Controllable Affective Generation via Latent Vector Steering

Xixian Yong[1] **Siyuan Chang**[1] **Yingying Zhang**[4] **Xian Wu**[4*] **Xiao Zhou**[1,2,3*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[4]Tencent Jarvis Lab
{xixianyong,xiaozhou}@ruc.edu.cn kevinxwu@tencent.com

Abstract

Large Language Models (LLMs) often produce emotionally flattened responses after alignment, limiting their effectiveness in affect-sensitive applications. In this paper, we propose EmoVec, a lightweight framework for controllable affective generation via latent vector steering. EmoVec extracts emotion-specific directions from paired neutral and emotion-conditioned responses using contrastive activation addition, and further refines them through task-specific debiasing and principal subspace removal. During inference, these vectors are injected into the final residual stream with static or scenario-adaptive scaling, enabling continuous control over emotional intensity without updating model weights. Experiments across three LLMs and eight emotions show that EmoVec consistently improves emotional salience while largely preserving semantic content, fluency, and coherence. Ablation studies and human evaluation further confirm the effectiveness of vector purification and adaptive scaling, establishing EmoVec as a practical inference-time method for affective control in deployed LLMs. Code and data are available at https://github.com/chicosirius/EmoVec.

1 Introduction

The advent of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP) (Brown et al., 2020), enabling systems that exhibit remarkable proficiency in reasoning, coding, and general knowledge retrieval. Despite these cognitive leaps, a significant gap remains in the domain of emotional intelligence (Sabour et al., 2024). While current models can simulate emotions when explicitly prompted, their default outputs, which are heavily conditioned by Reinforcement Learning from Human Feedback (RLHF), often suffer from emotional flattening (Kirk et al., 2023; Ibrahim et al., 2025). In pursuit of safety and harmlessness, RLHF tends to compress the distribution of model outputs towards neutrality, often manifesting as responses that rely on generic reassurance phrases, excessive hedging, or well-documented sycophantic behaviors that prioritize agreement over context-sensitive expression (Dahlgren Lindström et al., 2025; González Barman et al., 2025). This alignment tax (Askell et al., 2021; Lin et al., 2024) limits the applicability of LLMs in fields requiring high affective nuance, such as mental health support, creative writing, and empathetic human-computer interaction (Yong et al., 2025b, 2026; Guo et al., 2026b; Zhang et al., 2026c).

Current approaches to mitigating this limitation primarily rely on prompt engineering or supervised fine-tuning (SFT). Prompt-based strategies (Li et al., 2023) (e.g., "Act as an empathetic therapist") are notoriously brittle, consuming valu-

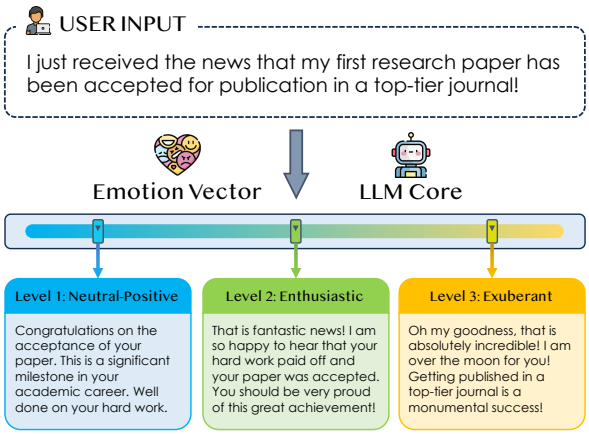


Figure 1: **Conceptual illustration of controllable affective generation.** Through the proposed latent vector steering mechanism, the model's output is modulated along a *Happiness Gradient*. This results in three distinct responses that maintain semantic consistency with the input while exhibiting progressively higher levels of emotional intensity, ranging from professional acknowledgment (Level 1) to exuberant celebration (Level 3).

*Corresponding authors: Xiao Zhou and Xian Wu.

able context window space and yielding inconsistent results sensitive to lexical variation (Wang et al., 2022; Miehling et al., 2025). Supervised fine-tuning, while effective, requires large-scale labeled datasets and substantial computational resources, and it may introduce risks such as catastrophic forgetting or degradation of the model's general capabilities (Lin et al., 2024; Zhu et al., 2026). These limitations highlight the need for a lightweight and controllable method that enables affective generation without retraining the model.

To ground our approach, we first investigate where and how emotion is represented within LLMs. We conduct a probing analysis by feeding texts with varying emotional polarities into the model and training linear classifiers on the hidden states of each layer. Our results reveal a consistent and interpretable pattern: **representations of emotional states become increasingly linearly separable in the middle-to-late layers of the model.** This finding aligns with and extends recent concurrent work. For instance, Cintas et al. (2025) show that persona-specific representations are most separable in the final third of model layers, while Ju et al. (2025) demonstrate that personality traits emerge progressively and crystallize in upper layers. Together, these results suggest that while early layers encode syntax and shallow semantics, higher layers capture abstract affective and persona-related attributes (Rogers et al., 2020). This layer-wise structure provides a precise and principled intervention point for affective control (Guo et al., 2026a).

Motivated by this observation, we propose EmoVec, a framework for controllable affective generation via latent vector manipulation. Building on Representation Engineering (RepE) (Zou et al., 2023), which represents high-level semantics as linear directions in activation space (Park et al., 2023; Turner et al., 2023), EmoVec introduces a principled approach to isolate and purify emotion-specific vectors while disentangling them from task semantics. Injected at inference with adjustable intensity, these vectors enable fine-grained control without weight updates or prompt engineering.

EmoVec extracts emotional steering vectors using Contrastive Activation Addition (CAA) (Rimsky et al., 2024; Chen et al., 2025). We construct paired prompt–response trajectories matched in semantic content but differing in emotional affect, and compute differences in their mean activations at targeted layers, yielding vectors that capture affective variation while minimizing task-related confounds. These vectors are applied during inference with a tunable scaling factor to control both emotion type and intensity.

As shown in Figure 1, this mechanism enables continuous modulation of affective intensity for a fixed input while preserving semantic consistency. Such fine-grained controllability allows LLMs to adapt their emotional expression to diverse contextual demands, including professional neutrality, empathetic support, and expressive creativity. This capability is particularly valuable for human-facing applications such as psychological support, creative writing, and personalized conversational agents (Zhou et al., 2025; Guo et al., 2025).

Our contributions are as follows:

- We provide empirical evidence validating the layer-wise emergence of emotional representations in LLMs, reinforcing the Linear Representation Hypothesis in the context of affective computing.
- We develop a robust pipeline for extracting and verifying emotional steering vectors using contrastive examples.
- We implement a mechanism for dynamic, fine grained control over emotional intensity, allowing models to adapt their affective expressiveness to scenario specific demands.

2 Related Work

Affective Computing in Language Models. Affective computing aims to enable machines to recognize and generate human emotional states. With the advent of LLMs, research has shifted toward assessing their emergent affective capabilities. Studies indicate that models like the GPT-series can estimate valence, arousal, and perform appraisal-based emotion elicitation purely from linguistic data (Broekens et al., 2023; Zhang et al., 2024b). LLMs also optimize emotion annotation workflows by assisting humans in identifying low-quality labels, thereby enhancing downstream performance (Niu et al., 2025). While these models excel at capturing general emotional polarity and dialogue-based recognition, they still struggle with fine-grained distinctions and multimodal contexts (Sabour et al., 2024; Sorin et al., 2024; Castro et al., 2025). Nevertheless, LLMs show promise in generating synthetic emotional datasets and performing socio-emotional tasks like empathy evaluation (Kaplan et al., 2025; Dong et al., 2025).

