# PychoAgent: Psychology-driven LLM Agents for Explainable Panic Prediction on Social Media during Sudden Disaster Events


**Mengzhu Liu[†1,3], Zhengqiu Zhu[†1,3], Chuan Ai[*1,3], Chen Gao[*2], Xinghong Li[1,3], Lingnan He[4], Kaisheng Lai[5], Yingfeng Chen [2], Xin Lu[3], Yong Li[2], Quanjun Yin[1,3]**

[1]State Key Laboratory of Digital Intelligent Modeling and Simulation, National University of Defense Technology

[2]BNRist, Tsinghua University [3]College of Systems Engineering, National University of Defense Technology

[4]Department of Psychology, Sun Yat-Sen University [5]School of Journalism and Communication, Jinan University

[†]Equal contribution {liumengzhu2001, zhuzhengqiu12, aichuan, lixinghong24}@nudt.edu.cn,

[*]Corresponding author {chgao96, chenying24, liyong07}@mails.tsinghua.edu.cn,

heln3@mail.sysu.edu.cn, kaishenglai@jnu.edu.cn, xin.lu.lab@outlook.com, yin_quanjun@163.com


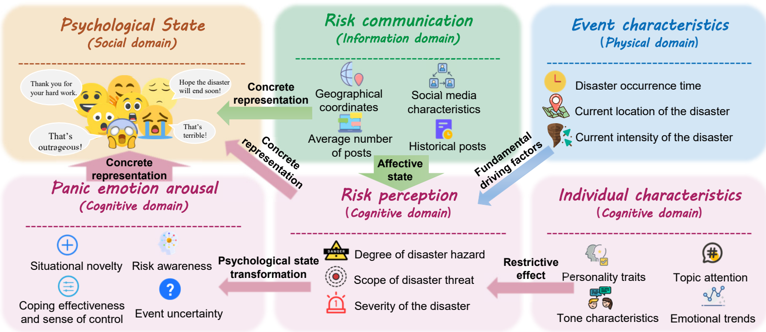


Figure 1: Path of panic formation. The physical domain supplies an objective benchmark, while the information domain dynamically adjusts cognitive inputs. The cognitive domain processes these inputs through psychological mechanisms including individual constraints, **risk perception**, and **emotional arousal theories**, triggering panic emotions that ultimately materialize as observable social media behaviors through the social domain.


## Abstract

Accurately predicting public panic sentiment on social media is crucial for proactive governance and crisis management. Current efforts on this problem face three main challenges: lack of finely annotated data hinders emotion prediction studies, unmodeled risk perception causes prediction inaccuracies, and insufficient interpretability of panic formation mechanisms limits mechanistic insight. We address these issues by proposing a **Psycho**logy-driven generative **Agent** framework (PsychoAgent) for explainable panic prediction based on emotion arousal theory. Specifically, we first construct a fine-grained panic emotion dataset (namely COPE) via human-AI (Large Language Models, LLMs) collaboration, combining scalable LLM-based labeling with human annotators to ensure accuracy for panic emotion and to mitigate biases from linguistic variations. Then, we construct PsychoAgent integrating cross-domain heterogeneous data grounded in psychological mechanisms to model risk perception and cognitive differences in emotion generation. To enhance interpretability, we design an LLM-based role-playing agent that simulates *individual psychological chains* through dedicatedly designed prompts. Experimental results on our annotated dataset show that PsychoAgent improves panic emotion prediction performance by 13%[1] to 21% compared to baseline models. Furthermore, the explainability and generalization of our approach is validated. Crucially, this represents a paradigm shift from opaque "data-driven fitting" to transparent "role-based simulation with mechanistic interpretation" for panic emotion prediction during emergencies. Our implementation is publicly available at: https://github.com/supersonic0919/PsychoAgent.


## 1 Introduction

Social media accelerates the spread of extreme emotions during sudden events (Organization et al., 2020), yet its real-time nature, rich information content, and spatiotemporal granularity offer unique opportunities for emotion prediction and management. Among various negative emotions, uncontrolled panic can exhaust emergency resources, trigger secondary disasters, even escalate isolated hazards into complex societal crises (Rune and Keech, 2023). For instance, during the 2011 Japan earth-

[1]All percentage symbols (%) in this paper denote percentage points (pp.) unless otherwise specified.




*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17116–17134

quake, panic-driven mass salt-buying in China exacerbated price surges and supply shortages (Pierson, 2011). Hence, timely prediction and monitoring of panic enable proactive interventions to mitigate its spread and impact (Wu et al., 2025).

Current research focuses on coarse-grained retrospective emotion classification using single-modality textual data (Mitrović et al., 2024; Weigang et al., 2024; Mitrovic and Kanjirangat, 2022), while dedicated studies on panic emotion prediction remain scarce. More details about emotion prediction research are in Appendix A.1. Due to the limitations of dataset annotation quality and inherent methodological constraints, precisely predicting public panic remains significant challenges.

Firstly, **current panic emotion annotation methods suffer from semantic deviations.** Existing datasets rely on static annotations and lack normal-state behavioral data, impeding predictive modeling of panic evolution and causing: (1) oversimplified categorization of panic as fear (Yang et al., 2022), ignoring disaster-specific risk perception; (2) semantic drift in keyword-based detection due to variations in linguistic nuances (Mitrović et al., 2024; Weigang et al., 2024), weakening cognition-semantics alignment.

Secondly, **current prediction models suffer from cross-domain feature fusion and dynamic modeling limitations.** Psychological studies indicate that public emotions are influenced by multi-domain features. While current models attempt to incorporate such features (Regan et al., 2024), they remain limited to simplistic combinations of text with single factors. Moreover, existing models often use static temporal modeling, fail to capture evolving feature interactions, significantly reduce predictive efficacy (Mitrović et al., 2024).

Lastly, **current studies suffer from explainability gap in panic formation mechanism.** Current studies focus on surface emotion detection and statistical correlations (Mitrović et al., 2024), neglecting psychological drivers like cognitive appraisal mechanisms (Hariharan et al., 2017). Though cognitive-emotion interactions are validated in psychological studies, existing works fail to computationally map the risk perception-to-arousal chain (Regan et al., 2024; Houlihan et al., 2023), leaving emotional evolution pathways unexplained.

To address the above-mentioned challenges, we propose a **Psycho**logy-driven generative **Agent** framework (PsychoAgent) for panic emotion prediction during sudden disasters. Specifically, we first adopt a human-LLM collaborative annotation method to construct a fine-grained panic emotion dataset (**C**ollaborative Fine-grained **O**pen **P**anic **E**motions Dataset, COPE), resolving semantic deviation and cross-cultural annotation biases. Secondly, we develop a mental modeling approach to fuse multi-domain features through the psychological mechanisms of panic formation (as illustrated in Figure 1). Finally, guided by theoretical lenses of risk perception and emotion arousal, we design a chain-of-thought (CoT)-driven LLM-based agent to simulate the full psychological chain of "disaster perception, risk cognition, panic emotion arousal, and posting behavior response". Our work shows significant implications for developing explainable AI systems capable of providing deeper insights into collective emotional dynamics during sudden disaster events.

Our contributions are summarized as follows:

- To support panic prediction, we pioneer a novel dual-phase panic emotion annotation dataset. We also fine-tune a discriminator to detect panic signals in generated texts.
- We innovatively propose a psychology-driven, multi-domain fusion mental model for public panic prediction during disasters. Our approach integrates multi-domain data via an LLM-based role-playing agent for interpretable psychological modeling of panic formation mechanisms.
- Experimental results show that the proposed framework achieves an accuracy of $\mathbf{86\%}$ ($\uparrow$ at least $\mathbf{13\%}$ versus SOTA) in panic emotion prediction task. Moreover, scalability studies and case studies further validate the explainability and generalization of our approach.

## 2 Related Work

We discuss the related works from three aspects: panic arousal theory, current panic analysis methods, and LLM-based role-playing approaches.

**Panic Emotion Arousal Theory.** Psychological studies establish panic as an uncertainty-driven fear linked to risk perception (Weigang et al., 2024), shaped by event-individual trait interactions (Davis et al., 2020; Syrdal and Briggs, 2018). Traditional methods (e.g., surveys) suffer from subjective bias and sampling limitations, failing to track real-time risk cognition in social media contexts. Our work bridges this gap by integrating psychological arousal theory with multi-domain data modeling, advancing computational panic analysis through mechanism-driven frameworks.

**Panic Emotion Analysis.** Current panic emotion research relies on event-driven static labeling and keyword filtering (Mitrović et al., 2024; Weigang et al., 2024), focusing on post-disaster data while lacking baseline behavioral data for prediction. Existing models, constrained to single-domain text analysis (Mitrović et al., 2024; Hariharan et al., 2017), exhibit limited interpretability due to insufficient multi-domain feature integration and inability to deconstruct panic formation pathways. We address these gaps via a dual-phase dataset and psychology-anchored LLM agent framework for interpretable panic pathway modeling.

**LLM-Based Personalized Role-Playing.** Personalized role simulation typically employs parameterized fine-tuning or prompt engineering. Prior work employs data-intensive training for fictional/historical role replication (Li et al., 2023; Wang et al., 2023; Ran et al., 2024), yet struggles with social media's data sparsity and privacy constraints (Abbasiantaeb et al., 2024). Nonparametric prompting methods (Park et al., 2024; Jiang et al., 2023) reduce data dependency but produce semantically shallow responses lacking psychological depth. We propose a CoT-driven LLM approach to simulate panic response logic via risk-emotion causal chains, enabling lightweight, mechanism-driven role modeling for social media.

## 3 Problem Definition and Dataset Construction

In this section, we first formalize the problem definition, then detail the methods for building a fine-grained panic emotion dataset and finally fine-tune a BERT-based model for panic recognition.

### 3.1 Problem Definition

The sentiment prediction task can be defined as:

$$\sigma_{t+1}^{i} = F\left(I_t^i\right), \tag{1}$$

where $\sigma_{t+1}^{i}$ is the sentiment label of user $i$ at future time $t+1$, and $I_t^i$ is post contents that user $i$ publishes on social media at time $t$.

