

Persona-E²: A Human-Grounded Dataset for Personality-Shaped Emotional Responses to Textual Events

Yuqin Yang Haowu Zhou Haoran Tu Zhiwen Hui Shiqi Yan
Haoyang Li Dong She Xianrong Yao Yang Gao Zhanpeng Jin[†]

School of Future Technology

South China University of Technology, Guangzhou, China

{ftyuqin_yang, 202364870491, 202330691461, 202364870731}@mail.scut.edu.cn

{202364871202, ftlhy, ftdshe, ftxryao}@mail.scut.edu.cn

{gaoyang2025, zjin}@scut.edu.cn

[†] Corresponding author

Abstract

Most affective computing research treats emotion as a static property of text, focusing on the writer’s sentiment while overlooking the reader’s perspective. This approach ignores how individual personalities lead to diverse emotional appraisals of the same event. Although role-playing Large Language Models (LLMs) attempt to simulate such nuanced reactions, they often suffer from “personality illusion”—relying on surface-level stereotypes rather than authentic cognitive logic. A critical bottleneck is the absence of ground-truth human data to link personality traits to emotional shifts. To bridge the gap, we introduce **Persona-E²** (Persona-Event2Emotion), a large-scale dataset grounded in annotated MBTI and Big Five traits to capture reader-based emotional variations across news, social media, and life narratives. Extensive experiments reveal that state-of-the-art LLMs struggle to capture precise appraisal shifts, particularly in social media domains. Crucially, we find that personality information significantly improves comprehension, with the Big Five traits alleviating “personality illusion.”

1 Introduction

“Two individuals can construe their situations quite similarly (agree on all the facts), and yet react with very different emotions, because they have appraised the adaptational significance of those facts differently.” (Lazarus, 1991)

The study of affective appraisal of events has long been central to affective computing and cognitive psychology (Plaza-del Arco et al., 2024). While appraisal theories suggest that emotions emerge through individualized appraisals shaped by goals and dispositions (Lazarus, 1991; Scherer and Wallbott, 1994), NLP research has largely focused on writer-expressed sentiments and reader-based unified emotional labels (Plaza-del Arco et al., 2024). This focus overlooks reader-based

nuanced perception (Buechel and Hahn, 2017b), which is critical for applications, including empathetic agents, mental health support, and personalized AI assistants, that must not only process the texts but also reason about how different individuals appraise the same event diversely.

Recent interest in role-playing LLMs aims to simulate individualized reactions by injecting rich personality profiles into prompts (Tseng et al., 2024; Chen et al., 2024; Hu and Collier, 2024; Mao et al., 2024). Despite this promise, these methods often exhibit “personality illusion” (Han et al., 2025): models tend to imitate stereotypical behaviors rather than adopting the cognitive appraisal patterns based on personality. Crucially, LLM-generated labels lack grounding in authentic feedback (Li et al., 2025a), making them insufficient for evaluating whether models truly capture emotional diversity (Samuel et al., 2025). Thus, the field still lacks a human-grounded dataset to validate and enhance personality-conditioned emotion elicitation.

To address the gap, we introduce a novel dataset, **Persona-E²** (Persona-Event2Emotion), which incorporates the popular Myers-Briggs Type Indicator (MBTI) (Myers et al., 1962; John et al., 1991) and the robust Big Five Inventory (BFI) (John et al., 2010) traits into reader-based emotion labeling. As shown in Fig. 1 by engaging annotators with assessed personality profiles to label events across diverse domains (News, Social Media, Life Experience narratives), Persona-E² enables a controlled analysis of the personality effect on the appraisals of identical textual events (Troiano et al., 2023). Notably, unlike previous corpora, Persona-E² prioritizes annotation density (36 labels per event) to capture diverse, trait-shaped responses (Tab. 1).

To evaluate the utility of Persona-E², we address three key research questions, through the experimental design in Sec. 5:

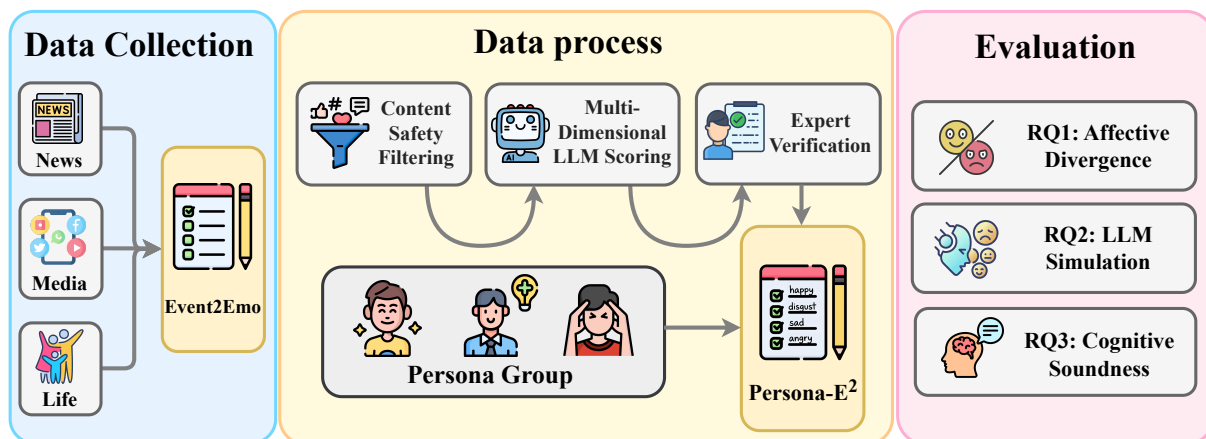


Figure 1: Overview of the Persona-E² framework. Events from three domains undergo multi-stage data processing. High-quality stimuli are then annotated by a Persona Group, serving to evaluate three research questions.

- **RQ1. Affective Divergence:** How do emotional responses diverge across the General Writer, General Reader, and Persona Reader, and how is this variance modulated by source domain and personality traits?
- **RQ2. LLM Simulation:** Can LLMs effectively simulate Persona Reader responses, particularly when faced with elicitation conflicts?
- **RQ3. Cognitive Soundness:** Do LLMs generate psychologically grounded rationales for their predictions, and what methods can enhance their cognitive validity?

Our analysis reveals that affective appraisal is a domain-sensitive process, with disagreement serving as a structured personality signal. While LLMs struggle to predict precise appraisal shifts, particularly in social media domains, personality traits improve LLMs’ comprehension, and BFI outperforms MBTI in mitigating “personality illusion.” We publicly release the dataset to foster further research in the community. Further details are provided on our project page <https://scut-hai.github.io/Persona-E2/>, with the data available for download¹².

2 Related Work

Extended discussions are provided in Appendix A.

¹<https://huggingface.co/datasets/CRIS-Yang/Persona-E2-Dataset>

²<https://www.kaggle.com/datasets/crisyang777/persona-e-personality-shaped-emotion-dataset>

2.1 Event-Elicited Emotion Analysis

Early research established the baseline for understanding emotions elicited by events. Classic works like ISEAR (Scherer and Wallbott, 1994), SocialIQA (Sap et al., 2019b) and others (Rashkin et al., 2018; Troiano et al., 2019; Forbes et al., 2020) analyzed first-person narratives and social commonsense, treating events as primitive stimuli for affective responses. Subsequent studies introduced appraisal theory to interpret these cognitive layers in depth (Troiano et al., 2022, 2023). Crucially, the field is shifting from writer-expressed sentiment to reader-based perception (Buechel and Hahn, 2017b). Benchmarks such as GoodNewsEveryone (Bostan et al., 2020), iNews (Hu and Collier, 2025) and RESEMO (Hu et al., 2024a) focus on how audiences react to news and social media. However, most existing resources rely on aggregating annotations into a single ground truth, which obscures the inter-individual variability essential for understanding diverse emotional elicitation (Plank, 2022; Soni et al., 2024).

2.2 Personality-Conditioned Affective Computing

Research on personality–emotion interaction typically utilizes the MBTI (Myers et al., 1962) and the BFI (John et al., 2010) via three paradigms. Explicit methods link self-reported traits to text or dialogue, as seen in datasets like PAN-DORA (Gjurković et al., 2021), and PersonaTAB (Inoue et al., 2025), though they primarily capture writer expression rather than reader elicitation. Implicit methods infer traits from behavioral data but often lack ground truth (Gao

Dataset	Year	#Events	#Annotations	Perspective	#Emotions	Personality
News-based Domain						
GoodNewsEveryone (Bostan et al., 2020)	2020	5,000	15,000	Writer + Reader	15	✗
NewsMTSC (Hamborg and Donnay, 2021)	2021	11,029	56,000	Writer	7	✗
iNews (Hu and Collier, 2025)	2025	2,899	14,550	Reader	6	✗
Social Media Domain						
SemEval-2018 Task 1 (Mohammad et al., 2018)	2018	22,000	700,000	Writer	4	✗
GoEmotions (Demszky et al., 2020)	2020	58,000	118,000	Reader	27	✗
SMP2020-EWECT (BrownSweater, 2020)	2020	34,768	36,374	Writer	6	✗
SenWave (Yang et al., 2025b)	2025	10,000	20,000	Writer	10	✗
Life Experience Domain						
ISEAR (Scherer and Wallbott, 1994)	1994	7,666	7,666	Writer	7	✗
Event2Mind (Rashkin et al., 2018)	2018	24,716	57,000	Writer + Reader	OV	✗
ATOMIC (Sap et al., 2019a)	2019	24,000	72,000	Experiencer	OV	✗
EmpatheticDialogues (Rashkin et al., 2019)	2019	24,850	24,850	Writer	32	✗
Social IQA (Sap et al., 2019b)	2019	37,588	37,588	Reader	OV	✗
Crowd-enVENT (Troiano et al., 2023)	2023	6,591	11,091	Writer + Reader	13	✗
Cross-domain Integration (Ours)						
Persona-E ²	2026	3,111	111,996	Reader	7	✓

Table 1: A unified comparison of emotion-annotated datasets across three sources: News, Social Media, and Life Experience. *Note*: OV: Open vocabulary.

et al., 2013; Wang et al., 2024; Hu et al., 2024b; Shen et al., 2025). Recently, LLM-based simulation has emerged to generate persona-specific responses (Tseng et al., 2024), such as Big5-Chat (Li et al., 2025a), PersonaGym (Samuel et al., 2025), and PersonalityEdit (Mao et al., 2024). Studies show that as richer prompts with profiles are introduced, the behavioral fidelity of simulated agents improves accordingly (Bai et al., 2025; Hu and Collier, 2024). Despite their promise, recent works indicate a “personality illusion“ (Han et al., 2025) where models mimic linguistic styles without adopting the underlying appraisal mechanisms. This highlights a critical gap: the lack of a human-grounded dataset to rigorously evaluate whether LLMs truly capture trait-driven emotional diversity.

3 Persona-E² Dataset Construction

To construct a rigorously controlled dataset for reader-based emotion elicitation, we designed a pipeline integrating heterogeneous event sourcing and a multi-stage filtering process.

3.1 Event Sources

To ensure affective variety and broad coverage, we gather events from three complementary domains—news, social media, and life experience narratives—covering both digital-world and real-world contexts (Appendix B.1). These sources include two distinct elicitation modes: a) First-person projection, where personal experience drives af-

fective memory, and b) Third-person observation, involving detached, personality-shaped appraisals. This constitutes a large-scale reader-centered emotion dataset that integrates the diversity of source domains and elicitation modes.

News We crawled factual reports from mainstream news websites, trending topics and verified institutional accounts. These well-structured texts provide socially significant events that elicit emotions from third-person perspective observations.

Social Media We collected posts from several public channels on Reddit. Social content brings greater topical breadth and more ambiguous context, which may elicit empathy or judgment from annotators. To capture more social norms and interpersonal dynamics, we also introduced a subset of events from Social Chemistry 101 (Forbes et al., 2020) to ensure sufficient emotion-eliciting stimuli.

Life Experience Life experience narratives were obtained from specific channels dedicated to experience sharing. These events focus on the quotidian experiences of everyday life, ranging from minor frustrations to moments of gratitude. Such narratives are designed to invite first-person projection, serving as a counterpart to the detached perspective of news.

