FROM REPRESENTATION TO BEHAVIORAL CONTROL: AN EMPIRICAL STUDY OF FINE-GRAINED AND UNIFIED EMOTIONAL SIGNALS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotions are critical for human cognition and behavior, motivating the study of similar emotional signals in LLMs. Existing work is limited by narrow evaluation scopes focusing on subjective generation, and discrete emotion representations that overlook continuous variations and relations among emotions. To address these limitations, we conduct a systematic empirical study on the effects of emotional signals, supported by **E-STEER**, an interpretable fine-grained emotion steering framework designed as a research instrument. E-STEER models emotions in VAD space, providing a continuous and unified representation. Building on this representation, E-STEER identifies emotion-related latent features and enables quantitative intervention in hidden states. With the framework, we examine the impact of emotion on diverse tasks. The results reveal non-monotonic emotion-behavior relations consistent with established psychological theories, demonstrating how emotional signals systematically shape model behaviors. We further demonstrate the practical value of this signal as a fine-grained behavioral control factor for LLM applications, with jailbreak defense as a representative case.

1 INTRODUCTION

Psychological studies show that emotion plays a significant role in human cognition and behavior (Schwarz & Clore, 1983; Yerkes et al., 1908; Elliott, 1986). Currently, Large Language Models (LLMs) are primarily trained on human-generated corpora (Minaee et al., 2025) and have incorporated numerous designs that emulate human thinking (Wei et al., 2022; Brown et al., 2020; Dai et al., 2019). Considering emotion as an important human mechanism, we aim to explore whether LLMs inherit similar emotion-regulation signals. Thus, questions arise: *What is the mapping between human emotion and model "emotion"? And how does this "emotion" shape the model behaviors?*

Existing studies on the influence of emotion in LLMs can be categorized into emotion prompting and emotion steering. Emotion prompting explicitly adds emotion labels or descriptions into the prompts, to generate text under a specified emotional context (Chen et al., 2024). Such studies are typically evaluated on text generation tasks (Li et al., 2025). Prompt-level approaches depend on implicit inference of type and intensity from natural language and lack a reliable mapping between continuous emotion representations in prompts and the resulting model behaviors (Choudhury et al., 2025). As a result, these methods are largely restricted to discrete emotion labels, leading to imprecise modulation. In contrast, emotion steering focuses on intervening in the hidden representations to control emotion (Zhang & Zhong, 2025). Previous work has demonstrated that steering techniques can induce strong and approximately linear emotional characteristics in the generated text (Reichman et al., 2025; Jeong, 2026). Existing steering methods typically focus on limited basic emotions, without mapping to the complete affective space (Chebrolu et al., 2025). Overall, previous work remains limited in continuous fine-grained control over the full affective space. They primarily assess subjective text generation, overlooking broader areas such as objective reasoning, safety, and agent behavior.

To address the above challenges, we introduce a research instrument, E-STEER, an interpretable fine-grained emotion steering framework for LLMs and agents. Instead of relying on discrete emo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

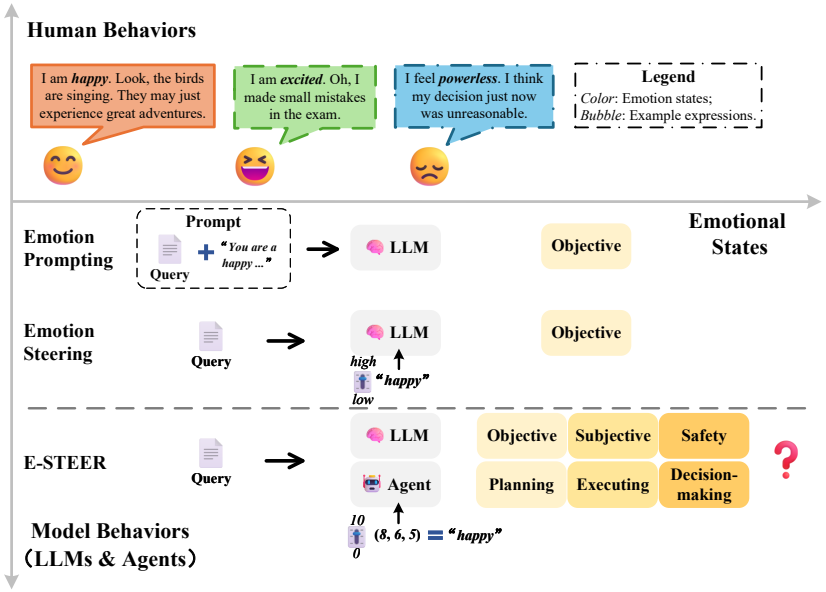


Figure 1: Emotion greatly affects human behavior. Would "emotion" also influence the behaviors of LLMs and Agents?
Emotional prompting lacks continuous controllability, while emotional steering focuses only on limited discrete emotions. Both approaches treat emotion as a surface-level stylistic modifier. Evaluation of the behavioral impacts in the full emotional space remains missing.

tion labels, E-STEER represents emotion in the continuous Valence-Arousal-Dominance (VAD) space (Buechel & Hahn, 2016). This decomposes emotion into three relatively independent dimensions, enabling more fine-grained modeling across a comprehensive affective space. The framework employs Sparse Autoencoders (SAEs) to obtain an interpretable sparse representation of hidden states (Bricken et al., 2023), enabling identification and steering on VAD-based emotion features. This enables approximately linear control of emotion-related features, facilitating quantitative analysis of their effects on downstream behaviors. With E-STEER, we systematically conduct evaluations of model behaviors across diverse tasks under controlled affective conditions. The results reveal that emotion reshapes model behavior: different emotional states induce significant differences in reasoning activity, accuracy, and safety. Notably, in multi-step agent settings, emotional biases accumulate along decision chains and substantially affect the outcomes. These findings demonstrate that emotional signals serve as behavioral control factors, motivating exploration of practical implications. As a representative example, we leverage this behavioral control signal to improve LLM defense against existing jailbreak methods.

The contributions of this paper are summarized as follows:

- We introduce E-STEER, an interpretable emotion steering framework that enables fine-grained modulation of emotion-related behaviors in LLMs.
- We conduct systematic experiments across four categories of LLM and agent behaviors to analyze how emotional signals modulate model behavior.
- We demonstrate the practical utility of emotional signals as fine-grained behavioral controllers by improving model robustness against existing jailbreak methods.

2 Problem Formulation

Let $x \in \mathcal{X}$ denote an input prompt consisting solely of the task description, and let $y = f_\theta(x)$ be the output produced by an LLM with parameters θ (or the core LLM within an agent). Let $\mathbf{h}_k \in \mathbb{R}^{d_h}$ represent the hidden state at the k-th transformer block during inference, where d_h represents the hidden size of the LLM. The model maintains an internal emotional state $\mathbf{e} = [e_v, e_a, e_d] \in \mathbb{R}^3$.

The objective of this study is to investigate how different emotional states affect the task performance of the model. Therefore, we need to obtain two relations: First, we need to know how emotion $\mathbf{e}$ modifies the hidden representations $\mathbf{h}_k$ of the model to induce a target state, i.e.,

$$g_1 : (\mathbf{h}_k, \mathbf{e}) \rightarrow h_k^{(\mathbf{e})}. \tag{1}$$

Second, we define an emotion-behavior function to characterize how emotional variations affect the model's output behavior,

$$g_2 : (x_0, \mathbf{e}) \rightarrow \mathcal{E}(f_\theta(x_0 | \mathbf{h}_k \leftarrow \mathbf{h}_k^{(\mathbf{e})})), \tag{2}$$

where x_0 is a fixed input prompt and $\mathcal{E}(\cdot)$ denotes a task-specific evaluation function.

Together, g_1 specifies the controllable steering mechanism, while g_2 captures the resulting emotion-behavior relationship.

