

RECAP: Transparent Inference-Time Emotion Alignment for Medical Dialogue Systems

Adarsh Srinivasan*, Jacob Dineen*, Muhammad Umar Afzal†, Muhammad Uzair Sarfraz†, Irbaz B. Riaz†, Ben Zhou*

*Arizona State University, Tempe, Arizona, USA
{asrini81, jdineen, xzhou202}@asu.edu

†Mayo Clinic, Tempe, Arizona, USA
{Afzal.MuhammadUmar, Sarfraz.MuhammadUzair, Riaz.Irbaz}@mayo.edu

Abstract

Large language models in healthcare often produce emotionally flat or opaque responses, failing to provide the transparent reasoning required for clinical trust. We present RECAP (Reflect–Extract–Calibrate–Align–Produce), an inference-time framework grounded in cognitive appraisal theory that decomposes patient input into auditable, appraisal-theoretic stages without retraining. Across multiple benchmarks and models from 8B to 120B parameters, RECAP improves alignment with human judgments, with gains inversely proportional to model scale. Intermediate outputs further reveal that models systematically underweight relational factors such as social support. In blinded evaluations, oncology fellows rated RECAP responses significantly higher than baselines with 76–88% win rates, demonstrating that principled prompting can enhance medical AI’s emotional intelligence while maintaining the transparency required for clinical deployment.

Keywords: emotional reasoning, medical dialogue, interpretability, appraisal theory, inference-time alignment

1. Introduction

Healthcare communication is fundamentally human-centered. When a patient receives difficult news or expresses anxiety, clinicians draw on training to recognize emotional cues, assess the patient’s psychological state, and calibrate their response accordingly. This emotional attunement is central to building trust, supporting patient agency, and improving outcomes (Chen et al., 2025). As conversational AI systems enter clinical settings, a fundamental question arises about how these systems process emotional content and whether we can make this reasoning transparent.

Current AI-assisted medical dialogue systems, while sophisticated in knowledge retrieval, leave emotional reasoning as a black box. A patient learning of a cancer diagnosis faces not only informational needs but profound emotional upheaval, including fear, uncertainty about treatment, and concerns about family. Two patients with identical diagnoses may require vastly different support depending on their sense of control, coping resources, and life circumstances (Ortony et al., 1988; Lazarus, 1991). Yet we cannot determine whether contemporary AI systems recognize these distinctions or simply pattern-match to empathetic-sounding language. In preliminary evaluations (Table 1), even state-of-the-art models consistently under-attend to affective dimensions that clinicians consider essential.

This opacity matters for both deployment and understanding. Prior work documents cases where AI responses felt “cold” or “robotic” despite being

medically accurate (Sorin et al., 2024; Armbruster et al., 2024). Clinicians express reluctance to deploy systems whose emotional reasoning remains opaque, because they need to understand why an AI produces a particular response (Chen et al., 2025). This need for interpretability distinguishes healthcare from domains where explainability is desirable but not essential. Beyond deployment concerns, this opacity limits our scientific understanding of how LLMs represent and reason about human emotion.

Cognitive appraisal theories suggest that humans evaluate events along key dimensions such as goal relevance, agency, and coping potential to determine emotional responses (Ortony et al., 1988; Lazarus, 1991; Huang et al., 2024). While LLMs do not experience emotions, recent work shows they can emulate aspects of this appraisal process when properly guided (Croissant et al., 2024). This presents an opportunity. By prompting models to produce appraisal-theoretic intermediates explicitly, we can both improve emotional alignment and examine how models reason about affect.

We introduce RECAP (Reflect–Extract–Calibrate–Align–Produce), an inference-time pipeline that externalizes emotional reasoning into inspectable stages. RECAP guides models through structured reasoning, abstracting the situation, identifying psychological factors, generating candidate emotions, quantifying their likelihood via Likert scales, and producing an emotionally aligned response. Each stage yields interpretable artifacts such as situational abstractions, psychological