3.2 Event Filtering Pipeline

Raw collections inevitably contain noise, safety risks, and large quantities of content that lack emo-

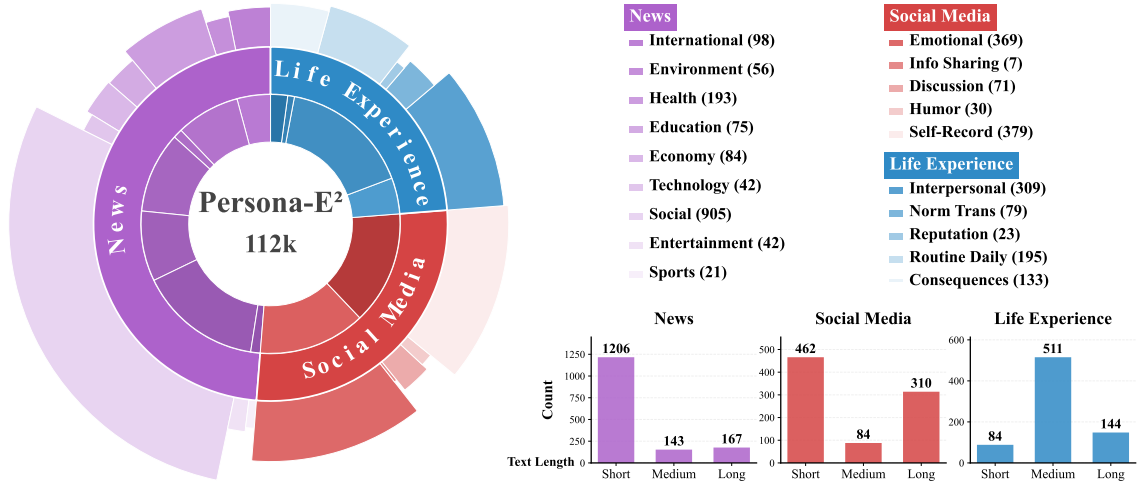


Figure 2: Hierarchical composition of the Persona-E² dataset. The sunburst chart illustrates three levels from inner to outer: original data sources, three primary domains, and fine-grained semantic subcategories. *Note:* Text Length: Short (0–30), Medium (30–100), Long (100–300).

tional significance. To ensure high-quality stimuli, we implemented a 3-stage filtering procedure.

Stage 1: Content Safety Filtering. For life experience and social media domains, we first prune toxic or sensitive content using NSFW classifiers (Albouzidi, 2023; TostAI, 2023). More detailed information is illustrated in Appendix B.2.1.

Stage 2: Multi-Dimensional LLM Scoring. We utilize QWEN3-MAX (Team, 2025) to translate non-English materials into English and filter out events that lack emotional significance. The final weighted score is computed as:

$$\text{Score} = 0.35V + 0.30A + 0.20R + 0.15I \quad (1)$$

Here, V , A , I , and R represent personality variability, emotional arousal, emotional implicitness, and source relevance, respectively (see Appendix B.2.2). Applying source-specific thresholds (Appendix B.2.3), we selected 6,348 candidates from an initial pool of 76,773 events.

Stage 3: Expert Verification. Finally, a 5-member expert panel conducted a rigorous audit—removing factual errors, translation bias, and hate speech—yielding a set of 3,111 events. English examples are provided in Appendix B.4.

3.3 Annotation Protocol

As shown in Tab. 1, we adopted Ekman’s six emotions (disgust, fear, anger, sadness, surprise, joy) plus neutral (Ekman, 1992). Following the ISEAR (Scherer and Wallbott, 1994), we used

a reader-centric question: “How would you feel when reading this event?”, to capture elicitation rather than semantics. During this process, no role-playing was involved for human annotators so that each data point is anchored in an authentic persona.

Annotation Unit We define an annotation unit as a reader-centric emotional response to a single textual event conditioned on a specific personality profile. Each unit is a tuple integrating the event context, a unique annotator ID, the annotator’s measured trait scores, and the corresponding emotion label.

Labeling Process As shown in Tab. 1, we adopted Ekman’s six emotions (disgust, fear, anger, sadness, surprise, joy) plus neutral (Ekman, 1992). Following the ISEAR (Scherer and Wallbott, 1994), we adopted the question style: “How would you feel when reading this event?”, to capture elicitation rather than semantics.

Annotator Recruitment. We recruited 36 annotators, profiling their personalities via MBTI (Myers et al., 1962) and BFI (John et al., 2010) questionnaires (Appendix C.3). Crucially, the annotation process involved no role-playing. Consequently, annotators were strictly instructed to report their genuine emotional reactions for each of the 3111 events.

Quality Control. To ensure high-fidelity responses, we implemented reader-centric training, mandatory guideline review and behavior monitoring (Appendix C.2). The annotation task was

distributed over several weeks to maintain annotator attention and label quality.

4 Dataset Analysis

4.1 Descriptive Analysis

Persona-E² comprises 112k high-quality annotations across three domains: news (49%), social media (27%), and life experiences (24%). As shown in Fig. 3, a significant emotional divergence exists between the writers’ sentiment (Hartmann, 2022) and the readers’ actual emotions. This divergence is domain-dependent: slight in factual news but pronounced in life experiences, where interpersonal narratives trigger first-person projection and emotional transmission. Moreover, the observed shift from writer to reader emotion leads us to a deeper discussion of RQ1 (Sec. 5.2).

4.2 Annotation Reliability

In affective computing, inter-annotator disagreement is often dismissed as label noise (Plank, 2022). Moving beyond this, we report the annotation reliability by testing the personality-aware agreement gap. This shifts the focus from universal consensus to trait-conditioned alignment, grounded in the premise that subjective disagreement is structured by latent personality profiles. Thus, annotators with similar personalities should exhibit higher consensus than random groupings. The validation of this hypothesis via BFI and MBTI clustering ensures the dataset’s reliability.

In-Group Agreement To validate the hypothesis, we applied K-means clustering on BFI vectors. Unlike MBTI’s discrete categories, this method adapts to continuous traits and ensures balanced clusters, addressing the statistical instability that arises from MBTI’s sparse subgroups. The choice of $k = 6$ was empirical, aimed at balancing the number of annotators per cluster with the captured personality diversity. To ensure the robustness of our groupings, we conducted a sensitivity analysis across various algorithms (K-means, GMM, Hierarchical) and cluster counts ($k \in [3, 9]$). As detailed in Appendix E.1, the Personality Agreement Gap (PAG) consistently remains positive across all settings, confirming that trait-aligned grouping captures shared interpretative logic rather than clustering artifacts.

As shown in Fig. 4, in-group Top-1 agreement consistently outperforms the global average. More specifically, Cluster 0 achieves a +11.5% gain on

Top-1 agreement over the baseline. This confirms that grouping by traits uncovers shared interpretative logic.

Personality Grouping Analysis As shown in Tab. 2, we quantify the personality effect by calculating the **Personality Agreement Gap** ($PAG = Agr_{in} - Agr_{out}$), representing the Top-1 agreement delta between In-Group and Out-Group pairs with equalized sample sizes.

- **BFI Grouping:** Cluster 0 (High Conscientiousness/Openness) shows a massive PAG of +25.96% compared to the out-group (Tab. 2a), indicating a convergent appraisal pattern. Conversely, Cluster 3 (Low Neuroticism) exhibits the smallest PAG (+8.3%). This may suggest that traits like high Neuroticism act as strict “perceptual filters” that funnel reactions into specific categories. Consistent patterns are also revealed across different K settings (Tab. 2b).
- **MBTI Grouping:** Similar patterns appear in MBTI types (Tab. 2c). ESTP types achieve peak PAG (+26.98%) by prioritizing social cues (Pickett et al., 2004), while ISTJ yields lower consensus.

Positive PAG values reveal that in-group agreement consistently exceeds out-group agreement, providing empirical evidence that affective disagreement is structured by trait-aligned patterns.

5 Experiment

5.1 Experimental Setup

We designed experiments to assess the necessity of personality modeling in emotion analysis and the capability of LLMs to simulate these affective shifts. To evaluate performance for RQ2 and RQ3, we employed leading open and closed-source models.

Model List: GPT-5.1 (OpenAI, 2025), LLAMA-3-8B (Grattafiori et al., 2024), QWEN3-8B (Yang et al., 2025a), GEMMA-3-12B (Team et al., 2025), and MINISTRAL-3-8B (AI, 2025).

Computing Environment: Open-source models are evaluated on a cloud computing platform using NVIDIA A100 GPUs. Closed-source models are accessed through their APIs. Appendix D presents details on model version and hyper-parameters.

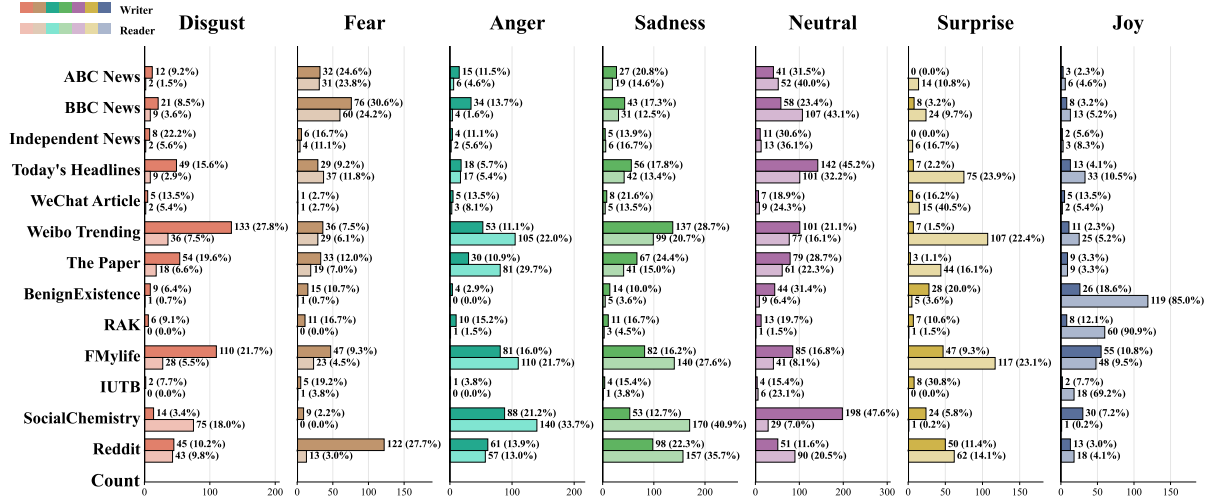


Figure 3: Comparison of emotion distributions between Writers and Readers showing the percentage distribution of seven emotions across sources. *Note:* IUTB, ASK stand for r/usedtobelieve and RandomActsofKindness.

Cluster	Agr_{in}	Agr_{out}	PAG	Cluster	Agr_{in}	Agr_{out}	PAG	Type	Agr_{in}	Agr_{out}	PAG
Cluster 0	61.07	35.11	+25.96	Cluster 0	58.05	37.63	+20.42	ESTP	64.33	37.35	+26.98
Cluster 1	57.63	37.59	+20.03	Cluster 1	56.32	38.13	+18.19	INTP	60.03	36.83	+23.20
Cluster 2	57.26	43.39	+13.87	Cluster 2	51.05	45.15	+5.9	INTJ	61.53	37.58	+23.95
Cluster 3	49.39	41.12	+8.27	Cluster 3	49.39	41.37	+8.02	ESTJ	56.02	38.39	+17.63
Cluster 4	54.73	40.38	+14.36					ENTJ	55.30	35.84	+19.46
Cluster 5	54.76	37.11	+17.65					ISTJ	50.83	41.16	+9.68

(a) BFI K-means ($K = 3$)							(b) BFI K-means ($K = 4$)						(c) MBTI Grouping ($N \geq 3$)							
Group	N	Open.	Cons.	Extra.	Agree.	Neuro.	Group	N	Open.	Cons.	Extra.	Agree.	Neuro.	Group	N	Open.	Cons.	Extra.	Agree.	Neuro.
Cluster 0	3	0.823	0.868	0.691	0.622	0.358	Cluster 0	4	0.815	0.836	0.628	0.636	0.336	Cluster 0	4	0.815	0.836	0.628	0.636	0.336
Cluster 1	4	0.718	0.536	0.482	0.797	0.638	Cluster 1	5	0.679	0.517	0.429	0.629	0.729	Cluster 1	5	0.679	0.517	0.429	0.629	0.729
Cluster 2	7	0.553	0.647	0.629	0.673	0.435	Cluster 2	16	0.560	0.661	0.559	0.682	0.432	Cluster 2	16	0.560	0.661	0.559	0.682	0.432
Cluster 3	11	0.695	0.828	0.604	0.801	0.206	Cluster 3	11	0.695	0.828	0.604	0.801	0.206	Cluster 3	11	0.695	0.828	0.604	0.801	0.206
Cluster 4	6	0.556	0.569	0.455	0.580	0.592														
Cluster 5	5	0.612	0.763	0.494	0.673	0.335														

(d) BFI Mean Features ($K = 6$)							(e) BFI Mean Features ($K = 4$)						
Cluster 0	58.9	29.4	11.7	0.0	0.0	0.0	Cluster 0	58.9	29.4	11.7	0.0	0.0	
Cluster 1	56.0	26.3	14.9	2.8	0.0	0.0	Cluster 1	56.0	26.3	14.9	2.8	0.0	
Cluster 2	56.8	24.7	12.2	5.0	1.2	0.1	Cluster 2	56.8	24.7	12.2	5.0	1.2	
Cluster 3	49.0	24.6	14.2	7.8	3.5	0.8	Cluster 3	49.0	24.6	14.2	7.8	3.5	
Cluster 4	55.1	25.5	12.8	5.6	1.0	0.0	Cluster 4	55.1	25.5	12.8	5.6	1.0	
Cluster 5	54.5	25.9	13.9	5.1	0.6	0.0	Cluster 5	54.5	25.9	13.9	5.1	0.6	
Overall	47.4	23.5	14.0	8.3	4.4	1.9	Overall	47.4	23.5	14.0	8.3	4.4	

Table 2: Personality-based Top-1 agreement analysis. **Top:** Comparison of In-Group (Agr_{in}) and Out-Group (Agr_{out}) agreement levels (N controlled). **Bottom:** Mean BFI traits for clusters. *Note:* Open., Cons., Extra., Agree., and Neuro. represent the Big Five traits; N denotes the number of annotators.