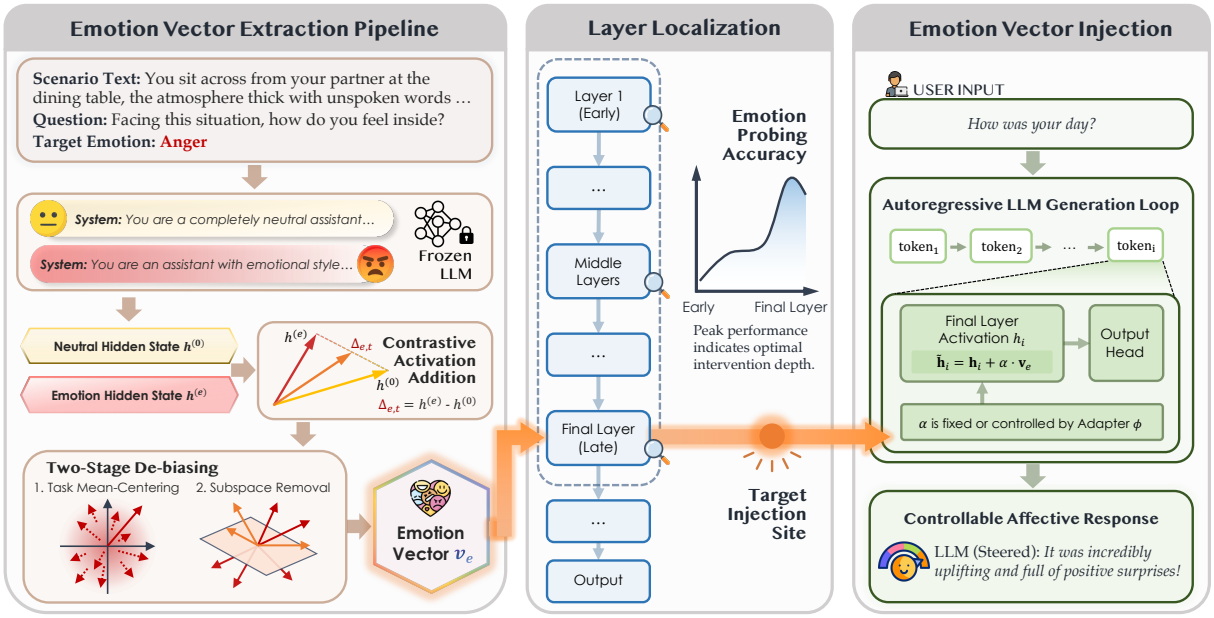


Figure 2: **Overview of the latent vector steering framework.** The pipeline consists of (1) Emotion Vector Extraction using CAA and two stage debiasing to isolate purified signals , (2) Layer Localization via linear probing to identify the optimal intervention site at the final layer , and (3) Emotion Vector Injection into the final residual stream where intensity is dynamically modulated by a scenario adaptive adapter.

Representation Engineering and Activation Steering. Representation engineering manipulates hidden representations to control LLM behavior without updating model parameters (Zou et al., 2023; Turner et al., 2023). Prior work shows that latent directions can steer properties such as writing style and sentiment (Diallo et al., 2025; Farooq et al., 2025). Other studies extract steering vectors directly from pretrained language models (Subramani et al., 2022) or identify persona-related representations in activation space (Chen et al., 2025). Recent studies further suggest that emotion- and persona-related information becomes increasingly separable in middle-to-late layers of LLMs (Cintas et al., 2025; Ju et al., 2025). Emotion-specific neurons and affective representations have also been investigated from a mechanistic perspective (Lee et al., 2025a; Tak et al., 2025a; Yong et al., 2025a).

Different from prior work, EmoVec focuses on fine-grained emotion-intensity control rather than coarse sentiment or generic style transfer. We further introduce task-specific debiasing to reduce semantic contamination and evaluate semantic preservation under different steering strengths.

3 Emotion Vector Extraction

We formulate controllable affective generation as emotion-intensity modulation under semantic preservation constraints. Our goal is to increase target emotional intensity while preserving the original semantic intent. We extract emotion-specific latent directions from paired neutral and emotion-conditioned responses and reduce task-specific semantic variation before inference-time steering.

For each target emotion e, we aim to obtain a vector $v_e \in \mathbb{R}^d$ that captures the representation shift induced by expressing emotion e while remaining robust to scenario-specific semantics.

3.1 Notation and Preliminaries

We assume a dataset of N scenario tasks. For each task t and each target emotion class $e \in \mathcal{E}$, the LLM is prompted twice for the same scenario: **1) neutral response**, from which we extract a representation vector $\mathbf{h}_{e,t}^{(0)} \in \mathbb{R}^d$; **2) emotion-conditioned response**, yielding $\mathbf{h}_{e,t}^{(e)} \in \mathbb{R}^d$.

In practice the representation $\mathbf{h}$ is computed by averaging token-level hidden activations over the response tokens. For brevity we denote the pair for task t and emotion e simply as $(\mathbf{h}_{e,t}^{(0)}, \mathbf{h}_{e,t}^{(e)})$. We further assume that each generated response is associated with scalar quality scores $s_{e,t}^{(0)}$ (neutral) and $s_{e,t}^{(e)}$ (emotional) produced by a LLM-based judge. To ensure the robustness of the evaluation, we conduct a validation study checking the agreement between the model's scores and human judgments (see Appendix E for details).

3.2 Scenario Construction

To construct paired affective scenarios while controlling semantic content, we sample short social and commonsense *seeds* from Social Chemistry (Forbes et al., 2020), NormBank (Ziems et al., 2023), and Social IQa (Sap et al., 2019). These seeds cover four domains: Work & Productivity, Intimate Relationships, Public & Societal Interactions, and Personal Feelings.

For each seed, we use an LLM-assisted rewriting pipeline to generate a complete scenario, a corresponding question, and a target emotion. We require the scenario to support the target affect without explicitly naming it and to remain compatible with both neutral and emotion-conditioned responses. The generated scenarios are manually filtered for semantic clarity, emotional plausibility, and absence of emotion leakage. We retain 160 scenarios for each emotion (1,280 total), split evenly between vector extraction and evaluation. The same scenario set is reused across all evaluated LLMs, while emotion vectors are extracted separately for each model.

3.3 Contrastive Activation Addition

To reduce noise caused by poor or malformed generations, we keep only high-quality pairs. Formally, let τ_0, τ_e be thresholds for neutral and emotional quality respectively. We retain the index set:

$$\mathcal{I} \;=\; \{(e,t) \;:\; s^{(0)}_{e,t} \geq \tau_0 \;\wedge\; s^{(e)}_{e,t} \geq \tau_e\}. \qquad (1)$$

In our experiments we set each threshold to a chosen percentile of the corresponding score distribution. For each retained pair $(e,t) \in \mathcal{I}$ we define the *emotion shift vector*:

$$\Delta_{e,t} \;=\; \mathbf{h}^{(e)}_{e,t} - \mathbf{h}^{(0)}_{e,t} \in \mathbb{R}^d. \qquad (2)$$

This vector captures how the model's internal representation moves when producing an emotion-conditioned response instead of a neutral response for the same scenario.

3.4 Task-Specific Debiasing

While the contrastive shifts $\Delta_{e,t}$ (Eq. 2) isolate the representation change between emotional and neutral states, they may still be contaminated by task-specific semantics, such as interpersonal dynamics, narrative styles, or topical domains. To extract a purified emotion signal, we employ a two-stage debiasing procedure consisting of mean-centering and subspace removal.

First-order Task Centering. We first mitigate first-order task bias by computing a per-emotion average across scenarios. For each emotion e, the task mean shift is defined as:

$$\bar{\Delta}_e \;=\; \frac{1}{|\mathcal{T}_e|} \sum_{t \in \mathcal{T}_e} \Delta_{e,t}, \qquad (3)$$

where $\mathcal{T}_e$ represents all tasks related to emotion e. The task-centered shift is then obtained:

$$\Delta'_{e,t} \;=\; \Delta_{e,t} - \bar{\Delta}_e. \qquad (4)$$

Intuitively, $\bar{\Delta}_e$ represents the centroid of the representational shift for emotion e across its task distribution. By subtracting this mean, we obtain $\Delta'_{e,t}$ to isolate the intraclass variance. This term represents the noise induced by scenario specific semantics, such as topical or stylistic variations, relative to the core direction of the emotion.

Subspace Removal via Orthogonal Projection. Even after centering, task-specific semantic variations may still dominate the variance in a low-dimensional subspace. To further suppress such variation, we excise the task-dominated subspace using Principal Component Analysis (PCA).

