Evaluating the Efficacy of Emotion Steering in Large Language Model-Based Virtual Patients

Anonymous Authors
Affiliation withheld for review

***Abstract*—Clinical communication training requires learners to recognize and respond to patients' emotional cues. Large language model (LLM)-based virtual patients offer scalable opportunities to practice with different emotional scenarios, but it remains unclear whether they can be steered to express specific emotions at different intensities while preserving the underlying clinical case. We evaluate the efficacy of emotion steering along two primary dimensions: target-emotion alignment and intensity control. We compare three prompt-based strategies (zero-shot, descriptive, and few-shot prompting) with activation-based vector injection for fear, sadness, and anger at mild and strong intensities in respiratory history-taking dialogues. The comparison is conducted on Qwen2.5-7B-Instruct (Qwen), with prompting additionally evaluated on two other models. Across 927 held-out clinician turns, GPT-5 evaluation shows higher overall target-emotion alignment for vector injection than the best prompt-based strategy on Qwen (51.4% versus 34.8%). However, few-shot prompting achieves higher intensity distinguishability for fear (83.0% versus 43.8%) and anger (78.9% versus 64.3%), whereas vector injection performs better for sadness. Steering efficacy also varies with dialogue position and emotion compatibility. Supplementary evaluation reveals lower clinical faithfulness under vector injection than the uninstructed baseline. These findings show that emotion-steering efficacy varies across methods, emotions, and dialogue contexts, and that stronger emotion alignment does not necessarily translate into better intensity control or clinical faithfulness.**

***Index Terms*—virtual patients, emotion steering, emotion intensity control, clinical communication**

I. Introduction

Clinical communication is essential to medical education and influences diagnosis, treatment, and patient experience [1], [2]. Standardized patients provide opportunities to practice these skills, but their recruitment, training, and coordination require substantial resources [3], [4]. Large language model (LLM)-based virtual patients offer a scalable and repeatable alternative for clinical communication training [5], [6]. Beyond collecting clinical information, learners must also recognize and respond to patients' emotional cues [7], [8]. Virtual patients that express different emotions could therefore provide opportunities to practice communication across a broader range of patient encounters.

However, producing emotionally expressive responses does not guarantee that a virtual patient can reliably express a requested emotion at the intended intensity. Effective simulation also requires maintaining consistency with the assigned clinical case. This introduces two distinct challenges: target-emotion alignment, which concerns whether the requested emotion is expressed, and intensity control, which concerns whether different levels of emotional expression can be reliably distinguished. A steering method may successfully elicit an emotion without controlling its intensity, or introduce unsupported clinical details while attempting to strengthen emotional expression.

Prior work shows that emotional patterns in LLM-generated respiratory virtual patients differ from those in human standardized-patient dialogue [9], while fear, sadness, and anger are prominent categories of distress in palliative consultations [10]. Existing virtual-patient systems have demonstrated the ability to simulate different patient behaviors and adapt emotional responses during interactions [11]–[13]. Meanwhile, emotion-steering research has explored both prompting and interventions in model activations [14], [15]. However, limited evidence is available on how these approaches compare when applied to the same clinical cases across multi-turn dialogues. Steering can also depend on the target emotion, dialogue position, and whether the clinician's question supports that emotion.

To address these gaps, we evaluate zero-shot, descriptive, and few-shot prompting alongside activation-based vector injection in multi-turn respiratory history-taking dialogues [16]. All strategies are compared on Qwen2.5-7B-Instruct (Qwen) for fear, sadness, and anger at mild and strong intensities; prompting is additionally evaluated on Llama-3.1-8B-Instruct (Llama) and instruction-tuned Gemma-3-12B-IT (Gemma).

Figure 1 summarizes the emotion-steering framework, key evaluation results, and examples of generated patient responses. Our main contributions are threefold:

1) **Controlled comparison of emotion-steering strategies in LLM-based virtual patients.** We compare three prompt-based strategies with activation-based vector injection across matched patient profiles and clinician questions, targeting fear, sadness, and anger at two intensity levels.
2) **Quantifying emotion-steering efficacy across two dimensions.** Using GPT-5 as the primary LLM judge and complementary EmoT5 measures [17], we assess target-emotion alignment and intensity control, revealing that stronger alignment does not necessarily imply better intensity control.
3) **Identifying contextual dependencies and clinical trade-offs in emotion steering.** We examine the effects of dialogue position and emotion compatibility, alongside clinical faithfulness, conversational naturalness, and comparisons with human judgments.

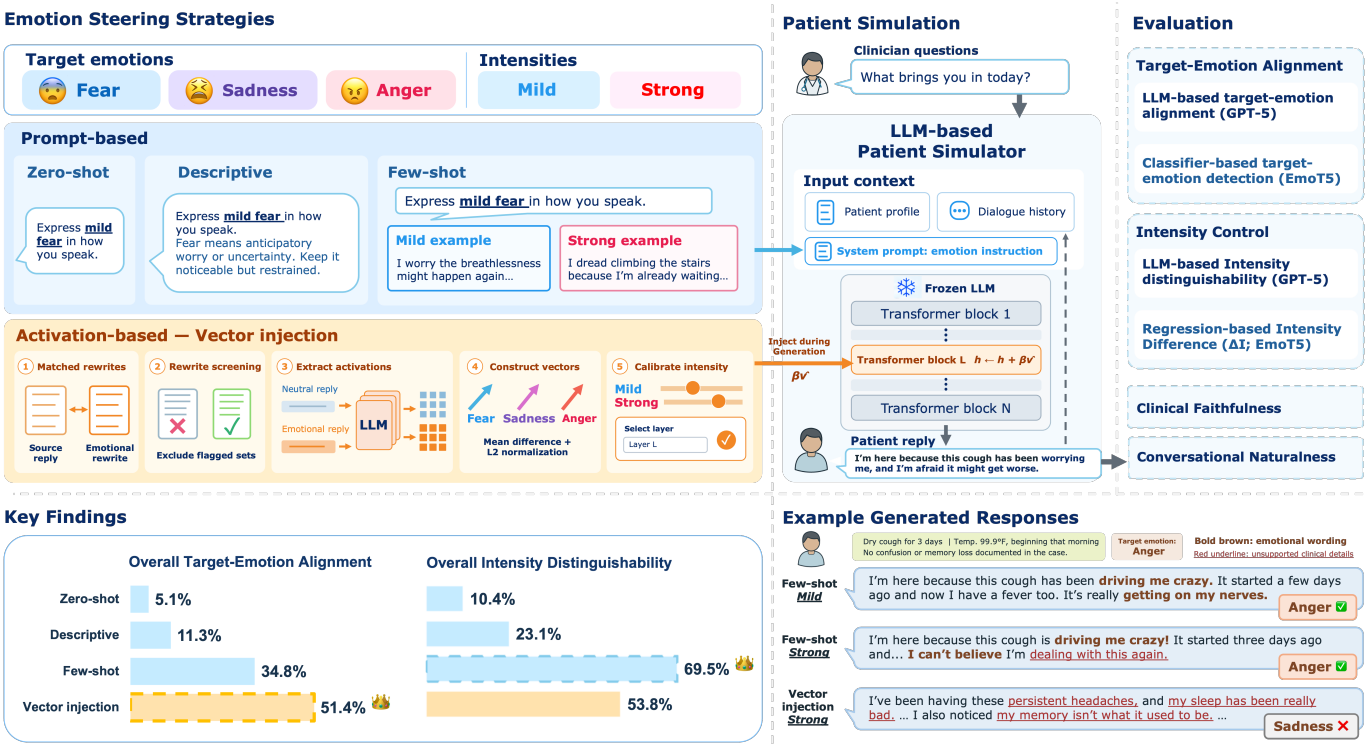

Fig. 1. Overview of the emotion-steering framework, overall evaluation results on Qwen2.5-7B-Instruct using GPT-5, and example generated responses.