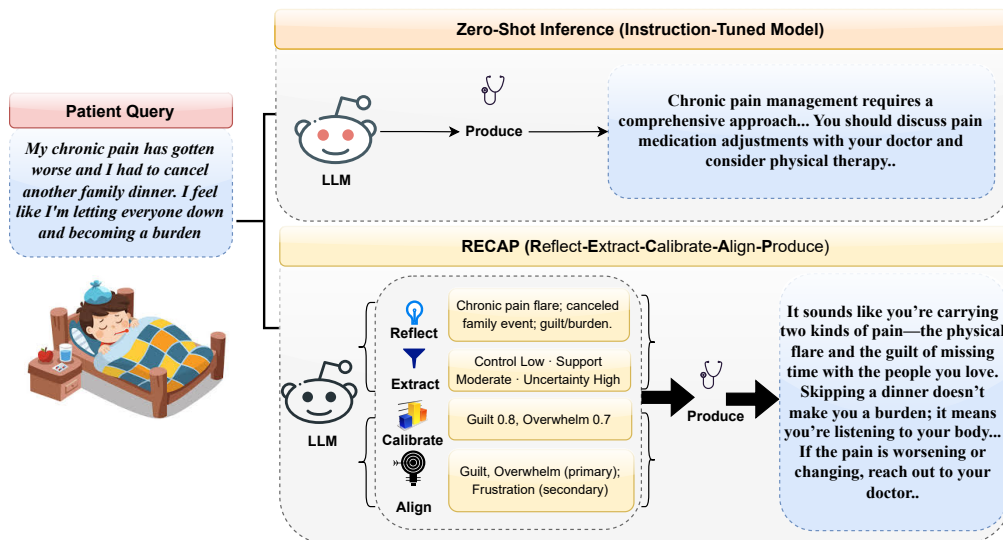


Figure 1: Patient input (left) is transformed into appraisal-theoretic intermediates with per-dimension Likert ratings (center), which condition generation to produce a targeted reply (right). Compared with a vanilla instruction-tuned model, RECAP explicitly acknowledges the patient’s emotions and provides concrete next steps. (*Scenario text is synthetic for illustration.*)

factors, and emotion candidates with intensity ratings. These artifacts serve two purposes, providing clinicians with auditable checkpoints and enabling analysis of how models represent emotional content.

Our experiments reveal three insights about emotional reasoning in LLMs. First, explicit calibration improves alignment with human judgments, as RECAP reduces distance from human consensus by 20% on SECEU. Second, structured reasoning disproportionately benefits smaller models (28% improvement for 8B vs 5% for 120B), suggesting explicit decomposition compensates for capabilities larger models have internalized. Third, examining intermediate outputs reveals systematic biases, with models consistently identifying control and uncertainty but underweighting relational factors like social support. Beyond these insights, RECAP improves practical performance. Oncology fellows rated RECAP responses significantly higher than baselines (3.41 vs 2.93–2.95, $p < .001$), with 76–88% win rates across clinical scenarios.

2. Background and Related Work

2.1. Empathy in Medical Dialogue

LLMs are increasingly used for health information and triage. Multiple studies report that chatbot replies can be perceived as highly empathetic, sometimes surpassing physicians in perceived warmth while maintaining clinical accuracy (Ayers et al., 2023; Chen et al., 2024; Luo et al., 2024; Franke Føyen et al., 2025). However, system-

atic reviews find empathy to be brittle and prompt-sensitive, with overuse of generic phrases that undermine genuine connection (Sorin et al., 2024; Armbruster et al., 2024; Ye et al., 2025a). This raises a fundamental question about whether LLMs reason about emotional states or rely on surface-level stylistic mimicry. Answering this requires methods that externalize emotional reasoning into inspectable components.

2.2. Approaches to Emotional Alignment

Training-time methods including RLHF, DPO, and supervised fine-tuning on empathetic corpora can raise perceived empathy to clinician-rated levels (Ouyang et al., 2022; Rafailov et al., 2023; Rashkin et al., 2019; Chen et al., 2023; Franke Føyen et al., 2025). However, these approaches embed empathetic behavior in model weights, making reasoning opaque and difficult to audit (Beede et al., 2020). We cannot determine whether a fine-tuned model recognized the user’s emotional state or simply learned empathetic-sounding language, a critical limitation for medical applications where transparency is not merely desirable but essential for safety, regulatory compliance, and clinical trust.

Inference-time approaches offer greater interpretability. Persona prompts like “supportive clinician” show mixed effectiveness (Zhang et al., 2018; Hu and Collier, 2024; Zheng et al., 2024; Luz de Araujo and Roth, 2025). Multi-step reasoning methods like Chain of Thought, Chain of Empathy, and Chain of Emotion improve emotional coherence through structured prompting (Wei et al., 2022;

Lee et al., 2024; Croissant et al., 2024). Recent work demonstrates that decomposing inference into explicit steps improves both performance and auditability (Zhou et al., 2022; Wolfson et al., 2020; RRV et al., 2025; Shen et al., 2025; Xu et al., 2024; Tafjord et al., 2022; Feng et al., 2024, 2023; Zhou et al., 2024; Ye et al., 2025b).

