusion. Fear achieves a correct number of classifications (71), with misclassification concentrated in Sadness, further confirming the feature coupling problem within the negative emotion cluster. Anger's outer ring misclassification scattering is relatively dispersed, reflecting the overlap between the linguistic features of anger and other emotions as the intensity changes. The Surprise model achieves 85 correct classifications, with misclassifications primarily falling into the Fear and Anger categories, indicating that the semantic distance between high-arousal emotions is relatively close at the feature level. Overall, cross-class confusion between positive and negative emotion clusters is minimal, and misclassification mainly occurs within the same polarity of emotion. The above confusion structure shows the training behavior of different classes, and the comprehensive performance of the model should be compared with some baseline methods to see. In Table 1, the proposed method is compared with three baseline methods, namely RoBERTa-base, PET, and GPT-

3.5 zero-shot methods, on macro F1 score, precision, recall, and accuracy.

TABLE 1. Performance comparison of different methods on the instantaneous emotion classification task

Method	Macro F1 value	Precision	Recall	Accuracy
RoBERTa-base	0.79	0.77	0.81	0.76
PET	0.81	0.80	0.82	0.78
GPT-3.5 zero samples	0.75	0.73	0.77	0.71
Proposed method	0.85	0.85	0.85	0.86

In Table 1, the proposed method achieves a macro F1 score of 0.85, significantly outperforming the RoBERTa-base baseline (0.79), with a macro F1 score of 0.85, precision and recall both at 0.85, and accuracy at 0.86, making it the best performing method among all comparisons. To verify the statistical significance of these improvements, a McNemar's test was conducted, yielding a p-value < 0.01, confirming that the algorithmic superiority is statistically significant and not due to random variation. RoBERTa-base has a macro F1 score of 0.79 and PET of 0.81. Both rely on the pre-trained language model's overall encoding of text semantics, and their ability to distinguish negative emotion cluster boundaries is limited by the fit between the general pre-training objective and emotion features, leading to a certain imbalance between recall and precision. GPT-3.5's zero-shot macro F1 score is 0.75, the lowest among the four methods. This is directly related to the lack of specific calibration for the current dataset's emotion boundaries under the zero-shot setting, making it particularly unstable in distinguishing overlapping feature categories such as Sadness and Fear. The high consistency between precision and recall in the proposed method indicates that the multi-dimensional instantaneous emotion features extracted by the input layer have a simultaneous effect on reducing both missed and misclassification errors. This balance, coupled with the structural pattern in the sunburst diagram where misclassification is concentrated within the same polarity of emotion and cross-polarity confusion is minimal, corroborates the synergistic effect of the feature system presented in this paper at both the emotion polarity discrimination and category segmentation levels.

B. PERFORMANCE OF DEEP MENTAL STATE PREDICTION

The output of the deep mental state prediction task is a continuous value. A single correlation coefficient index is insufficient to reflect the distributional correspondence between predicted and true values and the spatial clustering pattern of errors. To visually present the overall fitting quality of the proposed method in continuous mental state prediction, this paper visualizes the predicted and true values of 2000 samples on the test set using hexagonal binning. The color depth encoding of sample density reveals the degree of deviation of the predicted distribution from the ideal prediction line, as shown in Figure 4.

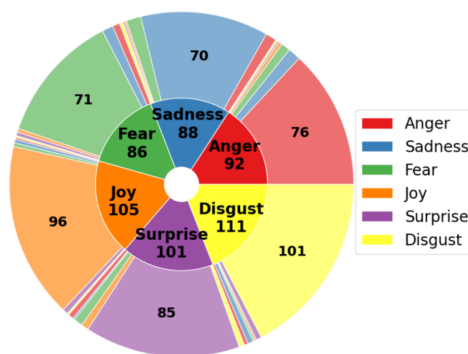


FIGURE 3. Confusion matrix sunburst plot of six emotion classification results

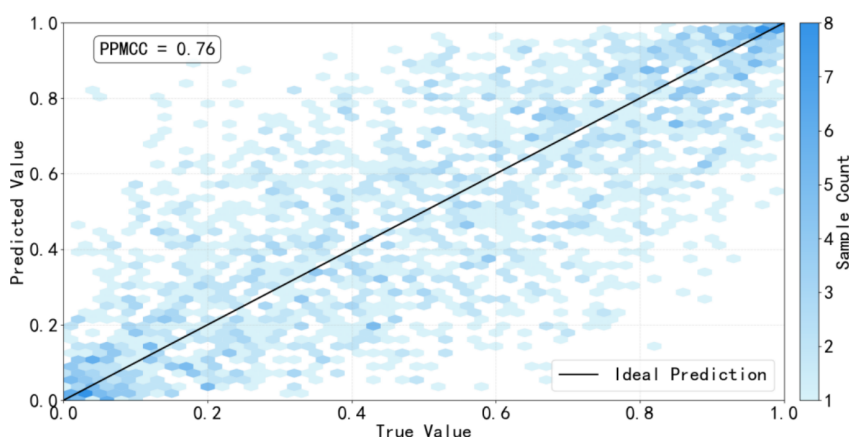


FIGURE 4. Hexagonal binning density distribution of predicted and actual deep psychological states