This study aims to predict users' panic emotion labels after disaster occurrence through deep fusion of multi-domain features from their pre-disaster social media posts, integrated with psychological theories. Formally, this task can be formulated as:

$$\sigma_{t+1}^{i} = F\left(I_t^i, P_{t+1}, C_t^i\right), \tag{2}$$

where $t+1$ is the time after the sudden disaster, $t$ is the time before the disaster, $\sigma_{t+1}^{i}$ is the post-disaster sentiment label of user $i$, $P_{t+1}$ is the post-disaster physical features, and $C_t^i$ is the personal traits of user $i$ from pre-disaster analysis.

### 3.2 COPE Dataset Benchmark

To address the gap in high-quality panic emotion resources, we develop the first fine-grained dataset COPE, spanning pre- and post-disaster phases, which fills the gap in standardized panic emotion datasets and provides a foundation for advancing research in emotion prediction and related fields. Its core innovation is the standardized human-LLM collaborative annotation process, structured as follows. For further details regarding the dataset construction process, specific procedures, and comprehensive data, please refer to Appendix A.2.

- **Multi-source data collection:** Focusing on Hurricane Sandy (2012), we used 52.25 million tweets from 13.75 million users (Oct. 15-Nov. 12) (Kryvasheyeu et al., 2015). After preprocessing, the dataset includes 9,065 users, yielding a final experimental corpus of 1,384,989 high-quality posts, including user IDs, follower/followee counts, timestamps, locations, and ternary emotion labels.
- **Panic annotation framework:** To label post-disaster texts, we employ a hybrid LLM–human annotation pipeline: (1) An LLM produces initial labels via semantic parsing using contextual and linguistic cues to reduce lexical simplification bias; (2) Human reviewers then refine the labels through a psychology-guided validation process, annotating 1,065 posts over three rounds with fine-grained rules to mitigate semantic and cross-cultural biases; (3) A fine-tuned BERT model, trained on the curated labels, automates the remaining annotations while maintaining bias-aware generalization. This model also serves as a reliable baseline for subsequent text sentiment classification tasks.
- **User-level panic identification:** We identify panicked users via a "one-veto" rule ($\geq$1 panic text).

## 4 Psychology-Driven LLM Agent Framework

We first delineate the architecture of PsychoAgent. Then we describe the methodology for individual feature extraction. Finally, we introduce the psychology-driven LLM-based agent to simulate posting behavior for predicting panic emotions.

### 4.1 Overview of PsychoAgent

Based on the studies of risk perception and emotion arousal in psychology, we summarize the panic emotion formation pathway as follows (Figure 1):

- In the physical domain, we transform spatiotemporal disaster intensity into interpretable data through risk communication;
- In the information domain, we amplify disaster severity through communication channels, modulating risk perception and refining physical data interpretation, establishing via physical-cognitive bidirectional feedback loops;
- In the cognitive domain, we utilize individual traits to mediate risk interpretation biases, triggering panic through emotional arousal to achieve the transformation from data to emotion;
- In the social domain, we materialize individual panic into collective psychological states through social media behavior patterns.

Based on this, we propose PsychoAgent (Figure 2), a multi-domain fusion-driven LLM agent framework for panic emotion prediction, which comprises four main parts: **(1) In the information domain**, we focus on collecting social media texts and risk communication channel features; **(2) In the physical domain**, we extract disaster characteristics via geospatial platforms, gathering key indicators to convert physical features into disaster information; **(3) In the cognitive domain**, we extract individual traits from pre-disaster social media posts. Then, we construct user profiles to drive the LLM agents, aiming at simulating "disaster perception, risk perception, panic arousal, and posting behavior" psychological chain; **(4) In the social domain**, we use the fine-tuned BERT model to verify the generated text. Then, we enforce a "one-vote" veto rule for user-level prediction. Subsequent sections detail cognitive domain implementations.

### 4.2 Psychological and Behavioral Features Extraction

Psychological studies establish that personality traits, sentiment dynamics, topical concerns, and linguistic patterns critically shape risk perception and emotional responses (Gross and John, 2003; Brosch et al., 2013; Syrdal and Briggs, 2018; Davis et al., 2020). We accordingly extract four main features from users' pre-disaster posts.

**Big Five Personality Traits** We employ the publicly available model bert-base-personality (Devlin et al., 2019) to analyze the user's personality traits. This process can be formally expressed as:

$$\rho^i = \psi_{\text{personality}}(I_t^i), \tag{3}$$

where $\rho^i$ is the personality vector of the $i$-th user, and $\psi_{\text{personality}}$ is the model. More details are provided in Appendix A.3.1.

**Sentiment Trend Features** We use the BERTweet model (Nguyen et al., 2020) to perform ternary sentiment analysis on pre-disaster posts, aggregating results to generate an emotional trend profile. This is formalized as:

$$\gamma^i = \psi_{\text{sentiment}}(I_t^i), \tag{4}$$

where $\gamma^i$ is the sentiment trend vector of the $i$-th user; $\psi_{\text{sentiment}}$ is the sentiment model.

**Topic-Aware Characteristic Features** We employ Latent Dirichlet Allocation (LDA) and LLM-guided merging to extract topical features from user historical posts. The process is expressed as:

$$\tau^i = \Gamma \cdot \Theta^i, \tag{5}$$

where $\Theta^i$ is the LDA-clustered topic vector of the $i$-th user; $\Gamma$ is the membership relationships; $\tau^i$ is the consolidated thematic focus vector summarized by LLM. Appendix A.3.2 offers additional details.

**Linguistic Features** We design prompts to capture users' tone features via LLM. The process is formalized as:

$$v^i = \text{LLM}_{\text{language}}(I_t^i, p_v), \tag{6}$$

where $v^i$ denotes the $i$-th user's tonal features, extracted by $\text{LLM}_{\text{language}}$ using prompt $p_\nu$. The detailed prompts employed for LLM-based tonal feature extraction can be found in Appendix A.3.3.

### 4.3 Building LLM Agent with Psychological Theories

We design a CoT-driven LLM-based agent to simulate users' perception and responsiveness through four stages, with key stages centering on risk perception and emotional arousal (see in Figure 3):

- Disaster perception stage (physical-information domain): We integrate psychological knowledge, disaster features, and user profiles to form agent's long-term memory;

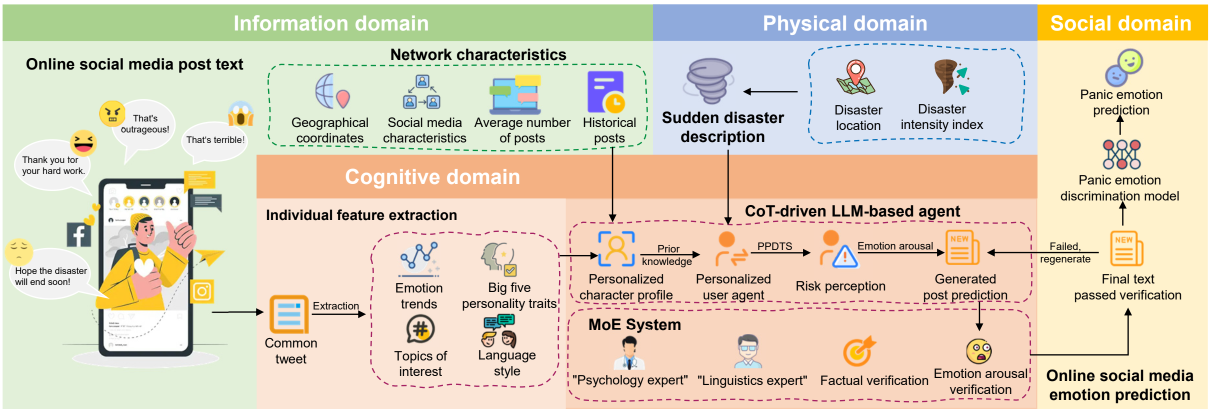


Figure 2: Overview of the proposed panic emotion prediction framework——PsychoAgent. It comprises four modules: (1) Information domain: collecting social media texts and risk communication features; (2) Physical domain: extracting geospatial disaster indicators; (3) Cognitive domain: simulating user perception and psychological chain with LLM agents; (4) Social domain: verifying generated text and performing user-level prediction.

- Risk perception stage (cognitive domain): We prompt the agent to engage with the psychological preparedness for disaster threats scale (PPDTS) (McLennan et al., 2020), which is a two-factor assessment tool designed to measure public knowledge awareness and anticipatory management capabilities for disasters;
- Panic arousal stage (cognitive domain): We prompt the agent to quantify panic probability through multi-dimensional analysis;
- Posting response stage (social domain): We generate tweets with Mixture of Experts (MoE) based consistency verification mechanism.

**Disaster Perception Stage** This stage equips the agent with essential prior knowledge by integrating psychological knowledge, disaster data, and user features into its long-term memory via structured prompts. The process is expressed as:

$$\mathrm{LLM}_U^i = \mathrm{LLM}_{\mathrm{RP}}(K, D, U^i, p_u), \tag{7}$$

where $\mathrm{LLM}_U^i$ denotes the user feature learning agent for user $i$; $\mathrm{LLM}_{\mathrm{RP}}$ is the role-playing LLM-based agent; $K$ is psychological knowledge constraints; $D$ is disaster data; $U^i$ is features of user $i$, encompassing both static traits and risk communication patterns; $p_u$ is the initialization prompt. Detailed implementation is in Appendix A.4.1.

**Risk Perception Stage** We construct a dynamic risk perception framework using disaster perception stage's memory and PPDTS (see Appendix A.4.2), emulating users' risk perception processes via structured psychological assessment. By leveraging LLM's reasoning capabilities with a human cognitive-inspired stepwise CoT prompting mechanism: question comprehension, memory retrieval, option mapping, and summary judgment, we prompt the agent to output risk perception scores (1-4 levels) for each question, formalized as:

$$S^i = \mathrm{LLM}_U^i(Q, p_{\mathrm{perception}}), \tag{8}$$

where $S^i$ is the PPDTS evaluation scores for user $i$; $Q$ is the question set comprising the PPDTS; $p_{\mathrm{perception}}$ is the PPDTS assessment prompt.