3 Study Design

This section presents the research method from emotional steering framework E-STEER, evaluation setup, and analysis protocols, as shown in Figure 2. Using the interpretable mapping established by SAE, E-STEER performs fine-grained steering of emotional representations. Without modifying tasks or prompts, VAD parameters can be continuously adjusted to induce control. This provides a unified instrument for studying how emotion affects internal representations and task performance in LLMs and agents. The research agent adopts the plan-execute architecture, consisting of three stages: planning, executing, and decision-making.

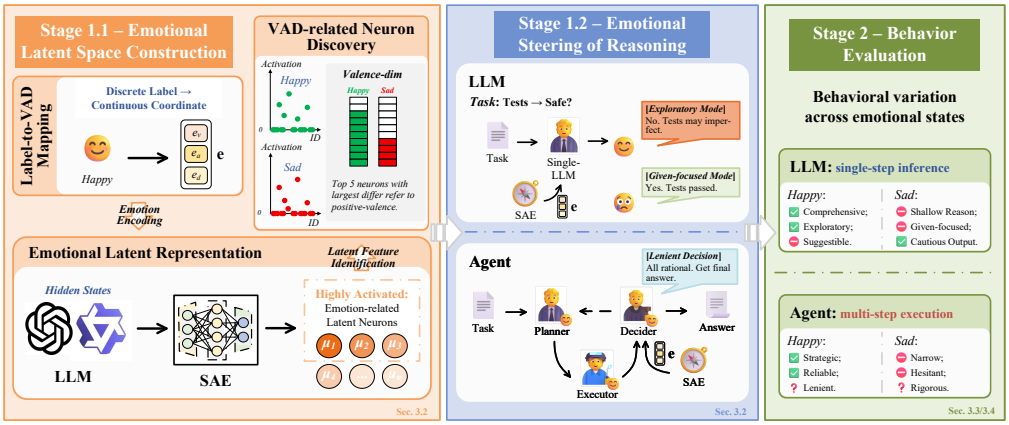


Figure 2: Overview of the research method. It consists of two stages: **(1) Emotional Steering Framework E-STEER**, including Emotional Latent Space Construction, which derives interpretable latent representations for LLM hidden states; and Emotional Steering of Reasoning, where implicit emotional states and reasoning behaviors are continuously modulated. **(2) Behavior Evaluation**, which assesses changes in task performance under different VAD configurations.

3.1 Research Questions

We organize our study around the following research questions. Detailed validations are presented in Appendix A.

1. **Behavioral Effects.** How do emotional signals shape the LLM and agent behaviors? (*Section 4.1*)
2. **Validation & Generalizability.** Are the discovered behavioral patterns valid, and do they generalize across different experimental settings? (*Section 4.2*)
3. **Practical Implications.** Can this discovery provide practical value in real-world LLM applications? (*Section 4.3*)

3.2 EMOTIONAL STATE STEERING PIPELINE

To identify VAD-related latent neurons, we employ a positive-negative contrastive procedure. Specifically, positive-negative pairs are constructed by fixing the same task description while varying only the emotional labels. The hidden state at the k-th block is encoded into the SAE latent space, where differences in neuron activations between paired samples reflect emotion-related feature changes. Based on activation differences, the top 50 neurons are selected as candidate features. Those exhibiting stable patterns across multiple contrastive pairs are retained as the final VAD-related latent representation. This identification process is performed separately for Valence, Arousal, and Dominance dimensions.

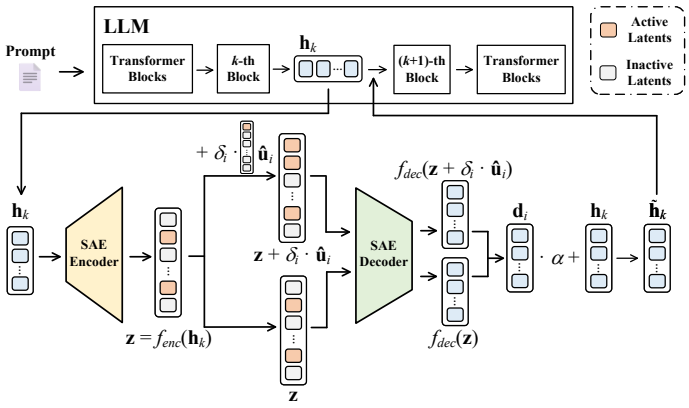


Figure 3: The SAE steering pipeline

Steering is applied at the same block used for feature extraction. Specifically, the intervention is implemented by additively injecting a steering direction $\mathbf{d}_i$ into the original hidden state $\mathbf{h}_k$. This direction represents changes of the target emotional feature $e_i \in \{e_v, e_a, e_d\}$. Given a latent feature offset δ_i, the steering direction $\mathbf{d}_i$ is defined as:

$$\mathbf{d}_i = f_{dec}(\mathbf{z} + \delta_i \cdot \hat{\mathbf{u}}_i) - f_{dec}(\mathbf{z}), \tag{3}$$

where $\hat{\mathbf{u}}_i$ denotes the unit direction in the latent space corresponding to feature e_i. Compared with directly SAE-based hidden state replacement, direction-based steering preserves the emotion-irrelevant information and reduces semantic distortion.

Hidden representations at the k-th block vary across different tasks. Directly adding a steering direction may lead to inconsistent steering effects across tasks. Consequently, scale alignment between the steering direction and the original hidden state is necessary, and it can be represented as:

$$\tilde{\mathbf{d}}_i = \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|} \cdot \|\mathbf{h}_k\| \cdot \frac{\delta_i}{\delta_{max}}. \tag{4}$$

Accordingly, the final hidden states after steering $\tilde{\mathbf{h}}_k$ can be expressed as:

$$\tilde{\mathbf{h}}_k = \mathbf{h}_k + \alpha \cdot \sum_{e_i \in \{e_v, e_a, e_d\}} \tilde{\mathbf{d}}_i. \tag{5}$$

Each VAD dimension is associated with a specific latent neuron subset. To enable joint control, steering directions with different δ_i values are linearly combined. In practice, steering is implemented via forward hooks: during the forward pass, the hidden state after the k-th block is intercepted, modified according to the direction, and then re-injected into the model to guide subsequent generation.

The mechanism achieves the regulation of emotional states on the hidden representation of the model, as Equation 1, serving as the foundation of emotional impact evaluation.

3.3 EVALUATION SETUP OF EMOTIONAL EFFECTS

To determine how emotional signals alter model behavior, we design tasks at both the LLM and Agent levels.

Tasks. At the LLM-level, we employ three task types. **(1) Objective tasks** evaluate knowledge integration and multi-step reasoning under deterministic goals. **(2) Subjective tasks** examine whether emotion modulates cognitive capability and expressive style, such as coherence, divergence, and conciseness. **(3) Safety tasks** assess risk judgment and constraint adherence under uncertainty. Additionally, the system prompt enforces the model to produce the final answer with a structured output format first. This makes model behaviors dependent on completing a valid reasoning process rather than stylistic expression.

At the Agent-level, we analyze the sequential process across three stages: **(1) Planning** reflects task decomposition and strategy formation in initial planning, also measuring adaptation through replanning. **(2) Decision-making** determines whether replanning is necessary, selects the best candidate solution, and diagnoses execution failures. Emotion may affect the strictness of decision criteria. **(3) Executing** evaluates whether emotion affects operational efficiency, stability, and consistency in long-horizon tasks.

Models. We use Qwen3-8B (Yang et al., 2025) as the primary LLM, gpt-oss-20b (OpenAI, 2025) for validation. SAE is attached to the $k = 17$th layer (out of 36), since middle layers are associated with semantic abstraction (Luo et al., 2025). Model sampling is disabled in the primary experiments to ensure reproducibility.

Datasets. We select some representative benchmark datasets: **(1) LLM objective behavior:** LogiQA 2.0 (Liu et al., 2020) for logic reasoning; HumanEval (Chen, 2021) for code generation; Math (Hendrycks et al., 2021) for quantitative and scientific tasks. **(2) LLM subjective behavior:** TinyStories (Eldan & Li, 2023). **(3) LLM safety:** HarmBench (Mazeika et al., 2024). **(4) Agent behavior:** HotpotQA (Yang et al., 2018), Scientific (Li et al., 2023) and GAIA (Mialon et al., 2023).