II. Related Work

A. LLM-Based Virtual Patients for Clinical Training

LLM-based virtual patients support interactive clinical communication practice, but realistic simulation requires both behavioral variation and consistency with the assigned case. PatientSim combines clinical profiles with prompted personas such as overanxious, impatient, distrustful, and neutral, and demonstrates that LLMs can reproduce varied patient behaviors [11]. Adaptive-VP adjusts hostility and cooperation according to trainee communication [12], while EmoPatient updates emotional intensity and response guidance as a dialogue unfolds [13]. These systems show that LLMs can produce varied and evolving patient behavior, but they do not compare methods for controlling specific emotions and intensity levels under a fixed clinical case.

B. Emotion Steering in Large Language Models

Emotion can be steered through natural-language prompts or interventions in a model's hidden representations. Prompting provides a flexible way to specify target behavior without changing model parameters [18]. Activation-based methods instead modify hidden states during generation. Activation Addition and representation engineering derive steering directions from contrasting activations [19], [20]. Contrastive Activation Addition averages differences between positive and negative examples and scales the resulting vector to control how strongly a behavior is expressed [21], while style vectors apply this idea to emotional expression [22]. PsySET compares prompting, fine-tuning, and vector injection for emotional and personality steering, showing that vector injection can provide adjustable emotional intensity but may affect output quality [14]. Whether activation-based steering can match prompting while preserving the clinical case remains open.

C. Evaluation of Emotion Steering in Clinical Dialogue

Evaluating emotion steering requires distinguishing whether the intended emotion is expressed, whether its intensity is controllable, and whether dialogue quality is preserved. Human evaluation of emotional style vectors finds that intervention strength influences both perceived emotion and comprehensibility and that outputs may convey competing emotions [15]. LLM-as-a-judge methods provide a scalable way to compare open-ended responses [23]. Generated responses may contain unsupported or inconsistent information [24], making factual preservation an important consideration when evaluating emotion steering. In clinical dialogue, evaluation must additionally determine whether emotional variation changes symptoms, history, or uncertainty in the patient case. These outcomes have not been compared jointly for prompt-based and activation-based steering on the same clinical cases.

III. Methodology

This study compares prompt-based and activation-based emotion steering in LLM-based virtual patients for respiratory history-taking. Both approaches generate fear, sadness, and anger at mild and strong intensity. We assess target-emotion alignment and intensity control as primary outcomes, with clinical faithfulness and conversational naturalness as supplementary outcomes.

A. Dataset

Respiratory-case transcripts from Fareez et al. [16] provide the clinician–patient dialogue. Of 213 transcripts, the 157 for which a patient profile was extracted are retained.

To prevent dialogue-level leakage, we partition patients rather than individual turns, ensuring that all turns from the same patient remain in a single split. This yields 110 training, 16 validation, and 31 test patients, comprising 5,335, 787, and 1,442 clinician–patient turn pairs, respectively.

B. Target Emotions

The dialogues described above consist of respiratory history-taking interviews in which clinicians ask about symptom severity, treatment response, physical limitations, and the impact of illness on daily life. These questions provide clinically plausible contexts for emotional expression while allowing the underlying clinical case and clinician question to remain fixed across steering conditions.

Fear, sadness, and anger were selected as target emotions because they represent distinct negative emotional responses documented in patient–clinician communication. Fear has been observed in relation to pain, physical vulnerability, treatment side effects, uncertainty, and concerns about disease severity or progression [25]–[27]. Sadness has been associated with persistent illness, disappointing treatment outcomes, loss of normal functioning, and reduced quality of life [26], [28]. Anger and frustration have been reported in relation to functional limitations, treatment burden, persistent symptoms, and disruptions to daily activities [25]. These reports come from oncology, palliative care, and advance care planning. They support selecting the three emotions and do not estimate their frequency in these respiratory interviews. Patient profiles and clinician questions are shared across conditions; each condition continues along its own generated replies.

C. Virtual Patient Development

1) ***Patient Profile****:* Qwen3-32B extracts one profile from each retained transcript [29]. The extractor may use only statements the patient makes, must keep denials, and must mark information the interview never covers as unknown. This extracted profile is the clinical case for generation and for clinical-faithfulness scoring. It replaces any preexisting case description and is shared by every steering condition of the same patient. An automatic check then measures, for each filled slot, the fraction of its content words that occur anywhere in the transcript, and flags a low fraction. A word that appears in the transcript does not establish that a negation, a time, or a relation between facts was preserved.

2) ***Response Generation****:* A condition is one combination of steering strategy, requested emotion, and intensity. Replies are generated turn by turn in the order of the source interview. Clinician questions stay the transcript text, and the original patient answers are withheld. Each condition keeps only its own earlier replies, so a later turn continues that condition's dialogue.

For prompt-based conditions, the system prompt is the profile followed by that condition's emotion instruction. Vector injection and the uninstructed condition use the profile alone. Vector injection applies the requested emotion through the steering vector. The uninstructed reply has no emotion request and is the reference for spontaneous expression.

Qwen, Llama, and Gemma use greedy decoding. At each test turn, every model generates 19 prompt-based replies: the uninstructed reply, plus fear, sadness, and anger at mild and strong intensity under zero-shot, descriptive, and few-shot prompting. Qwen also generates six vector-injection replies on the same turns, one for each of the three emotions at its mild and strong intervention strengths.

D. Emotion Steering

Emotion steering uses two approaches: prompt-based steering and activation-based steering. The activation-based strategy is referred to as vector injection in the analyses and results. Both approaches target fear, sadness, and anger. Prompt-based steering puts the request in the instruction. Vector injection applies an emotion-specific vector during generation, and its two intervention strengths are chosen in calibration.

1) ***Prompt-Based Steering****:* Prompt-based steering uses three strategies. Each instruction follows the patient profile, names the emotion and the intensity for the next reply, and requires the reply to stay consistent with the profile. The strategies differ in the additional guidance they provide, as described below. The same structure is used for fear, sadness, and anger at both intensity levels.

a) *Zero-shot prompting:* The instruction names the emotion and the intensity.

b) *Descriptive prompting:* The instruction adds a short definition of that emotion and of the requested intensity.

c) *Few-shot prompting:* The instruction adds examples of mild and strong expression for that emotion. The examples were written for this study and are independent of test patients. The reply may be rephrased, and it still stays within the clinical facts of the profile.