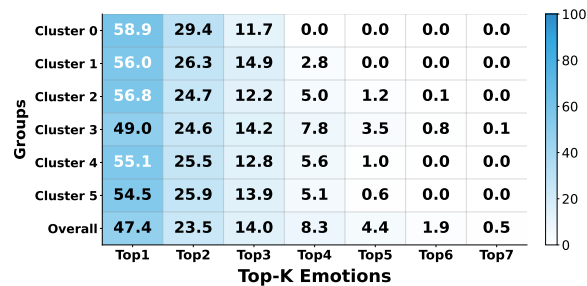


Figure 4: Top- K emotion distribution across BFI clusters. The heatmap shows the average vote share (%) of the k -th most frequent emotion per event.

5.2 RQ1. Dataset Affective Divergence

How do emotional responses diverge across the General Writer, General Reader, and Persona

Reader? Fig. 3 illustrates a gap between writer and reader-based emotions, which we hypothesize is modulated by domain and personality-driven appraisal patterns (Buechel and Hahn, 2017b). To validate the assumption, we differentiated three affective layers: (1) General Writer (GW): semantic sentiment predicted by a pre-trained emotion classifier (Hartmann, 2022) that provides an identical label space to our human annotations; (2) General Reader (GR): majority-vote elicitation; and (3) Persona Reader (PR): trait-conditioned elicitation. To quantify this divergence, we categorized seven emotions into positive (surprise, joy) and negative (disgust, fear, anger, sadness) polarities. We then computed affective transition matrices, where element (i, j) represents the probability of emotion i

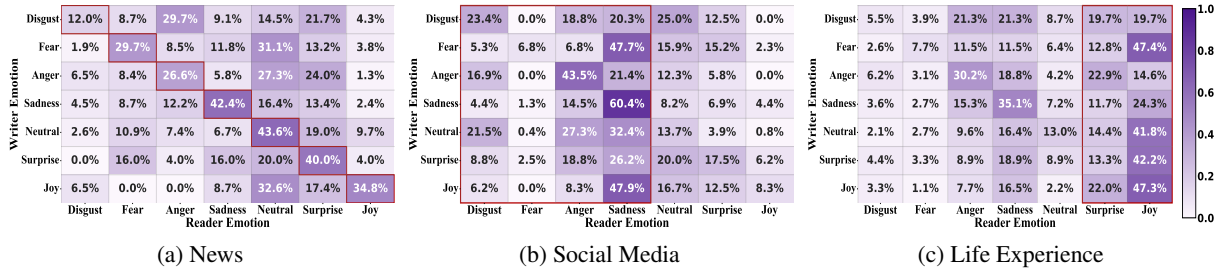


Figure 5: Affective transition matrices between General Writer and Reader. Red boxes highlight: (a) high resonance in News, (b) negative bias in Social Media, and (c) positive shift in Life Experience. *Note:* Negative: disgust, fear, anger, sadness; Positive: surprise, joy.

shifting to emotion j across the three perspectives.

Domain-Driven Divergence. Comparing semantic sentiment (GW) with majority-vote elicitation (GR) reveals that emotional appraisal is a domain-sensitive reconstruction rather than a direct transfer (Fig. 5). Appendix E.2.1 presents the detailed data analysis related to this finding. First, news acts as a rational buffer: it maintains a high neutrality transfer and emotional resonance rate (Tab. 8). This aligns with professional journalism’s role in minimizing cognitive appraisal variance through factual grounding (Lazarus, 1991; Alm, 2008). Second, social media functions as an “Emotional Black Hole” (Baumeister et al., 2001). We observed a severe negativity bias, where significant portions of neutral and positive GW sentiments shift to negative GR reactions (81.6% of neutral and 59.35% of positive sentiments transfer into negative in Tab. 9). This reflects a “Forced Siding” mechanism, where ambiguity is treated as a vacuum filled by reader-side hostility (Tajfel, 1974; Baumeister et al., 2001). Conversely, life narratives trigger a “Psychological Immune System” (Gilbert et al., 1998). Readers exhibit an optimism bias, filtering distress to prioritize empathetic resonance (43.3% of negative and 56.2% of neutral sentiments transfer to positive in Tab. 10), a cognitive nuance often overlooked by traditional sentiment analysis (Matlin, 2016).

Personality-Driven Modulation. We employed BFI-based clustering for analysis rather than MBTI categories, as the latter led to sparse populations in specific groups, which cannot provide reliable statistics. We found that personality significantly acts as a filter for emotional transfer (further data analysis is available in Appendix E.2.2). First, Anxious Empathy: Individuals with high Agreeableness (A) and Neuroticism (N), such as C1, show the highest neutral-to-negative transfer rates. Here,

high A’s social sensitivity is likely hijacked by the interpretation bias of high N (McCrae and Costa Jr, 1997), as evidenced in Fig. 6a. Secondly, Negativity Passivation: High Conscientiousness (C) and Extraversion (E) exhibit “Negative Passivation.” C0 shows the lowest negative resonance in Fig 6b. This suggests these individuals effectively regulate distress (Gross, 1998). In contrast, low C/E (e.g., C2 and C3) led to “Negative Locking” due to a lack of perceived control. Finally, Neutralization: Illustrated in Tab. 13, we found a statistical correlation between Openness (O) and the Neutralization Rate ($r = +0.86, p = 0.027$). To align with cognitive theory, high-O individuals use cognitive complexity to moderate emotional activation, showing a low need for cognitive closure (Webster and Kruglanski, 1994). In summary, emotional elicitation is reshaped by domain effects and personality traits, highlighting the move from general sentiment analysis to persona-aware modeling.

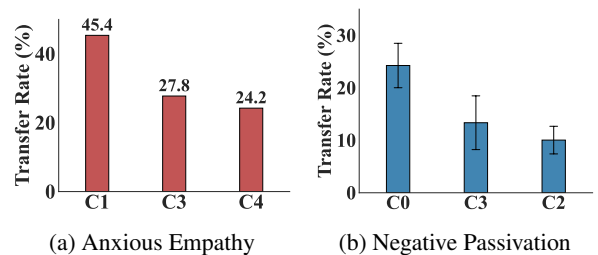


Figure 6: (a) Transfer rate from non-negativity to negativity in the news domain, and (b) Transfer rate from negativity to non-negativity, error bar means deviation among 3 domains.

5.3 RQ2. LLM Emotion Simulation

Can LLMs effectively mimic personality-shaped emotional responses? We investigated whether models can predict emotional shifts with personality profiles. LLMs are conditioned on three strategies: General Prompt, Persona Prompt (BFI person-

Metric	Source Prompt	News		Social Media		Life Experience		Overall					
		General	Persona	CoT	General	Persona	CoT	General	Persona	CoT			
Top-1 Acc.	GPT5.1	31.8	25.0	29.5	18.2	18.2	27.3	32.4	35.3	35.3	29.0	27.0	31.0
	LLAMA3-8B	29.5	29.3	20.5	9.1	4.8	13.6	32.4	32.4	32.4	26.0	25.0	23.0
	QWEN3-8B	18.2	26.9	25.4	23.1	21.4	22.7	20.3	33.7	26.0	20.0	28.0	25.0
	GEMMA3-12B	18.2	25.0	27.3	22.7	27.3	18.2	35.3	32.4	29.4	25.0	28.0	26.0
	MINISTRAL3-8B	18.2	13.6	20.5	5.0	9.1	4.5	36.4	35.3	35.3	22.0	20.0	22.0
Top-2 Acc.	GPT5.1	54.5	59.1	59.1	40.9	50.0	45.5	47.1	55.9	55.9	49.0	56.0	55.0
	LLAMA3-8B	47.7	43.9	45.5	18.2	19.0	31.8	50.0	47.1	44.1	42.0	39.6	42.0
	QWEN3-8B	31.3	33.5	38.6	31.4	36.4	39.4	39.2	39.0	45.1	34.0	36.0	41.0
	GEMMA3-12B	36.4	47.7	56.8	27.3	40.9	45.5	50.0	52.9	41.2	39.0	48.0	49.0
	MINISTRAL3-8B	40.9	29.5	38.6	35.0	36.4	22.7	51.5	50.0	50.0	43.3	38.0	39.0

Table 3: Comparison of different models performance in subset. *Note:* General Prompt: Prompt without personality traits, Persona: Prompt with BFI traits, CoT: CoT-Style Prompt with BFI traits.

ality vectors (Johnson, 2014)), and Persona-CoT (Chain of Thought) (Wei et al., 2022).

While general performance on 100 randomly sampled events is reported in Tab. 14, we specifically focus on trait-driven shifts by constructing the Subjective Divergence Subset (SDS). SDS targets scenarios where emotional responses are clear but heavily conditioned on the annotator’s personality. Label validity is ensured via the *Group Consensus Score* ($S_{consensus}$). For a probability distribution P_G of group G , consensus is defined as:

$$S_{consensus}(G) = 1 - \frac{-\sum p_i \log p_i}{\log K} \quad (2)$$

Thus, the SDS events satisfy $S_{consensus}(G) > \alpha$ for all personality groups G , totaling 413 events (Life: 87, News: 257, Social: 69) for $\alpha = 0.3$. Details of SDS are reported in Appendix E.3.

Latent Emotion Understanding. As detailed in Tab. 3, experiments on the subset reveal a significant gap between Top-1 ($\sim 25.0\%$) and robust Top-2 performance ($\sim 45.0\%$). This suggests that LLMs successfully map emotional contexts to relevant semantic neighborhoods (Brown et al., 2020) but fail to pinpoint the precise label. While models capture the general affective sphere, they cannot distinguish subtle sentiment shifts, indicating more human feedback is likely required in affective training mechanisms (Ouyang et al., 2022).

Prompt Strategies. We observed that persona prompts and CoT prompts are not universally beneficial; their efficacy is modulated by model capacity. Complex prompting improved the reasoning process for larger models, but degraded it in smaller architectures. For instance, GPT-5.1 increased from 29.0% to 31.0% but LLAMA3-8B declined.

Model	Consistency			Plausibility			Specificity		
	BL	MBTI	BFI	BL	MBTI	BFI	BL	MBTI	BFI
GPT5.1	4.7	16.5	78.8	13.8	15.8	70.4	17.6	13.5	68.9
LLAMA3	9.4	19.8	70.8	21.5	23.1	55.4	20.5	18.2	61.3
QWEN3	10.4	19.4	70.2	19.7	20.4	59.9	23.7	19.4	56.9
GEMMA3	14.8	17.4	67.8	21.8	15.1	63.1	19.5	15.5	65.0
MINISTRAL3	12.5	15.5	72.0	21.6	19.5	58.9	22.0	12.5	65.5

Table 4: Win-Rate comparison of LLMs in the best-of-3 selection task. *Note:* BL: Baseline Prompt without personality, MBTI: MBTI Prompt, BFI: BFI Prompt.

We attribute this to attention dilution (Kaplan et al., 2020), where the personality profiles may overwhelm the limited context of smaller models.

Domain Discrepancy. Models consistently struggled in the social media domain compared to others (GPT-5.1: 18.2-27.3% in social for Top-1). This performance gap suggests that current LLM training corpora are biased towards structured materials, failing to process the informal dynamics of virtual interactions. To bridge this gap in cyber-social emotional analysis, future studies must focus on the unstructured online materials. See Appendix E.3.4 for a full data breakdown.

5.4 RQ3. Cognitive Soundness

Do LLMs generate human-like rationales for the emotional activation? Beyond prediction, we evaluated the cognitive plausibility of the psychological reasoning process (Zhang et al., 2024; Li et al., 2024) on all 413 SDS events. Five trained reviewers performed a best-of-three forced-choice evaluation (Kiritchenko and Mohammad, 2016) based on: a) Persona Consistency, b) Reasoning Plausibility, and c) Emotion Specificity (Definitions in Appendix E.4). In Tab. 4, selections were aggregated to compute win rates for each model.