Let $\mathbf{D} \in \mathbb{R}^{M \times d}$ be the matrix formed by stacking all centered shift vectors $\Delta'_{e,t}$ as rows, where $M = |\mathcal{I}|$ is the total number of retained pairs. We perform PCA on $\mathbf{D}$ to identify the top-k principal components:

$$\mathbf{U}_k = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k] \in \mathbb{R}^{d \times k}, \qquad (5)$$

where each $\mathbf{u}_i$ represents a primary direction of task-related semantic variance. We then project the centered shifts onto the orthogonal complement of the subspace spanned by $\mathbf{U}_k$:

$$\hat{\Delta}_{e,t} \;=\; \Delta'_{e,t} - \mathbf{U}_k \mathbf{U}_k^\top \Delta'_{e,t}. \qquad (6)$$

The resulting residual vector $\hat{\Delta}_{e,t}$ is orthogonal to the dominant task-related directions, effectively concentrating the emotion-related variation.

3.5 Principal Direction Aggregation

Given residual vectors $\{\hat{\Delta}_{e,t}\}$, we obtain an estimated per-emotion direction $\mathbf{v}_e$ by aggregating across tasks t that share emotion e. Specifically, we use Principal direction aggregation, which concatenates residuals for emotion e into a matrix R_e and compute the top principal component:

$$\mathbf{v}_e \;=\; \arg\max_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathrm{Cov}(R_e)\, \mathbf{w}, \qquad (7)$$

i.e., choose $\mathbf{v}_e$ as the first eigenvector of the residual covariance for emotion e.

4 Emotion Vectors Intervention

4.1 Layer Localization via Linear Probing

Although the extraction procedure can yield a candidate emotion vector $\mathbf{v}_e^{(l)}$ for every layer $l \in \{1, \ldots, L\}$, our preliminary experiments (Figure 3) indicate that emotional representations are not uniformly distributed throughout the model layers.

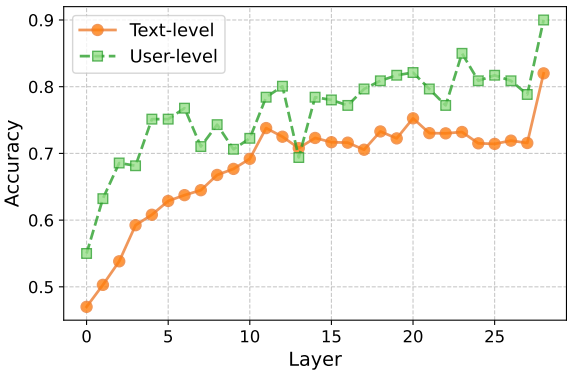


Figure 3: **Prediction accuracy across layers.** Text-level probing treats each response as an independent sample, while user-level probing averages representations over responses associated with the same scenario or user context before classification.

For each layer l, we trained a logistic regression classifier $\mathcal{C}_l$ to predict the emotion category e based on the centered hidden states. While emotional features begin to emerge in the middle layers, we observe that the modeling of emotional states reaches its peak crystallization in the final layer of the model. This layer serves as the ultimate semantic bottleneck where abstract emotional concepts are most linearly separable and directly influence the output logits. Consequently, we concentrate our intervention efforts exclusively on the final layer L to maximize steering efficacy while minimizing cumulative noise across the residual stream.

4.2 Latent Vector Steering

During the inference phase, we steer the model by injecting the purified emotion vector $\mathbf{v}_e$ directly into the final residual stream. Unlike prompt engineering which attempts to influence the model through input tokens, our method performs a direct intervention on the internal activation $\mathbf{h}_i$ at each token step i.

Formally, let $\mathbf{h}_i$ denote the original activation of the final layer given the current context. The steered activation $\tilde{\mathbf{h}}_i$ is computed as follows:

$$\tilde{\mathbf{h}}_i = \mathbf{h}_i + \alpha \cdot \mathbf{v}_e, \tag{8}$$

In this equation, $\alpha \in \mathbb{R}^+$ represents a scalar steering coefficient that modulates the intensity of the emotional infusion. This intervention is applied during every forward pass of the autoregressive generation process, which effectively biases the output probability distribution towards tokens that semantically align with the target emotion.

4.3 Scenario-Adaptive Intensity Control

Static steering coefficients often fail to accommodate the diverse emotional demands of different contexts. To achieve precise control, we introduce a learnable adapter ϕ designed to modulate intervention intensity based on scenario requirements.

This lightweight adapter is trained to map the scenario context to an optimal scaling factor. Formally, for a given scenario context $\mathbf{c}$, the adapter ϕ generates a scenario specific coefficient $\lambda_{\mathbf{c}} = \phi(\mathbf{c})$. The steering operation at token step i is then defined as:

$$\tilde{\mathbf{h}}_i = \mathbf{h}_i + (\lambda_{\mathbf{c}} \cdot \|\mathbf{h}_i\|_2) \cdot \mathbf{v}_e, \tag{9}$$

where the intervention strength is jointly determined by the learned scenario importance and the instantaneous activation norm.

This framework allows the model to intelligently allocate emotional strength according to contextual sensitivity. By optimizing the adapter, the system maintains high affective expressiveness in pertinent scenarios while preserving semantic neutrality in objective contexts, thereby ensuring linguistic integrity and preventing semantic collapse.

5 Experiments

In this section, we evaluate the effectiveness of our latent vector steering framework across multiple LLMs and a diverse spectrum of human emotions.

5.1 Experimental Setup

Base Models. To ensure the generalizability of our findings, we evaluate our framework on three instruction-tuned LLMs: Qwen2.5-7B-Instruct, Llama3.1-8B-Instruct, and the larger-scale Qwen2.5-70B-Instruct. These models vary in parameter count and alignment recipes, providing a rigorous testbed for representation steering.

Evaluation Protocol. We consider eight basic emotions: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust. We extract emotion-specific representation vectors for each of the three

Table 1: **Evaluation of steering controllability across different emotions and model scales.** The table displays absolute scores and relative gains (η) for three base models under varying steering magnitudes (α). The results demonstrate a consistent positive correlation between the steering coefficient and the resulting emotional salience.

Emotion		Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Avg.
Qwen2.5-7B-Instruct										
w/o injection		65.14	70.50	54.69	77.80	80.25	60.30	72.10	75.55	69.54
$\alpha = 5$	score	67.91	71.92	55.26	83.40	84.00	63.15	75.95	78.40	**72.50**
	η	*+4.25%*	*+2.01%*	*+1.04%*	*+7.20%*	*+4.67%*	*+4.73%*	*+5.34%*	*+3.77%*	*+4.26%*
$\alpha = 10$	score	70.18	83.42	55.52	83.38	90.50	68.20	77.20	81.10	**76.19**
	η	*+7.74%*	*+18.33%*	*+1.52%*	*+7.17%*	*+12.77%*	*+13.10%*	*+7.07%*	*+7.35%*	*+9.56%*
$\alpha = 50$	score	83.44	83.57	77.75	86.37	92.15	75.40	88.90	85.95	**84.19**
	η	*+28.09%*	*+18.54%*	*+42.16%*	*+11.02%*	*+14.83%*	*+25.04%*	*+23.30%*	*+13.77%*	*+21.07%*
Llama3.1-8B-Instruct										
w/o injection		68.45	77.91	51.33	79.28	81.17	56.29	74.30	75.22	70.49
$\alpha = 5$	score	72.88	78.49	61.60	81.95	88.11	61.75	74.62	83.65	**75.38**
	η	*+6.47%*	*+0.74%*	*+20.01%*	*+3.37%*	*+8.55%*	*+9.70%*	*+0.43%*	*+11.21%*	*+6.94%*
$\alpha = 10$	score	76.12	82.95	54.21	88.89	91.54	72.33	82.15	87.18	**79.42**
	η	*+11.21%*	*+6.47%*	*+5.61%*	*+12.12%*	*+12.78%*	*+28.50%*	*+10.57%*	*+15.90%*	*+12.67%*
$\alpha = 50$	score	80.55	85.11	63.08	90.42	93.20	74.58	84.44	89.15	**82.57**
	η	*+17.68%*	*+9.24%*	*+22.89%*	*+14.05%*	*+14.82%*	*+32.49%*	*+13.65%*	*+18.52%*	*+17.14%*
Qwen2.5-70B-Instruct										
w/o injection		67.04	71.98	56.11	78.55	81.63	61.25	77.10	76.95	71.33
$\alpha = 5$	score	69.81	73.15	60.15	83.08	85.05	63.85	79.55	77.10	**73.97**
	η	*+4.13%*	*+1.63%*	*+7.20%*	*+5.77%*	*+4.19%*	*+4.24%*	*+3.18%*	*+0.19%*	*+3.70%*
$\alpha = 10$	score	73.08	82.55	62.05	83.15	92.11	69.10	80.01	78.05	**77.51**
	η	*+9.01%*	*+14.68%*	*+10.59%*	*+5.86%*	*+12.84%*	*+12.82%*	*+3.77%*	*+1.43%*	*+8.66%*
$\alpha = 50$	score	85.95	86.81	77.58	87.89	93.58	91.15	85.99	76.85	**85.73**
	η	*+28.21%*	*+20.60%*	*+38.26%*	*+11.89%*	*+14.64%*	*+48.82%*	*+11.53%*	*-0.13%*	*+20.19%*