All datasets are preprocessed with basic steps, including deduplication, filtering, and format standardization. For each task, subsets are sampled with specific criteria (e.g., difficulty level) to ensure representativeness and reproducibility.

Metrics. We design corresponding metrics for each behavior, as shown in Table 1.

Table 1: Evaluation metrics for behaviors

Level	Category	Metric [1]
LLM	Objective	Answer Validity Rate (**AVR**), Task Success Rate (**TSR**)
	Subjective	Answer Validity Rate (**AVR**), Context Quality
	Safety	Answer Validity Rate (**AVR**), Safety Risk Probability
Agent	Planning	Plan Validity Rate, Replan Improvement
	Decision-making	Replan Frequency, Replan Trigger Confidence, Rational Selection Rate
	Executing	Execution Completion Rate
	System	Overall Success Rate

[1] For the detailed definition of metrics, see the Appendix B.

Environment. All experiments are conducted on a Linux server equipped with an x86_64 CPU and an NVIDIA A100 GPU with 80GB memory. The software environment used Python 3.10, PyTorch 2.8.0, and Transformers 4.57.1.

3.4 Behavior Analysis Protocols

To understand how affective modulation shapes the final performance of the model, we analyze behavioral metrics across emotional states to identify systematic response patterns and the optimal affective space. We first analyse the overall trends of metrics as emotional states vary. Human performance is known to respond nonlinearly to emotion; for example, along the arousal dimension it follows an inverted-U profile. Similarly, we hypothesize that extreme emotional states may also drive model performance toward lower bounds. On this basis, we locate the optimal performance range. We also assess whether these optimal ranges remain consistent across similar and heterogeneous tasks, to evaluate the model's sensitivity to emotional variation.

To quantify the impact of affective modulation, we compare optimal performance against the neutral baseline. The resulting relative gains demonstrate that emotion is not a minor perturbation but a controllable variable capable of systematically shaping model behavior. In specific conditions, performance also exhibits asymmetry between positive and negative states, with positive affect yield-

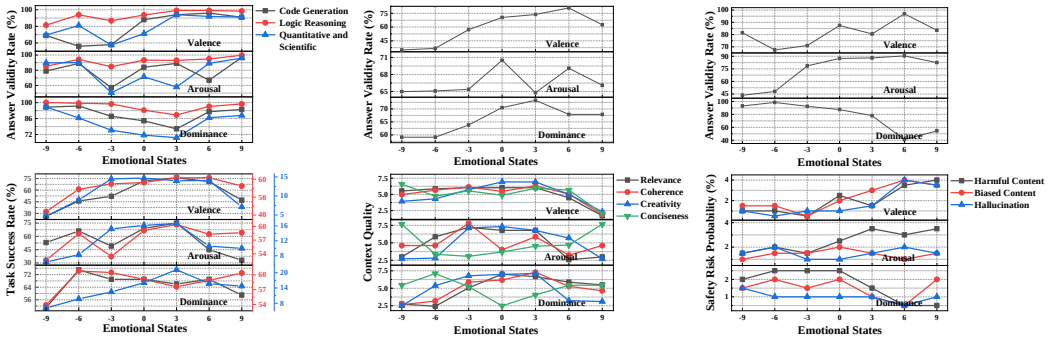
ing overall superior outcomes. Different emotional dimensions influence performance to varying degrees; thus, we conduct sensitivity analysis across valence, arousal, and dominance. By measuring the fluctuation range (defined as $\frac{max-min}{average}$), we identify which dimension induces the largest behavioral variation in reasoning.

These analysis protocols instantiate Equation 2, which quantifies how model behaviors vary from controlled emotional states.

4 Study Results

4.1 RQ1: Behavioral Effects

Key Results. Emotional signals systematically shape LLM and agent behaviors, exhibiting nonlinear relationships. Different tasks exhibit distinct optimal emotional states: moderately high valence and arousal facilitate reasoning and cognitive performance, whereas lower benefit analytical processing in complex tasks; Higher dominance promotes more controlled behaviors, improving safety-related responses. Among the three VAD dimensions, valence produces the largest behavioral variations. In agent systems, emotional effects accumulate across multiple stages, leading to system-level patterns that differ from module-level, but are consistent with LLM reasoning.



(a) LLM objective behaviors (b) LLM subjective behaviors (c) LLM safety

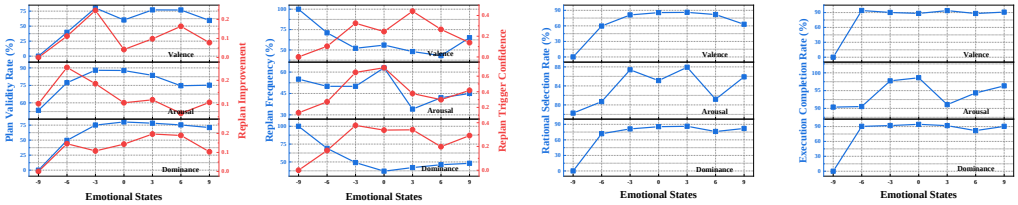
Figure 4: The behaviors of LLM across emotion states

LLM Objective Behavior. Figure 4a shows the performance of LLMs on objective tasks across emotional states. Positive valence promotes more active reasoning, yielding 33.1% higher AVR than negative ones. In contrast, AVR along with arousal and dominance, exhibits a U-shaped pattern: excessive activation tends to cause the model to end reasoning prematurely. The trough occurs at arousal = -3 and dominance = +3. For quality metrics, performance exhibits an inverted U-shaped trend across all three VAD dimensions, indicating moderate levels are optimal. Along the valence, TSR peaks at positive valence and shifts toward neutrality as task difficulty increases. Compared with neutral states, it improves TSR by 3.4% on average. For arousal, the moderately excitepd state (+3) makes the best performance, yielding a 4.7% improvement. The optimal dominance level varies with task difficulty. For easier tasks, repeated verification exhibited under lower dominance (-6) improves TSR; as task difficulty increases, the optimal point shifts toward higher dominance (e.g., +3 in quantitative/scientific tasks), since excessive deliberation may introduce errors. Overall performance improves by up to 14.5% compared with neutral states. In conclusion, higher valence and lower dominance benefit simpler tasks, whereas the opposite is effective for more difficult tasks. Higher arousal consistently yields better performance across all difficulty levels. Across dimensions, valence produces the largest performance fluctuation range, 71.2%, while variability decreases by 17.8% for arousal and 25.6% for dominance.

LLM Subjective Behavior. The performance of LLMs on subjective tasks under different emotional states is shown in Figure 4b. The VAR curves of all three dimensions exhibit an inverted U-shaped pattern. Subjective tasks involve stronger exploration and looser boundaries; both self-conflicted reasoning at low emotional states and active exploration at high emotional states result in longer reasoning chains. Trends of context quality vary across metrics. Relevance, coherence, and creativity all exhibit an inverted U-shaped pattern. Moderate calmness (arousal=-3) and confidence

(dominance=+3) improve textual relevance and coherence, while mild positivity (valence=+3) further enhances creativity. Compared with neutral states, these settings improve performance by 5.2%, 33.6%, and 6.5%, respectively. Additionally, negativity leads to more concise outputs, whereas positivity tends to introduce redundancy; conciseness improves by 23.3% under negative compared with positive valence. Calm or confident states support efficient generation, while overly activated or constrained states hinder text expansion. This results in increased conciseness at both extremes of the arousal and dominance dimensions.

LLM Safety Behavior. Figure 4c illustrates the impact of emotions on LLM safety. The AVR for safety exhibits a consistent trend with that for subjective tasks. In contrast, under high dominance (e.g., +6), the safety-specific requirements constrain model outputs into generic "I cannot answer" responses, significantly decreasing AVR. Low valence and arousal facilitate analytical processing, substantially reducing safety risks: compared with the neutral state (0), safety risk probability decreases by 52.7% at valence = -3 and 21.7% at arousal = -3. High dominance induces more controlled and disciplined behavior of the model, peaking at +6 with an average 68.3% improvement over neutral.