In Figure 4, the high-density regions of the sample points are distributed in a band along the ideal prediction diagonal. The darker hexagonal grids are concentrated near the diagonal, indicating that the predicted values of most samples maintain a high degree of numerical consistency with the actual values. The overall fitting trend is consistent with the quantization result of PPMCC 0.76. In the low value range of 0.0 to 0.3, the dispersion of the predicted values is relatively large; the binning distribution is sparse; the proportion of samples deviating from the diagonal increases. This phenomenon stems from the sparse dialogue history signal in the low psychological state value range, where the accumulated semantic features are insufficient to support accurate continuous value regression, resulting in a higher prediction variance in this range. In the high value range of 0.6 to 1.0, the high-density binning fits the diagonal relatively well, indicating that the model's prediction results in this range have better consistency with the actual values. The above distribution pattern reveals the spatial structure of the prediction error from the perspective of sample density. The overall performance gain of the proposed method compared to baseline methods needs to be confirmed through a systematic comparison of multiple quantitative indicators. Table 2 summarizes the comparison results of the proposed

method with XGBoost, Transformer sequence models, and ablation variants on four indicators: PPMCC, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 .

TABLE 2. Performance comparison of different methods on deep mental state prediction tasks

Method	PPMCC	RMSE	MAE	R2
XGBoost	0.65	0.18	0.14	0.42
Transformer sequence model	0.71	0.15	0.12	0.51
Ours (w/o emotional-temporal features)	0.69	0.16	0.13	0.48
Proposed method	0.76	0.13	0.10	0.58

Table 2 shows that the proposed method achieves a PPMCC of 0.76, an RMSE of 0.13, a MAE of 0.10, and an R^2 of 0.58, leading all four methods. XGBoost's PPMCC is only 0.65, with RMSE and MAE of 0.18 and 0.14 respectively. This is directly related to its static feature modeling approach, which fails to capture the temporal evolution of topic coherence and language habits in dialogue history. The dynamic psychological signals accumulated through dialogue turns are largely lost in this framework. The Transformer sequence model improves its PPMCC to 0.71, reflecting that sequence modeling utilizes temporal

features better than static methods. However, its R^2 is only 0.51, indicating a deficiency in its ability to interpret variance of continuous values. This is related to its lack of a specific mechanism for constructing psychological state features. The ablation variant of the proposed method, which removes temporal emotion features, achieves a PPMCC of 0.69, lower than the 0.76 of the complete method. This indicates that the instantaneous emotion features extracted from the input layer contribute significantly to deep psychological state prediction, and the temporal patterns of instantaneous emotions provide effective supplementary signals for semantic layer psychological state inference. Overall comparison of the four metrics shows that the performance improvement of the proposed method on continuous psychological state prediction tasks stems from the synergistic effect of the dual-layer feature system, rather than the local optimization of a single module.

C. UNCERTAINTY ESTIMATION CALIBRATION EFFECT

Whether the confidence level output by the uncertainty estimation module truly reflects the reliability of the model's prediction directly determines the effectiveness of the interaction avoidance mechanism. If there is a systematic deviation between the confidence level and the actual accuracy, the threshold-driven avoidance rule loses its statistical basis. To examine the calibration quality of the confidence level output of the proposed method, the test set samples are merged for both tasks and binned them according to confidence level intervals, presenting the overall calibration quality of the dual-task uncertainty estimation within a unified confidence-accuracy framework. The continuous value output of the deep psychological state prediction task is binarized and then included in the accuracy statistics. The actual accuracy and its 95% confidence interval in each interval are statistically analyzed, and the sample distribution of the confidence level is presented simultaneously, as shown in Figure 5.

In Sub-figure 5(1), the actual accuracy curve is generally below the perfect calibration line, indicating that the model has a certain degree of overconfidence tendency in each confidence interval, that is the output confidence is systematically higher than the true accuracy of the corresponding interval. The Expected Calibration Error (ECE) value of 0.118 reflects the overall magnitude of this bias. In the interval with a confidence level below 0.5, there is a certain deviation between the actual accuracy and the perfect calibration line, and the 95% confidence interval is relatively wide. This is directly related to the scarcity of samples in the low confidence interval in Subfigure 5(2). Insufficient samples lead to a large variance in the accuracy estimate in this interval, and the statistical stability of the calibration curve is low. Within the confidence level range of 0.6 to 0.9, the actual accuracy monotonically increases with confidence, and the curve trend remains basically parallel to the perfect

