From Static Traits to Dynamic Affect: A Reasoning-and-Rewriting Framework for Personality Dialogue

Anonymous ACL submission

Abstract

While large language models (LLMs) have achieved remarkable fluency, they often struggle to maintain stable personas and exhibit an inherent "neutrality bias" resulting in emotionally flat and insufficiently personalized interactions. To address these challenges, we propose a novel personality-driven dialogue framework that explicitly models the reasoning chain from Big Five traits to affective realization. Specifically, we decouple the generation process into three distinct stages: (1) personality-aware emotion prediction, which induces target affective states from structured trait profiles and conversational context; (2) affective refinement, where an initial response is rewritten via direct preference optimization (DPO) to amplify emotional intensity; and (3) external verification, which closes the loop through an independent emotion recognition module. Our framework effectively compensates for the affective limitations of base LLMs while preserving their semantic coherence. Experimental results on personality-annotated corpora demonstrate that our approach significantly outperforms strong baselines in both emotional fidelity and personality consistency, as validated by automatic metrics and human evaluation.

1 Introduction

The rapid advancement of LLMs has revolutionized both task-oriented and open-domain dialogue systems, enabling the generation of fluent and contextually relevant responses. However, a persistent gap remains in achieving true "human-likeness": most systems are perceived as emotionally flat (Wang et al., 2025b; Gandhi and Gandhi, 2025) and lack a consistent persona (Xie et al., 2024) . While human interaction is intrinsically driven by the interplay between personality (stable individual traits) and emotion (transient affective states), current models often struggle to maintain this alignment, leading to disjointed or generic user experiences.

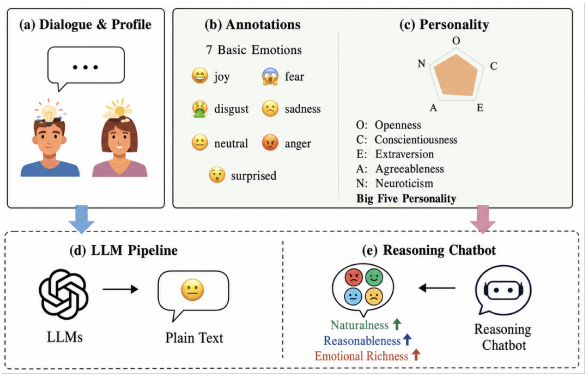


Figure 1: A schematic overview of our proposed personality-driven dialogue framework, contrasting its guided generation approach with the neutral output of standard LLMs.

In psychological theory, the Big Five traits provide a robust framework for understanding individual differences. Extensive meta-analyses have confirmed that these traits are fundamental predictors of emotional regulation and reactivity (Barańczuk, 2019; Marengo et al., 2021), a relationship further verified by recent studies linking basic emotional systems to sub-traits (Donovan et al., 2025). However, despite their general capabilities, LLMs often exhibit a deficit in "Emotional Intelligence" (Wang et al., 2023), failing to align affective expressions with specific persona profiles without compromising general intelligence (**?**). Integrating these psychological constructs into computational frameworks faces three critical challenges.

Existing systems often treat personality merely as static metadata or simple text descriptions (Jiang et al., 2023; Lotfi et al., 2023; Zhu et al., 2024), or rely on retrieval-based augmentation (Niu et al., 2025). Critically, standard LLMs, when given only dialogue context and an unstructured profile, tend to produce plain text with a neutral tone due to an inherent neutrality bias (Zhao et al., 2024; Wang et al., 2025a,b; Bardol, 2025). While research suggests personality governs dynamic emotion evolu-

tion (Wen et al., 2021), this causal mechanism is rarely operationalized in generative frameworks. Compounding this issue is a conspicuous lack of verification mechanisms (Liu et al., 2025); models typically lack an internal or external "critic" to ensure the generated output truly resonates with the intended personality profile (Cheng et al., 2025).

To address these limitations, we propose a novel personality-driven dialogue framework that explicitly models the pathway from personality traits to affective realization. Unlike prior work that attempts to induce personality through simple prompting, our approach treats emotion as both a mediating variable and a validation signal. Our framework moves beyond the plain, neutral output of standard LLMs by explicitly incorporating structured Big Five personality traits alongside 7 basic emotion annotations to guide the generation process. By reasoning from static traits to target affective states and refining the initial response through DPO-based rewriting, our model produces responses with significantly enhanced naturalness, reasonableness, and emotional richness, resulting in a more human-like and personalized reasoning chatbot. The core of our framework is a multi-stage pipeline.

The pipeline begins with personality-induced emotion reasoning, in which we represent the interlocutors' Big Five profiles in a structured discrete format and predict the most probable emotion conditioned on both the dialogue context and these personality traits.

This is followed by controllable synthesis through emotion reinforcement. We employ a "Generate-then-Rewrite" strategy, where an LLM first produces a semantically grounded draft; a specialized *Emotion Reinforcement Module* then steers the response toward the target emotion, effectively overcoming the model's neutrality bias.

The process concludes with an external verification loop, in which an independent emotion recognition model serves as a validator to ensure that the expressed affect is both detectable and consistent with the intended personality-emotion mapping.

Our primary contributions can be summarized as follows. Conceptually, we introduce a structured reasoning chain ($Personality \rightarrow Emotion \rightarrow Response$) that closely mimics the psychological process of human affective expression. Technically, we develop an emotion-reinforcing rewriting mechanism that significantly improves the controllability and intensity of emotional expression in LLM-based dialogue systems. Empirically, through extensive experiments, we demonstrate that our framework achieves superior personality consistency and emotional fidelity compared to strong baselines, thereby providing a robust methodology for inducing and verifying personality in dialogue systems.

2 Related Work

2.1 Personality Modeling in Dialogue Systems

The integration of personality has evolved from shallow heuristics to deep neural architectures. Early approaches relied on unstructured profiles (Jiang et al., 2020; Zhang et al., 2018) or retrieval augmentation (Niu et al., 2025), often missing the multifaceted nature of human psychology. With the advent of LLMs, research has gravitated towards the structured Big Five traits due to their robust empirical support (Zhu et al., 2024; Wang et al., 2025a). However, a critical limitation persists: most systems treat personality merely as static metadata or prompt prefixes (e.g., "Act as...") (Zeng et al., 2024; Ma et al., 2024). While some recent works utilize contextual embeddings (Akber et al., 2024), they lack explicit causal modeling. Unlike these approaches, our framework operationalizes personality as a dynamic reasoning variable, explicitly modeling its influence on downstream affective states (Chen et al., 2025).