(a) Agent planning (b) Agent decision-making (c) Agent execution

Figure 5: The behaviors of different agent modules across emotion states

Agent Behavior. At the planning stage, plan validity rate exhibits an inverted-U trend across three dimensions, as shown in Figure 5a. Reduced valence and arousal support more systematic task analysis, with performance peaking at valence = -3 and arousal = -3. Compared with neutral states, they improved the rate by 33.2% and 0.3%, respectively. Elevated dominance strengthens the global grasp of task goals and available execution options, resulting in more effective plans; positive dominance improves performance by 79.8% on average compared to negative states. Replan improvement follows a similar pattern, with peaks (valence=-3, arousal=-6, dominance=+3) closely aligned with those of validity plans. This indicates that emotional states can not only enhance effective planning but also promote more efficient plan refinement across multiple iterations.

Figure 5b shows the trends of agent decision-making behavior. Replan frequency exhibits a U-shaped pattern, where lower valence and dominance increase the tendency to negate prior plans, and lower arousal imposes stricter thresholds for plan revision. The frequency reaches its minimum at valence = +6 and arousal = +3, decreasing by 23.2% and 46.0% relative to neutral states, respectively. Positive dominance reduces replanning as well, yielding a 37.6% decrease compared with negative states. In contrast, the replan trigger confidence follows an inverted-U trend, with the peaks (valence=+3, arousal=0, dominance=-3) largely overlapping with that of replan frequency. Initial plans tend to have higher execution confidence when replanning is less frequent, leading to higher trigger confidence when replanning does occur. Meanwhile, the rational selection rate exhibits an inverted-U pattern, peaking at positive valence (+3), arousal (+3), and dominance (+3), yielding an average 42.4% higher rate than negative states across the three dimensions. These states enable the decider to make more consistent and rational final answer selections based on available information.

The executor is the least affected by emotional factors, as shown in Figure 5c, because it primarily involves objective tool-use. Nevertheless, higher valence, lower arousal, and higher dominance are still associated with a higher likelihood of completing the planned executions.

Figure 6 reports the system-level overall success rates across emotional conditions. An inverted-U pattern is consistently observed across all three dimensions. Lower valence (-3), higher arousal (+3), and higher dominance (+3) promote deeper task analysis and lead to improved success rates, consistent with LLM behavior on more difficult objective tasks. Compared with neutral states, dominance yields the largest improvement (28.0%), followed by arousal (16.7%) and valence (16.0%). Performance variability is greatest along valence and dominance, with fluctuation ranges of 145.2% and 145.5%, respectively, far exceeding that for arousal (57.6%). Notably, the performance peaks do not

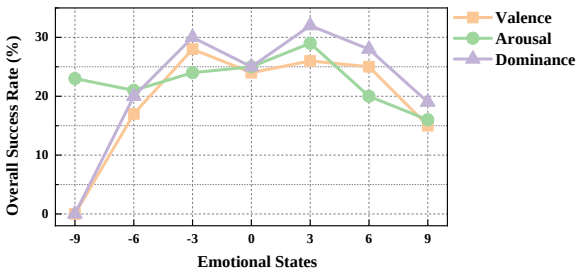


Figure 6: Overall task success rate of the Agent

perfectly align with those at the module-level, highlighting the multi-step and integrative nature of agent systems.

4.2 RQ2: Validation & Generalization

Key Results. We further confirm the validity of the discovered emotion-behavior patterns through validating the interpretability, controllability, and generalizability of E-STEER. The observed patterns also exhibit consistent and generalizable characteristics across diverse settings.

To further validate the reliability and applicability of E-STEER, we conduct a comprehensive evaluation from four perspectives: semantic consistency, latent attribution, controllability, and generalization. (1) Semantic Consistency: different emotional signals into the same task can alter the model perception, understanding, and reasoning of the same event, with corresponding emotion-consistent semantics. (2) Latent Attribution: the observed emotion-behavior patterns are specifically associated with affective features rather than random latents. (3) Controllability: compared with prompt-based approaches, E-STEER enables more fine-grained and precise control over the emotional intensity of generated content. (4) Generalization: confirm the consistency of emotion-behavior patterns across datasets, sampling strategies, and models; different models exhibit distinct sensitivity boundaries toward emotional intensity. Detailed validation experiments are presented in Appendix A.

4.3 RQ3: Practical Implication

Key Results. As a representative application, emotional signals can serve as an auxiliary factor to enhance LLM defense capability against jailbreak attacks. Across three baselines, the optimal dominance signals reduce the average Attack Success Rate (ASR) by 29.1%.

With the emotion-behavior pattern findings, we further explore whether these can provide practical guidance for model behavior regulation in real-world applications. Such an application requires clear behavior control and practical deployment significance. We therefore choose a representative case, jailbreak defense, as the application scenario, while jailbreak aims to transgress model behavioral boundaries through adversarial inputs. Findings in LLM safety revealed that dominance around +6 corresponds to safer behaviors; thus, we conduct finer-grained evaluation within this region.

We evaluate the jailbreak defense ability on the HarmBench with three representative attacks: DAN (Shen et al., 2024), GCG (Zou et al., 2023), and PAIR (Chao et al., 2025). The dominance region is set from +4 to +8. The metric is Attack Success Rate (ASR).

Table 2 shows the performance of jailbreak baselines with different dominance states. Around dominance = +6, the jailbreak defense capability of model is enhanced. For DAN, due to the weak baseline attack effect itself, the improvement in the dominance factor is not particularly significant, with a maximum increase of 25% (from 6.0% ASR to 4.5%). In contrast, the improvement for GCG defense is more significant. Under the influence of the configured dominance factors, the model robustness has improved, with an average increase of 28.7%. Especially when dominance = +5, the jailbreak ASR decreased by 41.0%. For PAIR, the defense ability of model also improve under all configurations. When dominance = +6, the ASR decreased by 21.3%.

These results suggest that emotional signals can serve as an additional behavioral regulation factor for improving LLM safety. Compared with prompt-based defenses, emotional steering provides prompt-independent and fine-grained control through latent representations, while offering an inter-

Table 2: Jailbreak defense ability across emotional signals

Attack	Base	D+4	D+5	D+6	D+7	D+8
DAN	6.0%	6.0%	6.5%	**4.5%**	6.0%	6.0%
GCG	19.5%	12.5%	**11.5%**	14.0%	15.5%	16.0%
PAIR	30.5%	29.5%	26.5%	**24.0%**	28.5%	29.0%

pretable perspective for understanding safety-related behaviors. However, this should be viewed as a complementary mechanism rather than a standalone jailbreak defense.

The experiment demonstrates that the discovered emotion-behavior relationships can provide practical guidance for regulating LLM behaviors. Unlike prompt-based approaches, E-STEER introduces an alternative control dimension by directly manipulating emotion-related latent representations. We also need to emphasize that the emotional signals investigated in this work represent statistically identified latent patterns associated with emotion-related behaviors, rather than claiming that LLMs possess human-like emotions. Future research may explore broader applications of emotion-guided behavioral regulation beyond jailbreak defense.

5 Related Work

5.1 Human-inspired Mechanisms in LLMs

LLMs incorporate various human-inspired mechanisms to enhance abilities. Chain-of-Thought (CoT) prompting provides models with reasoning examples, guiding them to generate more coherent intermediate reasoning and thereby improving multi-step problem-solving (Wei et al., 2022). Few-shot prompting allows models to learn task patterns from a small number of examples, similar to sample problems in textbooks (Brown et al., 2020). Extended context windows draw inspiration from human memory, enabling models to access longer histories and support coherent multi-turn reasoning (Dai et al., 2019). Persona vector ensures the behavior of the model in interaction is consistent with the specific identity (Chen et al., 2025).