**Panic Arousal Stage** To bridge risk perception and emotion arousal, we reparameterize the discrete PPDTS scores into panic propensity drivers. Specifically, we prompt the agent to analyze four core factors (Risk awareness, coping effectiveness, event uncertainty, and situational novelty) through a cognitive chain reasoning mechanism, and infer the panic propensity probability based on prior outputs. This process can be represented as:

$$C^i = \mathrm{LLM}_U^i(S^i, p_c), \tag{9}$$

$$P_{\mathrm{panic}}^i = \mathrm{LLM}_U^i(S^i, C^i, p_{\mathrm{panic}}), \tag{10}$$

where $C^i$ is the summary of four factors for user $i$; $P_{\mathrm{panic}}^i$ is the panic propensity probability; $P_C$ is the prompt directing the agent in summarizing the four factors; $p_{\mathrm{panic}}$ is the prompt guiding the agent to evaluate user's panic probability.

**Posting Response Stage** In this stage, we prompt the agent to integrate knowledge across stages and predict the post content the user is most likely to

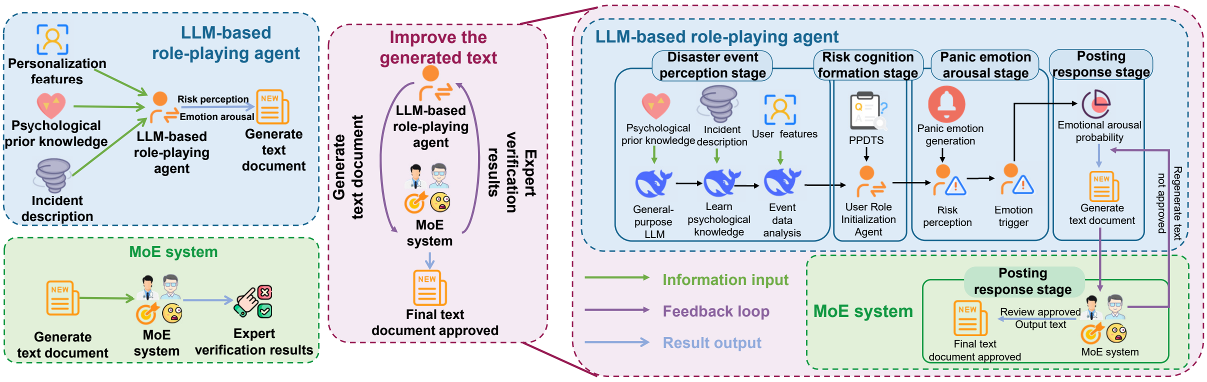


Figure 3: Detailed stages for building LLM-based agent driven by CoT in the cognitive domain module. Left: framework hierarchy with the LLM-based agent (blue), MoE system (green), and their iterative interaction (red). Right: expanded view of the interaction module (red) showing internal structures and feedback loops. Colors consistently represent modules across panels.

generate under the disaster event. This process can be represented as:

$$T^i = \text{LLM}_U^i(S^i, C^i, P^i_{\text{panic}}, p_T), \tag{11}$$

where $T^i$ is the set of tweets most likely published by the $i$-th user; $P_T$ is the prompt guiding the agent to generate predictive texts.

To mitigate subjective bias and randomness risks in evaluation, we introduce a MoE system with psychological, linguistic, factual consistency, and emotional alignment models, which assesses the generated text's multi-dimensional consistency. The details are provided in Appendix A.4.3. This process can be represented as:

$$\varepsilon^i_k = \text{LLM}^K(U^i, D, P^i_{\text{panic}}, T^i, p_k), \tag{12}$$

where $\varepsilon^i_k \in 0, 1$ is the evaluation result of the $k$-th expert for the $i$-th user's generated text; $\text{LLM}^K$ is the $k$-th expert model induced via prompt $p_k$.

The final evaluation result is determined by all expert models. The generated text $T^i$ is only validated if all expert models give a positive evaluation. Otherwise, the system adjusts and retries based on the experts' feedback.

Moreover, we specifically evaluate the role-playing accuracy of the LLM-Agent from the perspective of Big Five personality consistency. Detailed results and analysis are provided in Appendix A.4.4. Further details of prompts and case demonstrations are in Appendix A.4.5.

# 5 Experiment

In this section, we conduct extensive experiments on COPE to validate the proposed framework, including performance comparison, ablation study, scalability study and case study, with more experimental details provided in Appendix A.5.

## 5.1 Experimental Setup

**Evaluation Metrics** To address the class imbalance in panic detection, we adopt five macro-averaged metrics (accuracy, AUC, precision, recall, F1-score) (Sokolova and Lapalme, 2009), which compute scores per class and average them to mitigate majority-class bias.

**Implementation Details** The experiments are driven by the DeepSeek-v3. During risk perception and emotion arousal analysis, the temperature is set to 0.4 for stable psychological reasoning; In tweet generation, it rise to 0.7 (with a repetition penalty of 0.4) for linguistic diversity; In text verification, the temperature reset to 0.4 for rigorous checks, with 3 adaptive retries. For baselines, we adapt models to the psychology prediction task while adhering to original technical specifications. All models use the same input features, data split (8:2), and evaluation metrics to ensure comparability.

**Baselines** Although research on panic emotion prediction remains exploratory, affective computing offers a robust methodological framework with extensive prior work. To validate our framework's effectiveness, we choose three mainstream types of emotion analysis and prediction methods. Details of these baselines can be found in Appendix A.5.

- Feature-engineered ML (machine learning) models: We select classic ML methods like SVM (Cortes and Vapnik, 1995), Logistic Regression (Hosmer Jr et al., 2013), Random Forest (Breiman, 2001), and XGBoost (Chen and

| Category | Method | | Accuracy | AUC | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|---|---|---|
| **Traditional ML Methods** | SVM | Panic | - | - | 0.56 | 0.70 | 0.62 | 630 |
| | | No Panic | - | - | 0.82 | 0.70 | 0.76 | 1183 |
| | | Average | 0.70 | 0.76 | 0.69 | 0.70 | 0.69 | 1813 |
| | Logistic Regression | Panic | - | - | 0.55 | 0.73 | 0.63 | 630 |
| | | No Panic | - | - | 0.83 | 0.68 | 0.74 | 1183 |
| | | Average | 0.70 | 0.75 | 0.69 | 0.70 | 0.69 | 1813 |
| | Random Forest | Panic | - | - | 0.61 | 0.62 | 0.62 | 630 |
| | | No Panic | - | - | 0.80 | 0.79 | 0.79 | 1183 |
| | | Average | <u>0.73</u> | <u>0.80</u> | <u>0.70</u> | 0.71 | <u>0.70</u> | 1813 |
| | XGBoost | Panic | - | - | 0.60 | 0.61 | 0.61 | 630 |
| | | No Panic | - | - | 0.79 | 0.78 | 0.79 | 1183 |
| | | Average | <u>0.73</u> | 0.79 | <u>0.70</u> | 0.70 | <u>0.70</u> | 1813 |
| **Deep Learning Methods** | Bi-LSTM | Panic | - | - | 0.56 | 0.60 | 0.58 | 648 |
| | | No Panic | - | - | 0.77 | 0.73 | 0.75 | 1165 |
| | | Average | 0.69 | 0.74 | 0.66 | 0.67 | 0.67 | 1813 |
| | Transformer | Panic | - | - | 0.58 | 0.73 | 0.65 | 648 |
| | | No Panic | - | - | 0.82 | 0.71 | 0.76 | 1165 |
| | | Average | 0.71 | 0.71 | <u>0.70</u> | <u>0.72</u> | <u>0.70</u> | 1813 |
| | PredNet | Panic | - | - | 0.56 | 0.59 | 0.57 | 648 |
| | | No Panic | - | - | 0.76 | 0.74 | 0.75 | 1165 |
| | | Average | 0.69 | 0.68 | 0.66 | 0.67 | 0.66 | 1813 |
| **LLM-based Methods** | Deepseek-v3 | Panic | - | - | 0.50 | 0.00 | 0.01 | 562 |
| | | No Panic | - | - | 0.65 | 1.00 | 0.79 | 1065 |
| | | Average | 0.65 | 0.50 | 0.60 | 0.65 | 0.52 | 1627 |
| | ChatGPT-4o | Panic | - | - | 0.67 | 0.01 | 0.02 | 630 |
| | | No Panic | - | - | 0.65 | 1.00 | 0.79 | 1183 |
| | | Average | 0.65 | 0.50 | 0.66 | 0.50 | 0.40 | 1813 |
| | THOR-ISA | Panic | - | - | 0.00 | 0.00 | 0.00 | 630 |
| | | No Panic | - | - | 0.65 | 1.00 | 0.79 | 1183 |
| | | Average | 0.65 | 0.50 | 0.33 | 0.50 | 0.39 | 1813 |
| **Our Method** | PsychoAgent | Panic | - | - | 0.74 | 0.90 | 0.81 | 581 |
| | | No Panic | - | - | 0.94 | 0.84 | 0.89 | 1130 |
| | | Average | **0.86** | **0.87** | **0.84** | **0.87** | **0.85** | 1711 |

**Note:** (a) The Deepseek-v3 method ultimately predicts only 1,627 users due to sensitive content in user historical texts; (b) PsychoAgent successfully predicts 1,711 users after Stage 4 questionnaire validity screening, excluding cases with invalid questionnaires containing fewer than 18 responses (invalidity criterion: returned questions $< 18$).

Table 1: The panic emotion prediction performance of PsychoAgent and the baselines, where the best performance is shown in **bold** and the second best is <u>underlined</u>.

Guestrin, 2016) as baselines. For each, we integrate user features including big five personality traits, sentiment trend features, linguistic features, and topic-aware characteristic features.