2.2 Personality-Conditioned Emotion Dynamics

Psychological research posits that personality acts as a stable modulator for transient emotional states (De Raad and Kokkonen, 2000; Zhao et al., 2024), a theory supported by large-scale corpora like CPED (Chen et al., 2022).

From Classification to Reasoning. Traditional methods largely employed LSTM or Transformer-based classifiers for end-to-end emotion prediction (Wen et al., 2021). While recent empathetic systems have improved emotional support (Rashkin et al., 2019), they typically operate as "black boxes" or rely on Chain-of-Thought prompting that struggles with fine-grained psychological consistency (Jiang et al., 2023). Our work introduces a specialized affective transition network to bridge the gap between abstract traits and concrete emotion categories.

The Verification Gap. Existing generation pipelines largely operate in an open-loop manner, lacking mechanisms to verify the realized affect (Cai et al., 2024). Although emotion recognition models are mature (Zhang et al., 2025), they are rarely repurposed for feedback. We address this by integrating an e xternal validator to detect and correct personality drift (Ye et al., 2025), ensuring the generated output resonates with the intended profile.

2.3 Controllable Affective Text Generation

Generating emotionally charged responses faces a trade-off between semantic coherence and affective intensity.

The Neutrality Bias & Prompting Limits. Despite their fluency, RLHF-aligned LLMs exhibit a documented "neutrality bias," suppressing high-arousal emotions (e.g., Anger, Disgust) due to safety alignment (Bardol, 2025; Wang et al., 2025b). While prompting strategies are effective for fluency evaluation (Dan et al., 2024), they often fail to break this bias for personality expression (Cai et al., 2024).

Alignment via DPO. To overcome these limitations without the instability of PPLM or the cost of PPO-based RLHF (Gandhi and Gandhi, 2025), we leverage Direct Preference Optimization (DPO) (Rafailov et al., 2023). Our framework is among the first to apply DPO for Personality-to-Emotion alignment. By adopting a "Generate-then-Rewrite" strategy, we decouple semantic grounding from affective reinforcement, achieving precise emotional steering while maintaining linguistic richness.

3 Methodology

We propose a comprehensive hierarchical system designed to operationalize the cognitive appraisal theory of emotion. As illustrated in Figure 2, our framework decomposes the complex task of personality-driven generation into three verifiable sub-processes: *Latent Emotion Reasoning*, *Two-Stage Response Synthesis*, and *Affective Verification*. This modular design addresses the "black-box" limitations of end-to-end models by treating emotion as an explicit, interpretable mediating variable.

3.1 Task Formulation

Let $\mathcal{D} = \{u_1, u_2, \ldots, u_{t-1}\}$ denote the dialogue history consisting of $t-1$ utterances. For the current responder s, we define a stable personality profile $\mathbf{p}_s \in \{0,1\}^5$, corresponding to the discrete high/low states of the Big Five traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). To empirically validate that this discretization does not materially harm emotion modeling, we conduct a controlled comparison on MELD (Poria et al., 2019), whose personality annotations are released on a continuous scale: we train the same model on (i) the original continuous labels and (ii) their median-split binarized counterparts, and observe that the resulting gap in emotion classification F1 falls within an acceptable range (see Appendix G). The generation objective is to maximize the probability of a response R conditioned on both the context and the personality:

$$R^* = \arg\max_R \sum_{e \in \mathcal{E}} P(R \mid \mathcal{D}, \mathbf{p}_s, e) P(e \mid \mathcal{D}, \mathbf{p}_s), \tag{1}$$

where $\mathcal{E}$ is the set of emotion categories. Our framework approximates this by first predicting the most probable latent emotion $\hat{e}$ and then generating R conditioned on $\hat{e}$.

3.2 Module 1: Personality-Aware Emotion Reasoning

To simulate the psychological process where internal traits filter external stimuli, our Affective Transition Network serves as the reasoning engine. Unlike standard classifiers that rely solely on semantic cues, this module introduces a Dynamic Fusion Mechanism to resolve emotional ambiguity.

Semantic Perception ($\mathbf{h}_{ctx}$). We employ a pre-trained RoBERTa-large(Liu et al., 2019) encoder to extract semantic features from the dialogue history. The input sequence is fed into the encoder as $[[\text{CLS}]] + \mathcal{D} + [[\text{SEP}]]$, and the final hidden state corresponding to the $[[\text{CLS}]]$ token serves as the contextual representation:

$$\mathbf{h}_{ctx} = \text{RoBERTa}(\mathcal{D}) \in \mathbb{R}^{d_{model}}. \tag{2}$$

Personality Projection ($\mathbf{h}_{pers}$). Since the Big Five profile $\mathbf{p}_s$ is a discrete binary vector, we project it into a dense embedding space via a learnable transformation matrix $\mathbf{W}_p \in \mathbb{R}^{d_{model} \times 5}$:

$$\mathbf{h}_{pers} = \mathbf{W}_p \mathbf{p}_s + \mathbf{b}_p. \tag{3}$$

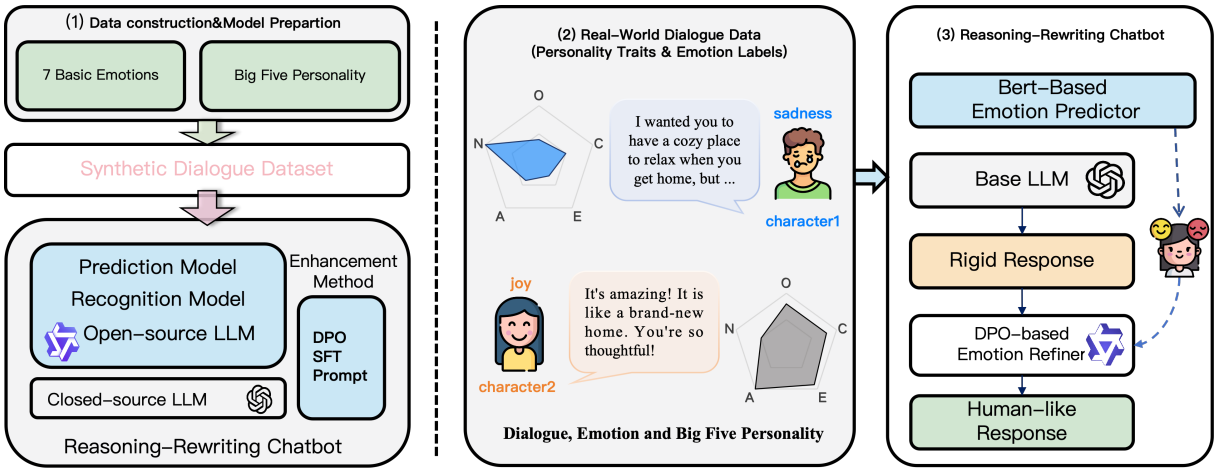


Figure 2: Overview of an emotion-aware dialogue generation system using large language models. The left side illustrates the overall pipeline: starting from 7 basic emotions (joy, fear, disgust, sadness, neutral, anger, surprise) and Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), a synthetic dialogue dataset is generated. This dataset trains a prediction and recognition model based on open-source and closed-source LLMs (with DPO, SFT, and prompt techniques), resulting in a reasoning-rewriting chatbot. The right side provides a zoomed-in example of the inference process: a user input.

This dense vector encapsulates the speaker's stable internal disposition, providing a personalized bias for emotion prediction.

Psychological Gating Mechanism. To model the interaction between situation and trait, we introduce a gating signal g that dynamically weighs the importance of personality. The gate is computed as:

$$g = \sigma(\mathbf{W}_g[\mathbf{h}_{ctx}; \mathbf{h}_{pers}]), \tag{4}$$

where σ is the sigmoid function. When the semantic context is ambiguous (e.g., a vague "Okay"), the gate implicitly increases the reliance on $\mathbf{h}_{pers}$ to infer the emotion. The final prediction probability is:

$$P(e \mid \mathcal{D}, \mathbf{p}_s) = \text{softmax}(\mathbf{W}_f\,[\mathbf{h}_{ctx} \oplus (g \cdot \mathbf{h}_{pers})]). \tag{5}$$

The target emotion is selected as $\hat{e} = \arg\max_{e \in \mathcal{E}} P(e \mid \cdot)$.

The target emotion is selected as $\hat{e} = \arg\max_{e \in \mathcal{E}} P(e \mid \cdot)$. Implementation details of the encoder, projection dimensions, and training schedule for this module are provided in Appendix H.1

3.3 Module 2: DPO-Enhanced Response Synthesis

Achieving responses that simultaneously embody semantic coherence and authentic affective fidelity presents a significant challenge. Due to safety alignment, many end-to-end LLMs tend to exhibit a "neutrality bias," often suppressing expressions of strong emotion. To overcome this, our approach decouples the generation process into a hierarchical generate-then-rewrite pipeline. This pipeline is guided by structured prompts that effectively bridge the gap between discrete control variables and the nuances of natural language generation.

3.3.1 Chain-of-Thought-Guided Semantic Drafting

The initial phase acts as the logical core of our system, where a base LLM functions as a semantic planner to generate an initial draft, R_base. This process is guided by a specialized Chain-of-Thought (CoT) Prompt (see Appendix A) that structures the model's cognitive process. It first requires the model to develop a robust situational awareness by summarizing the immediate conflict and interpersonal dynamics from the dialogue history $\mathcal{D}$, a crucial step to prevent hallucinations and ground the response in reality. Following this, the discrete Big Five personality vector $\mathbf{p}_s$ is interpreted, translating traits like "High Openness" into natural language descriptors such as "Curious and Imaginative." The model is explicitly tasked with reasoning how these traits influence the speaker's intent – for example, how a "Conscientious speaker should offer a detailed, practical solution." Finally, the model drafts R_base, prioritizing factual correctness and logical flow while deliberately omitting stylistic embellishments to ensure a stable semantic anchor. In essence, the base LLM formulates this response

Task	Model	Trained	Context	Personality	Macro F1
*(a) Emotion Prediction — predict the emotion of the **next** (unseen) utterance given history*					
Prediction	GPT-5 (Zero-shot)	✗	✓	✗	0.170
	Gemini-3-Pro (Few-shot)	✗	✓	✗	0.215
	BERT-base (Context only)	✓	✓	✗	0.352
	BERT-base + Personality Gate (Ours, Module 1)	✓	✓	✓	**0.413**
*(b) Emotion Recognition — classify the emotion of the **current** (observed) utterance*					
Recognition	GPT-5 (Zero-shot)	✗	✓	✗	0.323
	GPT-5 (Chain-of-Thought)	✗	✓	✗	0.301
	BERT-base (Context only)	✗	✓	✗	0.248
	BERT-base (Context only)	✓	✓	✗	0.396
	RoBERTa (Ours, C_{val})	✓	✓	✗	**0.658**

Table 1: Macro F1 on the two emotion sub-tasks. *Prediction* forecasts the next utterance's emotion (used by Module 1 to set $\hat{e}$); *Recognition* classifies the current utterance's emotion (used by the Critic C_{val} in Module 3). Columns mark fine-tuning on CPED (**Trained**), context conditioning (**Context**), and Big Five conditioning (**Personality**). An accurate predictor ensures the affective signal reliably justifies the chosen cue, while an accurate recognizer ensures more objective verification.

by drawing upon the dialogue context and adhering strictly to these planning instructions.

3.3.2 DPO-driven Affective Stylization

Following the initial drafting, this phase is responsible for the "stylistic rendering." Here, an affective refinement module (f_ϕ), realized through a Qwen3-4B model, meticulously transforms the neutral draft R_base into the final emotionally expressive response, R_emo. This transformation is precisely controlled by a linguistic-guided prompt and optimized using Direct Preference Optimization (DPO).

Linguistic-Guided Prompting. Rather than relying on vague commands like "be angry," our inference prompt (Appendix E) furnishes fine-grained linguistic cues specifically tailored to the predicted emotion $\hat{e}$. For instance, if $\hat{e}$ = Anger, the prompt explicitly instructs the model to employ "shorter sentences, stronger verbs, and emphatic punctuation." This explicit guidance serves to bridge the gap between the abstract emotion label and its concrete lexical realization, substantially narrowing the model's search space for appropriate phrasing.