RECAP extends decomposition to emotional reasoning, grounding our pipeline in cognitive appraisal theory (Ortony et al., 1988; Lazarus, 1991). Unlike prior work, we expose per-dimension ratings that make emotional reasoning inspectable, enabling examination of whether LLMs apply consistent theories of emotion causation or exhibit systematic biases (Amershi et al., 2019; Buçinca et al., 2021).

2.3. Transparency in Clinical AI

Transparency is critical for AI systems making emotional assessments in healthcare. High technical accuracy alone is insufficient, and systems fail when reasoning is opaque or misaligned with clinical practices (Beede et al., 2020). Systems with interpretable reasoning are more likely to be trusted and adopted (Cai et al., 2019; Yang et al., 2019; Dinneen et al., 2024, 2025; Dolk et al., 2020). RECAP addresses this by exposing intermediate artifacts, including situation abstractions, psychological factors, emotion candidates, and calibrated ratings, that clinicians can inspect, verify, or override.

3. RECAP

Understanding how LLMs process emotional content remains largely opaque. When a model produces an empathetic response, we cannot determine whether it recognized the user’s emotional state, inferred relevant psychological context, or simply pattern-matched to training examples. RECAP addresses this by decomposing emotional reasoning into five explicit stages, each producing an inspectable intermediate artifact. This design serves two purposes. First, it improves response quality through structured reasoning. Second, it creates observation points that reveal how models represent and reason about affect.

We ground our decomposition in cognitive appraisal theory (Ortony et al., 1988; Lazarus, 1991), which posits that emotions arise from evaluations along dimensions such as goal relevance, agency, and coping potential. Our five-stage structure corresponds to theorized components of human emotional appraisal (Scherer and Moors, 2019), namely situation encoding, relevance assessment, emotion categorization, intensity estimation, and response selection. Unlike chain-of-thought prompting, which produces unstructured traces that re-

sist quantitative analysis, RECAP’s fixed schema enables aggregation across examples. We can compute statistics over factor distributions, emotion candidate frequencies, and calibration curves that reveal population-level patterns in model behavior. The pipeline requires no fine-tuning and executes in a single conversation turn. Figure 2 illustrates the complete pipeline with a representative clinical example.

To demonstrate each stage, we use a running example where a patient expresses the following.

“I was just diagnosed with cancer last month and the chemo is making me so sick. I’m scared this won’t get any better. My sons keep asking me if I’m okay and I don’t know what to say to them...”

3.1. Situation Understanding (Stages 1–2)

Stage 1, Situation Abstraction. The pipeline first distills raw input into a concise description of the core situation, stripping lengthy narratives to capture essential circumstances (Sweller, 2023; Feng et al., 2024). The abstraction reveals what the model considers emotionally salient. Abstractions that omit key triggers (e.g., family concerns in Figure 2) signal gaps in emotional scene understanding.

Example: “The patient has been diagnosed with a serious illness and is undergoing difficult treatment, which makes them feel overwhelmed and anxious about the future.”

Stage 2, Latent Factor Induction. The model identifies three psychological factors that shape emotional responses, assigning values and justifications. These constructs directly map to core appraisal theory dimensions, specifically *primary appraisal* (assessing goal relevance and congruence) and *secondary appraisal* (evaluating agency, control, and coping potential) (Scherer and Moors, 2019; Lazarus, 1991). Factor induction exposes the model’s implicit theory of emotion causation. The factors selected, and those not selected, reveal which dimensions the model treats as emotionally relevant, enabling comparison with human appraisal patterns and identifying systematic biases that propagate downstream. For the running example, the induced factors are as follows.

- **Control (Secondary Appraisal)** Low, as the patient has little control over illness or treatment.
- **Support (Coping Potential)** Moderate, as family is present but the patient feels isolated.

Patient Input: "I was just diagnosed with cancer last month and the chemo is making me so sick. I'm scared this won't get any better. My sons keep asking me if I'm okay and I don't know what to say to them..."