- End-to-end DL (deep learning) models: We select classic DL models like Bi-LSTM (Graves and Schmidhuber, 2005), Transformer (Vaswani et al., 2017), and PredNet (Lotter et al., 2016) as baselines, with inputs similar to ML models.
- Advanced pre-trained language models: We use few-shot prompting on DeepSeek-v3 (Liu et al., 2024) and ChatGPT-4o (Achiam et al., 2023) to directly generate panic prediction results via few-shot prompting. We also evaluate the THOR-ISA (Fei et al., 2023) three-step prompting framework for panic prediction and include it as a baseline.

### 5.2 Performance Comparison with Baselines

As shown in Table 1, our approach significantly outperforms the SOTA baselines across all evaluation metrics, demonstrating the following strengths:

**Psychological theory-driven framework design:** The PsychoAgent framework achieves an absolutely $\mathbf{7\%}$ AUC improvement (0.87 vs 0.80 for Random Forest). Its core advantage lies in the task-specific design for emotion prediction, which integrates psychological priors to model cognitive-emotional chains. Unlike static feature-based ML models, PsychoAgent dynamically simulates user psychological states, enhancing minority-class detection under class imbalance.

**Temporal dynamic modeling of multi-domain fusion:** The PsychoAgent framework demonstrates superior temporal dynamic modeling capabilities, surpasses Bi-LSTM (0.74), Transformer (0.71), and PredNet (0.68) by $\mathbf{13\%-19\%}$ AUC. The PsychoAgent overcoming Bi-LSTM's fixed time window limitations for abrupt psychological shifts and PredNet's static feature constraints which can't integrate physical and cognitive features effectively. Its phased memory-augmented architecture enables precise capture of both disaster dynamics and abrupt psychological mutations, validating robust multi-domain dynamic modeling.

**Limitations of LLMs and our strategy:** While LLMs show strong semantic understanding, their direct classification suffers from bias toward negative classes (precision: 0.66, recall: 0.50), due to misalignment between generative probabilities and

| Method | | Acc. | AUC | Prec. | Recall | F1 | supp. |
|---|---|---|---|---|---|---|---|
| **Full** | Panic | – | – | 0.74 | 0.90 | 0.81 | 581 |
| | No Panic | – | – | 0.94 | 0.84 | 0.89 | 1130 |
| | Avg. | **0.86** | **0.87** | **0.84** | **0.87** | **0.85** | 1711 |
| w/o RS, EA, MEA | Panic | – | – | 0.77 | 0.03 | 0.05 | 630 |
| | No Panic | – | – | 0.66 | 1 | 0.79 | 1183 |
| | Avg. | 0.66 | 0.51 | 0.72 | 0.51 | 0.42 | 1813 |
| w/o EA, MEA | Panic | – | – | 0.61 | 0.06 | 0.10 | 595 |
| | No Panic | – | – | 0.66 | 0.98 | 0.79 | 1105 |
| | Avg. | 0.66 | 0.52 | 0.63 | 0.52 | 0.44 | 1700 |
| w/o MEA | Panic | – | – | 0.56 | 0.76 | 0.64 | 605 |
| | No Panic | – | – | 0.85 | 0.69 | 0.76 | 1163 |
| | Avg. | 0.71 | 0.72 | 0.70 | 0.72 | 0.70 | 1768 |

Table 2: Ablation study, where the best performance is shown in **bold**. **RS** denotes *Risk Sensing*, **EA** denotes *Emotion Arousal*, and **MEA** denotes *Multi-Expert Assessment*.

hard decision boundaries. PsychoAgent addresses this by decoupling LLM-based semantic parsing and risk simulation from direct label generation, achieving balanced precision (0.84, +**18**%) and recall (0.87, +**37**%), demonstrating the efficacy of indirect LLM use.

Overall, these results demonstrate the superior multi-domain dynamic modeling and minority class recognition capabilities of PsychoAgent.

### 5.3 Ablation Studies

In this section, we conduct ablation studies to analyze the contributions of key components in the PsychoAgent framework. By progressively removing key components: risk sensing (RS), emotion arousal (EA), and multi-expert assessment (MEA), we evaluate their individual and synergistic impacts on performance. These three components form a three-stage reasoning chain within the psychological CoT. As shown in Table 2, the full framework excels across all metrics. Ablating any component notably reduces performance, underscoring their vital role in the overall effectiveness.

**The synergy of RS, EA, and MEA:** Removing all three stages (w/o RS,EA,MEA) leads to severe degradation (accuracy: 0.66, -**20**%; AUC: 0.51, -**36**%), This essentially severs the reasoning chain and reduces the model to a simple end-to-end predictor without explicit cognitive processes, indicating their synergistic effects are critical, which severely impairs the model's ability to capture users' psychological traits and behavioral patterns.

**The foundational role of RS:** Removing EA and MEA (w/o EA,MEA) yields similar declines (accuracy: 0.66, -**20**%; AUC: 0.52, -**35**%), underscoring RS's critical role in initializing context-aware risk perception for psychological state recognition and its importance as input to later stages.

| Model | Scale | Method | Class | Acc. | AUC | Prec. | Recall | F1 |
|---|---|---|---|---|---|---|---|---|
| Qwen 2.5-14B-Instruct | 14B | Direct Pred. | Panic | - | - | 0.00 | 0.00 | 0.00 |
| | | | No Panic | - | - | 0.65 | 0.98 | 0.78 |
| | | | Avg. | 0.64 | 0.49 | 0.32 | 0.49 | 0.39 |
| | | psycho-Agent | Panic | - | - | 0.95 | 0.28 | 0.43 |
| | | | No Panic | - | - | 0.74 | 0.99 | 0.84 |
| | | | Avg. | 0.76 | 0.64 | **0.84** | 0.64 | 0.64 |
| Qwen 2.5-32B-Instruct | 32B | Direct Pred. | Panic | - | - | 0.00 | 0.00 | 0.00 |
| | | | No Panic | - | - | 0.65 | 1.00 | 0.79 |
| | | | Avg. | 0.65 | 0.50 | 0.33 | 0.50 | 0.39 |
| | | psycho-Agent | Panic | - | - | 1.00 | 0.14 | 0.25 |
| | | | No Panic | - | - | 0.68 | 1.00 | 0.81 |
| | | | Avg. | 0.70 | 0.57 | **0.84** | 0.57 | 0.53 |
| Qwen 2.5-72B-Instruct | 72B | Direct Pred. | Panic | - | - | 0.00 | 0.00 | 0.00 |
| | | | No Panic | - | - | 0.65 | 1.00 | 0.79 |
| | | | Avg. | 0.65 | 0.50 | 0.33 | 0.50 | 0.39 |
| | | psycho-Agent | Panic | - | - | 0.86 | 0.31 | 0.45 |
| | | | No Panic | - | - | 0.81 | 0.98 | 0.89 |
| | | | Avg. | 0.81 | 0.64 | 0.83 | 0.65 | 0.67 |
| DeepSeek-V3 | 671B | Direct Pred. | Panic | - | - | 0.50 | 0.00 | 0.01 |
| | | | No Panic | - | - | 0.65 | 1.00 | 0.79 |
| | | | Avg. | 0.65 | 0.50 | 0.60 | 0.65 | 0.52 |
| | | psycho-Agent | Panic | - | - | 0.74 | 0.90 | 0.81 |
| | | | No Panic | - | - | 0.94 | 0.84 | 0.89 |
| | | | Avg. | **0.86** | **0.87** | **0.84** | **0.87** | **0.85** |

Table 3: Performance Comparison of Different Model Sizes, where the best performance is shown in **bold** and the second best is underlined.

**The optimization role of MEA:** Removing only MEA (w/o MEA) results in smaller declines(accuracy: 0.71, -**15**%; AUC: 0.72, -**15**%), suggesting its optimization role in refining predictions, although its absence alone has a smaller impact compared to the joint absence of RS and EA.

Overall, these results validate the effectiveness of the PsychoAgent design. RS, EA and MEA work exhibit synergistic interactions, enabling comprehensive psychological modeling.

### 5.4 Scalability Studies

In this section, we investigate the interplay between model scale and performance under resource constraints by benchmarking four model sizes (14B, 32B, 72B, 671B) with/without the PsychoAgent framework. The experimental results are shown in Table 3, and our key findings are as follows.

- Pure parameter scaling fails to address data imbalance. All direct-prediction models achieve near-zero F1 on the Panic class (max 0.01 F1 for 671B), with macro-F1 stagnating at 0.39 from 14B to 72B, indicating inherent distributional bias toward majority classes;
- PsychoAgent empowers small models to surpass native large counterparts. The 14B model achieves 0.76 accuracy (+**12**%) and 0.43 Panic F1 (from zero), outperforming non-framework 32B/72B/671B models, demonstrating smaller models can transcend parameter limitations via task-driven reasoning under this framework;
- Framework efficacy exhibits scale-dependent synergy. The 671B model achieves peak ac-

| User Personalized Profile | | |
|---|---|---|
| **User id** | 574718143 | |
| **Panic Emotion Label** | Non-Panic Users | |
| **Event Attitude** | **Neuroticism** | **Language Features** |
| negative, indifferent, resigned | 0.532422089 | casual, conversational, weather-focused |
| **Emotional Trend** | [1, 1, -1, 1, 1, 1, 1, -1, -1, 0, -1, 1, -1, 1, -1, -1, 1, 0, 0, 1, 1, 1] | |
| **Topics of Interest** | Natural Disasters and Weather, Society and News etc. | |

| Panic Emotion Arousal Results | | |
|---|---|---|
| **Influencing Factors** | **Score** | **Reason** |
| Awareness | 3/5 | User shows basic risk recognition with weather-focused interests but lacks detailed understanding of specific hurricane dangers like flooding or wind damage. |
| Coping | 3/5 | User's neuroticism suggests moderate confidence in handling stress, but emotional stability is fluctuating, indicating balanced but not strong crisis management. |
| Uncertainty | 3/5 | User has some understanding of risks due to weather awareness but lacks explicit knowledge of hurricane mechanisms or response strategies. |
| Novelty | 3/5 | User's weather-focused tone suggests limited prior exposure to hurricanes but not complete unfamiliarity. |
| **Probability of Panic Emotion Activation** | | **50%** |

| Generate predictive text and predict panic user results | |
|---|---|
| **Generated Prediction Text** | The wind is picking up and the rain is relentless. Staying cautious but hopeful. #WeatherUpdate #StaySafe |
| **The detection of Panic Emotion in prediction Text** | Non-Panic Emotion Text |
| **Is the User a Panic User?** | **The user is a non-panic user.** |