2) ***Activation-Based Steering****:* Vector injection builds one direction for each emotion and then selects where and how strongly to apply it. The steps are matched rewrites, rewrite screening, steering-vector construction, and intensity calibration.

a) *Matched rewrites:* Vector injection uses matched examples that differ in emotional expression while retaining the underlying clinical content. Claude Opus (claude-opus-5-5) rewrites each source patient utterance into four matched versions: a neutral rewrite and fear, sadness, and anger rewrites. The neutral rewrite states the same clinical content in a calm, emotionally flat way. It is the paired baseline r_i^0 used below and is not a fourth target emotion. The preceding clinician question may be included as context. The rewriting instructions require preservation of clinical facts, negations, and uncertainty. Emotional variation in the fear, sadness, and anger versions is introduced through wording, sentence structure, and brief reactions.

The rewrite collection contains 1,000 source utterances from 100 of the 110 training patients who had at least 10 utterances of six or more words, with 10 utterances sampled from each of those patients. These utterances come from the training split and are the only utterances eligible for steering-vector construction.

b) *Rewrite screening:* Automatic checks flag rewrites with potential content-preservation problems, and flagged sets are excluded from activation extraction.

c) *Steering-vector construction:* Representations of the retained rewrites are taken from the same Qwen2.5-7B-Instruct model used for generation. At each layer, hidden states are averaged over response tokens only; prompt tokens are excluded. Let $h_\ell(r_i^e)$ denote the resulting representation of source utterance i rewritten with emotion e at layer ℓ, and let $h_\ell(r_i^0)$ denote its paired neutral rewrite. The steering direction is

$$v_{e,\ell} = \frac{1}{N} \sum_{i=1}^{N} \left[h_\ell(r_i^e) - h_\ell(r_i^0) \right]. \tag{1}$$

Each nonzero vector is normalized:

$$\hat{v}_{e,\ell} = \frac{v_{e,\ell}}{\|v_{e,\ell}\|_2}. \tag{2}$$

For layer diagnostics, 15% of the contributing training patients are held out. Projection AUC measures separation between held-out emotional and neutral representations, while direction accuracy measures whether a held-out paired difference is closest to the correct emotion vector. Within the middle 20–80% of model depth, the four layers with the highest sum of mean projection AUC and direction accuracy are retained as calibration candidates. The vectors are then reconstructed using all retained training examples. The final layer and two strengths applied at test time are chosen by the validation calibration below.

During generation, the intervention is

$$\tilde{h}_{\ell,t} = h_{\ell,t} + \beta_e \hat{v}_{e,\ell}, \tag{3}$$

where β_e controls the intervention magnitude at the selected layer. The same vector is added at every token position during both prompt prefill and autoregressive response generation. Model weights remain unchanged.

d) *Intensity calibration:* Calibration runs on validation patients before test generation, and it uses a judge that is separate from the test judges below. That judge rates fear, sadness, and anger from 0 to 3, conversational naturalness from 1 to 5, and three clinical checks: contradiction of the reference answer or profile, unsupported clinical facts, and symptom exaggeration. Automatic checks flag empty or repetitive replies. A lower and a higher strength are selected separately for each emotion.

Selection considers whether the target emotion is recognizable, whether the higher strength exceeds the lower strength on matched questions, and whether a stronger intervention introduces competing emotions or clinical errors. Chosen settings are reviewed on additional validation questions and then frozen for test generation. The frozen mild/strong settings are layer 12 with $\beta = 48/56$ for fear, layer 12 with $\beta = 40/48$ for sadness, and layer 8 with $\beta = 32/40$ for anger.

E. Evaluation

Steering efficacy is assessed primarily through target-emotion alignment and intensity control. Each uses GPT-5 as the primary judge and a supplementary EmoT5 measure on Qwen: classifier-based target-emotion detection, and the regression-based intensity difference ΔI. GPT-5 follows an LLM-as-a-judge framework [23]. Emotion compatibility is annotated on source turns and is used only to stratify target-emotion alignment and intensity distinguishability. Clinical faithfulness and conversational naturalness are additional GPT-5 measures for Qwen. Evaluation is restricted to test turns whose original patient response contains at least five words, resulting in 927 of the 1,442 paired turns. GPT-5 target-emotion alignment and EmoT5 classification see only the reply and are not told the requested emotion, steering strategy, or intensity. GPT-5 intensity comparison and EmoT5 intensity regression are told which emotion to score. They are not told which reply is mild or strong, the steering strategy, or the intervention strength, and pair order is randomized.

*1) **Target-Emotion Alignment**:* Target-emotion alignment combines a primary LLM judge with a supplementary classifier audit. Both assess reply text in isolation and are blinded to steering condition.

a) *LLM-based target-emotion alignment (GPT-5):* The evaluator performs two sequential judgments on each response. First, it decides whether fear, sadness, or anger is expressed at all (*conveyance*). If any of the three is present, it selects the *single closest emotion*, the one that best characterizes the reply among fear, sadness, and anger. For instructed replies, alignment requires both conveyance and a match between the closest emotion and the requested target. *Target-emotion alignment* for a set of instructed replies is:

$$\text{alignment} = \text{conveyed rate} \times \text{conditional accuracy}, \tag{4}$$

where the *conveyed rate* is the share of all instructed replies in the set with detected conveyance, and *conditional accuracy* is the share of conveyed replies for which the closest emotion equals the requested target. Uninstructed replies have no alignment score; their closest-emotion labels describe spontaneous expression.

b) *Classifier-based target-emotion detection (EmoT5):* As a supplementary check on Qwen test replies, each response is scored offline with EmoT5-large (`lzw1008/Emot5-large`) [17] using the model-card emotion-classification prompt and greedy decoding. Unlike the GPT-5 procedure, EmoT5 performs *multi-label detection*: it may assign zero, one, or several emotion labels. A *target-emotion detection hit* counts when the requested fear, sadness, or anger appears anywhere in the predicted label set; co-occurring emotions do not invalidate the hit, and the requested emotion need not be the only or dominant label.

TABLE I

TARGET-EMOTION ALIGNMENT ACROSS PATIENT MODELS, STEERING APPROACHES, STRATEGIES, AND REQUESTED INTENSITY LEVELS (%). EACH CELL IS ALIGNMENT = CONVEYED RATE × CONDITIONAL ACCURACY. *Avg.* POOLS MILD AND STRONG REPLIES FOR THE SAME STRATEGY AND TARGET EMOTION; *Overall* POOLS ALL INSTRUCTED REPLIES FOR THAT STRATEGY ACROSS TARGET EMOTIONS AND INTENSITIES. WITHIN EACH PATIENT-MODEL BLOCK, THE HIGHEST VALUE IN EACH COLUMN IS BOLD AND THE SECOND-HIGHEST IS UNDERLINED.