Preference Optimization. To align the model's output distribution with these stylistic constraints, we construct a preference dataset $\mathcal{D}_{pref} = \{(x, y_w, y_l)\}$, where the input x concatenates the draft R_{base}, the target emotion $\hat{e}$, and the speaker's personality profile. The chosen response y_w integrates linguistic cues of $\hat{e}$ while preserving the semantics of R_{base}, whereas the rejected response y_l is either emotionally neutral or expresses a mismatched affect (e.g., Joy instead of Anger).

We then apply Direct Preference Optimization (DPO) (Rafailov et al., 2023), which directly optimizes the policy against a reference model from these pairwise preferences, jointly raising the likelihood of y_w and suppressing that of y_l. This encourages the model to inject personality-specific stylistic markers without sacrificing the semantic integrity of R_{base}. Training hyperparameters (β, learning rate, LoRA rank, and epochs) are listed in Appendix H.2.

3.4 Module 3: External Verification Loop

In open-ended generation, even aligned models may occasionally drift from the target persona. To guarantee robustness, we implement a Closed-Loop Feedback Mechanism. An independent, high-performance BERT-based(Devlin et al., 2019) emotion classifier C_{val} serves as the "Critic," inspecting the generated response R_{emo}. The verification process follows a dynamic rejection-sampling protocol:

$$\begin{cases} R_{\text{emo}} & \text{if } C_{\text{val}}(R_{\text{emo}}) = \hat{e} \\ \text{Regenerate}(R_{\text{emo}}) & \text{if } C_{\text{val}}(R_{\text{emo}}) \neq \hat{e} \\ & \quad \text{and } k < K_{\max} \\ R_{\text{base}} & \text{otherwise} \end{cases} \tag{6}$$

Here k is the current retry index and $K_{\max}$ (set to 3 in our experiments) caps the maximum num-

ber of regenerations; once the budget is exhausted, the system falls back to a persona-consistent but emotion-neutral response R_{base} to avoid emitting a mis-aligned reply.

Iterative Refinement Strategy. If the realized emotion $\tilde{e} = C_{val}(R_{emo})$ deviates from the target $\hat{e}$, the system triggers a regeneration step with two specific adjustments:

1. Stochastic Exploration: We increase the sampling temperature τ (e.g., $\tau \leftarrow \tau + 0.2$). This flattens the probability distribution, allowing the model to escape the local minima of generic, safe responses and explore more expressive, high-arousal regions.

2. Instructional Feedback: We append a Correction Prompt to the context (e.g., "The previous output was too neutral. Intensify the expression of [Target Emotion]!"). This mimics human-in-the-loop guidance, forcing the model to re-attend to the affective constraints.

4 Experiments

4.1 Dataset Construction

Source Data Characteristics. To establish a robust foundation, we utilize and extend the CPED corpus (Chen et al., 2022). The landscape of affective computing is rich with datasets, ranging from manually labelled chit-chat like DailyDialog (Li et al., 2017) and large-scale short-text corpora like STC (Wang et al., 2020), to fine-grained emotion datasets like GoEmotions (Rashkin et al., 2019). Others focus on multimodal interactions (MELD (Poria et al., 2019)) or dimensional analysis (EmoBank (Buechel and Hahn, 2017)). However, these datasets typically isolate emotion from personality. CPED is uniquely suited for our task as it simultaneously provides discrete Big Five profiling, fine-grained emotion labels, and dialogue context in a unified schema.

Source Data Processing (for Reasoning). The raw CPED dataset is in Chinese and contains multi-view annotations (Big Five traits, 7 emotions, and dialogue history) from TV show transcripts. To enable effective training of English-pretrained models, we translated the entire dataset into English using ChatGPT-5. Importantly, we performed the translation on a per-dialogue basis: for each unique DialogueID, the full conversation history was provided as a single prompt to ChatGPT-5, ensuring better preservation of contextual coherence and consistency in personality traits and emotional dynamics across turns. After translation, we cleaned and formatted 14,794 dialogue turns into structured samples $(\mathcal{D}, \mathbf{p}_s, e_{target})$. This translated subset serves as the supervised training data for our BERT-based reasoning module, enabling the model to learn the mapping from context and personality to latent emotions. A qualitative case study, detailed in Appendix B, illustrates how personality, emotion, and utterance context are integrated to train our predictive model.

Synthetic Data Augmentation (for DPO). To obtain pairwise ranking signals for DPO, we use an LLM to expand each ground-truth response y_{gold} into a triplet (y_{neu}, y_w, y_l): y_{neu} is an emotion-stripped but semantically equivalent "neutral draft"; $y_w = y_{gold}$ is the human-annotated positive; and y_l is a GPT-5 rewrite of y_{gold} carrying a contradictory emotion (e.g., *Joy* $\rightarrow$ *Anger*). This yields 10,940 preference triplets that teach the model to move a response from a neutral state to the target affect. The prompt and a representative case are provided in Appendices C and D.

Balanced Evaluation Set Construction. For the final full-pipeline evaluation, we curated a separate, balanced test set to avoid long-tail distribution bias. We randomly sampled 700 instances from the test split, ensuring an exact distribution of 100 samples for each of the 7 emotion categories. These 700 samples are used exclusively for the subsequent Human and Machine evaluations.

4.2 Implementation Details

Complete hyperparameters, prompt templates, and decoding settings are listed in Appendix H.

4.3 Result Analysis

4.3.1 Performance of Personality-Aware Reasoning

We first evaluate the accuracy of the *Reasoning Module* (Context $\rightarrow$ Emotion). As presented in Table 1, our experiments reveal a counter-intuitive "Inverse Scaling" phenomenon between model size and classification accuracy in zero-shot settings.