Figure 4: Non-panic User Prediction Case.

| User Personalized Profile | | |
|---|---|---|
| **User id** | 781886935 | |
| **Panic Emotion Label** | Panic Users | |
| **Event Attitude** | **Emotional Trend** | **Topics of Interest** |
| indifferent, unrelated, neutral | [0, 0, -1, 0, 0, 0, 1, 1, 1, 0, 0] | Technology and Innovation, Sports and Entertainment, Society and News |
| **Conscientiousness** | **Neuroticism** | **Openness** |
| 0.301453 | 0.54772021 | 0.517733 |

| Panic Emotion Arousal Results | | |
|---|---|---|
| **Influencing Factors** | **Score** | **Reason** |
| Awareness | 3/5 | User has basic risk recognition but lacks deep awareness of specific hurricane dangers like flooding or wind damage. |
| Coping | 2/5 | User's low conscientiousness and neuroticism above average suggest limited confidence and strategies for crisis management. |
| Uncertainty | 4/5 | User's openness below norm and fluctuating emotional stability indicate partial understanding but significant confusion about risks and responses. |
| Novelty | 3/5 | User has likely seen similar events before but lacks extensive prior hurricane experience. |
| **Probability of Panic Emotion Activation** | | **60%** |

| Generate predictive text and predict panic user results | |
|---|---|
| **Generated Prediction Text** | OMG the wind is HOWLING outside!! This hurricane is SCARY AF!!! Stay safe everyone #StormAlert #StaySafe #EmergencyPrep |
| **The detection of Panic Emotion in prediction Text** | Panic Emotion Text |
| **Is the User a Panic User?** | **The user is a panic user.** |

Figure 5: Panic User Prediction Case.

curacy (0.86, **+21**%), while the framework-enhanced 72B model (0.81 accuracy achieves **124.6**%[2] efficiency over the 671B baseline (0.65) at lower cost, proving mid-sized models offer cost-efficient alternatives. Notably, the 32B model underperforms 14B in Panic F1 (0.25 vs. 0.43), revealing framework benefits emerge only when the model size exceeds a certain threshold.

Overall, the results indicate that model size and performance lack a straightforward linear relationship. PsychoAgent enables compact models to outperform larger counterparts in resource-constrained scenarios through psychological modeling.

[2]Here % denotes relative percentage.

### 5.5 Case Studies

To show our framework's superior mechanistic interpretability over traditional data-fitting methods, we present two representative prediction cases misclassified by Random Forest (Figure4 and 5).

For a non-panic user, traditional methods misclassify via negative emotional fluctuations, while our framework identifies dominant cross-domain contexts (e.g., weather/daily dialogue). Additionally, a coping capacity score (3/5) and dynamic simulations reveal stabilizing risk cognition → self-regulation loops, validated by generated text ("Staying cautious but hopeful"). Whereas traditional models, lacking temporal psychological modeling, overfit transient local features.

For a panic user, surface-level neutrality masks latent vulnerability (neuroticism: 0.548; conscientiousness: 0.301), which indicating weak emotional stability and inadequate crisis coping. Dynamic simulation reveals high uncertainty (4/5) and a cognitive-emotional cascade from risk misperception to panic, evident in expressions like "SCARY AF". Traditional models, relying on static features, fail to capture such implicit mechanisms.

Overall, our method offers superior mechanistic interpretability by dynamically simulating users' psychological processes, providing interpretable criteria for panic detection.

## 6 Conclusions

We introduce PsychoAgent, a novel framework for predicting dynamic panic emotion in social media users during disasters, grounded in psychological emotion arousal theory. PsychoAgent uniquely integrates a human-LLM collaboratively annotated dataset for fine-grained emotion analysis, a psychology-driven feature fusion mechanism to model public panic dynamics, and LLM-based agent simulations to trace panic formation pathways, moving beyond traditional reliance on explicit features. Experiments demonstrate that PsychoAgent significantly outperforms existing baselines in both prediction accuracy and interpretability. Our findings have important implications for developing explainable AI systems that offer deeper insights into collective emotional dynamics during crises. Future work could expand this framework to investigate panic propagation dynamics across social networks and to develop real-time intervention strategies to mitigate emotional contagion during emergencies.

## Limitations

This work primarily focus on individual-level panic prediction by simulating psychological mechanisms. While this approach effectively models cognitive-emotional chains and yields explainable predictions, it is not without limitations.

First, LLM hallucinations may cause deviations from psychological priors in modeling risk perception and emotion arousal. Future work should integrate more robust self-correction mechanisms within the agents to mitigate these discrepancies.

Second, stylistic differences between LLM-generated panic texts and authentic user posts can create semantic divergence, potentially leading to false negatives in our self-trained BERT classifier. Enhancing the linguistic realism of generated text to better align with human expression is an important next step.

Furthermore, mainstream LLMs (e.g., Deepseek, GPT) often impose political correctness guardrails that suppress negative emotional expressions, potentially underdetecting panic users, suggesting our framework is better suited for models without such constraints.

## Ethics Statement

The datasets used in this work are all publicly available, so there is no ethical concern. Beyond data provenance, we also consider ethical implications of the PsychoAgent system itself.

**First, regarding model bias risks.** As noted in the Limitations section, hallucinations in the underlying large language models (LLMs) may cause deviations from psychological priors when modeling user risk perception and emotion arousal. Spurious correlations and demographic biases in emotion recognition systems can lead to erroneous assessments of user states—for instance, misclassifying non-panic emotional expressions as panic, or failing to recognize panic in underrepresented user groups due to limited diversity in training data. We acknowledge this risk and propose future integration of robust self-correction mechanisms (e.g., iterative alignment with psychological benchmarks) to mitigate such biases.

**Second, regarding potential misuse and applicability risks.** The Limitations section discusses how "political correctness guardrails" in mainstream LLMs suppress negative emotional expressions, leading to underdetection of panic users. While PsychoAgent is more suitable for models without such over-constraints to improve detection accuracy, we recognize the ethical risk of unregulated deployment: removing guardrails without clear boundaries could enable excessive sensitivity to negative emotions (e.g., over-flagging mild anxiety as panic, triggering unnecessary emergency responses) or misuse in contexts where panic detection is weaponized (e.g., unjustified surveillance during public crises). Conversely, retaining overly strict constraints risks critical underdetection of genuine panic, undermining the system's core purpose of supporting emergency management.

All ethical risks discussed above have been substantively addressed in the Limitations section. This statement consolidates these considerations to explicitly align with ethical guidelines for AI in sentiment analysis and emergency management, ensuring transparency about PsychoAgent's ethical boundaries and mitigation strategies.

## Acknowledgments

We would like to thank all team members for their invaluable contributions to this project. This work is supported by the National Natural Science Foundation of China (72501291, 72025405, 72421002, 92467302), and the Major Program of Xiangjiang Laboratory (24XJJCYJ01001). It is also partially supported by the Hunan Provincial Natural Science Foundation (2025JJ60477). We thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

## References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdullah Alsaedi, Stuart Thomason, Floriana Grasso, and Phillip Brooker. 2022. Transfer learning model for social emotion prediction using writers emotions in comments. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 396–400. IEEE.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Tobias Brosch, Klaus Scherer, Didier Grandjean, and David Sander. 2013. The impact of emotion on perception, attention, memory, and decision-making. *Swiss medical weekly*, 143(1920):w13786–w13786.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Lloyd S Davis, Bienvenido León, Michael J Bourk, and Wiebke Finkler. 2020. Transformation of the media landscape: Infotainment versus expository narrations for communicating science in online videos. *Public Understanding of Science*, 29(7):688–701.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Ge Gao, Jongin Kim, Sejin Paik, Ekaterina Novozhilova, Yi Liu, Sarah T Bonna, Margrit Betke, and Derry Tanti Wijaya. 2024. Enhancing emotion prediction in news headlines: Insights from chatgpt and seq2seq models for free-text generation. *arXiv preprint arXiv:2407.10091*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

James J Gross and Oliver P John. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348.

Anuja Hariharan, Verena Dorner, Christof Weinhardt, and Georg W. Alpers. 2017. Detecting panic potential in social media tweets. In *25th European Conference on Information Systems, ECIS 2017, Guimarães, Portugal, June 5-10, 2017*.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Sean Dae Houlihan, Max Kleiman-Weiner, Luke B Hewitt, Joshua B Tenenbaum, and Rebecca Saxe. 2023. Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A*, 381(2251):20220047.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Yury Kryvasheyeu, Haohui Chen, Esteban Moro, Pascal Van Hentenryck, and Manuel Cebrian. 2015. Performance of social network sensors during hurricane sandy. *PLoS one*, 10(2):e0117288.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Dayu Li, Yang Li, and Suge Wang. 2020. Interactive double states emotion cell model for textual dialogue emotion prediction. *Knowledge-Based Systems*, 189:105084.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

William Lotter, Gabriel Kreiman, and David Cox. 2016. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Robert R McCrae and Paul T Costa. 1992. Revised neo personality inventory (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual. *Odessa, FL: Psychological Assessment Resources*.

Jim McLennan, Mathew D Marques, and Danielle Every. 2020. Conceptualising and measuring psychological preparedness for disaster: The psychological preparedness for disaster threat scale. *Natural Hazards*, 101(1):297–307.