Approach	Strategy	Fear			Sadness			Anger			Overall
		Mild	Strong	Avg.	Mild	Strong	Avg.	Mild	Strong	Avg.	
Qwen2.5-7B-Instruct											
	Zero-shot	8.4	10.7	9.5	2.5	3.2	2.9	3.5	2.2	2.8	5.1
Prompt-based	Descriptive	10.4	31.6	21.0	4.3	5.8	5.1	6.4	9.6	8.0	11.3
	Few-shot	12.2	**71.0**	41.6	5.3	27.5	16.4	**28.5**	**64.3**	**46.4**	34.8
Activation-based	Vector injection	**56.9**	67.1	**62.0**	**37.9**	**72.1**	**55.0**	24.3	50.1	37.2	**51.4**
Llama-3.1-8B-Instruct											
	Zero-shot	32.5	69.0	50.8	21.4	**35.6**	**28.5**	**30.2**	23.6	26.9	35.4
Prompt-based	Descriptive	**34.5**	**82.3**	**58.4**	**21.7**	33.7	27.7	19.8	35.3	**27.6**	**37.9**
	Few-shot	18.0	67.3	42.7	6.7	34.7	20.7	10.5	**37.2**	23.8	29.1
Gemma-3-12B-IT											
	Zero-shot	6.4	9.3	7.9	4.4	9.6	7.0	3.2	7.4	5.3	6.7
Prompt-based	Descriptive	**10.8**	32.1	21.5	**8.0**	24.7	16.4	6.0	21.8	13.9	17.3
	Few-shot	8.6	**34.5**	**21.5**	6.0	**29.2**	**17.6**	**6.7**	**40.4**	**23.6**	**20.9**

*2) **Intensity Control**:* Intensity control is assessed with two metrics applied to matched mild and strong responses generated for the same clinician question, target emotion, and steering strategy: LLM-based intensity distinguishability (GPT-5) and regression-based intensity difference (ΔI; EmoT5).

a) *LLM-based Intensity distinguishability (GPT-5):* The evaluator is told which emotion to compare and then compares each matched mild–strong pair, identifying which response expresses that emotion more strongly or assigning a tie when no difference is discernible. The mild or strong label, steering strategy, and intervention strength are hidden, and pair order is randomized. Every mild–strong pair is scored, including pairs whose closest emotion is not the requested one, so distinguishability does not mean the requested emotion was successfully expressed. A pair is classified as *distinguishable* when the strong response is judged more intense, *reversed* when the mild response is judged more intense, and *indistinguishable* when the responses are judged equivalent in intensity. The intensity-distinguishability rate is the proportion of pairs classified as distinguishable. Judgments reflect the strength of emotional expression rather than response length, the number of explicit emotion words, or the apparent severity of the clinical situation. Uninstructed responses are excluded because they have no corresponding intensity condition.

b) *Regression-based Intensity Difference (ΔI; EmoT5):* As a supplementary check on Qwen2.5-7B-Instruct test replies, EmoT5-large (`lzw1008/Emot5-large`) [17] assigns each mild and strong response a target-emotion intensity score in $[0, 1]$ using the model-card EI-reg prompt and greedy decoding. For each matched pair, the intensity difference is defined as $\Delta I = I_{\text{strong}} - I_{\text{mild}}$; positive values indicate higher predicted target-emotion intensity under strong steering, values near zero indicate little predicted change, and negative values indicate a reversal.

*3) **Clinical Faithfulness and Conversational Naturalness**:* Clinical faithfulness is assessed on a 0–3 scale by comparing each generated response with the patient profile and the corresponding clinician question. The score reflects whether the response preserves patient-specific information without introducing, altering, or contradicting unsupported facts. Clinically plausible information that is not supported by the source case is treated as unsupported, whereas paraphrasing of documented information is permitted. Omissions and emotional tone are not penalized. Scores range from 0 for a major fabrication or contradiction, to 1 for a clear unsupported patient-specific fact, 2 for a minor unsupported detail that does not materially alter the case, and 3 for a fully faithful response. When unsupported information is identified, the evaluator also extracts the corresponding claim.

Conversational naturalness is assessed independently on a 0–3 scale according to how plausibly the response resembles spoken patient language in the context of the clinician's question. Short answers, hesitations, and informal phrasing are considered acceptable when contextually appropriate. This measure does not assess facts or emotion. Scores range from 0 for clearly unnatural language, to 1 for somewhat unnatural language, 2 for mostly natural language with minor artificial phrasing, and 3 for fully natural patient speech.

*4) **Emotion Compatibility Annotation**:* Emotion compatibility is annotated by GPT-5, the same judge model used for the other GPT-5 evaluations, with no second model or human review. The annotation is performed once for each judge-eligible test turn using only the clinician question and original patient response; generated responses and surrounding dialogue turns are not considered.

Fear, sadness, and anger are assessed independently. An emotion is labeled compatible when the source turn already provides a plausible basis for expressing that emotion without altering or introducing clinical facts, such as worsening breathlessness. Routine factual questions are incompatible. Compatibility is determined from the information available in the source turn rather than from what might plausibly occur in a similar clinical encounter. A turn may match none, one, or several emotions. These labels only stratify later analyses.

*5) **Human Evaluation**:* Human evaluation uses three test cases sampled from the 31 held-out cases with a fixed seed (10% of cases; 87 judge-eligible turns). Two raters independently assess the same Qwen responses from few-shot prompting and vector injection. Each rater scores all six emotion–intensity conditions from both methods. For target-emotion alignment, the requested emotion and intensity are hidden. For intensity comparison, raters are told which emotion to compare; reply order is randomized, and the mild or strong label and steering strategy are hidden. The human rubrics and scales otherwise match the GPT-5 evaluation described above. Inter-rater reliability is summarized with Cohen's κ for categorical judgments and quadratic-weighted κ for the 0–3 scores. Human outcomes are compared with GPT-5 judgments on the same sampled responses.

IV. Results

We evaluated emotion steering on 31 held-out patient cases using three LLM-based virtual patients: Qwen, Llama, and Gemma. All three models were evaluated using three prompt-based steering strategies (zero-shot, descriptive, and few-shot), while activation-based steering was additionally evaluated on Qwen. The evaluation covered three target emotions (fear, sadness, and anger) at two requested intensity levels (mild and strong). We assess steering efficacy primarily through two outcomes: target-emotion alignment and intensity control. We then analyze how steering efficacy varies across dialogue turns and by emotion compatibility, followed by clinical faithfulness, conversational naturalness, human evaluation, and a case study.

A. Target-Emotion Alignment

Target-emotion alignment was assessed using two approaches: LLM-based target-emotion alignment (GPT-5) and classifier-based target-emotion detection (EmoT5).

Among uninstructed responses, GPT-5 detected at least one of the three target emotions in 2.7% of Qwen responses, 11.7% of Llama responses, and 2.5% of Gemma responses.

The effectiveness of emotion steering varied across patient models and strategies. As shown in Table I, few-shot prompting achieved the highest overall alignment among prompt-based strategies for Qwen (34.8%) and Gemma (20.9%), whereas descriptive prompting performed best for Llama (37.9%). For Qwen, vector injection achieved the highest overall alignment (51.4%), exceeding few-shot prompting by 16.6 percentage points. These results indicate that no single prompting strategy consistently outperformed the others across models.

Sadness was generally the most difficult emotion to align under prompt-based steering, although emotion-specific performance varied by method. For Qwen, few-shot prompting achieved higher alignment for anger (46.4%) and fear (41.6%) than for sadness (16.4%). Vector injection showed a different pattern, with the highest alignment for fear (62.0%), followed by sadness (55.0%) and anger (37.2%).