1. REFLECT	2. EXTRACT	3. CALIBRATE	4. ALIGN	5. PRODUCE
Situation Abstraction "The patient has been diagnosed with a serious illness and is undergoing difficult treatment, causing them to feel overwhelmed and have anxiety about the future."	Factor Induction Control: Low <small>(little control over illness)</small> Support: Moderate <small>(family present)</small> Uncertainty: High <small>(unsure of outcomes)</small>	Candidate Emotions <ul style="list-style-type: none"> Anxiety Fear Overwhelm/Stress Sadness Hope Guilt Helplessness 	Likert Rating Anxiety: 0.95 Fear: 0.95 Overwhelm: 0.95 Sadness: 0.50 Hope: 0.25 Guilt: 0.75 Helplessness: 0.75	Aligned Response <i>"I'm really sorry you're feeling so overwhelmed and anxious. That's completely understandable—facing serious illness can be frightening..."</i>

Figure 2: **RECAP Pipeline for Emotional Alignment.** Model-agnostic inference-time prompting that externalizes emotional reasoning into interpretable stages: (1) abstraction, (2) factor identification, (3) emotion enumeration, (4) Likert assessment, and (5) aligned response generation.

- **Outcome Uncertainty (Primary Appraisal)**
High, as the patient is unsure about treatment success.

By limiting to three factors, we ensure focus on the most salient dimensions while maintaining adaptability across situations. Open-vocabulary induction avoids dependence on any fixed ontology, allowing factors to vary naturally (e.g., "financial strain" for one patient, "parental responsibility" for another) while remaining recognizable to clinicians familiar with appraisal-based reasoning (Amershi et al., 2019).

3.2. Emotion Modeling (Stages 3–4)

Stage 3, Candidate Emotion Extraction. With situation and factors established, RECAP generates 5–7 plausible candidate emotions rather than committing to a single label, reflecting that real-world interactions evoke mixed affect (Hepburn and Potter, 2023). The candidate set reveals the model's emotion ontology. Systematic patterns such as consistently generating "anxiety" for medical contexts regardless of patient agency indicate stereotyped associations rather than context-sensitive reasoning. For the running example, candidates include anxiety, fear, overwhelm, sadness, hope, guilt, and helplessness.

Stage 4, Likert-Based Calibration. Each candidate is rated on a five-point scale (very-unlikely to very-likely, mapped to probabilities 0.05–0.95), capturing partial belief and mixed affect rather than forcing exclusivity. This addresses the failure mode where LLMs guess rather than express uncertainty (Kalai et al., 2025; Kadavath et al., 2022). The resulting profile is a compact, clinician-reviewable vector. Well-calibrated models should assign high ratings to human-endorsed emotions and mid-range ratings for ambiguous situations. Miscalibration patterns reveal reasoning deficits that benchmark accuracy alone cannot detect.

3.3. Response Generation (Stage 5)

Guided by the emotion-likelihood profile, the model acknowledges the highest-probability feelings, validates them, and adjusts tone before offering guidance, following clinical frameworks that prioritize emotions before information (Francis and Robertson, 2023). Conditioning on explicit profiles also enables counterfactual probing, allowing us to test whether varying Likert input predictably changes response tone. This causal testing of emotion-to-language mappings is impossible with end-to-end generation. The modular design additionally supports clinical override, because emotional guidance can be adjusted based on external input.

"I'm really sorry you're feeling so overwhelmed and anxious about your diagnosis and treatment. That's completely understandable—facing serious illness can be frightening and stressful. You're not alone in this..."

4. Experiments

4.1. Datasets

We evaluate RECAP on benchmarks spanning single-turn reasoning and multi-turn dialogue.

Single-Turn Datasets

- **EmoBench** (Sabour et al., 2024): Complex scenarios requiring inference of latent affect and causal emotional reasoning. Unlike simple emotion classification, EmoBench emphasizes reasoning about multi-party emotional dynamics and implicit cues.
- **SECEU** (Wang et al., 2023): Emotion intensity distribution task where models allocate relative intensity across candidate emotions. Reports EQ scores (higher is better) and raw distance from human consensus (lower is better).

Multi-Turn Datasets

- **EQ-Bench3** (Paech, 2023): Dialogue-centric emotional inference with Elo-style ratings, testing emotional reasoning in conversational contexts with social dynamics.
- **HealthBench** (Arora et al., 2025): 5,000 multi-turn healthcare conversations scored against physician-written rubrics spanning accuracy, completeness, context awareness, communication quality, and instruction following.