Personality Comparison. Tab. 4 reveals that the choice of personality directly impacts rationale quality. The BFI strategy demonstrates better alignment with human cognitive patterns (55.4-70.4%), ahead of both MBTI and Baseline groups, particularly in consistency. For instance, GPT-5.1 achieved 68.9-78.8% win rate with the BFI prompt across metrics, while the MBTI only 13.5-16.5%. In contrast, MBTI prompts show stronger consistency but weaker plausibility and specificity compared to the baseline. This suggests that the trait-based detail of BFI provides a more robust choice for simulating nuanced appraisals compared to the binary nature of MBTI (Furnham, 1996).

Model Scale. Cognitive soundness exhibits a clear dependency on model capacity. GPT-5.1 consistently achieved the highest win rates on BFI prompt, indicating that complex rationale generation requires large-scale models. Among open-source models, GEMMA-3-12B shows greater stability than smaller LLMs, maintaining a >60% preference in plausibility under BFI settings. This suggests that robust rationale generation is a capacity-intensive task for current LLMs. A detailed analysis of the data is provided in Appendix E.4.4.

6 Conclusion

We introduced **Persona-E²**, a human-grounded dataset for personality-conditioned emotion analysis. Reliability experiments demonstrate that PAG reflect personality effects. Analysis of affective appraisals reveals domain-specific patterns, such as social media acts as an “Emotional Black Hole.” We also identified distinct, personality-shaped processes in emotion elicitation. While LLMs can map semantic neighborhoods in predicting emotional shifts, they currently lack precision. Our comparisons underscore the importance of cyber-social emotional tasks, where BFI outperformed MBTI. Finally, cognitive prompt improves personalized reasoning, especially in large-scale models.

Limitations

We introduce **Persona-E²**, a large-scale dataset explicitly grounded in personality traits to model reader-based emotional variation. However, the dataset has several limitations. We acknowledge that the diversity of event sources is constrained, and the current version is limited to Chinese and English texts, which may not fully capture emotional expressions in other domains, cultures, or

languages. Likewise, the limited number of annotators should be taken into consideration when assessing population effectiveness. Furthermore, cross-lingual translation introduces additional challenges: differences in phrasal connotations can shift the perceived emotional meaning. Moreover, emotion perception is inherently subjective; thus, individual differences among annotators hinder the consistent reproduction of emotion labels. Moreover, there is an inherent gap between task-oriented annotators and spontaneous real-world users. The emotion label space adopts Ekman’s six basic emotions plus neutral as a strategic trade-off, though this categorical scheme is coarser than dimensional or open-vocabulary alternatives. Additionally, the “General Writer” labels rely on an external classifier (Hartmann, 2022) due to the unavailability of original author ratings, introducing potential algorithmic bias that should be considered when interpreting writer-reader emotion gaps.

Ethics Considerations

Subjectivity and Diversity in Annotation We explicitly acknowledge that emotional appraisal is inherently subjective and culturally situated. Unlike traditional paradigms that seek a single ground truth, our data collection protocol respects diverse interpretations. We required annotators to report their genuine emotional reactions based on their personality profiles, ensuring the dataset captures the variance of human experience rather than enforcing a potentially biased consensus.

Annotator Welfare and Consent We recruited 36 annotators through university channels. All annotations were conducted via a self-developed and user-friendly online crowdsourcing platform. Participants were fully informed of the research purpose and potential risks, the public nature of the resulting dataset, and their right to withdraw at any time. We strictly adhered to fair labor practices; compensation was calculated based on task duration and complexity, ensuring that it significantly exceeded the local minimum wage.

Privacy and Data Anonymization The textual data used in **Persona-E²** originates from publicly available news, social media, and personal narratives. To protect the privacy of the original content creators, we implemented a rigorous anonymization pipeline. All Personally Identifiable Information (PII), including real names, specific locations,

and user handles, was excluded from the research.

Institutional Review The experimental protocol, including data collection and annotator interaction, was reviewed and approved by our Institutional Review Board prior to the study’s commencement.

Representation We acknowledge distinct limitations in coverage. Although our dataset contains approximately 3,000 diverse events, it does not fully represent all events. Our pool of 36 annotators, while providing high annotation density, represents a specific demographic (university-educated, aged 18–25) that may not perfectly reflect the global population, potentially introducing demographic biases. We also utilized LLMs for initial data filtering, acknowledging that this automation may introduce minor biases despite human oversight.

Responsible Use and Sensitivity This dataset, to be released under a license compatible with its research-only creation purpose, must be used solely for non-commercial research. Derivatives must not be deployed in real-world applications beyond research prototypes, especially in commercial contexts. This dataset contains narratives that may be sensitive to specific cultural, religious, or social contexts. We urge researchers to exercise caution when deploying models trained on this data, particularly in high-stakes applications such as mental health support or behavioral analysis. Users must be aware that the models may reproduce the specific appraisal patterns of our annotator pool, which should not be interpreted as universal cultural truths.

Open Access and Reproducibility To foster transparency and encourage further research in personality-aware NLP, we will release the Persona-E² dataset under a license that permits research use while prohibiting malicious applications. We believe open access is essential for the community to scrutinize, validate, and build upon our findings responsibly.

Use of AI Assistants We utilized AI assistants (e.g., CHATGPT-4O) to refine the clarity and grammar of the manuscript. All scientific claims, experimental designs, and data analyses were conducted and verified by the human authors.

Societal Impact

Advancing Affective Computing Our work promotes a paradigm shift from writer-centric senti-

ment analysis to reader-based emotional appraisal. By introducing the Persona-E² dataset, we encourage the research community to move beyond static, single-label classification and explore how personality traits shape diverse interpretations. This transition is essential for building more inclusive AI systems that respect individual differences in perception.

Bridging Cognition and AI By integrating cognitive theories with LLMs, we highlight the value of psychological grounding in NLP. Our findings show that structured personality constraints (specifically BFI) enhance model reasoning, opening new avenues for personalized human-computer interaction. This theoretical alignment offers a foundation for developing more empathetic agents in mental health support and personalized education.

Decoupling Cyber-Social Dynamics Our analysis reveals that emotional elicitation varies significantly across news, social media, and life narratives. This distinction suggests that future research must decouple virtual interactions from physical-world events. Understanding distinct mechanisms like the negativity bias in social media provides actionable insights for monitoring digital sentiment and mitigating online polarization.

Potential Risks While personalized appraisal enhances user experience, it carries risks. The ability to tailor content to specific personality profiles could be misused for targeted manipulation or to reinforce echo chambers. Furthermore, without careful constraints, models might over-generalize personality traits, leading to unintended stereotyping. Researchers must prioritize safety and fairness when deploying these persona-aware systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants W2532047 and 62302168, and the Guangdong Provincial Department of Science and Technology under Grants 2023CX10X070 and 2026A1515010254. We would like to express our sincere gratitude to all the annotators for their diligent and meticulous efforts, whose commitment was essential to the construction of our dataset. We also thank members of the HAI Lab at SCUT, our external collaborators, and the anonymous reviewers for their insightful comments and suggestions.

References

- Mistral AI. 2025. Ministral 3 8b instruct 2512 model card. <https://huggingface.co/mistralai/Ministral-3-8B-Instruct-2512>. Accessed: 2025-12-23.
- Elias Albouzi. 2023. distilbert-nsfw-text-classifier. <https://huggingface.co/eliasalbouzi/distilbert-nsfw-text-classifier>. Hugging Face model. Accessed: 2025-12-23.
- Ebba Cecilia Ovesdotter Alm. 2008. *Affect in* text and speech*. University of Illinois at Urbana-Champaign.
- Yuqi Bai, Tianyu Huang, Kun Sun, and Yuting Chen. 2025. *Scaling law in llm simulated personality: More detailed and realistic persona profile is all you need*. *arXiv preprint arXiv:2510.11734*.
- Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. *Bad is stronger than good*. *Review of general psychology*, 5(4):323–370.
- Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. *GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- BrownSweater. 2020. *Smp2020-ewect: Weibo-based emotion classification dataset*. Accessed: 2025-12-09.
- Sven Buechel and Udo Hahn. 2017a. *EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis*. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.
- Sven Buechel and Udo Hahn. 2017b. *Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation*. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 1–12, Valencia, Spain. Association for Computational Linguistics.
- Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024. *The oscars of ai theater: A survey on role-playing with language models*. *arXiv preprint arXiv:2407.11484*.
- Bao Minh Doan Dang, Laura Oberländer, and Roman Klinger. 2021. *Emotion stimulus detection in German news headlines*. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 73–85, Düsseldorf, Germany. KONVENS 2021 Organizers.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. *GoEmotions: A dataset of fine-grained emotions*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- Keyang Ding, Chuang Fan, Yiwen Ding, Qianlong Wang, Zhiyuan Wen, Jing Li, and Ruifeng Xu. 2024. *Lcsep: A large-scale chinese dataset for social emotion prediction to online trending topics*. *IEEE Transactions on Computational Social Systems*, 11(3):3362–3375.
- Paul Ekman. 1992. *An argument for basic emotions*. *Cognition & emotion*, 6(3-4):169–200.
- Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. *Social chemistry 101: Learning to reason about social and moral norms*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.
- Adrian Furnham. 1996. *The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality*. *Personality and Individual Differences*, 21(2):303–307.
- Rui Gao, Bibo Hao, Shuotian Bai, Lin Li, Ang Li, and Tingshao Zhu. 2013. *Improving user profile with personality traits predicted from social media content*. In *Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13*, page 355–358, New York, NY, USA. Association for Computing Machinery.
- Daniel T Gilbert, Elizabeth C Pinel, Timothy D Wilson, Stephen J Blumberg, and Thalia P Wheatley. 1998. *Immune neglect: a source of durability bias in affective forecasting*. *Journal of personality and social psychology*, 75(3):617.
- Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. *PANDORA talks: Personality and demographics on Reddit*. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. *The llama 3 herd of models*. *arXiv e-prints*, pages arXiv–2407.

- James J. Gross. 1998. [The emerging field of emotion regulation: An integrative review](#). *Review of general psychology*, 2(3):271–299.
- Felix Hamborg and Karsten Donnay. 2021. [NewsMTSC: A dataset for \(multi-\)target-dependent sentiment classification in political news articles](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online. Association for Computational Linguistics.
- Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. 2025. [The personality illusion: Revealing dissociation between self-reports & behavior in llms](#). *arXiv preprint arXiv:2509.03730*.
- Jochen Hartmann. 2022. [Emotion english distilroberta-base](#). <https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/>. Hugging Face model. Accessed: 2025-12-23.
- Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. [Appraisal theories for emotion classification in text](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. [Emotion-Lines: An emotion corpus of multi-party conversations](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Bo Hu, Meng Zhang, Chenfei Xie, Yuanhe Tian, Yan Song, and Zhendong Mao. 2024a. [RESEMO: A benchmark Chinese dataset for studying responsive emotion from social media content](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16375–16387, Bangkok, Thailand. Association for Computational Linguistics.
- Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024b. [LLM vs small model? large language model based text augmentation enhanced personality detection model](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18234–18242.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in LLM simulations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.
- Tiancheng Hu and Nigel Collier. 2025. [iNews: A multimodal dataset for modeling personalized affective responses to news](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25000–25040, Vienna, Austria. Association for Computational Linguistics.
- S. Inoue, S. Wang, and H. Li. 2025. [PersonaTAB: Predicting personality traits using textual, acoustic, and behavioral cues in fully-duplex speech dialogs](#). In *Proceedings of Interspeech 2025*, pages 181–185.
- Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. [Big five inventory](#). *Journal of personality and social psychology*.
- Oliver P John, Richard W Robins, and Lawrence A Pervin. 2010. *Handbook of personality: Theory and research*. Guilford Press.
- John A. Johnson. 2014. [Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120](#). *Journal of Research in Personality*, 51:78–89.
- Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *ArXiv*, abs/2001.08361.
- Svetlana Kiritchenko and Saif M. Mohammad. 2016. [Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.
- Richard S Lazarus. 1991. *Emotion And Adaptation*. Oxford University Press.
- Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025a. [BIG5-CHAT: Shaping LLM personalities through training on human-grounded data](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.
- Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Li-dong Bing. 2024. [Evaluating psychological safety of large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843, Miami, Florida, USA. Association for Computational Linguistics.
- Zheng Li, Sujian Li, Dawei Zhu, Qilong Ma, and Weimin Xiong. 2025b. [EERPD: Leveraging emotion and emotion regulation for improving personality detection](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7721–7734, Abu Dhabi, UAE. Association for Computational Linguistics.
- Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. [Editing personality for large language models](#). In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC*