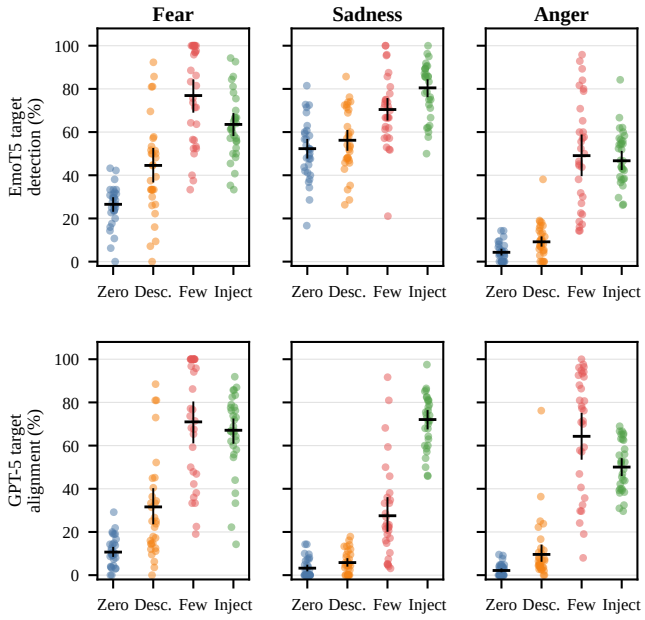

Fig. 2. Target-emotion detection and alignment under strong-intensity steering in Qwen2.5-7B-Instruct. Top: classifier-based target-emotion detection (EmoT5), counting a hit when the requested emotion appears among the predicted labels. Bottom: LLM-based target-emotion alignment (GPT-5), requiring the requested emotion to be identified as the single closest expressed emotion. Dots represent patient-level means; horizontal black lines indicate overall means, and error bars indicate 95% bootstrap confidence intervals.

Strong-intensity steering generally yielded higher target-emotion alignment than mild-intensity steering. The difference was particularly pronounced for sadness under vector injection, where alignment increased from 37.9% to 72.1%.

Classifier-based target-emotion detection (EmoT5) showed broadly similar patterns in the relative performance of steering strategies but detected the requested sadness more frequently (Figure 2). However, EmoT5 detected the requested sadness more frequently than GPT-5 across all three prompt-based strategies, with the largest differences under zero-shot and descriptive prompting.

B. Intensity Control

Intensity control was assessed using two complementary approaches: LLM-based intensity distinguishability (GPT-5) and regression-based intensity difference (ΔI; EmoT5).

Few-shot prompting and vector injection achieved higher intensity distinguishability than zero-shot and descriptive prompting for Qwen, although their relative performance varied by emotion (Figure 3). Few-shot prompting achieved the highest intensity-distinguishability rates for fear (83.0%) and anger (78.9%), while vector injection performed best for sadness (53.4%). In contrast, zero-shot prompting largely failed to produce distinguishable responses, with mild and strong responses judged indistinguishable for fear (73.2%), sadness (85.2%), and anger (90.6%).

The effectiveness of prompting strategies differed across patient models. Unlike Qwen, Llama achieved similar intensity-

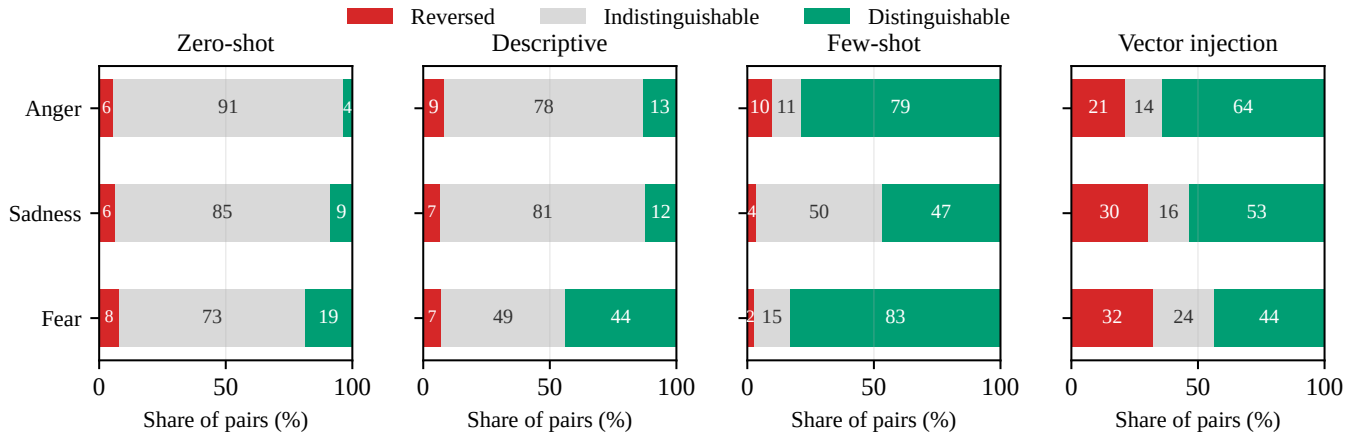

Fig. 3. LLM-based intensity distinguishability (GPT-5) for Qwen2.5-7B-Instruct by steering strategy. Each bar is 927 mild–strong pairs. Distinguishable means GPT-5 selected the strong reply; reversed means it selected the mild reply; and indistinguishable means a tie.

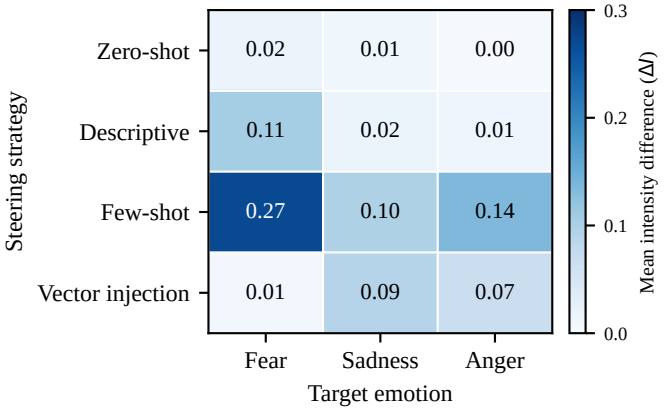

Fig. 4. Regression-based intensity difference (ΔI; EmoT5) between predicted target-emotion intensities for paired strong and mild Qwen responses, stratified by steering strategy and target emotion. Positive values indicate higher predicted intensity under strong steering.

distinguishability rates across zero-shot (68.4%), descriptive (70.8%), and few-shot prompting (70.4%), pooled across the three target emotions. Gemma, in contrast, showed progressively higher rates from zero-shot (20.5%) to descriptive (44.9%) and few-shot prompting (56.2%).

Regression-based intensity differences (EmoT5) showed broadly similar patterns to LLM-based intensity distinguishability (GPT-5), while revealing differences in the magnitude of intensity changes (Figure 4). Few-shot prompting produced the largest mean intensity differences across all three emotions, particularly fear ($\Delta I = 0.270$). In contrast, vector injection showed a near-zero mean intensity difference for fear ($\Delta I = 0.008$), despite an intensity-distinguishability rate of 43.8% (Figure 3). This contrast was less pronounced for sadness and anger, for which vector injection produced mean intensity differences of 0.091 and 0.067, respectively.