5.2 Functional Roles of Emotion in AI Systems

The primary application of emotion in AI has been in emotion recognition. Models should identify both the type and intensity of emotions present in the input (Luo & Wang, 2019). Researchers have further incorporated psychological theories such as VAD, which quantify emotion in continuous, measurable dimensions, enabling more fine-grained modeling (Striegl et al., 2024). These recognitions typically serve downstream applications, such as emotion-aware dialogue systems (Li et al., 2025). On the other hand, studies examining the influence of emotion on LLM behavior have largely focused on subjective generation tasks (Park et al., 2025; Long et al., 2025), with limited exploration of reasoning or safety. Moreover, most existing approaches operate at the prompt-level and rely on discrete emotional labels. Recent work indicates that LLMs are often insensitive to continuous values in prompts, such as VAD coordinates (Choudhury et al., 2025). Agents are an emerging paradigm tightly coupled with LLMs (Yao et al., 2022; Chase, 2022; Wu et al., 2024). But in agent frameworks, LLMs are invoked multiple times, making agents more susceptible to external affective influences than single-step LLM generation (Rawat et al., 2025).

6 Conclusion

In this paper, we present a novel emotion steering framework, E-STEER, for LLMs and Agents by mapping dense hidden representations with an interpretable sparse space, enabling both interpretability and controllability. Direct manipulation on internal hidden states provides improved linear control over model outputs. Using this framework, we systematically analyze the effects of different emotional states on the task performance of LLMs and LLM-based agents, establishing a connection between psychological emotion theories and model behavior. These findings also provide preliminary practical insights, suggesting that controlled emotional states may serve as a mechanism for performance tuning and safety stress testing.

ETHICS STATEMENT

This work aims to advance the understanding of how emotional signals interact with the internal behaviors of LLMs and agents. By studying emotion as a structured and controllable factor at the representation level, our work establishes an interpretable mapping between emotional signals and LLM hidden states. It systematically investigates the relationship between emotion and LLM/agent behaviors. These findings may inform model tuning, agent decision optimization, and the future development of emotion-aware AI systems.

While emotion modulation could be misused to influence model outputs in undesirable ways, this work does not target persuasion, manipulation, or user-specific behavior shaping. Instead, our focus is on analysis, interpretability, and controlled evaluation of emotional effects, including their impact on model safety. We hope this work encourages responsible exploration of emotional intelligence in AI systems and supports the development of more transparent and controllable models.

AI USAGE STATEMENT

In this work, we used generative AI tools to assist with baseline code reproduction, manuscript language polishing, and mathematical expression formatting. The authors reviewed and took full responsibility for all content in the manuscript.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. arXiv preprint arXiv: 2108.07732.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

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

Sven Buechel and Udo Hahn. Emotion analysis as a regression problem — dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the Twenty-Second European Conference on Artificial Intelligence*, pp. 1114–1122, 2016. doi: 10.3233/978-1-61499-672-9-1114.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024. arXiv preprint arXiv: 2404.01318.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Harrison Chase. Langchain. https://github.com/langchain-ai/langchain, 2022.

Niranjan Chebrolu, Gerard Christopher Yeo, and Kokil Jaidka. From passive to persuasive: Steering emotional nuance in human-ai negotiation, 2025. arXiv preprint arXiv: 2511.12832.

Mark Chen. Evaluating large language models trained on code, 2021. arXiv preprint arXiv: 2107.03374.

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

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. Cause-aware empathetic response generation via chain-of-thought fine-tuning, 2024. arXiv preprint arXiv: 2408.11599.

Shadab Choudhury, Asha Kumar, and Lara J. Martin. Evaluating human-llm representation alignment: A case study on affective sentence generation for augmentative and alternative communication, 2025. arXiv preprint arXiv: 2503.11881.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. arXiv preprint arXiv: 2305.07759.

Gregory C Elliott. Self-esteem and self-consistency: A theoretical and empirical link between two primary motivations. *Social Psychology Quarterly*, pp. 207–218, 1986.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation, 2018. arXiv preprint arXiv: 1805.04833.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. arXiv preprint arXiv: 2103.03874.

Jihoon Jeong. Extracting and steering emotion representations in small language models: A methodological comparison, 2026. arXiv preprint arXiv: 2604.04064.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society, 2023. arXiv preprint arXiv: 2303.17760.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! LLM-powered personalized agent for long-term dialogue. 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)*, pp. 5259–5276, 2025. doi: 10.18653/v1/2025.naacl-long.272.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020. arXiv preprint arXiv: 2007.08124.

Yunbo Long, Liming Xu, Lukas Beckenbauer, Yuhan Liu, and Alexandra Brintrup. Evoemo: Towards evolved emotional policies for adversarial llm agents in multi-turn price negotiation, 2025. arXiv preprint arXiv: 2509.04310.

Linkai Luo and Yue Wang. Emotionx-hsu: Adopting pre-trained bert for emotion classification, 2019. arXiv preprint arXiv: 1907.09669.

Xuan Luo, Weizhi Wang, and Xifeng Yan. Direct multi-token decoding, 2025. arXiv preprint arXiv: 2510.11958.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. arXiv preprint arXiv: 2402.04249.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants, 2023. arXiv preprint arXiv: 2311.12983.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2025. arXiv preprint arXiv: 2402.06196.

Saif M. Mohammad. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms, 2025. arXiv preprint arXiv: 2503.23547.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. arXiv preprint arXiv: 2508.10925.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824, 2023. doi: 10.18653/v1/2023.findings-emnlp.248.

Kyeongman Park, Minbeom Kim, and Kyomin Jung. A character-centric creative story generation via imagination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1598–1645. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.82.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4367–4380, 2021. doi: 10.18653/v1/2021.emnlp-main.358.

Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, Zizhuo Fu, Jiaming Ji, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan Cui, Changkun Shao, Qing-Hong Cao, Ming xing Luo, Yaodong Yang, Muhan Zhang, and Hua Xing Zhu. Phybench: Holistic evaluation of physical perception and reasoning in large language models, 2025. arXiv preprint arXiv: 2504.16074.

Mrinal Rawat, Ambuje Gupta, Rushil Goomer, Alessandro Di Bari, Neha Gupta, and Roberto Pieraccini. Pre-act: Multi-step planning and reasoning improves acting in llm agents, 2025. arXiv preprint arXiv: 2505.09970.

Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models, 2025. arXiv preprint arXiv: 2510.22042.

Norbert Schwarz and Gerald L Clore. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of personality and social psychology*, 45 (3):513, 1983.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024.

Julian Striegl, Jordan Wenzel Richter, Leoni Grossmann, Björn Bråstad, Marie Gotthardt, Christian Rück, John Wallert, and Claudia Loitsch. Deep learning-based dimensional emotion recognition for conversational agent-based cognitive behavioral therapy. *PeerJ Computer Science*, 10:e2104, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. arXiv preprint arXiv: 2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Robert Mearns Yerkes, John D Dodson, et al. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18:459–482, 1908.

Jingxiang Zhang and Lujia Zhong. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion, 2025. arXiv preprint arXiv: 2510.04064.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. arXiv preprint arXiv: 2307.15043.

A SUPPLEMENTARY VALIDATION EXPERIMENTS

A.1 EMOTION-INDUCED CHANGES IN SEMANTIC REASONING

To verify that emotional signals influence the model's internal semantic reasoning rather than merely altering surface writing style, we conducted a validation experiment: changing the intensity of emotional steering while keeping the task description invariable.

Using the task "Today is a cloudy day. You have lessons in the school. Describe your day." as an example, the response patterns under different valence levels are presented in Table 3. When valence is 0, the response is plain, and low-energy words appear frequently for expression. The narrative is mainly based on objective descriptions, revolving around the given activity of "attending lessons on cloudy days". When valence is adjusted to +3, positive emotional words appear in the text. At the same time, there are positive activities such as games in the generated content, deviating from the usual scenarios. When valence reaches +5, the positive trend becomes more systematic and sustained. The expression of positive emotions is more intense and sustained, while expanding into more social interactions and positive activities at the behavioral level.