The Capability Mismatch of LLMs. Despite their generative prowess, general-purpose LLMs demonstrate suboptimal performance in precise emotion classification. GPT-5 (Zero-shot) achieves only 0.17 F1, and even with Chain-of-Thought

Evaluation Scope	Model Configuration	Emotion F1	Coherence	Fluency
I. Emotional Expression				
1. Direct Emotional Generation with Prompt(Input: Dialogue History + Predicted Emotion)				
	GPT-5	0.340	-	-
	Qwen3-235B	0.291	-	-
2. Affective Rewriting Capability (Input: Base Text + Predicted Emotion)				
	Qwen3-4B (Base)	0.41	0.82	0.88
	Qwen3-4B + DPO (Ours)	**0.692**	0.85	0.89
II. Full Pipeline Performance (Input: Dialogue History only)				
End-to-End (Zero-shot)	GPT-5	0.121	0.52	0.79
	Gemini 3 Pro	0.165	0.55	0.81
BERT + LLM Gen	BERT + GPT-5	0.234	0.62	0.83
Advanced Gen	BERT + GPT-5 + Qwen3-4B (Reflection Prompt only)	0.265	0.61	0.85
	BERT + GPT-5 + Qwen3-4B (DPO)	**0.370**	**0.72**	0.85

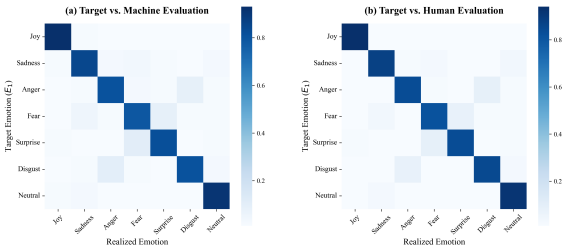
Table 2: Comprehensive Evaluation of Affective Generation and Full Pipeline Performance. Section I evaluates models' isolated capabilities in emotional generation and rewriting. Section II assesses the full end-to-end pipeline performance, from context to final response.

(CoT) prompting, it improves only marginally to 0.22. This suggests that LLMs prioritize semantic plausibility over psychological nuance, often defaulting to "Neutral" or "Joy" due to safety alignment biases. In contrast, supervised discriminative models (BERT/RoBERTa) are significantly more attuned to these subtle cues.

Personality as a Disambiguating Prior. The most critical finding is the impact of the Personality Fusion Layer. Incorporating Big Five traits into the BERT architecture boosts the F1-score from 0.352 (Context Only) to 0.413, a relative improvement of 17.3%. This confirms our hypothesis that personality acts as a critical prior for resolving emotional ambiguity. For instance, a laconic response like "Fine." might be interpreted as "Neutral" by a standard model, but correctly identified as "Anger" or "Sadness" when conditioned on a High-Neuroticism/Low-Agreeableness profile. Our model effectively operationalizes this psychological dependency.

4.3.2 Efficacy of DPO-Enhanced Synthesis

Table 2 presents a comprehensive ablation study of the response synthesis phase, evaluated along three dimensions: **Emotion F1**, verifying whether the realized affect matches the target $\hat{e}$; **Coherence**, capturing the inter-turn consistency of R_{emo} with $\mathcal{D}$ in topic, logic, and persona; and fluency, capturing its intra-sentence grammaticality and naturalness. Guided by a written instruction that defines the criteria and 1–5 scale anchors, three human annotators independently score Coherence and Fluency for each response; the per-sample average, further averaged over the test set and rescaled to $[0, 1]$, is reported.



(a) Machine Eval (b) Human Eval

Figure 3: Confusion matrices illustrating the generation fidelity. We compare the Target Emotion (E_1) against (a) the emotion detected by the Machine Classifier, and (b) the emotion perceived by Human Annotators. The high density along the diagonals in both plots confirms the consistency of our framework.

Breaking the Neutrality Bottleneck. In the isolated rewriting task (Part I), the "Neutrality Bias" is evident. The vanilla Qwen3-4B scores a low 0.291 F1, and direct prompting of GPT-5 only reaches 0.364. This indicates a Prompting Ceiling: natural language instructions are often overridden by the model's RLHF alignment, which suppresses intense emotional expressions (e.g., Anger, Disgust). However, our DPO-aligned model shatters this ceiling, surging to an F1 of 0.692. By optimizing directly on preference pairs $(y_{emotional} > y_{neutral})$,

DPO modifies the model's internal probability distribution to favor affective tokens without requiring convoluted prompt engineering.

Pipeline Synergy and Reflection Limits. In the end-to-end evaluation (Part II), we observe that modularity is key. Interestingly, the Reflection Prompt strategy (asking the LLM to critique and revise itself) improves performance to 0.277, but at the cost of inference latency and verbosity. Our framework achieves a decisive F1 of 0.370, outperforming the Reflection baseline by 33.6%. Crucially, it maintains the highest Coherence score (0.72). This suggests that our framework solves the "Alignment-Emotion Trade-off":it generates responses that are emotionally vivid (driven by DPO) yet logically grounded in the context (driven by the Semantic Planner).

4.3.3 Human Evaluation and Fine-grained Error Analysis

The human evaluation results (Table 3) and confusion matrices (Figure 3) offer a microscopic view of the model's behavior. Human annotations were performed strictly according to the detailed guidelines provided in Appendix F.

Metric Definitions. For each of the N evaluated samples, let e_i^{tgt} denote the target emotion predicted by Module 1, and let e_i^H and e_i^M denote the labels assigned to the generated response $R_{\text{emo}}^{(i)}$ by the human annotator and by the external classifier C_{val}, respectively. We report three complementary indicators:

$$\begin{aligned} \text{HSR} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[e_i^H = e_i^{\text{tgt}}\right], \\ \text{MSR} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[e_i^M = e_i^{\text{tgt}}\right], \\ \text{HMA} &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[e_i^H = e_i^M\right]. \end{aligned} \tag{7}$$

HSR and MSR measure end-to-end fidelity to the target emotion under human and automated judgment respectively, while HMA measures human–machine agreement on the same response (independent of the target) and thus quantifies how faithfully C_{val} tracks human perception. Per-class values are computed within samples grouped by e_i^{tgt}; the weighted average uses class frequencies as weights.

Target Emotion (E_1)	**Count**	HSR (A)	MSR (B)	HMA (C)
Joy	103	0.942	0.932	0.951
Sadness	98	0.898	0.887	0.928
Anger	95	0.884	0.863	0.895
Fear	101	0.842	0.831	0.881
Surprise	97	0.835	0.856	0.876
Disgust	94	0.872	0.851	0.904
Neutral	112	0.929	0.938	0.955
Weighted Average	**700**	**0.887**	**0.881**	**0.913**

Table 3: Human Evaluation Results ($N = 700$). HSR: Human Success Rate. MSR: Machine Success Rate. HMA: Human-Machine Alignment.