calibration line, with the deviation amplitude tending to stabilize. This indicates that the model has a strong ranking ability in the medium-to-high confidence level range, and the confidence level has a consistent characterization of the prediction reliability. Subfigure 5(2) shows that the samples are mainly concentrated in the 0.6 to 0.9 range, with a very low proportion of samples in the low confidence level range. This is consistent with the statistical characteristic in Table 1 that both types of confidence level distributions are skewed towards the high value range, indicating that the model tends to output medium-to-high confidence judgments on most samples. Combining the distribution patterns of the two subfigures, the confidence level has a stable monotonically corresponding relationship with the prediction accuracy in the medium-to-high value range, providing a calibration-level support basis for the avoidance trigger rule with a threshold $\tau = 0.6$.

V. CONCLUSION

The proposed dual-task parallel algorithm framework based on pure text interaction, instant emotion perception and deep psychological state inference share the common processing space. The explicit control over cognitive boundary of even the model itself is enabled by a quantifiable uncertainty management mechanism, which offers a possible algorithmic route of human computer interaction with both dynamic response capability and risk control in privacy preserving applications. Quantization of emotion dictionary and topic model are adopted for feature construction in current framework. The coverage of dictionary and number of topics are limitations for feature quality. The generalization performance of proposed framework on language general or domain transfer tasks remains to be evaluated. The computational overhead of multiple forward propagation in Monte Carlo Dropout is still reduced in real time interaction scenario. Meanwhile, the manual evaluation cost of deep psychological state label hinders the expansion of dataset. The cost of labeled data is alleviated by introducing active learning strategy into the uncertainty framework. Besides, it is also possible to extend this pure text framework to joint modeling of speech prosody or lightweight physiological signals based on the condition that more gradual access of multimodal signals is available.

AUTHOR CONTRIBUTIONS

Hu Yong designed, collected and analyzed the data, and drafted the manuscript. Hu Yong conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Hu Yong participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

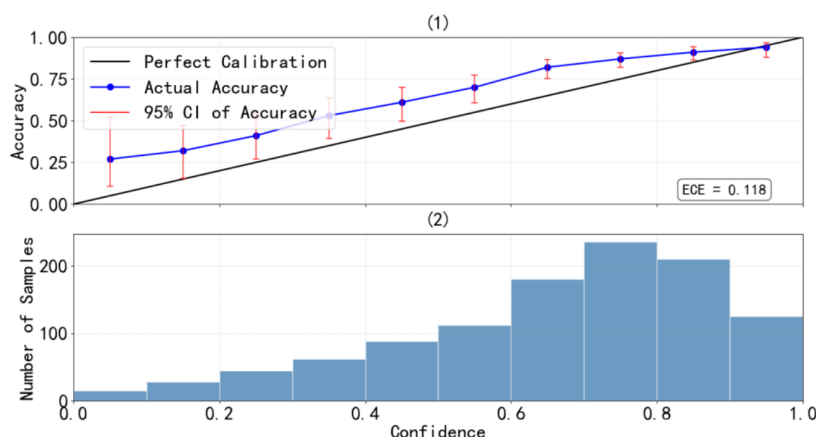


FIGURE 5. Confidence calibration curve and sample distribution histogram; Figure 5(1) Confidence-accuracy calibration curve; Figure 5(2) Sample size distribution in confidence intervals