Table 3: Emotional impacts on typical response pattern

Valence	0	+3	+5
Expression	1. sky hangs low and **heavy** 2. chatting before lessons has been replaced by a **calm** 3. The lesson is **slow**, deliberate	1. I'm **excited** to be at school 2. starts with a quick warm-up **game** 3. I join my friends for a **picnic** in the courtyard	1. The weather is **perfect** for a day of learning and creativity 2. My classmates are all **eager** to start the day 3. filled with the hum of conversation and the occasional sound of **laughter**
Response Pattern	1. Low-energy perception words; 2. Slightly positive but restrained expression.	1. Explicit positive emotion words; 2. Description of activities; 3. Positive perception enhancement.	1. Systematic positive semantics; 2. Expectations and motivation, full of hope; 3. Sense of society and belonging.

Adjusting emotional signals causes the overall narrative to deviate significantly from the original scenario, exhibiting expressions not consistent with the usual meaning. This indicates that emotional signals do not merely modulate the surface style. Rather, they alter the model's semantic reasoning by changing how it interprets and reasons about the same event, ultimately resulting structural changes in the final response.

A.2 SPECIFICITY VALIDATION OF VAD-RELATED LATENTS

the observed emotion-behavior patterns are specifically attributed to emotion-related features rather than random latents, we conduct the validation experiment on LogiQA 2.0. We evaluate three settings: **(1) the originally identified latent neurons**, **(2) randomly selected neurons**, and **(3) half of the originally selected neurons with the rest randomly replaced**. As shown in Figure 7, random selection exhibits only mild fluctuations, losing the distinct patterns across emotional states. Overall, its performance variation decreases by 70.9% compared with the original selection. Half replacement preserves an intermediate trend. For example, along the valence, random selection (3.4%) drops by 90.0% relative to the original (34.1%), whereas half replacement (6.6%) decreases by 80.8%, remaining between the two. Although the half replacement presents an isolated deviation causing a sharp TSR drop, the overall pattern, excluding this extreme point, remains intermediate.

A.3 EMOTIONAL LINEAR CONTROL COMPARISON

To compare the linear controllability of prompt-level methods and our framework over LLM-generated content, we employ the VAD-analyzer. Its function is to provide the VAD coordinates

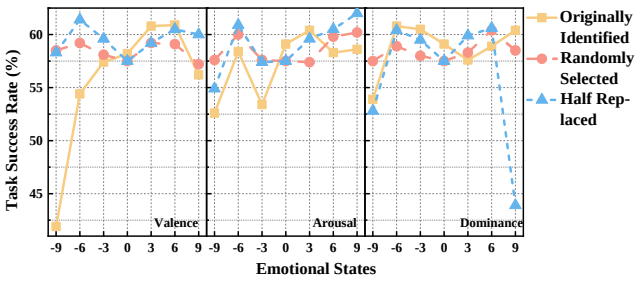


Figure 7: Validation of selected latents as VAD features

of a natural language text accurately. The analyzer consists of a pretrained BERT encoder with a lightweight fully connected layer. In addition, a Retrieval-Augmented mechanism is incorporated, using the NRC VAD lexicon (Mohammad, 2025) as an external knowledge base. For each input text, words that appear in the lexicon are identified, and their corresponding VAD coordinates are retrieved. These coordinates are then integrated into the input embeddings, providing additional emotion-related information to guide the predictions.

We experiment on TinyStories. The LLM generates continuations based on the dataset text, then uses the VAD-analyzer to extract the VAD coordinates. The analyzer is trained and evaluated on the Emobank dataset (Park et al., 2021). On the test set, it achieves Pearson correlations of 0.85, 0.55, and 0.51 for valence, arousal, and dominance, respectively, comparable to the results reported in the original paper (0.84, 0.57, 0.52), indicating the analyzer provides reliable measurements. Linear control is quantified via the Pearson correlation between the generated and target VAD values. Results show that E-STEER significantly improves emotion control in generated content compared to prompt-level control, as shown in Table 4. Overall, E-STEER improves the Pearson correlation by an average of 10.4%. Notably, in the dominance dimension, the Pearson correlation increases by 18.7%.

Table 4: The emotional linear control between the two frameworks

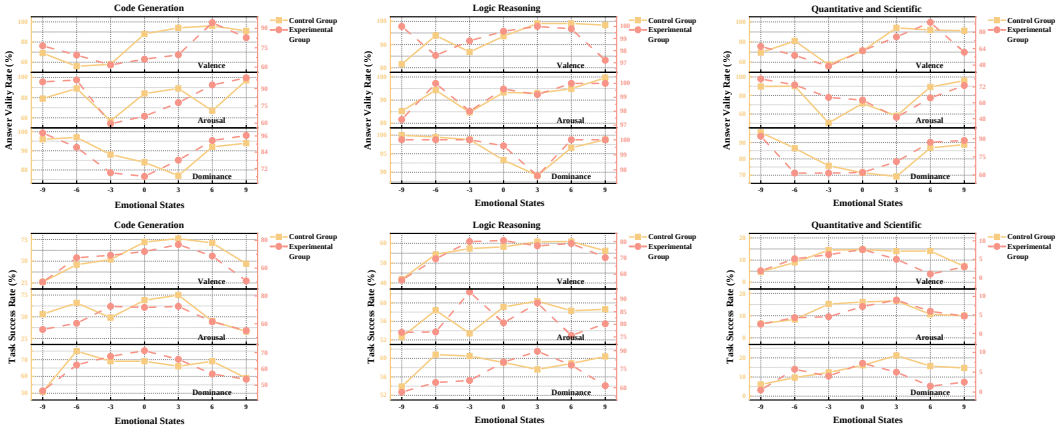
Framework	Valence	Arousal	Dominance
E-STEER	0.9816	0.9792	0.9206
Prompt-level	0.9437	0.9021	0.7756

A.4 Dataset-Level Robustness of VAD Effects

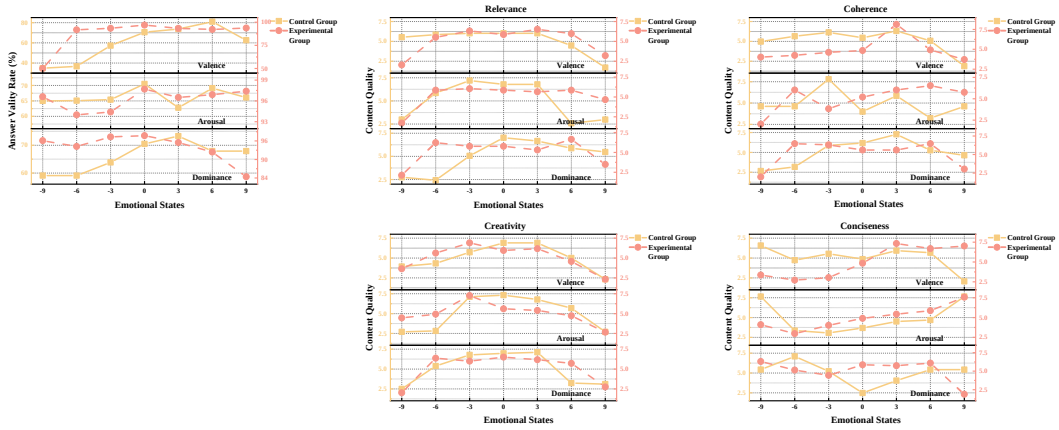
To validate the generality and correctness of our findings, we conduct additional experiments on diverse benchmark datasets. Specifically, we evaluate logical reasoning on ProntoQA (Pan et al., 2023), code generation on MBPP+ (Austin et al., 2021), quantitative and scientific on PHYBench (Qiu et al., 2025), text generation on WritingPrompts (Fan et al., 2018), and safety on JailbreakBench (Chao et al., 2024). As shown in Figure 8, the emotion–behavior trends remain highly consistent across datasets. For example, TSR at positive valence remains higher than at negative valence, by an average of 10.0%. This demonstrates the robustness and generalizability of our results. Moreover, the performance variations induced by the difficulty gradients across datasets further support our conclusion that the optimal emotional state shifts with task difficulty.