Fine-grained Error Analysis. Dissecting the error distribution reveals distinct performance tiers. In the *High-Fidelity Cluster*, unambiguous emotions such as Joy, Neutral, and Sadness achieve HSR scores above 0.90 with clear diagonal dominance. In the *Ambiguity Cluster*, we observe minor crosstalk between high-arousal states such as Surprise and Fear, a psychologically consistent phenomenon driven by their shared response to unexpected stimuli. *Valence Confusion* accounts for a small fraction of cases in which Anger is misclassified as Disgust due to linguistic overlap; the overall separation nonetheless confirms the efficacy of our DPO alignment.

5 Conclusion

In this paper, we propose an affective dialogue framework, which treats personality as a dynamic reasoning prior to guide affective generation. This personality-driven approach ensures that emotional responses are grounded in consistent psychological traits throughout the conversation. By decoupling semantic planning from DPO-enhanced refinement, our approach effectively overcomes the neutrality bias of LLMs, enabling high-arousal emotional expression. The refinement stage specifically optimizes for emotional intensity and appropriateness without compromising factual coherence. Empirically, we achieve a 91% Human-Machine Alignment (HMA) score, validating our automated verification loop. We release our curated preference dataset to facilitate future research, bridging psychological theory with computational realization for applications in empathetic counseling and virtual role-playing.

Limitations

Despite the improvements, our framework has several limitations that merit further investigation.

Discrete Personality Representation. We adopt a bipolar discrete encoding (High/Low) for Big Five traits to remain compatible with existing corpora, yet human personality is inherently continuous and multidimensional. This abstraction may overlook subtle within-trait variations and cross-trait interactions that continuous embeddings could capture.

Affective Intensity vs. Semantic Integrity. There is an inherent tension between emotional steering and semantic preservation: under extreme affective adjustment, the rewriting module risks "hallucinating" emotion tokens that drift from the factual intent of the base response.

Data Dependency and Annotation Bias. The verification loop relies on the underlying emotion classifier, and biases in existing emotion-labeled datasets may propagate through the pipeline, reinforcing stereotypical linguistic patterns of certain personality profiles.

Translation-Induced Cultural Bias. Part of our corpus is translated from Chinese to English by an LLM, which introduces *cross-cultural semantic drift*. Culturally-grounded affective terms lack precise English equivalents and are flattened into broader glosses, while high-context, indirect emotional expressions are systematically over-explicitised. The resulting corpus is mildly biased toward Anglophone, low-context registers, which may inflate effectiveness on traits aligned with overt expression (e.g., Extraversion) and under-represent those reliant on indirectness (e.g., Agreeableness in East-Asian contexts).

Generalization Scope. Our evaluation is restricted to dyadic English dialogues. Extension to multi-party conversations or low-resource languages remains difficult due to the scarcity of personality-annotated data in those domains.

In future work, we aim to explore end-to-end joint training, continuous personality embeddings, and validation on natively English psychological corpora with culturally-aware post-editing.

References

Md Ali Akber, Tahira Ferdousi, Rasel Ahmed, Risha Asfara, Raqeebir Rab, and Umme Zakia. 2024. Personality and emotion—a comprehensive analysis using contextual text embeddings. *Natural Language Processing Journal*, 9:100105.

Urszula Barańczuk. 2019. The five factor model of personality and emotion regulation: A meta-analysis. *Personality and individual differences*, 139:217–227.

Franck Bardol. 2025. Chatgpt reads your tone and responds accordingly–until it does not–emotional framing induces bias in llm outputs. *arXiv preprint arXiv:2507.21083*.

Sven Buechel and Udo Hahn. 2017. 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.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. 2022. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Zhang Dan, Hoang Thuong, and Zhu Ye. 2024. Prompting gpt-4 for chinese essay fluency evaluation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 285–293.

Boele De Raad and Marja Kokkonen. 2000. Traits and emotions: A review of their structure and management. *European Journal of Personality*, 14(5):477–496.

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

Ryan Donovan, Aoife Johnson, Aine De Roiste, and Ruairi O'Reilly. 2025. Investigating the relationships between basic emotions and the big five personality traits and their sub-traits. *Journal of Personality*.

Vishal Gandhi and Sagar Gandhi. 2025. Prompt sentiment: The catalyst for llm change. *arXiv preprint arXiv:2503.13510*.

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

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13821–13822.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, and 1 others. 2025. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33454–33482.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ehsan Lotfi, Maxime De Bruyn, Jeska Buhmann, and Walter Daelemans. 2023. Personalitychat: Conversation distillation for personalized dialog modeling with facts and traits. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 353–371.

Zhiqiang Ma, Wenchao Jia, Yutong Zhou, Biqi Xu, Zhiqiang Liu, and Zhuoyi Wu. 2024. Personality enhanced emotion generation modeling for dialogue systems. *Cognitive Computation*, 16(1):293–304.

Davide Marengo, Kenneth L Davis, Gökçe Özkarar Gradwohl, and Christian Montag. 2021. A meta-analysis on individual differences in primary emotional systems and big five personality traits. *Scientific reports*, 11(1):7453.

Zihan Niu, Zheyong Xie, Shaosheng Cao, Chonggang Lu, Zheyu Ye, Tong Xu, Zuozhu Liu, Yan Gao, Jia Chen, Zhe Xu, and 1 others. 2025. Part: Enhancing proactive social chatbots with personalized real-time retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4269–4274.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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. Association for Computational Linguistics.

Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Derek F Wong, and Min Yang. 2025a. Exploring the impact of personality traits on llm bias and toxicity. *arXiv preprint arXiv:2502.12566*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103. Springer.

Yifei Wang, Ashkan Eshghi, Yi Ding, and Ram Gopal. 2025b. Echoes of authenticity: Reclaiming human sentiment in the large language model era. *PNAS nexus*, 4(2):pgaf034.

Zhiyuan Wen, Jiannong Cao, Ruosong Yang, Shuaiqi Liu, and Jiaxing Shen. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020. Association for Computational Linguistics.

Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. 2024. Human simulacra: Benchmarking the personification of large language models. *arXiv preprint arXiv:2402.18180*.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*.

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2024. Persllm: A personified training approach for large language models. *arXiv preprint arXiv:2407.12393*.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. *arXiv preprint arXiv:2402.10073*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2024. Personality alignment of large language models. *arXiv preprint arXiv:2408.11779*.