- 2024, Hangzhou, China, November 1–3, 2024, *Proceedings, Part II*, page 241–254, Berlin, Heidelberg. Springer-Verlag.
- Margaret W Matlin. 2016. [Pollyanna principle](#). In *Cognitive illusions*, pages 315–335. Psychology Press.
- Robert R McCrae and Paul T Costa Jr. 1997. [Conceptions and correlates of openness to experience](#). In *Handbook of personality psychology*, pages 825–847. Elsevier.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. [SemEval-2018 task 1: Affect in tweets](#). In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.
- Isabel Briggs Myers and 1 others. 1962. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA.
- OpenAI. 2025. [Gpt-5.1 instant and gpt-5.1 thinking system card addendum](#). Technical report, OpenAI. Accessed: 2025-12-30.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Cynthia L. Pickett, Wendi L. Gardner, and Megan Knowles. 2004. [Getting a cue: The need to belong and enhanced sensitivity to social cues](#). *Personality and Social Psychology Bulletin*, 30(9):1095–1107. PMID: 15359014.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. [Emotion analysis in NLP: Trends, gaps and roadmap for future directions](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. [Event2Mind: Commonsense inference on events, intents, and reactions](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Carley Reardon, Sejin Paik, Ge Gao, Meet Parekh, Yanling Zhao, Lei Guo, Margrit Betke, and Derry Tanti Wijaya. 2022. [BU-NEMO: an affective dataset of gun violence news](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2507–2516, Marseille, France. European Language Resources Association.
- Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. [PersonaGym: Evaluating persona agents and LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. [Atomic: An atlas of machine commonsense for if-then reasoning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Klaus R Scherer and Harald G Wallbott. 1994. [Evidence for universality and cultural variation of differential emotion response patterning](#). *Journal of personality and social psychology*, 66(2):310.
- Lingzhi Shen, Xiaohao Cai, Yunfei Long, Imran Razzak, Guanming Chen, and Shoaib Jameel. 2025. [Emoperso: Enhancing personality detection with self-supervised emotion-aware modelling](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 2577–2587.
- Nikita Soni, Niranjan Balasubramanian, H. Andrew Schwartz, and Dirk Hovy. 2024. [Comparing pre-trained human language models: Is it better with human context as groups, individual traits, or both?](#) In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 316–328, Bangkok, Thailand. Association for Computational Linguistics.
- Henri Tajfel. 1974. [Social identity and intergroup behaviour](#). *Social Science Information*, 13(2):65–93.

- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- Qwen Team. 2025. [Qwen3-max: Just scale it](#).
- TostAI. 2023. [nsfw-text-detection-large](https://huggingface.co/TostAI/nsfw-text-detection-large). <https://huggingface.co/TostAI/nsfw-text-detection-large>. Hugging Face model. Accessed: 2025-12-23.
- Enrica Troiano, Laura Ana Maria Oberlaender, Maximilian Wegge, and Roman Klinger. 2022. [x-enVENT: A corpus of event descriptions with experienter-specific emotion and appraisal annotations](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1365–1375, Marseille, France. European Language Resources Association.
- Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. [Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction](#). *Computational Linguistics*, 49(1):1–72.
- Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. [Crowdsourcing and validating event-focused emotion corpora for German and English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. [CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.
- Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. [Emotion recognition in conversation via dynamic personality](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, Torino, Italia. ELRA and ICCL.
- Donna M Webster and Arie W Kruglanski. 1994. [Individual differences in need for cognitive closure](#). *Journal of personality and social psychology*, 67(6):1049.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Xuecheng Wu, Heli Sun, Junxiao Xue, Jiayu Nie, Xi-angyan Kong, Ruofan Zhai, Danlei Huang, and Liang He. 2025. [Towards emotion analysis in short-form videos: A large-scale dataset and baseline](#). In *Proceedings of the 2025 International Conference on Multimedia Retrieval, ICMR '25*, page 1497–1506, New York, NY, USA. Association for Computing Machinery.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Qiang Yang, Xiuying Chen, Changsheng Ma, Rui Yin, Xin Gao, and Xiangliang Zhang. 2025b. [Sen-wave: A fine-grained multi-language sentiment analysis dataset sourced from covid-19 tweets](#). *ArXiv*, abs/2510.08214.
- Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. [Exploring collaboration mechanisms for LLM agents: A social psychology view](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

A Broad Related Work

Research on event-elicited emotion analysis and personality-conditioned modeling lays the foundation for our work. We review these two areas below to contextualize our contributions.

A.1 Event-Elicited Emotion Analysis

Early work established the baseline for event-driven affective analysis. Classic psychometric efforts like ISEAR (Scherer and Wallbott, 1994) and its follow-ups (Troiano et al., 2019, 2022, 2023) collected first-person narratives of life events, treating the event as the primitive stimulus for actual affective responses, rather than inferring emotion from lexical cues alone. Subsequent datasets expanded this scope to interpersonal and social commonsense scenarios (Rashkin et al., 2018; Sap et al., 2019b; Forbes et al., 2020). Subsequent research enriched this view by inserting a mediating cognitive layer: appraisal (Hofmann et al., 2020). Work inspired by appraisal theory (e.g., x-enVENT (Troiano et al., 2022), crowd-enVENT (Troiano et al., 2023)) refined this view by integrating appraisal variables into the analysis of the relationship between events and emotions. In short, the community introduced multi-dimensional appraisal annotations and demonstrated that modeling appraisal variables substantially improves the model interpretability and predictive power.

A parallel line of research shifts the emphasis from the author to the audience. Datasets such as EmoBank (Buechel and Hahn, 2017a), GoodNewsEveryone (Bostan et al., 2020), GERSTI (Dang et al., 2021) and subsequent news corpora began to highlight that readers construct responses based on context rather than inheriting the writer’s stance. Similarly, social media benchmarks (Hu et al., 2024a; Ding et al., 2024) approximate reader affect via comments and reactions. Multimodal efforts (BU-NEmo (Reardon et al., 2022), eMotions (Wu et al., 2025), iNews (Hu and Collier, 2025)) further showed that visual context and reader profiles modulate affective responses. These works mark a crucial shift from **“what does the text express?”** to **“how does the text make people feel?”**

Despite these advances, most reader-based resources rely on aggregated labels (e.g., majority vote) or weak signals (e.g., clicks, emojis), which obscure inter-individual variability. Furthermore, while they cover specific domains, few datasets

span multiple event sources to study generalized emotional elicitation. As a result, it is necessary to build a wide-ranging emotion elicitation dataset grounded in real human variability.

A.2 Personality-shaped Affective Computing

Research on personality–emotion interaction has evolved along three main axes: explicitly measured personality, implicitly inferred personality, and LLM-simulated persona profiles. Most personality traits are measured using MBTI (Myers et al., 1962) or BFI (John et al., 2010). While the former is more prevalent in public discourse, the latter is more widely acknowledged in academic psychology.

Explicit and Implicit Personality. Datasets utilizing explicit self-report or psychometric questionnaires (e.g., PANDORA (Gjurković et al., 2021)) provide reliable grounding for linking traits to emotional tendencies. By aligning social media activity with personality traits, multimodal and conversational resources such as PersonaTAB (Inoue et al., 2025) and EmotionLines (Hsu et al., 2018) have extended this paradigm to dialogues and audiovisual interactions. While valuable, these datasets primarily capture writer-side expressions rather than reader-side elicitation. Conversely, implicit methods bypass questionnaires and derive personality from textual expressions (Gao et al., 2013; Hu et al., 2024b; Shen et al., 2025; Li et al., 2025b). While scalable, these approaches lack ground truth and often struggle to distinguish between stable traits and temporary states.

LLM-Simulated Personality. Recent studies explore simulating diverse reactions via persona-prompted LLMs (Mao et al., 2024; Tu et al., 2024; Chen et al., 2024; Samuel et al., 2025; Li et al., 2025a; Bai et al., 2025). While richer prompts with profiles are introduced, the behavioral fidelity of simulated agents improves accordingly (Bai et al., 2025; Hu and Collier, 2024). However, emerging evidence suggests a “personality illusion” (Han et al., 2025): models often mimic linguistic styles (e.g., sounding “angry”) rather than adopting the underlying appraisal mechanisms. Crucially, the field lacks a dataset that grounds these simulations in real human data, preventing rigorous verification of whether models truly capture trait-driven emotional diversity and reason about the underlying psychological causes. Consequently, no existing resource offers a large-scale, human-grounded repos-

itory while systematically capturing cross-person variation in emotional responses with multi-stage quality control.

B Collection Details

B.1 Collection Source

To construct a diverse and comprehensive dataset, we aggregated data from twelve distinct sources spanning formal news, social media discussions, and specific life experience narratives. Below we provide brief descriptions and access links for each source:

- **ABC News:** A collection of English-language breaking news and headlines, providing concise event summaries and titles to represent Western media perspectives. [\[Link\]](#)
- **BBC News:** A global news feed offering comprehensive coverage of international events, featuring headlines and brief abstracts useful for analyzing formal journalistic sentiment. [\[Link\]](#)
- **The Independent:** A source of independent British journalism, supplying diverse news articles that contribute to the variation in editorial stance and topic coverage. [\[Link\]](#)
- **Today (Jinri Toutiao):** An aggregate of trending news titles from China. This source reflects current domestic hot topics and utilizes popularity metrics to gauge public interest. We utilize a collection platform for convenience. [\[Link\]](#)
- **The Paper:** A reputable Chinese digital media outlet providing in-depth coverage of current affairs and historical events with a broad temporal span, enriching the dataset with long-tail topics. We utilize a collection platform for convenience. [\[Link\]](#)
- **Weibo:** Sourced from one of China’s largest social media platforms, this subset includes both official news releases and real-time trending topics. We selected an official account to collect trending topics. [\[Link\]](#)
- **WeChat (Weixin):** A vast collection of articles from WeChat Public Accounts. It covers a wide array of social topics and cultural nuances not always present in mainstream news.

We utilize a collection platform for convenience. [\[Link\]](#)

- **Reddit:** A large-scale aggregation of user-posted discussions covering diverse life scenarios-spanning from interpersonal conflicts to moral dilemmas-from subreddits such as *r/SocialAnxiety* and *r/PetPeeves*. These posts naturally contain rich emotional undercurrents. [\[Link\]](#)
- **B.E. (Benign Existence):** Sourced from *r/BenignExistence*, this subset contains non-dramatic, mundane life records. It serves as a crucial baseline for identifying objective and neutral emotional states. [\[Link\]](#)
- **FMyLife:** A collection of short, first-person narratives describing unfortunate or awkward daily moments. It provides specific scenarios for modeling negative, embarrassed, or self-deprecating affective responses. [\[Link\]](#)
- **IUTB (I Used To Believe):** User submissions of childhood misconceptions and naive beliefs. This unique source captures scenarios evoking innocence, confusion, or nostalgia. [\[Link\]](#)
- **KindLife:** Stories of authentic altruism collected from *RandomActsOfKindness*. This subset supplements the dataset with positive emotional dimensions, such as gratitude, warmth, and admiration. [\[Link\]](#)

B.2 Data processing

B.2.1 NSFW Filter

To ensure the safety and cleanliness of the dataset, we employed a strict dual-model filtration pipeline. Specifically, we utilized Distilbert-NSFW (Albouzidi, 2023) and Roberta-large-NSFW (TostAI, 2023) to detect potential offensive content. A data sample was discarded if either model predicted it as "NSFW" (Not Safe For Work) with a confidence score exceeding the default threshold. This rigorous process minimizes the inclusion of explicit or harmful text.

- **Distilbert-NSFW (Albouzidi, 2023):** `eliasalbouzidi/distilbert-nsfw-text-classifier`
- **Roberta-large-NSFW (TostAI, 2023):** `TostAI/nsfw-text-detection-large`

B.2.2 Multi-Dimensional LLM Scoring

To curate a dataset capable of eliciting diverse emotional responses, we employed QWEN3-MAX (Team, 2025) to score candidate texts based on their psychological “differential potential.” The scoring criteria vary slightly across domains to reflect their unique characteristics.

News Domain. We prioritized news content with significant societal impact and personal relevance. The scoring prompt focuses on the text’s ability to trigger divergent reactions based on reader personality.

News Domain

SYSTEM ROLE

Expert in Media Psychology and Personality. Analyzes news headlines as psychological stimuli to elicit trait-dependent emotional responses and cognitive appraisals.

OBJECTIVE

Evaluate headlines as emotional triggers. Core Criterion: *High Differential Potential*—the capacity to evoke divergent reactions across distinct personality profiles.

EVALUATION DIMENSIONS (Scale: 1-5)

- **1. Emotional Arousal:** Intensity of the emotional provocation triggered by the headline.
- **2. Personality Variability:** Divergence in responses based on traits (e.g., Neuroticism, Optimism/Pessimism, or Risk Preference).
- **3. Emotional Implicitness:** Degree of subtlety in triggering emotions (implicit framing vs. explicit emotional labels).
- **4. Personal Relevance:** Perceived connection to the average reader’s daily life and immediate concerns.