Sandra Mitrovic and Vani Kanjirangat. 2022. Enhancing bert performance with contextual valence shifters for panic detection in covid-19 tweets. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pages 89–92.

Sandra Mitrović, Oscar William Lithgow-Serrano, and Carlo Schillaci. 2024. Comparing panic and anxiety on a dataset collected from social media. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 153–165.

Xu Mou, Qinke Peng, Zhao Sun, Muhammad Fiaz Bashir, and Haozhou Li. 2024. Multi-document influence on readers: augmenting social emotion prediction by learning document interactions. *Neural Computing and Applications*, 36(12):6701–6719.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

World Health Organization and 1 others. 2020. Managing the covid-19 infodemic: promoting healthy behaviours and mitigating the harm from misinformation and disinformation. joint statement by who, un, unicef, undp. *UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC. September*, 23:2020.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

D Pierson. 2011. Japan radiation fears spark panic salt-buying in china. *Los Angeles Times March*, 18:2011.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Deqing Yang, and Yanghua Xiao. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. *arXiv preprint arXiv:2406.18921*.

Ciaran Regan, Nanami Iwahashi, Shogo Tanaka, and Mizuki Oka. 2024. Can generative agents predict emotion? *arXiv preprint arXiv:2402.04232*.

Karina T Rune and Jacob J Keech. 2023. Is it time to stock up? understanding panic buying during the covid-19 pandemic. *Australian Journal of Psychology*, 75(1):2180299.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models. *arXiv e-prints*, arXiv:2307.00184.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Holly A Syrdal and Elten Briggs. 2018. Engagement with social media content: A qualitative exploration. *Journal of marketing theory and practice*, 26(1-2):4–22.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Gong Weigang, Zhu Meng, and Chen Hao. 2024. The evolution mechanism of social emotions during the covid-19 pandemic: an analysis based on big data from twitter and gdelt. *The Journal of Chinese Sociology*, 11(1):25.

Wangqiang Wu, Ying Zhang, and Hongda Liu. 2025. Analyzing crowd emotional contagion in metro emergencies through the lens of the weber–fechner law: Predictions based on computational techniques applied to science. *Applied Sciences*, 15(3):1244.

Yixin Yang, Yingying Zhang, Xiaowan Zhang, Yihan Cao, and Jie Zhang. 2022. Spatial evolution patterns of public panic on chinese social networks amidst the covid-19 pandemic. *International Journal of Disaster Risk Reduction*, 70:102762.

Nutchanon Yongsatianchot, Tobias Thejll-Madsen, and Stacy Marsella. 2023. What's next in affective modeling? large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7. IEEE.

# A Appendix

## A.1 Comparison of Current Emotion Prediction Works

We summarize the research landscape of existing emotion prediction approaches and draw a comparison with our proposed framework by examining aspects including dataset construction, application scenarios, theoretical foundations, multi-domain feature fusion, generative model architectures, interpretability, and granularity of emotion prediction. The comparative results are presented in Table 4.

## A.2 Dataset Labeling and Self-trained BERT Model

Focusing on panic prediction during sudden disasters, our dataset centers on post-disaster texts with panic emotion annotations. We develop a hybrid annotation framework combining LLM-assisted labeling (1,065 entries), human crowdsourcing, Easy Data Augmentation (EDA)-based data augmentation, and BERT fine-tuning for automated tagging. Statistical details of the dataset is presented in Table 5. Specific implementation details are provided in subsequent sections.

### A.2.1 Data preprocessing procedures

Our human-LLM collaboratively annotated Hurricane Sandy dataset innovatively integrates panic emotion labels across pre-disaster and post-disaster phases. Before annotation, we preprocessed the collected dataset as follows:

1. Data Filtering: Removed malformed, highly redundant (cosine similarity $>0.85$), or invalid entries (posts with $<5$ meaningful tokens);
2. Irrelevant Content Removal: Leveraged human-LLM collaboration to exclude non-Sandy-related texts from the keyword-crawled public dataset(similar to Section 3.2);
3. User Selection: Retained users who posted in both phases, with $\geq 10$ pre-disaster tweets;
4. Text Sanitization: Stripped URLs, retweets ("RT @xxx:"), and non-alphanumeric characters;
5. Temporal Stratification: Split the dataset into pre-disaster and peri-disaster subsets based on timestamps.

### A.2.2 LLM-Assisted Annotation Prompts

The ChatGPT-4o assisted annotation prompts used in our dataset labeling process are shown as follows. Since LLMs are trained on extensive corpus data, they inherently contain general knowledge about Hurricane Sandy and panic emotions. Our approach simply required prompting the LLM to leverage its existing knowledge to determine whether input texts: (1) relate to Hurricane Sandy, and (2) exhibit user panic emotions.

**Prompt for LLM-assisted annotation (Hurricane Related)**

**Question prompt:** You are a helpful assistant that identifies text relevance to Hurricane Sandy using specific keywords and related terms. Evaluate the following text for relevance to Hurricane Sandy. Focus only on the text provided and check for keywords such as "Sandy", "fierce winds", "heavy rainfall", "flooding", "long power outages", "population centers", "target area", "economic loss", "hurricane", "storm", "superstorm", "tropical cyclone", and other disaster-related terms. If you find any relevant keywords, respond with "Yes" and provide an explanation. If not, respond with "No" and explain why it's not relevant. Input text is: Creepy clouds. Bad weather here in Chi...See the face?? Go back!!!!
**LLM answer:** Yes, the text is relevant to Hurricane Sandy. It contains the keyword "Hurricane" and specifically mentions "Sandy", which directly relates to Hurricane Sandy itself. Additionally, it provides information about the hurricane's location (26.3N 76.9W) and its movement (NW at 11 mph), which are pertinent details regarding the storm's status and trajectory. This aligns well with the context of discussing Hurricane Sandy and its impact.

**Prompt for LLM-assisted annotation (Panic Related)**

**Question prompt:** You are a helpful assistant that evaluates text for emotional content, particularly focusing on panic emotions. You will analyze the input text based on the context and linguistic cues to determine if it reflects panic emotions from the text's author. If panic emotions are present, respond with "Yes" and provide an explanation. If panic emotions are not present, respond with "No" and explain why. Input text is: Creepy clouds. Bad weather here in Chi...See the face?? Go back!!!!

**LLM answer:** Yes. The text reflects panic emotions through several linguistic cues. Phrases like "Creepy clouds" and "Bad weather" indicate a sense of unease and fear about the current situation. The exclamation "See the face??" suggests a heightened emotional response, ponssibly idicating a perceived threat or danger. The command "Go back!!!!" further emphasizes urgency and panic, as it implies a strong desire to escape or avoid a potentially harmful situation. Overall, the combination of unsettling descriptions and urgent commands conveys a clear sense of panic from the author.

### A.2.3 Human Crowdsourcing Verification Rules

Based on psychological studies of panic emotions and the specific context of Hurricane Sandy, we establish the crowdsourced validation rules for panic emotion annotation as shown in Figure 6. The rules operationalize panic emotion manifestations through five key dimensions.

To ensure annotation quality and reliability, we further report inter-annotator agreement (IAA) metrics from two independent rounds of human review. Detailed results are shown in Table 6.

The high simple agreement rate (0.9797) indicates strong consensus on the vast majority of sam-

| Ref. | Dataset | Scenario | Psychology | Features | Generative Architecture | Interpretability | Granularity |
|---|---|---|---|---|---|---|---|
| (Yongsatianchot et al., 2023) | / | Sports competition | / | Contextual | ChatGPT-4 | Partial (Output) | – |
| (Regan et al., 2024) | EmotionBench | Multi-scenario stories | Emotion assessment | Contextual | GPT-3.5-Turbo | Partial (Context) | Binary |
| (Mou et al., 2024) | / | News | / | Cross-doc | / | / | Ekman-6 |
| (Li et al., 2020) | / | Text dialogues | / | / | / | / | Ekman-6 /4-class |
| (Gao et al., 2024) | / | News | / | / | RoBERTa+ChatGPT | Partial (Explanation-Assisted) | 8-class |
| (Houlihan et al., 2023) | Game data | "Split or Steal" | Intuitive psychology | Preferences | / | Full-chain | 20-class |
| (Alsaedi et al., 2022) | / | Social media | / | / | / | / | 5-class |
| Ours | Disaster | Emergencies | Risk perception + Emotional arousal | Multi-domain fusion | LLM+MoE | Full-chain | Panic-specific |

Table 4: Comparative summary of emotion prediction research status. Dedicated dataset (whether creating task-specific data), Scenario (application in specific contexts), Psychology (integration of psychological theories), Features (multi-domain feature incorporation), Generative architecture (use of generative models), Interpretability (model transparency), and Granularity (emotion classification specificity). *Cells marked with a slash (/) indicate the absence of specified design components.*

| Category | Quantity |
|---|---|
| Panic User Count | 3,240 |
| Non-Panic User Count | 5,825 |
| **Total Users** | **9,065** |
| Pre-Disaster Tweet Count | 242,363 |
| Post-Disaster Tweet Count | 1,142,626 |
| **Total Tweets** | **1,384,989** |

Table 5: Details of the constructed dataset.

| Metric | Value |
|---|---|
| Simple Agreement Rate | 0.9797 |
| Cohen's kappa($k$) | 0.7059 |
| Disagreement Rate | 0.0203 |
| Agreement on Panic Samples | 0.6042 |
| Agreement on Non-Panic Samples | 0.9963 |

Table 6: Inter-Annotator Agreement (IAA) Metrics for Human Annotation

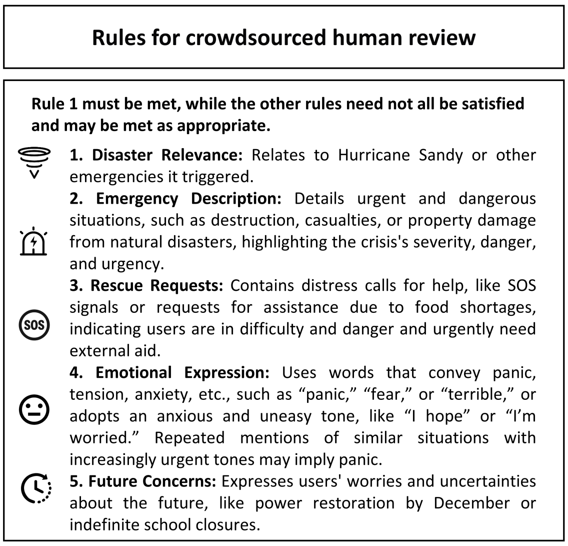


Figure 6: Crowdsourced validation protocol for panic emotion annotation.

ples, which is further supported by a low disagreement rate (0.0203). Cohen's Kappa ($k = 0.7059$), a stricter measure accounting for chance agreement, falls into the "Good" agreement range (>0.6) according to Landis and Koch (Landis and Koch, 1977), signifying substantial agreement between annotators. Notably, while agreement on non-panic samples is exceptionally high (0.9963), the lower agreement on panic samples (0.6042) aligns with the inherently greater difficulty and contextual ambiguity in annotating such emotional expressions. These results confirm the reliability of our annotated dataset as a benchmark for evaluation.