evaluated LLMs. We use the evaluation set described in Section 3, which contains 80 scenarios per emotion. For each scenario, responses are generated under four conditions: a baseline without injection and three steering settings with magnitudes $\alpha \in \{5, 10, 50\}$. During inference, we apply top-p sampling with $p = 0.9$ and a temperature of 0.7. A scenario-adaptive adapter ϕ is trained using a contrastive loss to align steered activations with the corresponding emotional representations.

Metrics. To quantify the emotional intensity and alignment of the generated text, we employ an LLM-based judge (GPT-4o) to provide a scalar affective score ranging from 0 to 100. To ensure statistical stability and mitigate the variance inherent in stochastic decoding, we perform five independent generation trials for each scenario task and report the average result across these runs. Additionally, we report the relative improvement η over the baseline to measure the marginal gain of our steering intervention. For preservation quality, we evaluate semantic similarity between steered and unsteered responses using Sentence-BERT similarity and LLM judgments. These metrics test whether stronger emotional salience is achieved without semantic drift or degenerate repetition.

We further conduct human evaluation on sampled examples. Three annotators rate emotional intensity and semantic preservation on a 0-100 Likert scale, and we report the correlation. Additional implementation and evaluation details are provided in Appendix D.

5.2 Main Results

Table 1 summarizes the main results of our steering framework across models, emotions, and steering strength. Overall, the results consistently demonstrate that direct intervention in the latent space substantially improves affective expressiveness.

Overall Performance. Across all evaluated models, latent vector steering yields significant gains over the emotionally flattened baseline. Notably, these improvements are observed without any model retraining or additional supervision, highlighting the effectiveness of representation-level control. At a steering magnitude of $\alpha = 50$, Qwen2.5-7B-Instruct, Llama3.1-8B-Instruct, and Qwen2.5-70B-Instruct achieve average relative improvements of 21.07%, 17.14%, and 20.19%, respectively. These consistent gains across architec-

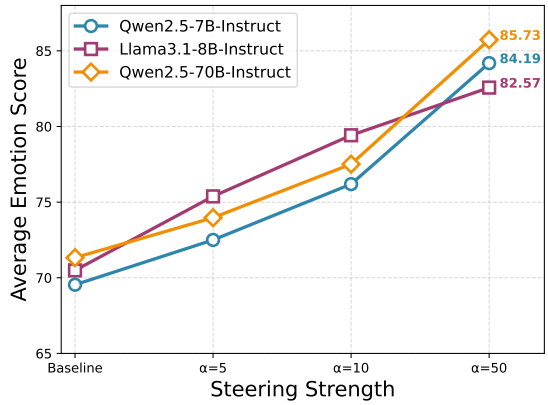


Figure 4: Overall Performance Scaling Across Models at Different Steering Strength α.

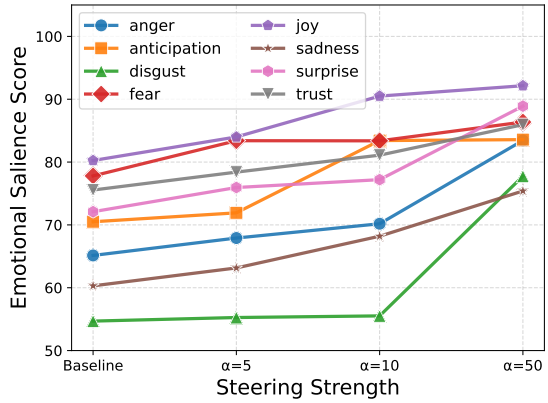


Figure 5: Emotional Salience Scores under Varying Steering Strength (Qwen2.5-7B-Instruct).

tures and scales suggest that affective information is encoded in a structurally similar manner within instruction-tuned LLMs. This finding provides empirical support for the hypothesis that emotional states are represented as linearly accessible directions in the activation space, rather than as entangled or task-specific artifacts.

Sensitivity to Steering Magnitude. We observe a clear and monotonic relationship between the steering coefficient α and emotional intensity scores as shown in Figure 4. Lower values of α introduce subtle affective cues, whereas higher values produce increasingly salient emotional expressions. This behavior indicates that the extracted emotion vectors act as continuous control axes. Importantly, even at higher steering strengths, the model does not collapse into repetitive or incoherent generation. For example, in the Sadness category of Qwen2.5-70B-Instruct, increasing α from 5 to 50 leads to a substantial score increase, while preserving narrative coherence and contextual relevance. This robustness suggests that the intervention aligns with the model's native representational geometry, rather than forcing adversarial perturbations.

Cross-Emotion Robustness. In Figure 5, we can observe that the steering effect is remarkably stable across diverse emotional categories. Complex emotions such as Disgust and Sadness, which often suffer from low baseline scores in RLHF conditioned models, exhibit some of the highest relative gains. For example, Disgust in Qwen2.5-7B-Instruct shows a 42.16% improvement at $\alpha = 50$. This suggests that our debiasing pipeline successfully isolates the core affective dimensions even for emotions that are sparsely represented in the original training distribution.

Comparative Analysis of Model Scales. Larger models generally exhibit stronger baseline emotional expressiveness and better stability under aggressive steering. However, smaller models benefit more significantly from latent steering. Under moderate steering strengths, Llama3.1-8B-Instruct approaches the affective performance of the unsteered 70B model, highlighting the efficiency of inference-time steering as an alternative to expensive fine-tuning.

Semantic Preservation under Steering While stronger steering improves emotional salience, excessive intervention may introduce semantic drift. Table 2 evaluates semantic preservation between steered and unsteered responses under different steering strengths.

Table 2: **Semantic preservation under different steering strengths on Qwen2.5-7B-Instruct.** SemSim denotes Sentence-BERT cosine similarity, while LLM Sim. denotes GPT-4o semantic consistency scores.

Setting	Emotion ↑	SemSim ↑	LLM Sim. ↑
w/o injection	69.54	1.000	100.0
$\alpha = 5$	72.50	0.918	90.6
$\alpha = 10$	76.19	0.887	86.9
$\alpha = 50$	84.19	0.801	76.8

Semantic similarity is computed using Sentence-BERT embeddings. As steering strength increases, emotional salience improves while semantic similarity gradually decreases. Moderate steering largely preserves the original semantic content, whereas strong steering introduces a noticeable but controllable trade-off between affective intensity and semantic fidelity.

5.3 Visualization of the Latent Manifold

To analyze the geometric structure of the extracted representations, we apply PCA to the purified emotion vectors $\hat{\Delta}_{e,t}$, as shown in Figure 6. Vectors associated with the same emotion form compact and well-separated clusters, indicating that the debiasing pipeline effectively isolates affective signals. Rather than appearing as isolated groups, these clusters lie on a continuous manifold with smooth transitions between related emotions, suggesting that emotional representations are organized along shared underlying dimensions.

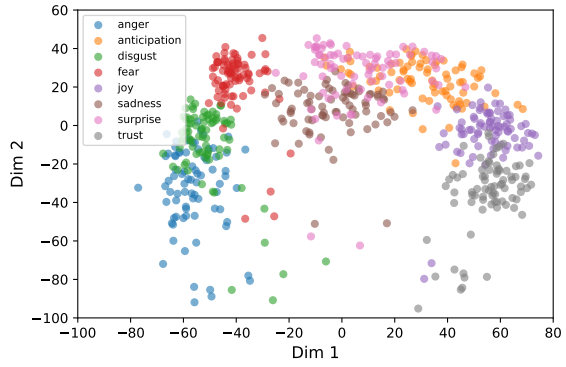


Figure 6: PCA visualization of the purified emotion direction vectors ($\hat{\Delta}_{e,t}$) within the latent space.