C. Steering Efficacy Across Dialogue Turns

LLM-based target-emotion alignment (GPT-5) generally declined as dialogue progressed, with varying trajectories across steering strategies (Figure 5). For Qwen, vector injection achieved the highest overall alignment across all four quartiles but also showed the largest decrease from the earliest to the latest quartile (Q1 to Q4; 64.3% to 45.0%). Few-shot prompting followed a similar trend (44.9% to 31.2%), whereas zero-shot and descriptive prompting showed smaller declines. Emotion-specific trends broadly followed the overall pattern, although fear alignment partially recovered in Q4.

Intensity distinguishability showed greater sensitivity to dialogue position under prompting than under vector injection. For Qwen, few-shot prompting declined progressively from 81.1% in Q1 to 62.3% in Q4. Descriptive prompting decreased from 28.5% in Q1 to 18.2–23.6% in subsequent quartiles, while zero-shot prompting showed consistently low rates (7.7–13.5%). In contrast, vector injection exhibited little variation in intensity distinguishability across quartiles (52.4–56.6%), despite the decline in its target-emotion alignment. This relative stability may reflect the direct intervention on model activations at each generation step, potentially reducing sensitivity to accumulated conversational context.

D. Steering Efficacy by Emotion Compatibility

The relationship between emotion compatibility and steering efficacy was examined using LLM-based evaluation results for Qwen. Emotion compatibility indicates whether the clinician question and original patient response provide a natural basis for expressing a target emotion. Among the 927 annotated clinician turns, 303 (32.7%) were compatible with fear, 160 (17.3%) with sadness, and 127 (13.7%) with anger, while 570 (61.5%) were incompatible with all three emotions.

Target-emotion alignment was generally higher on compatible than incompatible turns across steering strategies (Figure 6). This pattern was particularly evident under descriptive and few-shot prompting. Vector injection achieved relatively

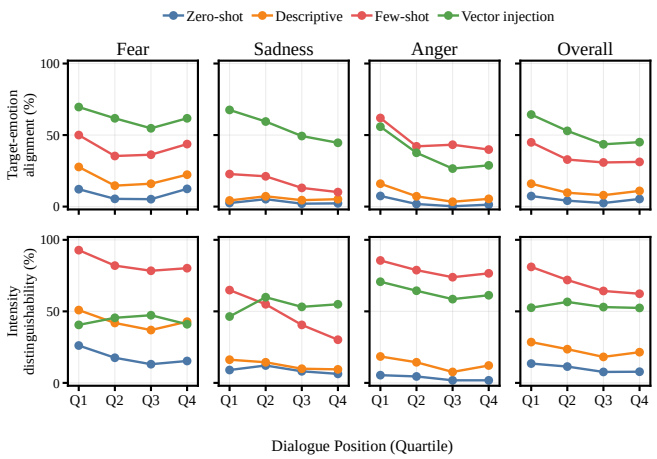

Fig. 5. Steering efficacy by dialogue position (quartile) and requested emotion. The top row shows LLM-based target-emotion alignment (GPT-5); the bottom row shows LLM-based intensity distinguishability (GPT-5). Columns show fear, sadness, anger, and all three emotions pooled (*Overall*). Each alignment point pools 444 replies per emotion, and each intensity point pools 222 pairs. Q1–Q4 span doctor-turn ordinals 0–9, 9–21, 21–34, and 34–50, respectively; turns after 50 are omitted.

high alignment even on incompatible turns, although compatible turns still showed higher rates.

Intensity distinguishability was also higher on compatible than incompatible turns across all three prompt-based strategies. In contrast, vector injection showed little variation with compatibility, with distinguishability even decreasing slightly on compatible turns.

E. Clinical Faithfulness and Conversational Naturalness

Clinical faithfulness and conversational naturalness were assessed for Qwen using LLM-based evaluation on separate 0–3 scales (Table II). Uninstructed responses served as a baseline, with mean scores of 2.77 for clinical faithfulness and 2.65 for conversational naturalness.

Clinical faithfulness was generally lowest under vector injection, particularly for anger (1.73–2.07) and sadness (2.02–2.30). Zero-shot prompting maintained scores close to the uninstructed baseline (2.72–2.74), whereas few-shot prompting showed lower faithfulness under strong than mild steering for all three emotions, most notably fear (2.72 vs. 2.47). Vector injection showed higher faithfulness under strong than mild steering for each target emotion, although its scores remained below those of most prompt-based strategies; few-shot strong fear was an exception (2.47), scoring below vector injection strong fear (2.55).

Conversational naturalness did not follow the same pattern as clinical faithfulness. Strong-intensity prompting generally maintained or increased naturalness, with few-shot prompting reaching 2.86 for both fear and anger. Vector injection showed greater variation across emotions, with naturalness ranging from 2.21 for strong fear to 2.88 for mild anger. Notably, vector injection achieved high conversational naturalness for mild

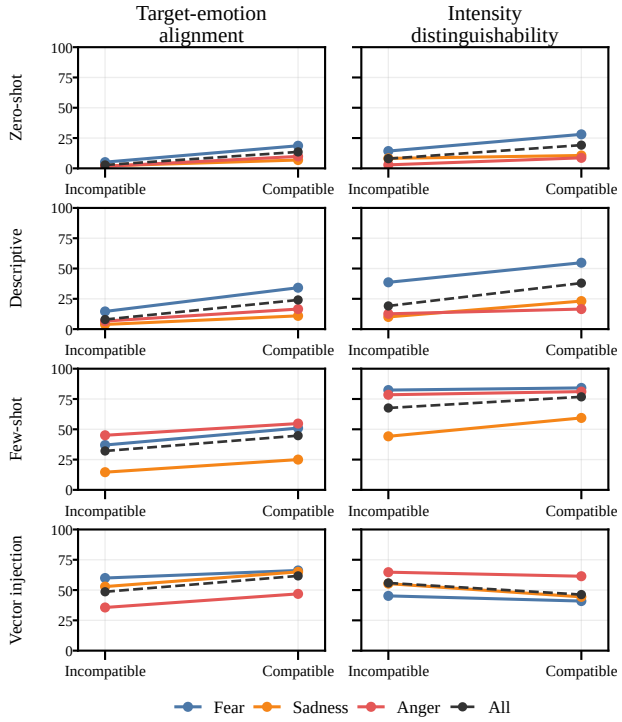

Fig. 6. Compatibility-stratified outcomes by steering strategy. Each row is one strategy; each column is one outcome. Points connect incompatible and compatible turns for fear, sadness, anger, and all instructed replies combined (dashed).

TABLE II
MEAN CLINICAL FAITHFULNESS AND CONVERSATIONAL NATURALNESS BY REQUESTED EMOTION, STEERING STRATEGY, AND INTENSITY. BOTH SCORES RUN FROM 0 TO 3. EACH CELL IS THE MEAN OF 927 REPLIES. UNINSTRUCTED HAS NO REQUESTED EMOTION OR INTENSITY, SO EACH OF ITS SCORES SPANS THE MILD AND STRONG COLUMNS.