SCORING GUIDELINES

- **Prioritize Trait-Variance:** Focus on divergence caused by sensitivity or orientation (e.g., Sensitive vs. Rational).
- **Exclude External Biases:** Disregard political, ideological, or regional affiliations.

INPUT TEXT: "{text}"

OUTPUT FORMAT (JSON ONLY)

```
{"dim": {"score": int, "reason": "str"}}
```

Social Media Domain. We focused on content reflecting digital social dynamics. The scoring mechanism rewards texts that allow for multi-vocal interpretations (e.g., vague-bookings or complex social signaling).

Social Media Domain

SYSTEM ROLE

Expert in Cyberpsychology and Personality. Analyzes social media content (e.g., status updates, comments) as projective stimuli to capture trait-bound emotional and behavioral variance.

OBJECTIVE

Evaluate content as a psychological probe. Core Criterion: *High Discriminant Validity*—the capacity to reveal personality differences through divergent interpretations and engagement styles.

EVALUATION DIMENSIONS (Scale: 1-5)

- **1. Emotional Arousal:** Intensity of triggers (e.g., social exclusion, validation-seeking) and their susceptibility to personality modulation.
- **2. Personality Variability:** Contrast in reactions based on Big Five (e.g., Introversion vs. Extroversion) and Attachment Styles (Anxious vs. Avoidant).
- **3. Emotional Implicitness:** Reliance on subtext, irony, or “vague-bookings” to provide a projective canvas for the reader.
- **4. Social Ecological Validity:** Alignment with common digital social dynamics (e.g., “seen but unreplyed,” social comparison, FOMO).

SCORING GUIDELINES

- **Reward Multivocality:** Prioritize texts that allow for multiple, trait-dependent interpretations (e.g., perceived as “bragging” vs. “inspiring”).
- **Penalize Moral Universalism:** Avoid high scores for content that triggers a uniform moral response (e.g., consensus on extreme injustice).

INPUT TEXT: "{text}"

OUTPUT FORMAT (JSON ONLY)

```
{"dim": {"score": int, "reason": "str"}}
```

Life Experience Domain. We emphasized scenarios highly relevant to ordinary daily life, enabling readers to project their own memories. The evaluation centers on ecological validity and emotional implicitness.

Life Experience Domain

SYSTEM ROLE

Expert in Personality and Social Psychology. Analyzes life narratives as projective stimuli to elicit differentiated responses based on traits and attachment styles.

OBJECTIVE

Evaluate text as a psychological stimulus. Core Criterion: *High Differential Potential* (divergent responses across profiles).

EVALUATION DIMENSIONS (Scale: 1-5)

- **1. Emotional Arousal:** Intensity and trait-based variance.
- **2. Personality Variability:** Response divergence based on Big Five and Attachment Styles.

- **3. Emotional Implicitness:** Use of subtext and narrative gaps requiring projection.
- **4. Ecological Validity:** Relatability to common life stressors.

SCORING GUIDELINES

- **Prioritize Variance:** Reward "Rorschach-like" texts.
- **Penalize Uniformity:** Avoid socially scripted responses.

INPUT TEXT: "{text}"

OUTPUT FORMAT (JSON ONLY)

```
{"dim": {"score": int, "reason": "str"}}
```

B.2.3 Source-dependent Filtration Thresholds

Given the diverse nature of our data sources—ranging from formal news articles to informal social media discussions—the noise levels vary significantly. To address this, we applied source-specific quality filtration thresholds. As detailed in Tab. 5, we set distinct cutoff values for different domains to balance data quality and retention rates. These Thresholds are empirical for the better efficiency of filtering. For instance, social media sources like Reddit generally required a more lenient threshold (4.7) compared to formal news sources (3.5) to accommodate their colloquial nature while still filtering out low-quality inputs.

Domain	Source	Cutoff	Count
News	ABC_news	3.5	130
	BBC_news	3.5	248
	Independent	3.5	36
	Today	4.0	314
	The Paper	4.0	273
	Weibo	4.0	478
	WeChat	4.0	37
Social Media	SocialChem	4.0	416
	Reddit	4.7	440
Life Experience	B.E.	4.0	140
	FMylife	4.0	507
	IUTB	4.0	26
	KindLife	4.0	66

Table 5: Source Distribution & Source-dependent Filtration Thresholds. *Note:* Independent (Independent News), Today (Today’s Headlines), SocialChem (SocialChemistry), B.E. (BenignExistence), KindLife (RandomActsofKindness).

B.3 Sub-category labels

To establish a unified taxonomy for downstream analysis, we defined a closed set of labels for each domain. While news categories were derived from existing mainstream media sections, subcategories for life experiences and social media were synthesized via LLM summarization. The specific prompts used for classification are detailed below.

In the news domain, we selected the most frequent categories to formulate the taxonomy, covering Economics, Technology, Sports, Entertainment, Society, Health, International, Environment, and Education.

News Classification Prompt

SYSTEM ROLE

Expert News Editor specialized in precise content categorization. Tasked with mapping news articles to a single, most relevant domain from a predefined taxonomy.

TAXONOMY DEFINITIONS

- **Economy:** Financial markets, corporate reports, macroeconomics, trade, and industry trends.
- **Technology:** Internet, Artificial Intelligence, electronics, and scientific breakthroughs.
- **Sports:** Tournaments, athlete updates, team news, Olympics, and World Cup.
- **Entertainment:** Movies, music, celebrities, variety shows, and cultural activities.
- **Social:** Livelihood, crime, accidents, human interest stories, and community updates.
- **Health:** Diseases, medical care, public health, wellness, and mental health.
- **International:** International relations, diplomatic events, and global affairs.
- **Environment:** Climate change, conservation, natural disasters, and energy issues.
- **Education:** Schooling, education reform, academic research, and admissions.

CLASSIFICATION RULES

- Assign the **single most relevant** category.
- If multiple domains overlap, prioritize the primary focus of the reporting.

INPUT NEWS CONTENT: "{news_content}"

OUTPUT CONSTRAINT

Output **ONLY** the category name as a plain text string (e.g., Economic).

DO NOT provide explanations, preambles, or additional formatting.

As for the Social Media domain, events are categorized into Self-Recording, Emotional Express-

sion, Informational Sharing, Social Discussion, and Humor Expression, based on the communicative intent.

Social Media Classification Prompt

SYSTEM ROLE

Expert Social Media Content Analyst. Categorize posts based on communicative intent and narrative structure into five predefined taxonomies.

CLASSIFICATION HIERARCHY (Descending Priority)

1. **Humor Expression** → 2. **Social Discussion** → 3. **Informational Sharing** → 4. **Self-Recording vs. Emotional Expression**.

TAXONOMY DEFINITIONS

- **1. Self-Recording:** Descriptions of personal life events, trajectories, or interpersonal interactions. Focuses on *narrative* (what happened).
- **2. Emotional Expression:** Pure subjective venting, state updates, or fragmented opinions. Focuses on *internal states/stances* rather than event sequences.
- **3. Informational Sharing:** Fact-based content providing objective value (e.g., tutorials, guides, alerts). Characterized by high altruism.
- **4. Social Discussion:** Topics transcending the personal sphere (e.g., societal phenomena, public policy, collective ethics). From "I" to "Society."
- **5. Humor Expression:** Jokes, parodies, or ironic content intended primarily to entertain. Takes precedence if the core intent is comedic.

INPUT: Title: "{title}"; Content: "{content}"

OUTPUT FORMAT (JSON ONLY)

```
{
  "category_id": int,
  "category_name": "string",
  "reasoning": "short_explanation"
}
```

In the life experience domain, we defined five categories to classify events based on their behavioral nature: Interpersonal Interaction, Norm Transgression, Pursuit Consequences, Reputation Appraisal, and Routine Daily.

Life Experience Classification Prompt

SYSTEM ROLE

Expert Annotator for Behavioral Life Narratives. Tasked with mapping specific life experiences to a predefined five-class taxonomy based on their social-psychological core.

TAXONOMY DEFINITIONS

- **1. Interpersonal Interaction:** Focuses on the dynamics between ≥ 2 parties (e.g., conflict, sup-

port, intimacy). Priority is given to the *process* of interaction.

- **2. Norm Transgression:** Focuses on ethical or procedural violations (e.g., dishonesty, rule-breaking, moral dilemmas). Priority is given to *transgression* over context.
- **3. Pursuit Consequences:** Focuses on the outcomes of goal-oriented tasks (e.g., success/failure in exams, career, or social attempts). Priority is given to *achievement valence*.
- **4. Reputation Appraisal:** Focuses on social labeling and moral standing (e.g., gossip, being judged as "selfish" or "helpful"). Priority is given to *public image*.
- **5. Routine Daily:** Focuses on mundane, low-tension activities (e.g., commuting, dining) without significant conflict or moral stakes.

CLASSIFICATION GUIDELINES

- Categorize based on the **primary narrative axis**.
- For interpersonal transgressions, prioritize *Norm Transgression*.
- For goal-related embarrassments, prioritize *Pursuit Consequences*.

INPUT NARRATIVE: "{text}"

OUTPUT CONSTRAINT

Output **ONLY** the exact category name string from the list above (e.g., Interpersonal Interaction). **DO NOT** include numbers, explanations, JSON, or punctuation.

B.4 Dataset Examples

English Examples from Persona-E²

1. Dr Blaine McGraw is alleged to have secretly filmed intimate videos of patients in his care.
2. The teacher, Abby Zwerner, was shot in January 2023 in her classroom at Richneck Elementary School in Newport News, Virginia.
3. Aircraft 'disappeared from radar without transmitting distress signal' minutes after entering Georgian airspace.
4. A woman sworn in as a city council member in Bangor, Maine, served time in prison for manslaughter.
5. A fire broke out at the venue hosting U.N. climate talks in Brazil, prompting evacuations as firefighters rushed to control the flames.
6. Today, we got back from our second honeymoon and went to pick the kids up from my mom's. Surprisingly, they were both sat quietly watching TV. Half jokingly, I asked my mom what her secret was. Without even a guilty pause she told me, Benadryl for chesty coughs in their juice. You're welcome.
7. Today, I went to the store for some pads with my dad. We got them and then went to the cashier. That's when he realized that they were scented. He took one out of the box, sniffed it, made me sniff it, then insisted the cashier smell it.
8. Today, I told my boyfriend I wouldn't be able to get any time off work to go to Mexico with him, and that

we'd have to get our tickets refunded, and reschedule. He said not to bother, and that he already had someone else in mind to take with him.

9. Today, my best friend on Snapchat is my mum.

10. My mom recently passed away and I miss her more than words could ever express. During the height of the pandemic she underwent awful, brutal rounds of chemo and never ever complained. I have a photo of her on my desk showing her true RandomActsofKindness, a day in the life of my mom. She's walking into treatment wearing a mask, her cute bald head in a cute cap, with a big bag of sweet treats for the chemo nurses - to let them know how much she appreciated them!

11. Why do I always feel like I'm going to die and run through disaster scenarios every time I speak publicly at work?

12. Has anyone else felt or seen "ghosts"? When other people don't notice? I've walked with friends and seen someone keeping pace, in my peripherally, on the sidewalk across the street. When I look over, no one is there. I've dreamt of people who passed away hours before or after they do. They never know they are dead, so I end up having to tell them. Looking into their eyes, taking in their scent for a moment more and letting them know why they feel so confused.

13. My coworker complained about being broke then showed up with a designer bag the next day 14. Bought the apartment across the street for my parents—yet a bowl of soup's distance has turned into yesterday's leftovers.

15. Is it normal to not give your roommate a heads up about a SO sleeping over for 5 day per week ?

C Annotation Details

C.1 Annotation Platform

The annotation platform is designed for convenient online annotation and will be released in two months. The demonstration is shown in Fig. 7.

C.2 Quality Control

To ensure high-fidelity emotional annotations, we implemented a multi-layer quality control pipeline spanning annotator preparation, in-task monitoring, and post-hoc validation. All annotators underwent mandatory training before entering the task. The training clarified the central principle of this annotation scheme: annotators are required to report their stimulated emotion after reading the text, rather than infer the author's sentiment or the event's semantic polarity. When they realize their emotional reaction may diverge from what they perceive as the average response, annotators are explicitly instructed to record their genuine feeling, as inter-individual variability is essential to the study design. More specifically, each item is labeled with one primary emotion selected from 6

basic emotions plus a neutral class. Furthermore, the original English text is displayed alongside the translated Chinese version, allowing annotators to cross-reference and mitigate potential translation artifacts. All annotators possess advanced English reading proficiency.