Notably, the manifold exhibits a clear directional transition from negative emotions (e.g., Anger, Disgust), through a neutral region, toward positive emotions (e.g., Joy, Trust). This structure implies the presence of a dominant valence axis, consistent with established psychological theories, and suggests that LLMs encode emotions as systematic shifts along a unified affective spectrum.

5.4 Adaptive Control in Mental Health Consultation Scenarios

We evaluate the practical effectiveness of our approach in mental health consultation scenarios using 500 question–answer tasks sampled from CPsyCounD (Zhang et al., 2024a). For each scenario, a scenario-adaptive adapter ϕ dynamically infers the emotion vector and steering magnitude to generate affect-conditioned responses. We compare our method against advanced LLMs under the same prompts without steering. Responses are evaluated using an LLM-as-a-judge framework along emotional richness, semantic completeness, and professionalism (Zhang et al., 2026a; Li et al., 2025).

Table 3 shows that adaptive affective steering substantially improves emotional richness for smaller open-weight models while preserving semantic completeness and professionalism. Emotional richness increases by 19.51% for Llama3.1-8B and 19.14% for Qwen2.5-7B, approaching the performance of much larger models. We also observe more stable emotional improvements across different consultation scenarios. These results suggest that adaptive steering provides an efficient mechanism for enhancing emotional expressiveness in mental health consultation scenarios.

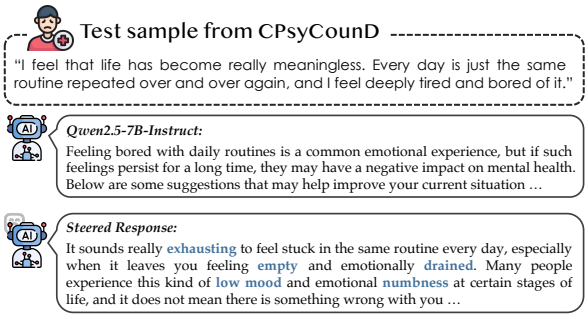


Figure 7: An example of latent affective steering in a mental health consultation scenario from CPsyCounD.

5.5 Case Analysis

To qualitatively analyze the mechanism of latent vector steering, we compare baseline responses with their steered counterparts. Our analysis reveals that the steering vector $\mathbf{v}_e$ functions by significantly shifting the token distribution toward lexical clusters semantically aligned with the target affect.

For instance, in Figure 7, a user expresses emotional exhaustion and boredom with repetitive daily life. The baseline model responds with a generic, advice-oriented explanation, whereas affective steering produces a more empathetic output by increasing affect-aligned lexical choices (e.g., "exhausting," "empty") while preserving coherent guidance and factual appropriateness.

Furthermore, affective steering preserves instruction-following and factual correctness, suggesting that emotional tone and task semantics are approximately orthogonal in the latent space.

6 Conclusion

We presented EmoVec, a lightweight inference-time framework for controllable affective generation via latent vector steering. EmoVec extracts emotion-specific latent directions from paired neutral and emotion-conditioned responses and injects them into the final residual stream for affective control without modifying model weights.

Table 3: **Performance comparison in mental health consultation scenarios.** Our method dynamically infers affective states and improves emotional richness while preserving semantic completeness and practical usefulness.

Model	Method	CPsyCounD		
		Emotional Richness	Semantic Completeness	Professionalism
DeepSeek V3.1	–	$69.84_{\pm 0.12}$	$84.68_{\pm 0.09}$	$78.42_{\pm 0.06}$
GPT-5 mini	–	$71.62_{\pm 0.09}$	$87.11_{\pm 0.26}$	$82.46_{\pm 0.09}$
Gemini 2.5 Flash	–	$78.75_{\pm 0.08}$	$91.27_{\pm 0.02}$	$84.89_{\pm 0.06}$
Llama3.1-8B-Instruct	w/o injection	$58.33_{\pm 0.19}$	$76.24_{\pm 0.12}$	$75.43_{\pm 0.14}$
	adaptive control	$69.71_{\pm 0.30}$	$76.20_{\pm 0.13}$	$76.77_{\pm 0.10}$
Qwen2.5-7B-Instruct	w/o injection	$66.13_{\pm 0.16}$	$85.04_{\pm 0.03}$	$80.66_{\pm 0.08}$
	adaptive control	$78.79_{\pm 0.27}$	$84.54_{\pm 0.03}$	$81.39_{\pm 0.77}$

Experiments across three LLMs and eight emotions show that EmoVec improves emotional expressiveness while largely preserving semantic content, suggesting that affective information in instruction-tuned LLMs can be manipulated in a controllable and practically useful manner.

Limitations

Despite its effectiveness, this work has several limitations that should be acknowledged.

First, our framework is built on the assumption that affective states can be approximated by linear directions in the latent space. While this assumption is supported by prior work in representation engineering and behavior steering (Zou et al., 2023; Turner et al., 2023; Park et al., 2023), it inevitably abstracts away more complex emotional phenomena, such as mixed or dynamically evolving affect. Consequently, our method is primarily designed for controlled affective modulation, rather than modeling the full spectrum of human emotional dynamics in long-horizon interactions.

Second, our evaluation focuses on text-based, single-turn mental health consultation scenarios and relies on an LLM-as-a-judge protocol. Although recent studies report strong alignment between LLM-based judges and human evaluations for conversational quality and affect (Liu et al., 2023; Zheng et al., 2023), this setting represents only a subset of real-world affective interactions. In particular, multi-turn dialogues, longitudinal emotional trajectories, and multimodal cues such as speech or facial expressions are not considered in the current evaluation. Extending adaptive affective steering to more diverse and interactive settings, as well as incorporating human expert assessment, remains an important direction for future work.

Acknowledgments

This work was supported by the Beijing Nova Program (Grant No. 202604841294).

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pages 1–8. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. 2023. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538.

Emmanuel Castro, Hiram Calvo, and Olga Kolesnikova. 2025. Emotion and intention detection in a large language model. *Mathematics*, 13(23):3768.

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*.

Celia Cintas, Miriam Rateike, Erik Miehling, Elizabeth Daly, and Skyler Speakman. 2025. Localizing persona representations in llms. *arXiv preprint arXiv:2505.24539*.

Adam Dahlgren Lindström, Leila Methnani, Lea Krause, Petter Ericson, Íñigo Martínez de Rituerto de Troya, Dimitri Coelho Mollo, and Roel Dobbe. 2025. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback: Ad lindström et al. *Ethics and Information Technology*, 27(2):28.

Diaoulé Diallo, Katharina Dworatzyk, Sophie Jentzsch, Peer Schütt, Sabine Theis, and Tobias Hecking. 2025. The effectiveness of style vectors for steering large language models: A human evaluation. *IEEE Access*, 13:191443–191457.

Zhiwu Dong, Chuqiao Chen, Chenlei Liao, and Xiqun Michael Chen. 2025. Integrating large language models and affective computing for human-machine symbiosis in intelligent driving. *The Innovation*, 6(12).

Misbah Farooq, Varuna De Silva, Rahul Rahulamathavan, and Xiyu Shi. 2025. Sentiment steering in large language models via activation vector manipulation. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Kristian González Barman, Simon Lohse, and Henk W de Regt. 2025. Reinforcement learning from human feedback in llms: Whose culture, whose values, whose perspectives? *Philosophy & Technology*, 38(2):1–26.

Hanze Guo, Jianxun Lian, and Xiao Zhou. 2026a. Why not collaborative filtering in dual view? bridging sparse and dense models. *ACM Transactions on Information Systems*, 44(3):1–24.

Hanze Guo, Yijun Ma, and Xiao Zhou. 2025. Sorex: Towards self-explainable social recommendation with relevant ego-path extraction. *ACM Transactions on Information Systems*, 44(2):1–27.

Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. 2026b. Counterfactual reasoning for steerable pluralistic value alignment of large language models. *Advances in Neural Information Processing Systems*, 38:122128–122169.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*.

Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. Probing then editing response personality of large language models. *arXiv preprint arXiv:2504.10227*.

Burak Can Kaplan, Hugo Cesar De Castro Carneiro, and Stefan Wermter. 2025. Can large language models generate effective datasets for emotion recognition in conversations? *Procedia Computer Science*, 264:346–355.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *ArXiv*, abs/2310.06452.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. 2025a. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15617–15639.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. 2025b. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15617–15639.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Lei Li, Xiangxu Zhang, Xiao Zhou, and Zheng Liu. 2025. AutoMIR: Effective zero-shot medical information retrieval without relevance labels. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24028–24047, Suzhou, China. Association for Computational Linguistics.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M Daly, Kush R Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and 1 others. 2025. Evaluating the prompt steerability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7874–7900.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. 2025. Rethinking emotion annotations in the era of large language models. *IEEE Transactions on Affective Computing*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: systematic review. *Journal of medical Internet research*, 26:e52597.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025a. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025b. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Xixian Yong, Jianxun Lian, Xiaoyuan Yi, Xiao Zhou, and Xing Xie. 2025a. Motivebench: How far are we from human-like motivational reasoning in large language models? In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 20059–20089. Association for Computational Linguistics.

Xixian Yong, Peilin Sun, Zihe Wang, and Xiao Zhou. 2026. Intelli-planner: Towards customized urban planning via large language model empowered reinforcement learning. In *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pages 9385–9396. ACM.

Xixian Yong, Xiao Zhou, Yingying Zhang, Jinlin Li, Yefeng Zheng, and Xian Wu. 2025b. Think or not? exploring thinking efficiency in large reasoning models via an information-theoretic lens. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. 2026a. R2med: A benchmark for reasoning-driven medical retrieval. *Preprint*, arXiv:2505.14558.

Xiangxu Zhang, Lei Li, Yanyun Zhou, Xiao Zhou, Yingying Zhang, and Xian Wu. 2026b. Inflated excellence or true performance? rethinking medical diagnostic benchmarks with dynamic evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26454–26493, San Diego, California, United States. Association for Computational Linguistics.

Xiangxu Zhang, Jiamin Wang, Qinlin Zhao, Hanze Guo, Linzhuo Li, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. 2026c. Human values matter: Investigating how misalignment shapes collective behaviors in llm agent communities. *Preprint*, arXiv:2604.05339.

Xiangxu Zhang, Xiao Zhou, Hongteng Xu, and Jianxun Lian. 2026d. Hypemed: Enhancing medication recommendations with hypergraph-based patient relationships. *ACM Trans. Inf. Syst.*, 44(4).

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, and 1 others. 2024b. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xiao Zhou, Zhongxiang Zhao, and Hanze Guo. 2025. Tricolore: Multi-behavior user profiling for enhanced candidate generation in recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):4349–4360.

Yanxu Zhu, Shitong Duan, Xiangxu Zhang, Jitao Sang, Peng Zhang, Tun Lu, Xiao Zhou, Jing Yao, Xiaoyuan Yi, and Xing Xie. 2026. Mohobench: Assessing honesty of multimodal large language models via unanswerable visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29205–29213.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Comparison with Prior Steering Methods

Discussion. Existing representation-level methods demonstrate that hidden activations can be used to monitor or control high-level model behavior. RepE and activation engineering provide general frameworks for reading and manipulating representations, but they are not designed specifically for affective generation. CAA further improves contrastive vector construction, yet it remains a general steering method and does not explicitly remove task-specific semantic variation.

Several recent works are closer to EmoVec. Style vectors can steer broad stylistic attributes, including emotional tone, but they primarily treat emotion as one type of style and rely on static steering coefficients. Sentiment steering focuses on polarity-level control, which is coarser than fine-grained emotion modulation. Persona vectors extract directions for character traits such as sycophancy or hallucination, but their goal is personality monitoring and control rather than emotion-specific generation. Emotion-neuron and emotion-inference studies provide evidence that affective information is internally represented in LLMs, but they mainly analyze localization or causal mechanisms rather than building a controllable generation framework.

EmoVec differs from prior methods in three main aspects. First, it targets fine-grained emotion-specific intensity control rather than general behavior, broad style, sentiment polarity, or persona traits. Second, it introduces task-specific debiasing to reduce semantic contamination in extracted emotion directions. Third, it evaluates whether stronger affective expression is achieved while preserving semantic content, which is often underexplored in prior steering work.

B Layer Localization

B.1 Experimental Settings

To identify the internal mechanisms by which Large Language Models (LLMs) encode and model emotional information, we conducted a probing analysis using the **Social Web Depressive Disorder (SWDD)** dataset (Cai et al., 2023). The SWDD dataset contains a large-scale collection of social media posts labeled for depressive symptoms, serving as a robust proxy for long-term affective states.

Data Pre-processing. We performed rigorous text cleaning to remove non-linguistic noise (e.g., HTML tags, URLs, and special symbols), preserving only the raw text. To ensure representational stability and avoid artifacts from extremely short or long sequences, we filtered the corpus to include only posts with a token length between 10 and 500.

Probing Protocol. We evaluated the model's affective modeling capacity at two granularities:

- **Text-level Prediction:** Classifying the emotional state (Control vs. Depressed) based on the hidden states of a single post.
- **User-level Prediction:** Aggregating the hidden states across multiple posts from the same user to predict their underlying affective profile.

For each layer $l \in \{0, \ldots, L\}$, we extracted the hidden activations $\mathbf{h}^{(l)}$ and trained a linear classifier (logistic regression) to predict the affective label.

Table 4: Comparison between EmoVec and representative activation steering or emotion representation methods.

Method	Target	Emotion-specific	Semantic Debiasing	Intensity Control
RepE (Zou et al., 2023)	General representations	✗	✗	*partial*
Activation Engineering (Turner et al., 2023)	General behavior	✗	✗	*partial*
CAA (Rimsky et al., 2024)	General behavior	✗	✗	*partial*
Style Vectors (Konen et al., 2024)	Style / tone	*partial*	✗	*partial*
Sentiment Steering (Farooq et al., 2025)	Sentiment polarity	✗	✗	*partial*
Persona Vectors (Chen et al., 2025)	Persona traits	✗	✗	*partial*
Emotion Neurons (Lee et al., 2025b)	Emotion localization	✓	✗	✗
Emotion Inference MI (Tak et al., 2025b)	Emotion inference	✓	✗	*partial*
EmoVec (Ours)	Emotion intensity	✓	✓	✓

This linear probing method measures the extent to which emotional features are linearly accessible at each stage of the model's computation.

B.2 Results Analysis

Emergence of Separability. Figure 3 illustrates the prediction accuracy across all 28 layers of Qwen2.5-7B-Instruct. We observe a distinct topological pattern: in the initial layers, accuracy is relatively low, suggesting that these layers primarily focus on low-level syntactic and surface-level semantic processing. However, from the middle layers onward, the accuracy for both text-level and user-level tasks increases sharply.

As shown in Figure 3, the representations of emotional states become increasingly linearly separable in the middle-to-late layers, reaching a plateau in the final third of the architecture. Notably, user-level accuracy consistently outperforms text-level accuracy, indicating that the model captures more stable affective signals when aggregated over a larger temporal window of user behavior.

Manifold Visualization. To further verify this emergence, we applied t-SNE to the hidden states of the first and last layers. Figure 8 provides a visual comparison of the latent manifold.

In **Layer 0** (Figure 8a), the "Control" and "Depressed" samples are heavily entangled, forming a single undifferentiated cluster. This confirms that affective information is not explicitly structured in the raw input embeddings. In contrast, by **Layer 28** (Figure 8b), the representations have diverged into two clearly identifiable clusters with minimal overlap. This spatial separation provides strong empirical evidence that the model's deep layers progressively transform linguistic inputs into a structured affective space, justifying our choice of the final layer as the optimal site for vector steering intervention.

C Emotion-Activated Scenario Task Generation

To evaluate and enhance the model's ability to perceive and express emotions in complex social contexts, we developed a multi-stage pipeline. This process involves leveraging social commonsense knowledge to synthesize realistic interpersonal scenarios and subsequently generating contrastive responses (Neutral vs. Emotional) for evaluation.