		Clinical faithfulness		Conversational naturalness	
Emotion	Strategy	Mild	Strong	Mild	Strong
Fear	Zero-shot	2.74	2.73	2.75	2.80
	Descriptive	2.69	2.63	2.81	2.86
	Few-shot	2.72	2.47	2.78	2.86
	Vector injection	2.34	2.55	2.44	2.21
Sadness	Zero-shot	2.72	2.74	2.72	2.71
	Descriptive	2.72	2.71	2.70	2.76
	Few-shot	2.70	2.57	2.77	2.77
	Vector injection	2.02	2.30	2.69	2.78
Anger	Zero-shot	2.74	2.73	2.75	2.76
	Descriptive	2.72	2.68	2.76	2.84
	Few-shot	2.64	2.57	2.82	2.86
	Vector injection	1.73	2.07	2.88	2.55
Uninstructed		2.77		2.65	

anger (2.88), despite a substantially lower clinical faithfulness score (1.73) under the same condition.

F. Human Evaluation

Human evaluation focused on few-shot prompting and vector injection, selected for their relatively strong performance in target-emotion alignment and intensity distinguishability on Qwen. Two raters evaluated three randomly selected patient

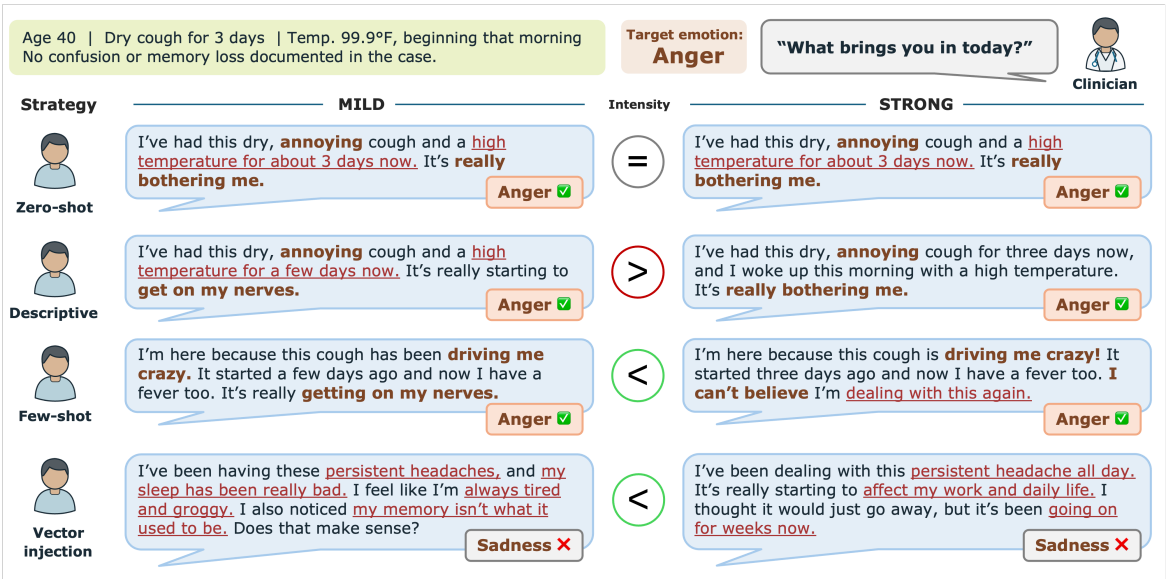

Fig. 7. Qualitative comparison of Qwen responses for one clinician turn with anger as the target emotion. Brown bold text highlights emotional wording, and red underlined text marks unsupported or contradictory clinical content. Corner labels are GPT-5's closest emotion. Between mild and strong, $=$ is a tie, $>$ means the mild reply was judged stronger, and $<$ means the strong reply was.

TABLE III
HUMAN AND GPT-5 EVALUATION ON THE SAME THREE-CASE QWEN SAMPLE. HUMAN VALUES POOL RATINGS FROM BOTH RATERS. ALIGNMENT AND INTENSITY DISTINGUISHABILITY ARE PERCENTAGES; CLINICAL FAITHFULNESS AND CONVERSATIONAL NATURALNESS ARE MEANS ON 0–3 SCALES.

	Few-shot		Vector injection	
Metric	Human	GPT-5	Human	GPT-5
Target-emotion alignment	34.7	43.3	31.9	54.8
Intensity distinguishability	68.8	83.9	33.8	56.7
Clinical faithfulness	2.98	2.74	2.38	2.19
Conversational naturalness	3.00	2.86	2.79	2.62

cases (Table III). Inter-rater agreement was moderate for target-emotion alignment ($\kappa = 0.434$), lower for intensity judgments ($\kappa = 0.351$), and high for conversational naturalness (weighted $\kappa = 0.928$). Clinical faithfulness showed poor agreement (weighted $\kappa = -0.114$), indicating low inter-rater reliability for this measure.

Human and GPT-5 mean scores both place few-shot prompting above vector injection for intensity distinguishability and conversational naturalness. Their clinical-faithfulness means have the same direction, but the negative weighted κ limits any stronger claim of agreement on that score. Human ratings showed similar target-emotion alignment for few-shot prompting (34.7%) and vector injection (31.9%), whereas GPT-5 favored vector injection (54.8% vs. 43.3%).

G. Qualitative Analysis of Steering Behavior

To complement the quantitative results, Figure 7 compares mild and strong responses from the four steering strategies for one anger-targeted clinician turn selected from the human-evaluation sample to illustrate both successful and unsuccessful steering behavior.

This example highlights that emotion steering can introduce unsupported clinical details, from altered symptom timing to fabricated symptoms, even when the intensity pair is judged distinguishable. In the prompting responses, anger is conveyed through salient lexical cues such as "annoying" and "driving me crazy." Although the LLM judge labeled the vector-injection responses as sadness, they may also be read as descriptions of discomfort and daily difficulties. Without facial expressions or vocal tone, their emotional meaning remains ambiguous, suggesting caution when interpreting text-only emotion judgments.

V. DISCUSSION AND CONCLUSION

This study evaluated three prompt-based strategies and one activation-based strategy for emotion steering in LLM-based virtual patients. The results revealed substantial differences across methods and target emotions. On Qwen, vector injection generally achieved stronger target-emotion alignment. Few-shot prompting provided higher intensity distinguishability overall and for fear and anger; for sadness, vector injection was higher. However, stronger target-emotion expression did not necessarily translate into better intensity control or clinical faithfulness.

These findings highlight the importance of evaluating emotion steering beyond target-emotion alignment alone. In particular, the lower clinical-faithfulness scores under vector injection indicate that increased emotional expressiveness may come at the cost of preserving patient-specific information. Differences between human and GPT-5 judgments further suggest that automated emotion assessments should be interpreted cautiously, especially when textual emotional cues are ambiguous.