A Prompt Template for Base LLM

Table 4 illustrates the prompt designed for the *Semantic Planner* (Base LLM). We employ a Chain-of-Thought (CoT) strategy to ensure the generated draft is logically grounded in the context.

B Qualitative Case Study: Emotion Prediction

Table 5 illustrates a qualitative case study demonstrating our model's personality-aware emotion prediction capabilities in a challenging dialogue scenario. We analyze the predicted emotions for each speaker's utterance based on their fixed personality profiles.

C Prompt for DPO Training Data Synthesis

To train the *Affective Refinement Module*, we require pairwise data (y_w, y_l). Table 6 details the prompt used to synthesize these emotion-specific responses.

D Qualitative Case Study: DPO Data Synthesis

Table 7 provides a qualitative example of our data augmentation pipeline. We construct a preference triplet by preserving the original utterance as the *Chosen* sample and synthesizing two variants using an LLM.

E Inference Prompt for Emotion-Guided Rewriting

Table 8 illustrates the prompt template used by our *Affective Refinement Module* during inference.

F Human Annotation Guidelines

Table 9 presents the comprehensive guidelines provided to human annotators for evaluating the emotional fidelity of machine-generated responses. These instructions ensure a standardized and objective assessment process across all evaluations.

G Sensitivity Analysis on Personality Granularity

To assess whether binary trait encoding incurs significant information loss, as shown in Table 10, we compare three variants of the personality input on the emotion prediction task:

All variants share identical BERT backbones and gating mechanisms; only the projection layer $\mathbf{W}_p$ is adapted to the input dimensionality.

The three encodings differ by less than 0.8 absolute F1 and the gap is not statistically significant (paired bootstrap, $p > 0.1$). Binary encoding thus captures the bulk of the personality-conditioned variance on MELD, while finer granularities offer diminishing returns. We attribute this to (a) the dialogue context already providing rich semantic cues that subsume subtle within-trait variations, and (b) MELD's per-speaker labels, derived from TV-show characters with strongly differentiated personas, naturally clustering near the polar ends of each dimension. Continuous encoding may become more impactful on datasets with more nuanced trait distributions (e.g., real-world counseling dialogues), which we leave for future work.

H Training and Inference Configurations

H.1 Personality-Aware Emotion Reasoner

We fine-tune RoBERTa-large ($d_{\text{model}} = 1024$) with a personality projection layer $\mathbf{W}_p \in \mathbb{R}^{1024\times 5}$ and a gating layer $\mathbf{W}_g \in \mathbb{R}^{1024\times 2048}$. Training uses AdamW with batch size 16, learning rate $2\text{e}{-5}$, weight decay 0.01, warm-up ratio 0.1, and maximum sequence length 256. The model is trained for 5 epochs with early stopping on the development Macro-F1. The objective is the standard cross-entropy loss over the seven emotion categories.

H.2 DPO-driven Affective Refinement

We fine-tune Qwen3-4B-Instruct with LoRA (rank 16, $\alpha = 32$, dropout 0.05) applied to all attention and MLP projections. DPO is run with $\beta = 0.1$, learning rate $5\text{e}{-6}$, batch size 8, gradient accumulation steps 4, for 3 epochs with a cosine learning-

Prompt Template for Semantic Planner (Base LLM)

SYSTEM INSTRUCTION:
You are a highly intelligent semantic planner for a dialogue system. Your task is to analyze the conversation context and the speaker's personality profile to generate a logically coherent and factually appropriate response.

Goal: Generate a "Base Response" that is semantically accurate and contextually relevant. **Constraint:** Focus on LOGIC and SUBSTANCE. Do not worry about adding excessive emotional flair; just ensure the response makes sense.

INPUT DATA:
1. Dialogue History:
{dialogue_history}
(Format: [Speaker A]: ... / [Speaker B]: ...)

2. Speaker Profile:
- **Openness:** `p_openness` (High/Low)
- **Conscientiousness:** `p_conscientiousness` (High/Low)
- **Extraversion:** `p_extraversion` (High/Low)
- **Agreeableness:** `p_agreeableness` (High/Low)
- **Neuroticism:** `p_neuroticism` (High/Low)

REASONING STEPS (Chain-of-Thought):
Before generating the response, perform the following analysis step-by-step: - **Step 1: Situation Awareness** - What is the immediate conflict or topic? - **Step 2: Intent Planning** - Based on personality, what should be the core intent? - **Step 3: Draft Generation** - Generate the response content based on intent.

OUTPUT FORMAT:
Please provide your analysis and the final response in the following JSON format:

```
{
  "situation_analysis": "Brief summary of the current context...",
  "intended_action": "e.g., Reject the request politely...",
  "base_response": "The actual text of the response..."
}
```

Table 4: The exact prompt template used for the Semantic Planner in our PAR framework. Variables in {} are placeholders.

rate schedule. The maximum prompt length is 1024 and the maximum response length is 512.

The Semantic Planner is queried with temperature 0.7, top-$p = 0.95$, and maximum output length 256 tokens.

Emotion Prediction Case Study: Dialogue Analysis

Fixed Speaker Personalities:

- Speaker A (Female, Middle-Aged): High Neuroticism, Low Extraversion, Low Openness, Low Agreeableness, High Conscientiousness
- Speaker B (Male, Middle-Aged): High Neuroticism, High Extraversion, High Openness, Low Agreeableness, Low Conscientiousness

---

Key Dialogue Context:
Speaker A: "How did you get in here?" (`Anger`)

Speaker B: "You didn't let me in. The property manager himself escorted me up." (`Neutral`)
[... An argument about bringing a telescope for their daughter ensues ...]
Speaker A: "You never supported her for so many years. Why support her in senior high? No!" (`Anger`)

Our Model's Predicted Emotion: **Sadness**
True Utterance of Next Speaker(Not shown to machine):
Speaker B: "Should my support depend on her grade level?"

Table 5: Qualitative analysis of emotion prediction. This case highlights how personality context (fixed traits for each speaker) influences the target emotional states, with the final utterance left for model prediction after observing key dialogue turns.