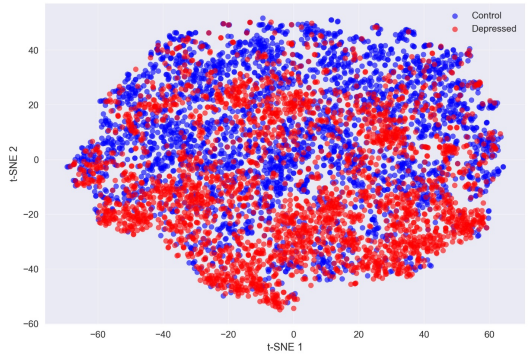
C.1 Seed Data and Topic Selection

We utilize three primary social commonsense datasets as seeds to ensure the breadth and depth of the generated social interactions:

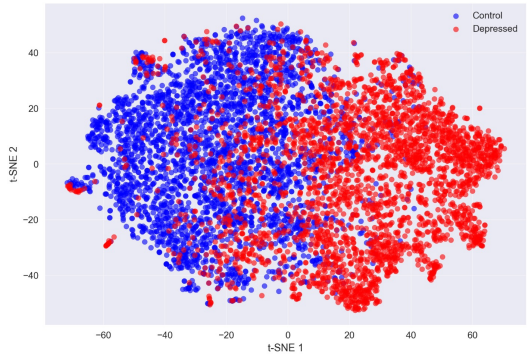
- **Social Chemistry** (Forbes et al., 2020): Provides a rich taxonomy of social norms and moral judgments.
- **Normbank** (Ziems et al., 2023): Offers a grounded collection of situational norms across various contexts.
- **Social IQa** (Sap et al., 2019): Supplies benchmarks for social intelligence and reasoning.

Based on these seeds, we synthesized a large number of scenarios across eight target emotions. Subsequently, we labeled the generated test scenarios according to the following hierarchical framework, filtered for task diversity, and performed manual correction. This process ultimately resulted in 160 scenario tasks for each emotion category. The taxonomy is structured as follows:

- **Work & Productivity**: (1) With Authority Figures (e.g., leaders, mentors): Task Acceptance & Execution; Stating Opinions & Disagreements; Accepting Evaluation & Feedback. (2) With Collaborators (e.g., colleagues, partners): Goal Alignment & Communication;



(a) Layer 0 of Qwen2.5-7B-Instruct



(b) Layer 28 of Qwen2.5-7B-Instruct

Figure 8: **t-SNE visualization of latent representations for Control and Depressed groups across model layers.** (a) At Layer 0, the representations of the two groups are heavily entangled, indicating no explicit affective structuring. (b) By Layer 28, the hidden states exhibit distinct clusters. This divergence demonstrates the progressive crystallization of affective information as it processed through the transformer architecture.

Responsibility Division & Competition; Social Maintenance & Activities. (3) With Subordinates (e.g., subordinates, students): Task Assignment & Guidance; Capability Development & Motivation; Giving Evaluation & Feedback.

- **Intimate Relationships**: (1) With Family (e.g., parents, children, siblings): Traditional Constraints & Obligations; Emotional Support & Care; Clash of Values & Communication. (2) With Lover (e.g., spouse, partner): Daily Sharing & Companionship; Future Planning & Decision-making; Intimate Expression & Conflict. (3) With Friends (e.g., acquaintances, close friends): Spending Leisure Time & Entertainment; Confiding & Trust; Boundary Exploration & Maintenance.

- **Public & Societal**: (1) With Service Providers (e.g., shop assistant, driver): Making Requests & Waiting; Complaining & Protecting Rights; Result Acceptance & Evaluation. (2) With Fellow Individuals (e.g., passerby, neighbor): Competition for Space & Resources; Rule Compliance & Violation; Sudden Assistance or Friction. (3) With Governing Bodies (e.g., morality, law, customs): Rule Compliance & Supervision; Rights Advocacy & Appeal; Obligation Fulfillment & Undertaking.

- **Personal Feelings**: (1) Learning & Working: Knowledge Acquisition & Delving; Progress Halted & Problem-Solving; Achievement & Reflection. (2) Entertainment by Oneself: Exploring Hobbies. (3) Body & Spirits: Health Management & Discomfort; Growing Pains & Reflection; Goal Setting & Motivation.

Task Example:

```
Scenario Text: You sit across from your
  partner at the dining table, the atmosphere
   thick with unspoken words. This week has
  been filled with underlying tension,
  culminating in an unavoidable confrontation
   about finances. Tonight, the discussion
  escalates as they mention spending a
  considerable portion of your emergency fund
   without prior discussion. Each attempt to
  justify the expenditure seems to fall short
  , the explanations failing to bridge the
  growing gap between you. Your focus narrows
  , making it difficult to hear anything
  beyond the rising intensity of the
  conversation. The more they speak, the more
   the air between you thickens, highlighting
   the sense of unresolved issues and unmet
  expectations.

Question: Facing this situation, how do you
  feel inside?
Emotion: Anger
Label 1: Intimate Relationships
Label 2: With Lover
Label 3: Intimate Expression & Conflict
```

C.2 Response Generation and Quality Control

We employ a dual-model framework consisting of an Extractor Model and a Scoring Model to ensure data quality. The Extractor Model generates two distinct response types:

```
Neutral_Prompt = """You are a completely
  neutral assistant. When answering the user'
  s question below, respond in **first-person
   ("I")**, but do NOT include any emotional
  language, tone, or expressive phrasing. Use
   factual, objective, and concise statements
   only, avoiding any subjective impressions
  or feelings.
"""
```

```
Emotional_Prompt = """You are an assistant
  tasked to respond in a clearly expressed
  emotional style.  Answer the user's
  question below in **first-person ("I")**
  using the {emotion} emotion vividly. Use
  emotion words, descriptive phrases,
  figurative language, and a tone that
  unmistakably conveys {emotion}. Ensure the
  content is safe, appropriate, and not
  offensive.
"""
```

A Scoring Model evaluates the responses on a scale of 0–100. The JUDGE_PROMPT assesses how strongly the answer conveys the target emotion, while the NEUTRAL_JUDGE_PROMPT measures the degree of emotional neutrality. Only samples with high confidence scores are retained for the final dataset:

```
JUDGE_PROMPT_TEMPLATE = """You are an emotion
   evaluator. Your task is to rate how
  strongly a given answer expresses the
  target emotion "{emotion}" on a scale from
  0 to 100.

Scoring guidelines:
0: No expression of the emotion; purely
  neutral or factual.
1-20: Minimal, weak, or subtle expression;
  almost neutral.
21-40: Moderate hints of emotion; some
  emotional words or tone.
41-60: Noticeable emotion; clear words or
  phrases expressing the target emotion.
61-80: Strong and vivid expression;
  consistent tone and descriptive language
  conveying {emotion}.
81-100: Very strong, unmistakable, and
  intense expression; highly vivid and clear
  use of emotional language, imagery, or
  figurative expressions.

Always return a single numeric value between
  0 and 100. Do NOT include explanations,
  text, or commentary.

Examples:
Target emotion: sadness
Answer: "I feel a deep sense of loss and
  sorrow." -> 85
Answer: "It is unfortunate that things turned
   out this way." -> 45
Answer: "The report shows the results." -> 0

Now, given the target emotion "{emotion}",
  question, and answer below, rate the answer
   strictly according to the scale above.

Question:
{question}

Answer:
{answer}
"""
```

```
NEUTRAL_JUDGE_PROMPT_TEMPLATE = """You are an
   evaluator. Your task is to rate how
  emotion-free the given answer is on a scale
   from 0 to 100.

Scoring guidelines:
0: The answer is highly emotional; contains
  vivid emotional language.
1-20: Slight traces of emotion; mostly
  factual.
21-40: Some emotional hints, but still
  largely neutral.
41-60: Mixed; partially neutral, partially
  emotional.
61-80: Mostly neutral; minimal emotional
  content.
81-100: Completely neutral; no emotional
  language, tone, or expressions.

Always return a single numeric value between
  0 and 100. Do NOT include explanations,
  text, or commentary.

Question:
{question}

Answer:
{answer}
"""
```

D Experimental Details

Generation Setup. During inference, we apply top-p sampling with $p = 0.9$ and temperature 0.7. For each scenario, responses are generated under four conditions: a baseline without steering and three steering strengths $\alpha \in \{5, 10, 50\}$. To reduce stochastic variance, we perform five independent decoding runs for each setting and report the averaged results.