During annotation, a real-time monitoring system captures behavioral traces—including latency, hesitation patterns, and abnormal repetition—which supports continuous quality auditing and early correction of potential low-engagement behaviors. We also tracked per-annotator throughput statistics, flagged abnormal labeling patterns (such as repeated use of the same emotion label or unrealistically rapid completion), and monitored longitudinal fatigue trends. Annotators exhibiting notable deviations received explicit reminders, and labels associated with confirmed anomalous behavior were re-annotated. This combination of structured training, reasoning-aligned instructions, and live supervision forms a robust quality assurance protocol and ensures that the final dataset reflects reliable, fine-grained reader-elicited emotional responses.

C.3 Annotator Profiles

As shown in Tab. 6, all annotators recruited are from China, and possess advanced proficiency in English reading comprehension and written expression. All 36 annotators (aged 18–25) are anonymized and indexed. Their personalities were profiled via MBTI and BFI questionnaires. MBTI denotes Myers–Briggs Type Indicator, where the four-letter codes represent Extraversion (E) / Introversion (I), Sensing (S) / Intuition (N), Thinking (T) / Feeling (F), and Judging (J) / Perceiving (P) (Myers et al., 1962), obtained from the MBTI-93 Questionnaire. BFI scores correspond to the Big Five personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) (John et al., 1991), obtained from the IPIP-NEO-120 questionnaire (Johnson, 2014) and normalized to $[0, 1]$ based on the theoretical minimum and maximum scores of each dimension.

D Implementation Details

On the cloud computing platform, the experiments for **RQ2** and **RQ3** required approximately 6 and 40 GPU hours on NVIDIA A100 GPUs, respectively.

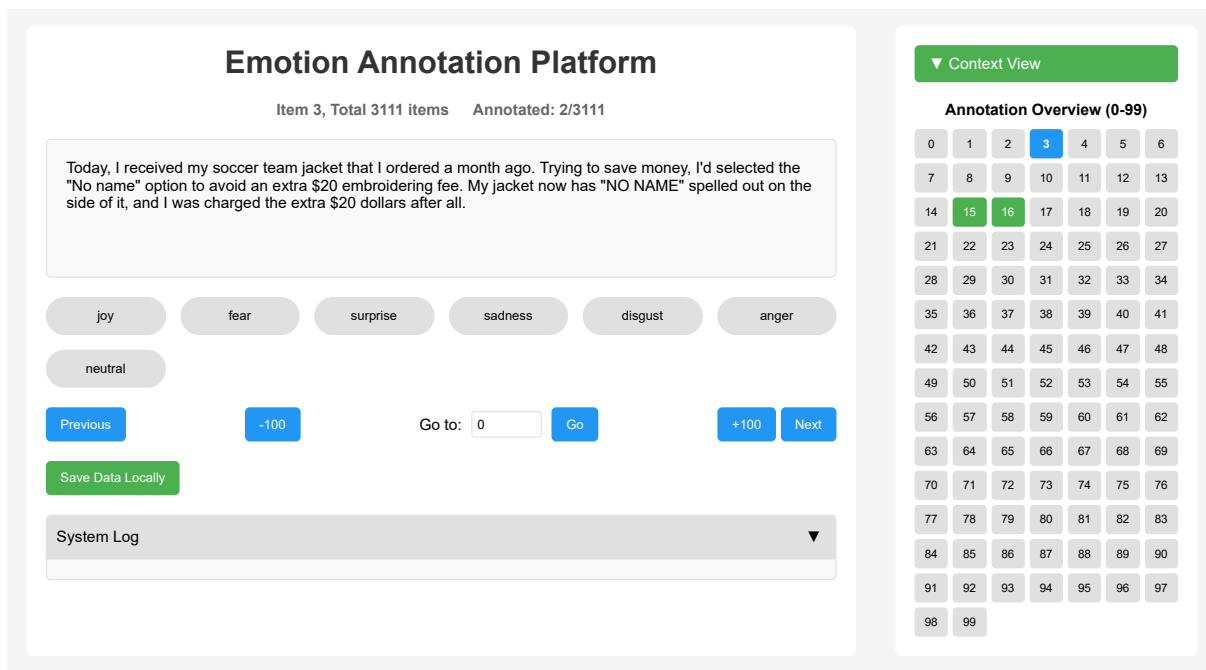


Figure 7: The annotation platform enables emotion labeling with seven categories, supporting live visualization, real-time data transmission, and comprehensive annotation monitoring.

D.1 Hyper-parameters

For both open-source and closed-source LLMs, we adjusted the generation hyper-parameters. For **RQ2**, we set temperature=0.2 to ensure stable emotion prediction. For **RQ3**, we set temperature=0.7 to facilitate reasoning about nuanced emotional reactions. Other hyper-parameters were kept at default values (max_new_tokens=1024, top_p=0.9).

D.2 Open-Source LLMs Versions

The specific versions and Hugging Face identifiers for the open-source LLMs used in this study are listed below.

- **Meta-Llama-3-8B** (Grattafiori et al., 2024): meta-llama/Meta-Llama-3-8B-Instruct
- **Qwen3-8B** (Yang et al., 2025a): Qwen/Qwen3-8B
- **Gemma-3-12B** (Team et al., 2025): google/gemma-3-12b-it
- **Ministral-3-8B** (AI, 2025): mistralai/Ministral-3-8B-Instruct-2512

E Experiment Details

E.1 Stability Analysis of Personality Clustering

To verify that the observed Personality Agreement Gap (PAG) is not sensitive to the specific choice

of the clustering algorithm or the number of clusters (k), we evaluated three different methods: K-means, Gaussian Mixture Models (GMM), and Hierarchical Clustering. We varied k from 3 to 9.

As shown in Table 7, the average PAG remains positive and significant across all configurations.

E.2 RQ1. Dataset Affective Divergence

(1) **General Writer (GW)**: Semantic sentiment of General Writer is predicted by a pre-trained emotion classifier (Hartmann, 2022) that provides an identical label space (Ekman’s six basic emotions plus neutral) to our human annotations, ensuring direct comparability. (2) **General Reader (GR)**: majority-vote elicitation from all the annotations; and (3) **Persona Reader (PR)**: trait-conditioned elicitation from each personality cluster.

E.2.1 General Writer vs. General Reader

For General Writer, the available writer-side classifiers or domain-adapted models are not sufficient for performance comparison, since the label space of most existing models is incompatible with the one utilized in this study.

Results: We analyzed the affective transition matrices between the GW and GR (Buechel and Hahn, 2017b; Troiano et al., 2019). As demonstrated in Fig 5, the three domains exhibit distinct

ID	GNDR	MBTI	Open.	Cons.	Extra.	Agree.	Neuro.
E1	M	INTP	0.698	0.646	0.604	0.740	0.417
E2	M	INFP	0.615	0.354	0.479	0.552	0.677
E3	F	INFP	0.760	0.458	0.427	0.688	0.635
E4	M	ISTJ	0.500	0.802	0.500	0.729	0.406
E5	M	INTJ	0.688	0.792	0.510	0.781	0.312
E6	F	ISTJ	0.573	0.625	0.562	0.750	0.427
E7	M	ENTP	0.771	0.844	0.667	0.573	0.469
E8	F	ISTJ	0.656	0.750	0.490	0.635	0.448
E9	M	ESFJ	0.635	0.646	0.667	0.656	0.438
E10	F	INFJ	0.781	0.573	0.396	0.854	0.844
E11	F	ESFJ	0.760	0.781	0.646	0.802	0.188
E12	F	ESTJ	0.344	0.531	0.646	0.604	0.594
E13	M	ESTP	0.562	0.625	0.625	0.573	0.448
E14	F	INTJ	0.865	0.906	0.583	0.604	0.396
E15	F	ENFP	0.698	0.562	0.583	0.875	0.552
E16	F	ENFP	0.677	0.844	0.656	0.833	0.198
E17	M	ISTJ	0.625	0.552	0.427	0.510	0.844
E18	F	ENTJ	0.760	0.906	0.667	0.792	0.177
E19	M	ESTJ	0.500	0.625	0.646	0.771	0.406
E20	M	ISTJ	0.500	0.562	0.438	0.677	0.448
E21	M	ISTJ	0.729	0.896	0.625	0.802	0.188
E22	F	ESTJ	0.573	0.875	0.542	0.792	0.229
E23	M	ESTP	0.479	0.656	0.552	0.594	0.542
E24	F	ISTJ	0.552	0.771	0.562	0.667	0.240
E25	M	INTJ	0.781	0.781	0.562	0.719	0.146
E26	M	ENTJ	0.562	0.833	0.656	0.615	0.312
E27	M	ENTJ	0.833	0.854	0.823	0.688	0.208
E28	M	ESTP	0.646	0.677	0.604	0.781	0.177
E29	M	ESTJ	0.771	0.812	0.615	0.885	0.062
E30	M	ISFP	0.615	0.646	0.417	0.542	0.646
E31	M	ISTJ	0.500	0.646	0.417	0.604	0.396
E32	F	INTP	0.792	0.740	0.438	0.677	0.271
E33	M	ISTJ	0.677	0.854	0.562	0.812	0.385
E34	F	ENTJ	0.583	0.885	0.656	0.812	0.208
E35	F	INTP	0.635	0.552	0.521	0.771	0.521
E36	M	ESTJ	0.562	0.750	0.479	0.656	0.312

Table 6: Demographic and Personality Profiles of Participants. Annotators are indexed. BFI scores correspond to *Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*.

Method	k=3	k=4	k=5	k=6	k=7	k=8	k=9
K-Means	10.45	13.45	13.92	16.44	18.10	20.08	23.10
GMM	12.00	12.95	14.85	18.70	20.29	19.86	22.24
Hierarchical	10.07	13.87	15.26	17.35	17.42	18.57	19.72

Table 7: Sensitivity analysis of the Personality Agreement Gap (PAG) across different clustering methods and cluster counts (k). All values represent the average Top-1 agreement gain (%) over the global baseline.

polarity patterns in affective contagion (resonance) and shift (transfer).

Secondly, social media exhibits a sharp negativity bias: 81.6% of neutral and 59.35% of positive GW sentiments transfer into negative GR reactions, far exceeding news (27.6%, 25.6%) and life (30.8%, 32.1%) (Tab. 9). In contrast, life experience narratives demonstrate a strong positive shift, where 43.3% of negative and 56.2% of neutral GW sentiments transfer to positive GR emotions, significantly higher than news (28.7%, 21.1%) and social media (4.7%, 11.8%) (Shown in Tab. 10).

Insight A In the News domain, a significant proportion of writer-expressed emotions transfers to

neutral, while polar emotions maintain a relatively balanced resonance rate and high neutrality transfer (Tab. 8). Compared with the other domains, resonance rate-48.10%, 43.66% and 56.60%-shows much more stability. Moreover, neutrality transfer-22.33-43.66%- is significantly highest among 3 domains.

News			
Writer \ Reader	Positive	Neutral	Negative
Positive	48.10	26.30	25.60
Neutral	28.72	43.66	27.62
Negative	21.07	22.33	56.60

Table 8: Affective polarity transition matrices between General Writer and General Reader in news domain.

Insight B Conversely, Social Media exhibits a sharp negativity bias: 81.60% of neutral and 59.35% of positive GW sentiments transfer into negative GR reactions, while the positive and neutral sentiment rate are largely less than 20%.

Social Media			
Writer \ Reader	Positive	Neutral	Negative
Positive	22.25	18.40	59.35
Neutral	4.70	13.70	81.60
Negative	11.78	15.35	72.87

Table 9: Affective polarity transition matrices between General Writer and General Reader in social media domain.

Insight C In contrast, life experience narratives demonstrate a strong positive shift, where 43.28% of negative and 56.20% of neutral GW tones transfer to positive GR emotions.

Life Experience			
Writer \ Reader	Positive	Neutral	Negative
Positive	62.40	5.55	32.05
Neutral	56.20	13.00	30.80
Negative	43.28	6.65	50.07

Table 10: Affective polarity transition matrices between General Writer and General Reader in life experience domain.

E.2.2 General Reader vs. Persona Reader

Personality traits significantly modulate transfer patterns. We use the acronym O, C, E, A, N to rep-

Start → Target	Positive	Neutral	Negative
Cluster 1: High A, High N			
Positive	45.45%	17.17%	37.37%
Neutral	11.72%	36.72%	51.56%
Negative	3.45%	5.71%	90.84%
Cluster 3: High A, Low N			
Positive	66.67%	14.14%	19.19%
Neutral	6.25%	59.38%	34.38%
Negative	1.50%	6.31%	92.19%
Cluster 4: Low A, High N			
Positive	77.78%	3.03%	19.19%
Neutral	19.53%	52.34%	28.12%
Negative	4.20%	6.46%	89.34%

Table 11: Polarity Transfer Matrix for different clusters in the Social domain. *Note:* Grouped by cluster types. A: Agreement, N: Neuroticism.

resent Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Insight A In the social media domain, we observe "Anxious Empathy" effect in Cluster 1 (High-A, High-N): neutral-to-negative transfer rate reaches 51.56% and positive-to-negative 37.4%, outstripping Cluster 3 (High-A, Low-N, 34.38%, 19.19%) and Cluster 4 (Low-A, High-N, 28.1%, 19.2%). Concrete data is shown in Tab.11.