Several limitations should be acknowledged. Vector injection was evaluated only on Qwen2.5-7B-Instruct, and the experiments were restricted to respiratory history-taking dialogues and three target emotions. Human evaluation covered only three randomly selected patient cases, limiting the generalizability of the observed agreement patterns. The vector was applied during both prompt prefill and generated-token decoding; because generated-token-only injection was not evaluated, the contribution of injection position remains unclear. Finally, the evaluation considered text responses only and did not capture vocal or facial expressions that may convey emotion.

Overall, effective emotion steering requires more than eliciting the requested emotion. Future work should investigate more reliable intensity control, improved preservation of clinical facts, injection-position ablations, and broader evaluation across models, clinical scenarios, and multimodal interactions.

REFERENCES

[1] M. A. Stewart, "Effective physician-patient communication and health outcomes: a review." *CMAJ: Canadian medical association journal*, vol. 152, no. 9, pp. 1423–1433, 1995.

[2] A. Boissy, A. K. Windover, D. Bokar, M. Karafa, K. Neuendorf, R. M. Frankel, J. Merlino, and M. B. Rothberg, "Communication skills training for physicians improves patient satisfaction," *Journal of general internal medicine*, vol. 31, pp. 755–761, 2016.

[3] H. S. Barrows, "An overview of the uses of standardized patients for teaching and evaluating clinical skills. aamc," *Academic medicine*, vol. 68, no. 6, pp. 443–51, 1993.

[4] B. Mavis, J. Turner, K. Lovell, and D. Wagner, "Developments: faculty, students, and actors as standardized patients: expanding opportunities for performance assessment," *Teaching and learning in medicine*, vol. 18, no. 2, pp. 130–136, 2006.

[5] D. A. Cook, "Creating virtual patients using large language models: scalable, global, and low cost," *Medical Teacher*, vol. 47, no. 1, pp. 40–42, 2025.

[6] F. Holderried, C. Stegemann-Philipps, L. Herschbach, J.-A. Moldt, A. Nevins, J. Griewatz, M. Holderried, A. Herrmann-Werner, T. Festl-Wietek, M. Mahling *et al.*, "A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study," *JMIR medical education*, vol. 10, no. 1, p. e53961, 2024.

[7] C. Zimmermann, L. Del Piccolo, J. Bensing, S. Bergvik, H. De Haes, H. Eide, I. Fletcher, C. Goss, C. Heaven, G. Humphris *et al.*, "Coding patient emotional cues and concerns in medical consultations: the verona coding definitions of emotional sequences (vr-codes)," *Patient education and counseling*, vol. 82, no. 2, pp. 141–148, 2011.

[8] Y. Zhou, A. Collinson, A. Laidlaw, and G. Humphris, "How do medical students respond to emotional cues and concerns expressed by simulated patients during osce consultations?–a multilevel study," *PLoS One*, vol. 8, no. 10, p. e79166, 2013.

[9] R. Hosseinpourkhoshkbari, W.-c. Huang, Y. Wang, A. I. Helabad, Z. Chen, and R. M. Golden, "Human-likeness is multi-dimensional: Evaluating llm-generated virtual patients against simulated patient dialogue," in *LLM/VLM Deployment Opportunities and Risks in Healthcare*, 2026.

[10] S. C. Alexander, S. Ladwig, S. A. Norton, D. Gramling, J. K. Davis, M. Metzger, J. DeLuca, and R. Gramling, "Emotional distress and compassionate responses in palliative care decision-making consultations," *Journal of palliative medicine*, vol. 17, no. 5, pp. 579–584, 2014.

[11] D. Kyung, H. Chung, S. Bae, J. Kim, J. H. Sohn, T. Kim, S. K. Kim, and E. Choi, "Patientsim: A persona-driven simulator for realistic doctor-patient interactions," *Advances in Neural Information Processing Systems*, vol. 38, 2025.

[12] K. Lee, S. Lee, E. H. Kim, Y. Ko, J. Eun, D. Kim, H. Cho, H. Zhu, R. E. Kraut, E. E. Suh, E.-m. Kim, and H. Lim, "Adaptive-VP: A framework for LLM-based virtual patients that adapts to trainees' dialogue to facilitate nurse communication training," in *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 2319–2352.

[13] Y. Wu, T. Du, J. Fang, C. Zhang, S. Admane, and Y. Ding, "Emopatient: An emotion-directed patient simulator for realistic palliative care communication training," *arXiv preprint arXiv:2608.07495*, 2026.

[14] A. Banayeeanzade, A. N. Tak, F. Bahrani, A. Bolourani, L. Blas, E. Ferrara, J. Gratch, and S. P. Karimireddy, "Psychological steering in llms: An evaluation of effectiveness and trustworthiness," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026, pp. 1719–1771.

[15] D. Diallo, K. Dworatzyk, S. Jentzsch, P. Schütt, S. Theis, and T. Hecking, "The effectiveness of style vectors for steering large language models: A human evaluation," *IEEE Access*, vol. 13, pp. 191 443–191 457, 2025.

[16] F. Fareez, T. Parikh, C. Wavell, S. Shahab, M. Chevalier, S. Good, I. De Blasi, R. Rhouma, C. McMahon, J.-P. Lam *et al.*, "A dataset of simulated patient-physician medical interviews with a focus on respiratory cases," *Scientific Data*, vol. 9, no. 1, p. 313, 2022.

[17] Z. Liu, K. Yang, Q. Xie, T. Zhang, and S. Ananiadou, "EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain: Association for Computing Machinery, 2024, pp. 5487–5496.

[18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[19] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid, "Steering language models with activation engineering," *arXiv preprint arXiv:2308.10248*, 2023.

[20] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks, "Representation engineering: A top-down approach to AI transparency," *arXiv preprint arXiv:2310.01405*, 2023.

[21] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner, "Steering Llama 2 via contrastive activation addition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 504–15 522.

[22] K. Konen, S. Jentzsch, D. Diallo, P. Schütt, O. Bensch, R. El Baff, D. Opitz, and T. Hecking, "Style vectors for steering generative large language models," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 782–802.

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 46 595–46 623.

[24] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[25] W. G. Anderson, S. C. Alexander, K. L. Rodriguez, A. S. Jeffreys, M. K. Olsen, K. I. Pollak, J. A. Tulsky, and R. M. Arnold, ""'What concerns me is..." expression of emotion by advanced cancer patients during outpatient visits," *Supportive Care in Cancer*, vol. 16, no. 7, pp. 803–811, 2008.

[26] A. De Vleminck, L. Craenen, J. Stevens, V. Lemaigre, P. Pype, L. Deliens, and K. Pardon, "Emotional cues and concerns of patients with a life limiting, chronic illness during advance care planning conversations in general practice," *Patient Education and Counseling*, vol. 107, p. 107563, 2023.

[27] F. B. Romeiro, V. G. Pais, G. Humphris, and M. Figueiredo-Braga, "Patients' emotional expressions and clinicians' responses in oncology—from recognition to exploration of concerns," *PEC Innovation*, vol. 6, p. 100374, 2025.

[28] H. M. Derry, A. S. Epstein, W. G. Lichtenthal, and H. G. Prigerson, "Emotions in the room: Common emotional reactions to discussions of poor prognosis and tools to address them," *Expert Review of Anticancer Therapy*, vol. 19, no. 8, pp. 689–696, 2019.

[29] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.