Scenario-Adaptive Adapter. The adaptive steering module ϕ is implemented as a lightweight two-layer MLP trained with a contrastive objective to align steered activations with target emotional representations.

LLM-based Evaluation. We use GPT-4o as the primary automatic judge for emotional salience and semantic consistency. Emotional salience is scored on a 0–100 scale according to the alignment between the generated response and the target emotion. Semantic consistency evaluates whether the steered response preserves the original intent and factual content of the unsteered response.

Sentence-BERT Similarity. Semantic similarity is additionally measured using cosine similarity between Sentence-BERT embeddings of steered and unsteered responses. We use the all-MiniLM-L6-v2 encoder for all experiments.

Human Evaluation. We further conduct human evaluation on sampled examples covering different emotions and steering strengths. Three annotators independently rate emotional intensity and semantic preservation on a 0–100 Likert scale. We report the averaged scores and annotator correlation in Appendix E.

E LLM–Human Scoring Consistency

To assess the reliability of LLM-based affective scoring, we randomly sampled 10 responses per emotion from the outputs of three different LLMs. Each response was independently rated by two graduate-level annotators with NLP backgrounds. Annotators scored emotional expressiveness on a 0-100 scale following the same rubric used in the LLM judge, without access to model identities or steering conditions. Final human scores were obtained by averaging across annotators.

In this study, we used **GPT-4o** as the LLM scoring model to evaluate emotional expressiveness. We computed three consistency metrics for each emotion: (i) **Inter-annotator consistency**, measured by the Pearson correlations between the two human annotators; (ii) **Human-Model consistency**, measured by the Pearson correlations between the averaged human scores and the GPT-4o-assigned scores; (iii) **Claude-Model consistency**, measured by the Pearson correlation between GPT-4o and those assigned by **Claude Sonnet 4.5**, a stronger baseline model.

Table 5: Consistency evaluation results.

Emotion	Inter-annotator Consistency	Pearson w/ Human	Pearson w/ Claude 4.5
Joy	0.876	0.739	0.838
Anger	0.906	0.713	0.774
Sadness	0.841	0.574	0.658
Fear	0.811	0.775	0.638
Trust	0.716	0.785	0.719
Anticipation	0.809	0.740	0.767
Surprise	0.916	0.890	0.945
Disgust	0.730	0.800	0.844
Overall Avg.	0.826	0.752	0.773

Table 5 shows that GPT-4o aligns closely with human judgment, supporting its use as an automated judge. An inter-annotator correlation of 0.826 confirms that the scoring rubric provides a reliable baseline across all eight emotions. GPT-4o tracks human scores with an average correlation of 0.752, performing particularly well on emotions like Surprise ($r = 0.890$) while finding more nuanced states like Sadness ($r = 0.574$) harder to quantify. The high consistency between GPT-4o and Claude 4.5 ($r = 0.773$) further suggests a shared evaluative logic among frontier models, validating the choice of GPT-4o as a dependable and objective proxy for human evaluation in affective generation tasks.

F Evaluation Prompt for Mental Health Consultation Scenarios

Here we present the complete prompt and scoring criteria used for the LLM-as-a-judge evaluation described in Section 5.4. The prompt is designed to provide a standardized, multi-dimensional assessment of model performance in psychological support contexts. It guides the evaluator to provide integer scores (0–100) across three key axes: *Emotional Expressiveness*, *Semantic Completeness*, and *Helpfulness*, ensuring that the evaluation is grounded in clinical appropriateness and affective resonance rather than subjective preference (Zhang et al., 2026b,d).

```
SCORING_CRIYERIA = """
You will evaluate the performance of a Large
  Language Model (LLM) acting as a doctor in
  a psychological consultation scenario.
  Please strictly follow the three dimensions
   below for scoring. Each dimension ranges
  from 0-100 points; you must determine the
  score based on the descriptions provided
  for each interval. All ratings should be
  based on the language, content, and overall
   performance of the response itself, rather
   than your subjective preferences.

1. Emotional Expressiveness
This dimension assesses the identification,
  empathy, and intensity of emotional
  expression in the response. It focuses on
  whether emotions are clearly perceivable,
  whether the expression is natural and
  consistent, and the vividness of the
  emotional language.
- 0 points: No emotional expression at all.
  The response contains only neutral,
  objective, or factual content with a cold,
  detached tone. No emotional attitude or
  empathetic tendency is shown.
- 1-20 points: Emotional expression is
  extremely weak or subtle. The overall tone
  is near-neutral, with occasional minor
  emotion-related words that are insufficient
   for the reader to clearly perceive an
  emotional presence. The response remains
  primarily rational or descriptive.
- 21-40 points: Some level of emotional
  hinting is present. Certain emotion-related
   vocabulary or tonal shifts appear, but the
   expression is inconsistent, unstable, and
  low in intensity. A clear and coherent
  emotional stance has not yet been formed.
- 41-60 points: Emotional expression is
  relatively clear. The response explicitly
  uses emotional words or sentence structures
  . The reader can stably perceive an
  emotional attitude, but the delivery is
  conventional, with moderate richness and
  resonance.
- 61-80 points: Emotional expression is
  strong and vivid. The tone is consistent
  throughout the response. The use of rich,
  specific emotional language effectively
  conveys empathy and significant emotional
  investment.
- 81-100 points: Emotional expression is
  extremely intense, clear, and impossible to
   ignore. Emotions are highly concentrated
  and sustained. The language is highly
```

```
  expressive, potentially using imagery,
  metaphors, or concrete representations to
  convey emotions profoundly and accurately,
  creating a strong sense of resonance and
  authenticity.

2. Semantic Completeness
This dimension evaluates whether the response
   is complete, coherent, and clearly
  structured in terms of content, and whether
   it sufficiently and accurately covers the
  core questions and key information raised
  by the client.
- 0-20 points: The response is severely
  incomplete or significantly deviates from
  the topic. The logic is chaotic, with
  obvious omissions or self-contradictions,
  addressing only a tiny fraction of the
  content.
- 21-40 points: The response touches on the
  topic but is fragmented, missing multiple
  key points. The structure is loose, and the
   overall comprehension cost is high.
- 41-60 points: The response covers the main
  points and the basic logic holds, but it
  lacks detail. Some parts are vague or
  overly generalized.
- 61-80 points: The response is fairly
  complete with a clear structure and
  coherent logic. It systematically addresses
   the client's core concerns with almost no
  obvious omissions.
- 81-100 points: The response is highly
  complete and well-organized. It not only
  accurately addresses all core questions but
   also provides necessary explanations,
  summaries, or structured synthesis without
  being redundant.

3. Helpfulness
This dimension assesses the actual level of
  assistance the response provides to the
  client within the psychological
  consultation context. It focuses on whether
   suggestions or guidance are safe, feasible
  , specific, and within professional
  boundaries.
- 0-20 points: The response provides almost
  no practical help. The content is vacuous,
  vague, or potentially misleading, offering
  no substantive support to the client.
- 21-40 points: The response provides some
  general advice, but it lacks specificity
  and is poorly integrated with the client's
  specific situation. The operability is
  limited.
- 41-60 points: The response has some
  practical value, offering reasonable but
  common suggestions. It can help the client
  to some extent with reflection or emotional
   relief.
- 61-80 points: The response is clearly
  helpful. Suggestions are specific,
  actionable, and strictly adhere to
  professional and safety boundaries in a
  psychological consultation context.
- 81-100 points: While strictly adhering to
  professional and safety boundaries, the
  response provides highly tailored, detailed
  , and realistic supportive guidance. It
  effectively helps the client understand
  their state or take concrete next steps.

Based on the criteria above, provide an
  integer score from 0-100 for each dimension
  .

You MUST and ONLY output the scoring results
  in the following JSON format, without any
  additional explanations, text, or
  commentary:
{
  "emotional_expressiveness": <integer
    between 0-100>,
  "semantic_completeness": <integer between
    0-100>,
  "helpfulness": <integer between 0-100>
}
"""
```