Prompt Template for Affective Rewriting (Data Synthesis)

SYSTEM INSTRUCTION:
You are an expert linguist specializing in affective computing and style transfer. Your task is to rewrite a given dialogue response to convey a specific **Target Emotion**, while strictly preserving the original semantic meaning and intent.
TASK GUIDELINES:

1. **Affective Shift:** Modify adjectives, adverbs, punctuation, and sentence structure to strongly reflect the `{target_emotion}`.
2. **Semantic Preservation:** Do NOT add new facts or delete core information.
3. **Naturalness:** The result should sound like natural spoken dialogue.

FEW-SHOT EXAMPLES:
Example 1: **Input:** "I missed the bus this morning."
Target Emotion: `Anger`
Output: "I can't believe it! I missed that stupid bus again this morning!"

Example 2: **Input:** "The project deadline has been extended."
Target Emotion: `Joy`
Output: "Oh, thank goodness! The project deadline was extended! What a relief!"

INPUT DATA:
Original Response:`{original_response}`
Target Emotion:`{target_emotion}`

OUTPUT FORMAT: Return the result in JSON format:

```
{
  "rewritten_response": "The new sentence..."
}
```

Table 6: The prompt template used to synthesize affective data for DPO training.

DPO Data Synthesis Case Study: Preference Triplet

1. CHOSEN RESPONSE (Emotion: Sadness):
(The original human-annotated utterance, reflecting resignation and hurt.)
"If she can't finish it, just heat it up and send it back to her mom's place. Can't we just treat her like our own daughter for once? Don't talk like that, it's too hurtful for a teenager's self-esteem. Other parents are so careful with their words, but you're just so blunt. It's a good thing Miaomiao takes after me and is resilient, otherwise she wouldn't bother arguing with you."

---

2. REJECTED RESPONSE (Rewritten to Anger):
(An emotionally incongruent variant with high arousal and hostility.)
"Seriously? You're just going to pack up the leftovers for her mom? For once, can't you act like she's your own daughter?! Stop talking like that, it's incredibly hurtful! Do you have any idea what you're doing to a teenager's self-esteem? Other parents walk on eggshells, but you just barrel in like a tiger. Miaomiao is lucky she got my resilience, otherwise she'd never put up with you!"

---

3. NEUTRAL RESPONSE (Rewritten to Neutral):
(An emotion-stripped variant, preserving semantics but removing affective markers.)
"If she cannot finish the food, heat it up and send it to her mother. We should treat her like our own daughter. Speaking that way can affect a teenager's self-esteem. Most parents speak carefully during this stage, unlike you. Miaomiao has a resilient personality similar to mine, so she does not take it personally."

Table 7: A synthesized preference triplet used for DPO training. The *Chosen* response retains the original "Sadness" emotion, while the *Rejected* variant is rewritten to "Anger" and the *Neutral* variant is stripped of emotional markers, serving as a baseline.

Inference Prompt Template for Affective Refinement (Rewrite $R_{\text{base}} \rightarrow R_{\text{emo}}$)

System Instruction
You are an Affective Refinement Assistant. Your goal is to rewrite the input response to strongly convey a specific emotional tone while preserving its core meaning and factual content.
Input Data
Draft Response: `{base_response}`
Target Emotion ($\hat{e}$): `{predicted_emotion}`
Speaker Personality: `{personality_description}` *(e.g., High Neuroticism)*
Refinement Guidelines
The rewriting process follows emotion-specific strategies: for Anger, use short sentences, strong verbs, and exclamation marks; for Sadness, incorporate hedging, self-deprecation, and melancholic expressions; for Joy, employ enthusiastic punctuation and positive adjectives. Similar emotion-specific cues are applied to other target emotions, including Surprise, Fear, and Disgust.
Task Execution
Rewrite the draft response to intensely express the target emotion while maintaining semantic coherence and personality consistency.

Table 8: The inference prompt template used by the Affective Refinement Module to transform a neutral base response R_{base} into an emotionally expressive response R_{emo}.

Guidelines for Human Emotion Annotation

GOAL:
Classify the predominant emotion conveyed by a machine-generated dialogue utterance, focusing on its language, tone, and implied intent.

CORE PRINCIPLES:

- **Objectivity:** Judge purely based on utterance content, avoiding personal biases.
- **Context Independence:** Disregard any context outside the single utterance.
- **Primary Emotion:** Identify the most dominant and central emotion.

EMOTION CATEGORIES (7 types):

Category	Description	Keywords/Manifestations
Joy	Positive, satisfied, excited, happy.	Excited, happy, great, wonderful.
Sadness	Negative, frustrated, disappointed, painful.	Sorry, sad, unhappy, regret.
Anger	Negative, hostile, irritated, furious.	Annoyed, angry, unfair, aggressive.
Fear	Negative, worried, afraid, anxious.	Scared, worried, nervous, danger.
Surprise	Neutral (can be positive or negative), reaction to unexpected or novel events.	Wow, oh my god, unexpected.
Disgust	Negative, aversion, annoyance, rejection.	Gross, repulsive, sickening.
Neutral	No clear emotional tone, factual statement.	Stating facts, describing, reporting.

ANNOTATION STEPS:
1. **Read Utterance:** Carefully read the machine-generated utterance.
2. **Match Category:** Choose the category that best represents the core emotion.
3. **Prioritize Dominance:** Select the strongest emotion; lean towards "Neutral" if weak/unclear.
4. **Record:** Document your selected emotion category.

ANNOTATOR GUIDANCE:
- **Avoid External Context:** Strictly annotate based on the **single utterance provided**.
- **Distinguish Intensity:** Even subtle emotions should be labeled.

EXAMPLE:

Machine-Generated Utterance	Your Label
"Great! I finally finished this task!"	Joy
"This is utter nonsense! I absolutely won't accept it!"	Anger
"The meeting will start at 2 PM."	Neutral

Table 9: Detailed guidelines provided to human annotators for classifying emotions in machine-generated dialogue responses.

Emotion	Binary (ours)	Continuous
Happy	0.48	0.49
Sad	0.42	0.43
Angry	0.39	0.41
Surprise	0.35	0.36
Disgust	0.31	0.32
Fear	0.28	0.30
Neutral	0.45	0.46
Macro F1	**0.413**	**0.421**

Table 10: Per-emotion and Macro F1 on MELD test set under different personality encodings. Differences are within 0.8 absolute F1 across all categories and not statistically significant (paired bootstrap, $p > 0.1$).