#### A.2.4 BERT Model Training Effect

We fine-tuned a BERT model on an augmented and balanced dataset for auto-labeling remaining data, which also provide a reliable benchmark tool for subsequent emotion discrimination in generated texts. To validate training efficacy, we tested the model on both the augmented dataset and the original human-verified dataset, with results (Table 7 and Table 8) demonstrating robust performance even on non-augmented data, confirming the effectiveness of our EDA enhanced training strategy.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| No Panic | 1.00 | 0.99 | 0.99 | 513 |
| Panic | 0.99 | 1.00 | 0.99 | 497 |
| Accuracy | | | 0.99 | 1010 |
| Macro Avg | 0.99 | 0.99 | 0.99 | 1010 |
| Weighted Avg | 0.99 | 0.99 | 0.99 | 1010 |

Table 7: Training Effect of BERT Model on Augmented Data.

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| No Panic | 1.00 | 0.99 | 1.00 | 1002 |
| Panic | 0.90 | 1.00 | 0.95 | 63 |
| Accuracy | | | 0.99 | 1065 |
| Macro Avg | 0.95 | 1.00 | 0.97 | 1065 |
| Weighted Avg | 0.99 | 0.99 | 0.99 | 1065 |

Table 8: Testing Effect of BERT Model on Original Imbalanced Data.

| **Floating** | **User Count/Proportion** | | **Total** |
|---|---|---|---|
| **range** | Consistent | Inconsistent | **Users** |
| 15% | 8880 (97.96%) | 185 (2.04%) | 9065 |
| 20% | 9032 (99.64%) | 33 (0.36%) | 9065 |

Table 9: Consistency Analysis of Bert-base-personality Personality Detection Model.

## A.3 Details of Individual Feature Extraction

### A.3.1 Extraction of the Big Five Personality Traits

The bert-base-personality model, fine-tuned on a curated personality dataset from the BERT-BASE-UNCASED backbone, predicts Big Five personality traits (Openness, Extraversion, Neuroticism, Agreeableness, Conscientiousness) via transfer learning, outputting a dictionary of normalized scores (0–1) for each dimension. While the model lacks disclosed accuracy/F1 metrics, we validated its consistency by testing it twice on 50% splits of users' pre-disaster texts. Aligned with widely accepted standards in personality psychology (McCrae and Costa, 1992; Serapio-García et al., 2023), we set acceptable fluctuation thresholds of 15% and 20% per personality dimension. Results (Table 9) show that $> 95\%$ intra-user trait alignment within acceptable variance thresholds, confirming its reliability for personality profiling.

### A.3.2 Focused Topic Feature Extraction

The LDA model, a generative Bayesian probabilistic algorithm with a three-layer structure (word, topic, corpus), includes two key parameters: the number of topics and keywords.

In this study, we configure the LDA model to cluster all users' pre-disaster posts into 25 distinct topics, with each topic characterized by 10 keywords. This process also identifies the most likely topic category for each post. This process is formalized as follows:

$$\theta^k,\ \phi_k = \psi_{\text{topic}}(I_t^i,\ k = 25,\ n = 10), \quad (13)$$

where $\theta^k \in \mathbb{R}^{25}$ is the 25 topics generated by LDA; $\phi_k$ is the keyword list for the $k$-th topic; $\psi_{\text{topic}}$ is the topic clustering model.

Subsequently, we prompt ChatGPT-4o to merge similar topics based on their summarized keywords, ultimately consolidating them into 8 common thematic categories (see Table 10 for more details). This process is formalized as:

$$\Gamma = LLM_{topic}(\{\phi_k\}_{k=1}^{25}, p_\Gamma), \quad (14)$$

where $p_\Gamma$ is the prompt used to instruct ChatGPT-4o for topic induction; $LLM_{topic}$ is the ChatGPT-4o-based topic consolidation model, which outputs 8 consolidated themes; $\Gamma$ is the membership relationships from the original 25 topics to the 8 consolidated themes.

Then, we aggregate each user's pre-disaster topical focuses based on the topic labels assigned to their historical posts prior to the disaster, which can be expressed as:

$$\tau^i = \Gamma \cdot \Theta^i, \quad (15)$$

where $\Theta^i$ is the topic vector of the $i$-th user on the LDA-clustered topics; $\Gamma$ is the membership relationships from the LDA-clustered topics to the LLM-consolidated themes; $\tau^i$ is the summarized thematic focus vector of the same user on the 8 consolidated themes, represented as a list of theme names.

| **Topic Category** | **Top Keywords** |
|---|---|
| Politics & Elections | debate, obama, presidential, governor, business |
| Natural Disasters & Weather | weather, wind, rain, hurricane, storm |
| Energy & Environment | power, solar, gas, energy, climate |
| Sports & Entertainment | york, giants, jets, nfl, yankees |
| Economy & Business | gas, prices, obama, economy, wall |
| Society & News | news, governor, business, china, friends |
| Technology & Innovation | power, solar, tech, play, technology |
| Miscellaneous | hurricane, tropical, newyork, storm, east |

Table 10: Social Media Users' Event Topics of Interest Before Disasters

### A.3.3 Tone Feature Extraction

Leveraging the capabilities of LLM in natural language processing (NLP) and text pattern recognition, we design a specialized prompt that extracts and condenses each user's linguistic tone features. Specifically, we instruct ChatGPT-4o to analyze the

**Prompt for Tone Feature Extraction**

**Question prompt:** You are a professional linguist expert who can identify the tone of voice in social media texts. A social media user has sent the following tweets: row["text"]. Describe this user's overall tone of voice on the social media with three words. Only output in the exact format: xxx, xxx, xxx."

**LLM answer:** Casual, Humorous, Restless

linguistic tone of each user's posts under normal conditions, returning three words to describe their tone features, which ensuring efficient and interpretable stylistic profiling. The prompt template for tone feature induction is detailed as follows.

## A.4 CoT-Driven LLM-Based Agent

### A.4.1 Psychological Knowledge Injection

(1) Psychological knowledge comprehension: we integrated a psychology knowledge system with six core modules, which covers public risk perception formation, personality traits and risk response (with a focus on the Big Five model), social media language style effects, content type emotional impacts, emotional stability mechanisms, social media network property roles, establishing the theoretical foundation and baseline cognitive judgment for psychology-driven prediction in subsequent phases (detailed in Table 11);

(2) Disaster event data analysis: Integrating real-time meteorological data to form a dynamic risk perception of disasters, analyzing the physical features and linking to the "risk event features" in psychology;

(3) User profile initialization: Using dual modeling for the agent's initialization. A basic profile combines user static traits (e.g., Big Five personality traits) with risk communication analysis (e.g., geographical location, network topology, and behavioral frequency). Additionally, the Contriever model retrieves the top 5 hurricane-related posts from users' pre-disaster posts as supplementary material, helping the agent infer latent behavioral tendencies for a more accurate individual simulation.

### A.4.2 Psychological Preparedness for Disaster Threats Scale (PPDTS)

The Perceived Preparedness for Disaster Threat Scale (PPDTS) exhibits a two-factor structure—Knowledge & Awareness (KA) and Anticipation, Awareness & Management (AAM)—with strong inter-component correlations and high internal consistency reliability for both the full scale and subscales. Its 18 scenario-based items align with specific theoretical constructs from psychological knowledge modules (see Table 12). For instance, the item "I am confident that I know what to do and what actions to take in a severe weather situation." activates the agent's long-term memory of the interplay between sense of control and user extraversion, enabling dynamic psychological computation.

### A.4.3 Tweet generation and multi-expert evaluation

To ensure generated tweets align with user personality traits while reflecting psychological dynamics in disaster events, we prompt the agent to perform the following reasoning steps in sequence, thus building a multi - source memory integration mechanism. Specifically:

(1) Extract psychological domain knowledge as constraints;

(2) Inject key situational parameters from real-time hurricane data;

(3) social network behavioral traits and personal traits from user profiles;

(4) Integrate risk perception features derived from Phase 2 assessments;

(5) Determine emotional tone via panic arousal probability values;

(6) Generate the top 3 most probable posts the user would publish.