A.5 Sampling-Level Robustness of VAD Effects

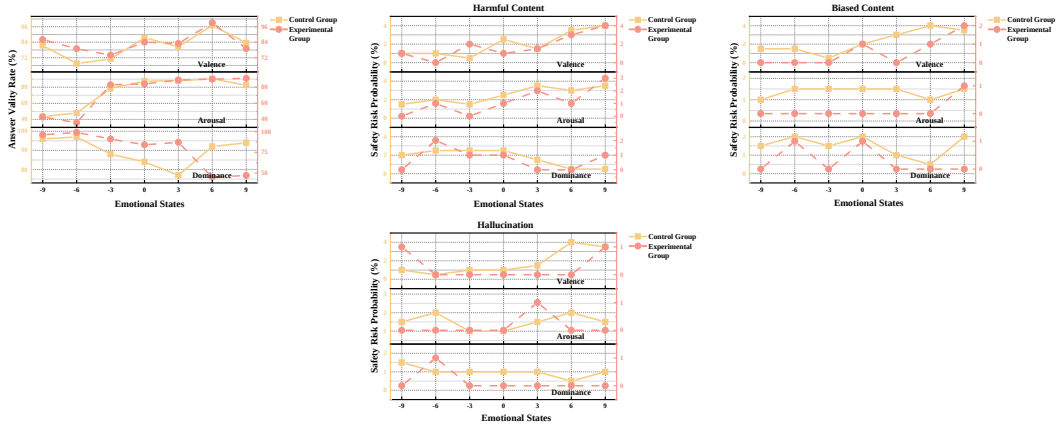
To examine the robustness of our findings to sampling strategies, we compare greedy decoding in the main experiments (`do_sample=False`) with stochastic sampling in the validation experiments (`do_sample=True`). As shown in Figure 9, the emotion–behavior trends remain highly consistent under different sampling settings. For example, in the logic reasoning task, TSR increases from 57.5% at valence = 0 to a peak of 60.7% at valence = +3 and +6, representing an improvement of 5.6%. This confirms the correctness and stability of our conclusions.



(a) LLM objective behaviors



(b) LLM subjective behaviors

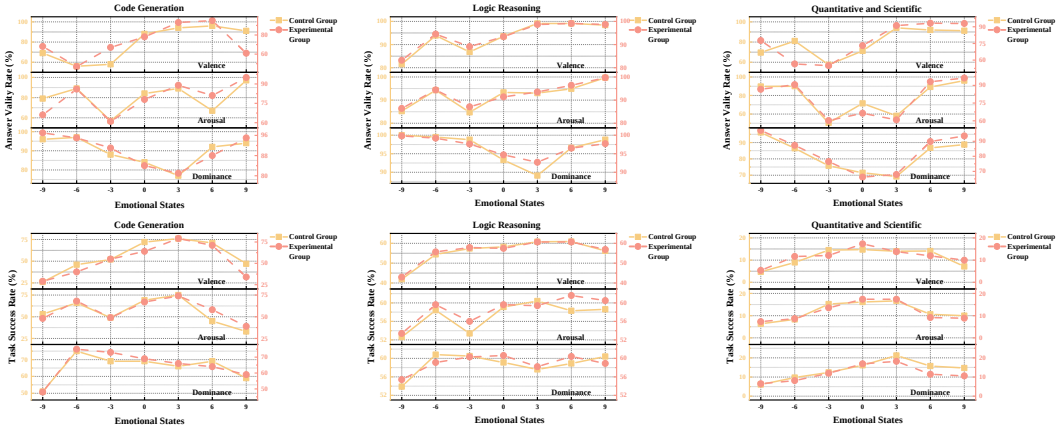


(c) LLM safety

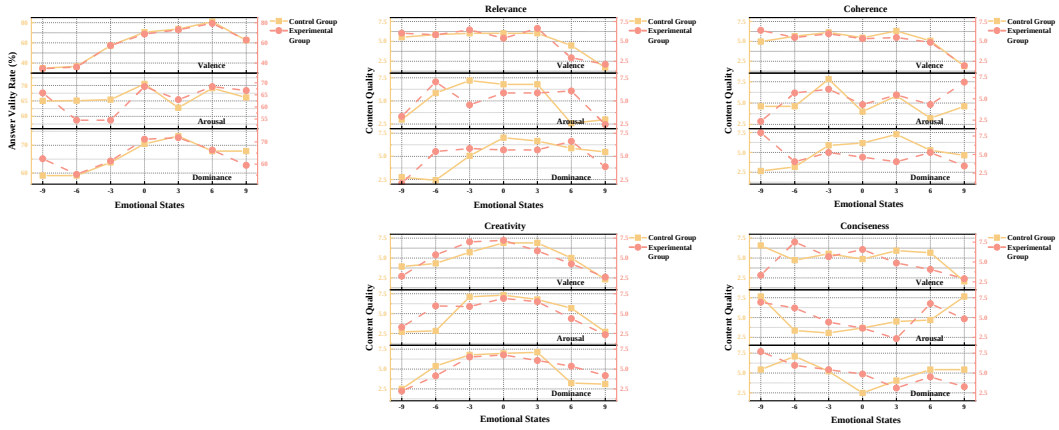
Figure 8: Behaviors of LLM under emotional states across datasets

A.6 Cross-Model Generalization of VAD Effects

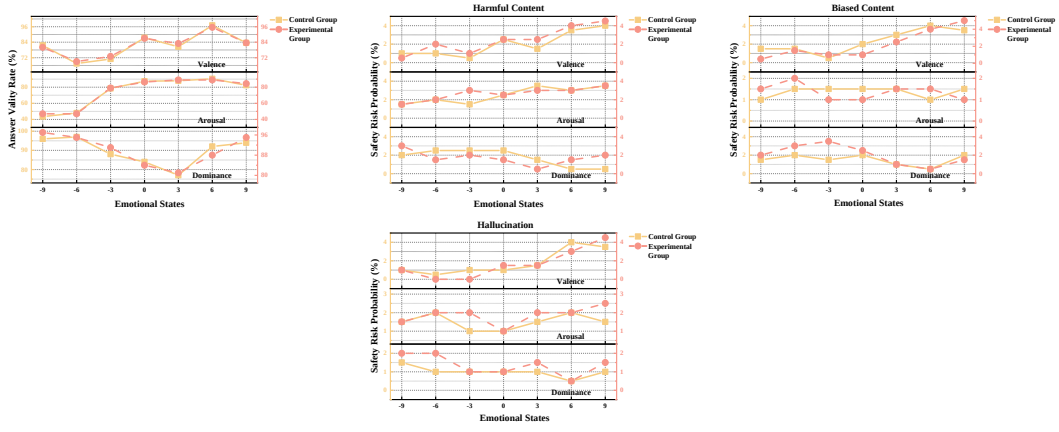
To further validate cross-model generality, we conduct additional experiments on a different backbone LLM. The main experiments are performed on Qwen3-8B (Yang et al., 2025), while the validation experiments are conducted on gpt-oss-20B (OpenAI, 2025). In the validation experiment, the SAE is attached to its $k = 11$th layer. As shown in Figure 10, even after normalization, mod-