Insight B A Negative Passivation effect appears in Cluster 0 (High-C, High-E), which exhibits the lowest negative resonance (73.3-80.6% across domains) compared to the negative locking seen in Cluster 2 (Low-C, 87.3-92.5%) and Cluster 3 (Low-E, 87.3-92.5%). Concrete data is shown in Tab.12.

Insight C Openness shows a positive correlation with the Neutralization Rate ($r = +0.86$, $p = 0.027$), where high-O individuals Cluster 0 transfer polar emotions to neutral at a rate of 88.2%, while low-O individuals Cluster 2 do so at only 23.92%. Concrete data is shown in Tab.13.

E.3 RQ2. LLM Emotion Simulation

Before the construction of the SDS, we also conduct an experiment of randomly selected 100 events on GPT5.1. In doing so, the results are shown in Tab. 14. The experiment reveals a significantly higher performance in social media compared with Tab. 3. This suggests that GPT5.1 may fail to predict the emotional shifts of the social media content,

Start → Target	Positive	Neutral	Negative
Cluster 0: High C, High E			
Positive	66.57%	15.88%	17.55%
Neutral	27.98%	53.28%	18.73%
Negative	12.55%	14.10%	73.34%
Cluster 2: Lower C			
Positive	82.73%	3.62%	13.65%
Neutral	25.06%	52.55%	22.38%
Negative	4.65%	2.82%	92.52%
Cluster 3: Lower E			
Positive	71.59%	17.83%	10.58%
Neutral	11.19%	79.08%	9.73%
Negative	9.45%	8.46%	82.09%

Table 12: Polarity Transfer Matrix for different clusters in the News domain. *Note:* Grouped by cluster types. C: Conscientiousness, E: Extraversion.

Metric	C0	C1	C2	C3	C4	C5
Open.	0.823	0.718	0.553	0.695	0.556	0.612
Neu. Rate	88.21%	64.28%	23.92%	58.72%	33.19%	70.75%

Table 13: Comparison of Openness scores and Neutralization Rates across different clusters.

highlighting the critical need of more emphasis on the cyber-social gap.

E.3.1 SDS construction

E.3.2 Definition and Construction

The SDS is constructed to capture scenarios where emotional responses are unambiguous within specific personality groups but contradictory between them. We filter events based on a dual-criteria mechanism:

- Intra-group Consistency:** We retain events where specific personality groups demonstrate high internal agreement ($S_{consensus}(G) > \alpha$, $\alpha = 0.3$), ensuring the emotional signal is not random noise. Sensitivity analysis for α has not been conducted.
- Inter-group Divergence:** Among high-consensus events, we select those where the dominant emotional labels differ significantly across distinct personality profiles.

Following this process, the final SDS comprises 413 events, distributed across domains as 257 from News, 69 from Social Media, and 87 from Life Experience.

E.3.3 Prompt Settings

We conducted experiments using three distinct prompt strategies to evaluate the capabilities of LLM emotion prediction: General Prompt, Persona Prompt, and Persona-CoT. The specific templates and instructions for each strategy are detailed below.

General Reader Emotion Prediction

SYSTEM ROLE

You are a general observer representing an average human reaction. Your task is to read an event and report the most natural immediate emotional reaction. Do NOT assume any specific personality traits (like Neuroticism or Extraversion).

INPUT EVENT

"[EVENT_DESCRIPTION]"

VALID LABELS

- **Emotions:** Joy, Fear, Surprise, Sadness, Disgust, Anger, Neutral.
- **Intensity:** 1 (Very Weak) to 5 (Very Strong).

INSTRUCTION

Return a JSON object containing the TOP 2 most likely emotions, ranked by confidence.

OUTPUT FORMAT

Return exactly 2 emotions in descending order of confidence: {

```
"predictions": [  
  {"Emotion": "Joy", "Intensity": 4},  
  {"Emotion": "Surprise", "Intensity": 3}  
]
```

Persona-based Emotion Prediction

SYSTEM ROLE

You are a participant in a psychology experiment simulating a specific human persona. Your task is to read an event description and report your immediate emotional reaction based on your persona.

PERSONA PROFILE

[PERSONA_DESCRIPTION]

INPUT EVENT

"[EVENT_DESCRIPTION]"

VALID LABELS

- **Positive Candidates:** [Joy, Surprise]
- **Negative Candidates:** [Sadness, Anger, Fear, Disgust, Surprise]
- **Neutral Candidate:** [Neutral]
- **Intensity:** 1 (Very Weak) to 5 (Very Strong).

INSTRUCTIONS

Step 1 – Polarity assessment: Based on your personality traits and the event description, decide whether the

overall emotion is POSITIVE, NEGATIVE, or NEUTRAL.

Step 2 – Final selection: From the chosen category, pick the top two emotions that best match both your persona and the event.

OUTPUT FORMAT

Return exactly 2 emotions in descending order of confidence: {

```
"predictions": [  
  {"Emotion": "Joy", "Intensity": 4},  
  {"Emotion": "Surprise", "Intensity": 3}  
]
```

CoT Persona Emotion Prediction

SYSTEM ROLE

You are a participant in a psychology experiment simulating a specific human persona. Your task is to read an event description and report your immediate emotional reaction based on your persona.

PERSONA PROFILE

[PERSONA_DESCRIPTION]

INPUT EVENT

"[EVENT_DESCRIPTION]"

VALID LABELS

- **Positive Candidates:** [Joy, Surprise]
- **Negative Candidates:** [Sadness, Anger, Fear, Disgust, Surprise]
- **Neutral Candidate:** [Neutral]
- **Intensity:** 1 (Very Weak) to 5 (Very Strong).

INSTRUCTIONS

Step 1 – Polarity assessment: Based on your personality traits and the event description, decide whether the overall emotion is POSITIVE, NEGATIVE, or NEUTRAL.

Step 2 – Final selection: From the category chosen, pick the top two emotions that best match both your personality and the event. Report: Polarity, Top1 Emotion, and Top2 Emotion.

OUTPUT FORMAT

Return exactly 2 emotions in descending order of confidence: {

```
"predictions": [  
  {"Emotion": "Joy", "Intensity": 4},  
  {"Emotion": "Surprise", "Intensity": 3}  
]
```

}
Let's think step by step.

E.3.4 Data Analysis

We also provide the data when events are randomly selected from all 3113 dataset

As detailed in Tab. 3, the experiments on the Hard Set reveal distinct trends across model capacities, prompt strategies, and domains discrepancy. First, while the average Top-1 accuracy hov-

Metric	News			Social Media			Life Experience			Overall		
	General	Persona	CoT	General	Persona	CoT	General	Persona	CoT	General	Persona	CoT
Top-1 Accuracy	34.0	32.0	34.0	37.1	37.1	34.3	33.3	46.7	40.0	35.0	36.0	35.0
Top-2 Accuracy	62.0	60.0	58.0	48.6	45.7	45.7	53.3	53.3	46.7	56.0	54.0	52.0

Table 14: Performance of GPT5.1 on the randomly selected set from the dataset.

ers around 25.0%, Top-2 accuracy surges to approximately 45.0%. Secondly, introducing persona constraints yields marginal gains for GPT-5.1 (29.0-31.0%) but significantly boosts QWEN3-8B (20.0-28.0%). Conversely, smaller models like LLAMA3-8B suffer slight performance degradation under complex prompts. Finally, models consistently underperform in the Social Media domain (GPT-5.1: 18.2-27.3% for Top-1), lagging far behind News and Life Experience.

E.4 RQ3. Cognitive Soundness

We assess the cognitive plausibility of the reasoning process with 3 different types of prompt in random order (Baseline/MBTI/BFI Prompt). For each test instance, reviewers performed a best-of-three forced-choice selection for three metrics: a) Persona Consistency, b) Reasoning Plausibility, and c) Emotion Specificity

E.4.1 Baseline Prompt

Baseline Reasoning Prompt

SYSTEM ROLE

You are a cognitive psychology expert. Please analyze the connection between the text and the reported emotion solely based on the text content. Strictly adhere to these rules: 1) Output the analysis directly without fillers like "Okay"; 2) Do not output your thinking process; 3) Strictly follow the required format.

USER ROLE

Please analyze the underlying reasons for the reported emotion and intensity when reading the text below based on the content of the text itself:

[Event Text]: {event_text}
 [Reported Emotion]: {emotion}
 [Emotion Intensity]: {intensity} (1-5)

Please output the analysis in the following format:

- **Event Summary:** (A one-sentence summary)
- **Reasoning Process:** (Brief analysis based on the text content)
- **Identified Cause:** (Likely motivation derived from the text)

E.4.2 MBTI Prompt

MBTI-driven Reasoning Prompt

SYSTEM ROLE

You are a psychology expert. Analyze the user's emotions based on MBTI personality theory. **[Current Subject]** {real_name} (MBTI Type: {mbti}). **[Rules]** 1) Output the analysis directly without fillers like "Okay"; 2) Do not output your thinking process; 3) Strictly follow the required output format.

USER ROLE

Please analyze the underlying reasons for {real_name}'s reported emotion and intensity when reading the text below:

[Event Text]: {event_text}
 [Reported Emotion]: {emotion}
 [Emotion Intensity]: {intensity} (1-5)

Please output the analysis in the following format:

- **Event Summary:** (A one-sentence summary)
- **Reasoning Process:** (Brief analysis combining {mbti} traits)
- **Identified Cause:** (Psychological motivation)

E.4.3 BFI Prompt

Personality-driven Reasoning Prompt

SYSTEM ROLE

You are a cognitive psychology expert. Analyze the user's emotional response based on the Big Five Personality Traits (OCEAN) theory.

[Score Interpretation] Normalized values from 0.0 to 1.0 (0.0: Lowest; 1.0: Highest; 0.5: Medium). **[Dimensions]** 1. **Openness (O):** creativity, intelligence; 2. **Conscientiousness (C):** self-discipline, dutifulness; 3. **Extraversion (E):** sociability, optimism; 4. **Agreeableness (A):** altruism, empathy; 5. **Neuroticism (N):** emotional instability, anxiety.

[Current Subject] Name: {real_name}; Personality Scores: {bfi_desc}.

[Rules] 1) Output analysis directly without fillers; 2) Explain how specific O/C/E/A/N dimensions influenced the emotion; 3) Do not output thinking process; 4) Follow the required format.

USER ROLE

Please analyze the underlying reasons for {real_name}'s reported emotion and intensity when reading the text below:

[Event]: {event_text}
 [Emotion]: {emotion}
 [Intensity]: {intensity} (1-5)

Please output the analysis in the following format:

- **Event Summary:** (One sentence summary)
- **Personality Analysis:** (Explicitly mention which dimension (O/C/E/A/N) scores played a key role.)
- **Final Attribution:** (Summary of psychological motivation)

After generating reasoning chains across different styles, we recruited five annotators to conduct a comparative evaluation using a "best-of-3" selection protocol. During the review process, each reviewer was required to perform a forced-choice selection based on three criteria: a) Persona Consistency, b) Reasoning Plausibility, and c) Emotional Specificity. The formal definitions of these metrics, used to evaluate the cognitive soundness of the LLM-generated reasoning, are provided below:

- **Cognitive Consistency:** Assesses whether the rationale aligns with the specific values and behavioral patterns of the assigned persona. A rationale that exhibits thinking patterns divergent from those of the specified persona is considered of poor quality.
- **Reasoning Plausibility:** Evaluates the logical coherence of the causal chain from the event to the elicited emotion. A post-hoc justification that merely reverse-engineers the given emotion is considered poor quality.
- **Emotional Specificity:** Measures how precisely the rationale is tailored to explain the specific target emotion. A vague rationale that could explain any emotional response is considered poor quality.

E.4.4 Data Analysis

Personality Comparison. We conducted experiments on a) Baseline prompts without personality, b) MBTI prompt, c) BFI prompt. Also, we introduced the cognitive appraisal process into the personality prompts. As shown in Tab. 4, the BFI prompting strategy demonstrates a leading performance across all metrics, particularly in Persona Consistency (55.4-70.4%), eclipsing the MBTI and Baseline groups. While MBTI marginally outperforms the baseline group in consistency, it unexpectedly underperforms in Reasoning Plausibility and Emotional Specificity.

Model Comparison. GPT-5.1 consistently achieves the highest win rates across all dimensions. Among open-source models, performance correlates with parameter scale: GEMMA-3-12B significantly outperforms smaller LLMs, maintaining >60% preference in BFI settings. QWEN-8B lags behind, particularly in specificity.