To assess the generated text's consistency and validity across psychological compatibility, linguistic coherence, factual reliability, and emotional rationality, we use prompt engineering to coordinate joint evaluation by four domain-specific expert models:

(1) The psychology expert model: Verifies alignment between generated texts and user psychological profiles (particularly Big Five personality traits);

(2) The linguistic expert model: Analyzes linguistic style coherence with historical posts;

(3) The factual consistency model: Ensures disaster-event relevance and accuracy;

| Knowledge Category | Content |
|---|---|
| Public Risk Perception Formation | • Risk perception is shaped by two factors and their interaction:<br>a) Characteristics of the risk event itself<br>b) Personal characteristics of the audience |
| Personality Traits and Risk Response | • High Psychoticism: Associated with overestimation of event controllability<br>• High Extraversion: Correlates with perceived understanding of emergencies (e.g., pandemic knowledge)<br>• High Neuroticism: Linked to lower emergency comprehension and higher fear levels<br>• Extraverts tend to adopt proactive measures<br>• Emotionally unstable individuals (high Neuroticism) prefer passive coping strategies |
| Social Media Language Style Effects | • Sarcasm/irony may amplify anxiety in crisis contexts |
| Content Type Emotional Impacts | • Disaster-related serious news increases situational awareness but may elevate stress |
| Emotional Stability Mechanisms | • Regular use of cognitive reappraisal strategies buffers acute stress during disasters |
| Social Media Network Property Roles | • Users with more follows/followers are more likely to be exposed to diverse and potentially conflicting information, which can increase cognitive load and anxiety<br>• Dense social networks (many friends) can lead to group polarization and echo chamber effects, amplifying panic through frequent interactions<br>• Social comparison on platforms with many users can weaken self-efficacy when others display superior coping resources |

Table 11: Psychological Knowledge Embedded Content.

| **Knowledge & Awareness (KA) sub-scale** | |
|---|---|
| 1 | I am familiar with the natural hazard/disaster preparedness materials relevant to my area. |
| 2 | I know which household preparedness measures are needed to stay safe in a natural hazard/disaster. |
| 3 | I know how to adequately prepare my home for the forthcoming fire/flood/cyclone season. |
| 4 | I know what to look out for in my home and workplace if an emergency weather situation should develop. |
| 5 | I am familiar with the disaster warning system messages used for extreme weather events. |
| 6 | I am confident that I know what to do and what actions to take in a severe weather situation. |
| 7 | I would be able to locate the natural hazard/disaster preparedness materials in a warning situation easily. |
| 8 | I am knowledgeable about the impact that a natural hazard/disaster can have on my home. |
| 9 | I know what the difference is between a disaster warning and a disaster watch situation. |
| 10 | I am familiar with the weather signs of an approaching fire/flood/cyclone. |
| **KA construct reliability** | |
| **Anticipation, Awareness & Management (AAM) sub-scale** | |
| 1 | I think I am able to manage my feelings pretty well in difficult and challenging situations. |
| 2 | In a natural hazard/disaster situation I would be able to cope with my anxiety and fear. |
| 3 | I seem to be able to stay cool and calm in most difficult situations. |
| 4 | I feel reasonably confident in my own ability to deal with stressful situations that I might find myself in. |
| 5 | When necessary, I can talk myself through challenging situations. |
| 6 | If I found myself in a natural hazard/disaster situation I would know how to manage my own response to the situation. |
| 7 | I know which strategies I could use to calm myself in a natural hazard/disaster situation. |
| 8 | I have a good idea of how I would likely respond in an emergency situation. |

Table 12: Psychological Preparedness for Disaster Threats Scale (PPDTS).

(4) The emotional alignment model: Maintains dynamic alignment between textual emotional intensity (via lexical analysis) and user panic probability.

#### A.4.4 LLM-Agent Personality Consistency Assessment

To evaluate the role-playing accuracy of our LLM-Agents, we specifically assessed their consistency with real users' Big Five personality traits. For 200 randomly selected users, we compared the personality traits reflected in LLM-generated predictive texts (extracted via bert-base-personality model) against the ground truth traits derived from users' actual pre-disaster posts. The consistency evaluation results under different fluctuation thresholds are summarized in Table 13.

The results quantitatively validates the consistency (reaching 88% at ±20%) between LLM-Agent role-playing and real users' Big Five personalities. These results indicate that the LLM-Agents within our framework successfully maintained core Big Five personality trait consistency for the majority of users during role-playing.. Cases not meeting consistency may stem from text-based prediction limitations (e.g., heightened emotional volatility during disasters), while MoE experts integrate multi-dimensional context.

#### A.4.5 Prompts and Case Demonstrations in Different Phases

Our proposed prompting methods are as shown below. Using a user near Florida as an example, we employ CoT to build an agent with long-term

## Stage 1: Disaster Perception Stage

**Question prompt:** You are a psychologist specializing in predicting public emotional trends during emergencies. Use these resources:
1. Psychological Principles: psychology
2. Hurricane monitoring data (Markdown): hurricane table
3. User Profile (JSON): user info)

Please always: 1. Directly output the final answer;
2. Disable any thought process;
3. Use plain text format.
**LLM answer:** Data understood.

## Stage 2: Risk Perception Stage

**Question prompt:** Answer the following questions. For each question below, follow this reasoning chain:
1. Question Comprehension;
2. Memory Retrieval;
3. Option Mapping;
4. Synthesis Questions:questions

Response Requirements:
1. Response Format: "Q[number]: [score](reason)" 2. Use 1-4 rating scale

**LLM answer:**
1. **Q1: 3** (User has interests in natural disasters and weather, indicating familiarity with preparedness materials, but emotional stability is not stable, suggesting some uncertainty.);
2. **Q2: 2** (User's conscientiousness is above baseline, suggesting some preparedness knowledge, but agreeableness is below mean, indicating potential skepticism about adequacy.);
3. **Q3: x** ......

## Stage 3: Panic Arousal Stage

**Question prompt:** Based on the previous stages, silently analyze these panic factors and assign a score (1-5) to each:
1. Awareness of Danger;
2. Coping Efficacy and Sense of Control;
3. Uncertainty of Risk;
4. Novelty of Risk
Each factor contributes $25\%$ weight. Calculate panic probability.

Response Format:
factors: [1-5]/5 (reason)
[panic probability (e.g., $50\%$)]

**LLM answer:**
**Awareness: 4/5** (User has interests in natural disasters and weather, indicating deep awareness of hurricane dangers, but emotional stability is not stable, suggesting some uncertainty.);
**Coping: 3/5** (User's extraversion and openness suggest moderate confidence in handling crises, but emotional stability is not stable, indicating some difficulty in managing responses.);
**Uncertainty: 3/5** (User's openness and interests suggest some understanding of risks, but emotional stability is not stable, indicating partial uncertainty.);
**Novelty: 3/5** (User's interests in natural disasters and weather suggest some prior exposure, but no explicit indication of extensive experience.)
**[55%]**

## Stage 4: Posting Response Stage (Predictive Text Generation)

**Question prompt:** After answering all questions, what text post would this user most likely publish during a hurricane? Generate tweet STRICTLY following these rules:
1. Generate EXACTLY 1 possible tweets;
2. [Response Format]: "[Tweet text with #hashtags]";
3. End with ### End;
4. Consider the user's panic probability panic probability%;
5. If panic probability $>51\%$, tweets should directly convey more panic, fear and anxiety, and tweets should include more emotional amplifiers, or more EMPHATIC capitalized words, or more repeated punctuation marks, or sensory details;
6. If panic probability $<49\%$, tweets should reflect more calmness and rationality, and tweets should show more composed language.

**LLM answer:** Stay safe everyone! The hurricane is getting stronger—winds up to 155 km/h now. I'm prepping supplies and staying informed. Let's all follow safety guidelines and look out for each other. #HurricaneAlert #StayPrepared #WeatherUpdate

## Stage 4: Posting Response Stage (Multi-Expert Evaluation)

**Question prompt:** You are a professional consistency evaluator. Please assess the user's newly generated text from the perspectives of psychology, linguistics, accuracy and emotion expression.
The user's new comment is: "tweet"
1. Generate EXACTLY 1 possible tweets;
2. Use exactly this format: "[Tweet text with #hashtags]";
3. Psychological Validation: Check if the tweet aligns with the user's psychological profile;
4. Linguistic Validation: Verify if the tweet's language style is consistent with the user's historical style;
5. Factual Validation: Confirm if the tweet is relevant to Hurricane Sandy and factually accurate;
6. Panic Probability Alignment: Use the user's panic probability value panic probability%
[Response Format]: Expert: YES/NO (reason)

**LLM answer: Psychological: YES** (The tweet aligns with the user's psychological profile, as the user has an interest in natural disasters and a conscientiousness score above the threshold, indicating a tendency to follow safety protocols.);
**Linguistic: YES** (The tweet's language style is consistent with the user's historical style, which includes informative and promotional tones.);
**Factual: YES** (The tweet is relevant to Hurricane Sandy and factually accurate, as the hurricane was indeed intensifying during the time period referenced.);
**Panic: YES** (The tweet aligns with the user's panic probability value of 55%, showing neutral concern without panic, which is appropriate for the given probability range.)

| Floating range | User Count/Proportion | | Total Users |
|---|---|---|---|
| | Consistent | Inconsistent | |
| 15% | 145 (72.5%) | 55 (27.5%) | 200 |
| 20% | 176 (88.0%) | 24 (12.0%) | 200 |

Table 13: LLM-Agent Personality Consistency vs. User Ground Truth (n=200 users)

memory. This agent simulates the same disaster scenario for the target user, aiming to predict potential panic emotion immediately before or during a hurricane's landfall.

### A.5 Baselines

In our experiments, we systematically benchmark our proposed framework against the SOTA mainstream methods for sentiment analysis and prediction. The baselines include:

- SVM: Constructs hyperplanes for classification via kernel functions to handle nonlinear separability;
- Logistic Regression: Maps linear outputs to [0,1] probabilities using the sigmoid function;
- Random Forest: Ensemble of decision trees trained on random feature subsets for diversity;
- XGBoost: A gradient boosted decision tree (GBDT) algorithm that iteratively adds trees to optimize prediction errors;
- Bi-LSTM: Bidirectional LSTM with gated mechanisms to capture long-term dependencies in sequences;
- Transformer: Models global feature dependencies via self-attention (Query-Key-Value interactions);
- PredNet:Predictive coding network minimizing prediction errors for dynamic feature learning;
- DeepSeek-v3: MoE language model using few-shot prompting with users' top 5 hurricane-related posts.
- ChatGPT-4o: Similar to DeepSeek-v3, also generates panic predictions via contextual prompting;
- THOR-ISA: Three-hop CoT framework simulating human reasoning for implicit sentiment analysis.