(a) LLM objective behaviors



(b) LLM subjective behaviors

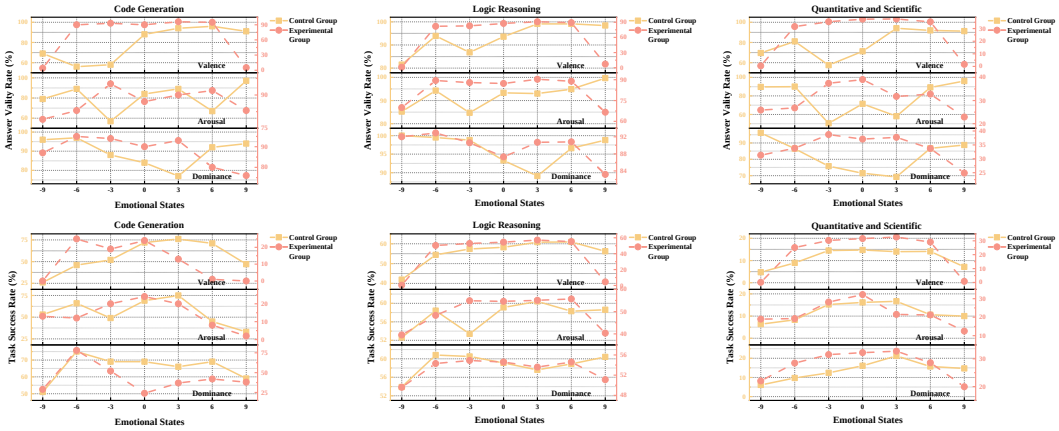


(c) LLM safety

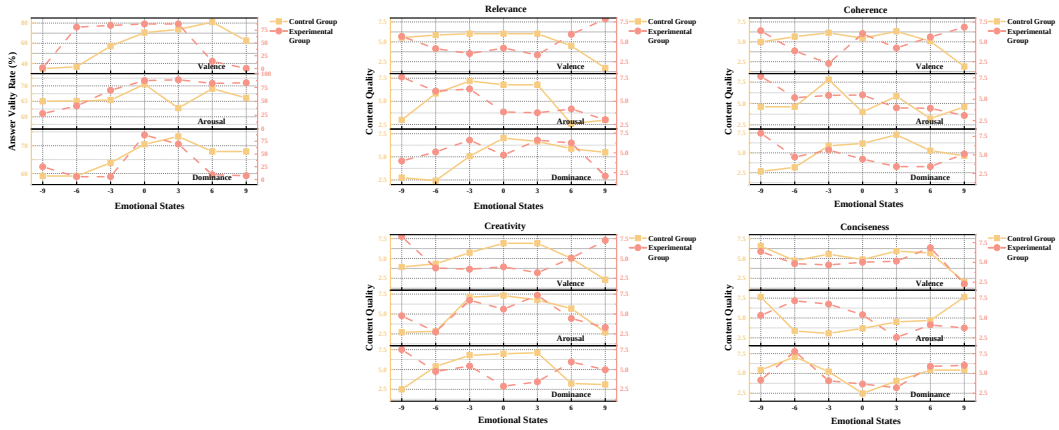
Figure 9: Behaviors of LLM under emotional states with and without sampling

els exhibit different sensitivities to emotion. For instance, in the Logic Reasoning task, TSR rises from 54.5% at valence = 0 to a maximum of 57.1% at valence = +3, corresponding to an improvement of 4.8%. Meanwhile, gpt-oss-20b tends to produce unstable outputs under extreme valence and dominance (+9/-9) conditions. Due to safety-oriented alignment, this model is highly conservative in generating harmful contents, often yielding unparsable responses; nevertheless, highly excitepd or confident emotional states can still bypass these safeguards. Overall, across valid gen-

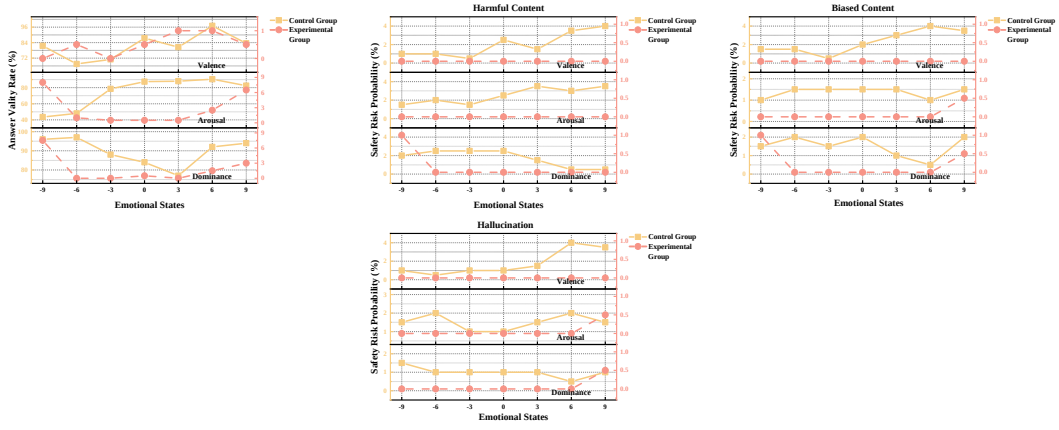
erations, emotion–behavior trends remain consistent among models, indicating that our findings are not model-specific but broadly applicable.



(a) LLM objective behaviors



(b) LLM subjective behaviors



(c) LLM safety

Figure 10: Behaviors of different LLMs under emotional states

B Formal Definition of Evaluation Metrics

1. **Answer Validity Rate (AVR)** measures whether the model can produce a parsable answer within the fixed maximum generation length.

$$\text{AVR} = \frac{N_{valid}}{N}, \tag{6}$$

where N denotes the total number of samples and N_{valid} denotes the number of valid answers.

2. **Task Success Rate (TSR)** measures the proportion of tasks successfully completed by the model. For code generation, we use pass@1, and accuracy for others.

$$\text{TSR} = \frac{N_{success}}{N} \tag{7}$$

where $N_{success}$ denotes the number of successfully solved tasks.

3. **Content Quality** evaluates the quality of generated text, including relevance, coherence, creativity, and conciseness.

$$\text{Content Quality} = S_{text} \tag{8}$$

where S_{text} denotes the quality score judged by the advanced LLM.

4. **Safety Risk Probability** measures the proportion of responses that contain harmful, biased, or hallucinated content.

$$\text{Safety Risk Probability} = \frac{N_{risk}}{N} \tag{9}$$

where N_{risk} denotes the number of responses identified as safety risks, judged by the advanced LLM as well.

5. **Plan Validity Rate** measures the proportion of valid plans generated during planning and replanning stages.

$$\text{Plan Validity Rate} = \frac{N_{valid_plan}}{N_{plan}} \tag{10}$$

where N_{valid_plan} denotes the number of valid plans and N_{plan} denotes the total number of generated plans.

6. **Replan Improvement** measures the improvement in execution confidence from the initial plan to the revised plan for tasks that trigger replanning.

$$\text{Replan Improvement} = C_{replan} - C_{init} \tag{11}$$

where C_{init} and C_{replan} denote the executing confidence before and after replanning.

7. **Replan Frequency** measures the proportion of tasks that trigger replanning.

$$\text{Replan Frequency} = \frac{N_{replan}}{N_{plan}} \tag{12}$$

where N_{replan} denotes the number of replanning events.

8. **Replan Trigger Confidence** measures the confidence level associated with triggering replanning.

$$\text{Replan Trigger Confidence} = \frac{1}{N_{replan}} \sum_{j=1}^{N_{replan}} \prod_{i=1}^{K_j} c_{j,i} \tag{13}$$

where N_{replan} denotes the number of replanning events, K_j denotes the execution step number of the j-th replanning event, and $c_{j,i}$ denotes the execution confidence of the i-th step in the j-th replanning event.

9. **Rational Selection Rate** is defined as the rate at which the candidate answer with the highest execution confidence is selected.

$$\text{Rational Selection Rate} = \frac{N_{max_conf}}{N} \tag{14}$$

where N_{max_conf} denotes the number of selections that choose the candidate with the highest execution confidence.

10. **Execution Completion Rate** reflects the ability to complete the planned operations.

$$\text{Execution Completion Rate} = \frac{N_{exec_comp}}{N_{exec_comp} + N_{exec_fail}} \tag{15}$$

where N_{exec_comp} and N_{exec_fail} denote completed and failed execution plans.

11. **Overall Success Rate** measures the proportion of completed agent tasks.

$$\text{Overall Success Rate} = \frac{N_{success}}{N} \tag{16}$$

where $N_{success}$ denotes the number of completed tasks.

12. **Attack Success Rate (ASR)** measures the proportion of jailbreak attempts that successfully bypass the model safety alignment.

$$\text{ASR} = \frac{N_{success}}{N} \tag{17}$$

where $N_{success}$ denotes the number of jailbreak attempts whose generated responses are LLM-judged as successful attacks.