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] P. Pereira, H. Moniz, and P. Carvalho J, "Deep emotion recognition in textual conversations: A survey," *Artificial Intelligence Review*, vol. 58, no. 1, pp. 10-18, 2024, doi: 10.1007/s10462-024-11010-y.
- [2] J. Jang, S. Yoon, G. Son, M. Kang, and H. Choeh J Y and Choi K, "Predicting personality and psychological distress using natural language processing: a study protocol," *Frontiers in Psychology*, vol. 13, no. 1, pp. 865541-865548, 2022, doi: 10.3389/fpsyg.2022.865541.
- [3] J. Schneider, "Mental model shifts in human-LLM interactions," *Journal of Intelligent Information Systems*, vol. 63, no. 5, pp. 1737-1752, 2025, doi: 10.1007/s10844-025-00960-6.
- [4] A. Sharma, W. Lin I, S. Miner A, and T. Atkins D C and Althoff, "Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support," *Nature Machine Intelligence*, vol. 5, no. 1, pp. 46-57, 2023, doi: 10.1038/s42256-022-00593-2.
- [5] J. Ren F, Y. Y. Zhou, J. W. Deng, K. Matsumoto, D. Feng, T. H. She, et al., "Tracking emotions using an evolutionary model of mental state transitions: Introducing a new paradigm," *Intelligent Computing*, vol. 3, no. 1, pp. 0075-0081, 2024, doi: 10.34133/icomputing.0075.
- [6] Y. Fu, S. Yuan, and J. Zhang C and Cao, "Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods," *Electronics*, vol. 12, no. 22, pp. 4714-4723, 2023, doi: 10.3390/electronics12224714.
- [7] H. Lian, C. Lu, S. Li, Y. Zhao, and Y. Tang C and Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, pp. 1440-1448, 2023, doi: 10.3390/e25101440.
- [8] S. Zhang, Y. Yang, C. Chen, X. Zhang, and X. Leng Q and Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, no. 1, pp. 121692-121700, 2024, doi: 10.1016/j.eswa.2023.121692.
- [9] S. Farquhar, J. Kossen, and Y. Kuhn L and Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625-630, 2024, doi: 10.1038/s41586-024-07421-0.
- [10] O. Shorinwa, Z. Mei, J. Lidard, and A. Ren A Z and Majumadar, "A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions," *ACM Computing Surveys*, vol. 58, no. 3, pp. 1-38, 2025, doi: 10.1145/3744238.
- [11] F. Mohammad, M. Khan, S. Nawaz, N. Jan, and A. Gohr N and Al-Rasheed, "Text augmentation-based model for emotion recognition using transformers," *Computers, Materials, & Continua*, vol. 76, no. 3, pp. 3523-3529, 2023, doi: 10.32604/cmc.2023.040202.
- [12] P. Ferré, I. Fraga, and A. Hinojosa J, "The interplay between language and emotion: a narrative review," *Cognition and Emotion*, vol. 39, no. 7, pp. 1418-1445, 2025, doi: 10.1080/02699931.2025.2549965.
- [13] M. A. Tabari, X. Lu, and Y. Wang, "Mapping the associations among task complexity, emotionality, and linguistic complexity in L2 writing," *Language awareness*, vol. 34, no. 2, pp. 345-372, 2025, doi: 10.1080/09658416.2024.2384754.
- [14] Q. Xu, P. Guo, F. Li, M. Zhang, and D. Ji, "Improving LLM-based opinion expression identification with dependency syntax," *Pattern Recognition Letters*, vol. 197, no. 1, pp. 81-87, 2025, doi: 10.1016/j.patrec.2025.07.012.
- [15] D. H. Solomon, X. Tian, and S. S. Li Y and Li, "Showcasing studies of conversational motifs within a dynamic dyadic systems perspective on interpersonal communication," *Journal of Language and Social Psychology*, vol. 44, no. 3-4, pp. 553-581, 2025, doi: 10.1177/0261927X251316863.
- [16] E. Džubur and V. Pejović, "Inferring a Mobile User's Valence and Arousal through On-Screen Text Analysis," *HCI SI Koper 2025*, vol. 1, pp. 171-180, 2025, doi: 10.26493/978-961-293-559-7.16.
- [17] J. Tian and R. Zhang, "Moral judgments influence emotional responses and comment lengths through the moderating role of linguistic style matching," *Scientific Reports*, vol. 15, no. 1, pp. 20972-20980, 2025, doi: 10.1038/s41598-025-07079-2.
- [18] Y. Yang, J. Cao, and P. Wen Y and Zhang, "Multiturn dialogue generation by modeling sentence-level and discourse-level contexts," *Scientific Reports*, vol. 12, no. 1, pp. 20349-20355, 2022, doi: 10.1038/s41598-022-24787-1.
- [19] K. Matsumoto, M. Sasayama, and T. Kiriara, "Topic break detection in interview dialogues using sentence embedding of utterance and speech intention based on multitask neural networks," *Sensors*, vol. 22, no. 2, pp. 694-702, 2022, doi: 10.3390/s22020694.
- [20] R. Ding, X. Luo, S. Chen, X. Wu, and L. Zhang C and Liu, "A cognitive topic modeling approach for detailed learner analysis in asynchronous discussions," *IEEE Access*, vol. 13, no. 1, pp. 94169-94185, 2025, doi: 10.1109/ACCESS.2025.3573873.
- [21] A. Hosseiny Marani and S. Baumer E P, "A review of stability in topic modeling: Metrics for assessing and techniques for improving stability," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1-32, 2023, doi: 10.1145/3623269.
- [22] J. Ryu, S. Heisig, C. McLaughlin, M. Katz, and X. Mayberg HS and Gu, "A natural language processing approach reveals first-person pronoun usage and non-fluency as markers of therapeutic alliance in psychotherapy," *IScience*, vol. 26, no. 6, pp. 106860-106867, 2023, doi: 10.1016/j.isci.2023.106860.
- [23] J. Piepers, A. Backus, and J. Swanenberg, "Sociopragmatic pronouns in Limburgian: inferring speakers' agency from self-reported automaticity, attitudes, and metalinguistic awareness," *Cognitive Linguistics*, vol. 36, no. 1, pp. 89-119, 2025, doi: 10.1515/cog-2023-0141.