4.2. Baselines

We compare RECAP against three prompting strategies.

1. **Zero-shot:** Task description and input only, without emotion-specific guidance. This tests the model’s inherent emotional understanding acquired during pretraining and instruction tuning.
2. **Emotion Priming:** Augments the zero-shot prompt with explicit instructions to analyze emotional states and cues, for example “Deeply analyze the provided story, focusing on the main character’s situation, actions, and any explicit or implicit emotional cues.” This tests whether directing attention to emotional analysis improves performance without structured reasoning.
3. **Supportive Clinician Persona:** Role-based conditioning as “an empathetic, supportive clinician” tasked with evaluating the emotional state of the person in the scenario. Unlike RECAP, this relies on implicit conditioning rather than explicit emotional inference.

We exclude fine-tuned models (e.g., SoulChat (Chen et al., 2023)) to isolate RECAP’s inference-time contribution under zero-training constraints.

4.3. Models and Configuration

Experiments use Llama-3.1-8B-Instruct (Grattafiori et al., 2024), OpenAI-OSS-20B, and OpenAI-OSS-120B (OpenAI, 2025). We focus on open-source models for three reasons. First, cost efficiency, as RECAP’s multi-step reasoning makes API costs prohibitive. Second, reproducibility. Third, competitive performance, since OSS-120B achieves near-parity with o4-mini on reasoning benchmarks. All experiments use temperature 0.6, max tokens 2048, with three trials per condition.

4.4. Clinical Evaluation Setup

We conducted blind evaluations with two board-certified oncology research fellows, comparing RECAP against all baselines for emotional support quality.

(a) Single-Turn Scenario

Dx De novo HER2+ metastatic breast cancer with gastric/cecal involvement

Tx Docetaxel + trastuzumab + pertuzumab q3w, limit docetaxel to 4–6 cycles, then maintenance

Pt 60-y/o grandmother, lost 30 lb in 3 mo. Fears losing her identity to her grandchildren.

Q “Everyone keeps saying this plan is palliative. How do I balance hope with honesty so my family doesn’t think I’m giving up?”

(b) Multi-Turn Scenario

Dx Stage IIA classical Hodgkin lymphoma

Tx ABVD chemotherapy q2w for 2–4 cycles with response-adapted PET imaging

Pt 20-y/o first-gen college student. Believes “I cannot fail. My parents sacrificed everything.”

Goals Refuse semester off, seek strategies to study through cognitive effects

Opening “I finally made the Dean’s List. I can’t take a semester off. My parents worked too hard for me to be here.”

Figure 3: Representative synthetic patient scenarios. (a) Single-turn evaluation assesses individual response quality. (b) Multi-turn evaluation tracks conversational dynamics across 3 turns.

Data Curation. We synthesized evaluation scenarios based on authentic clinical interactions. Source material consisted of anonymized patient charts and questions from chemotherapy consultations, compiled by oncology fellows through medical records review, chemotherapy educational documentation, and safety resources. Using this corpus, GPT-5 generated structurally similar scenarios preserving clinical complexity, and all underwent clinician review and approval.

Figure 3 shows representative examples. For single-turn evaluation, we constructed 25 scenarios comprising a diagnosis, a treatment plan, a clinical narrative, and a patient question. For multi-turn evaluation, 10 scenarios included synthetic patient personas with demographics, personality traits, and hidden conversational goals guiding realistic behavior across three turns. GPT-5 simulated patient responses, with persona and goals enabling dynamic adaptation. Both datasets were stratified across cancer types and emotional states. Our design aligns with work on LLM-based simulated patients (Wang et al., 2024; Lee et al., 2025; Bodonhelyi et al., 2025), embedding persona-level beliefs that support coherent emotional trajectories.

Annotation Process. Clinical annotators had no prior knowledge of RECAP’s design. Responses were presented in randomized, blinded order unique to each annotator. Each response was rated independently on Likert scales (1–5) with optional qualitative feedback. We conducted two evaluation tasks.

(a) Aggregate emotional-intelligence benchmarks

Method	EmoBench (Acc.↑)			SECEU (EQ↑)			EQ-Bench3 (Elo↑)		
	L3-8B	O-20B	O-120B	L3-8B	O-20B	O-120B	L3-8B	O-20B	O-120B
Zero-Shot	.35±.1	.61±.1	.78±.1	58.3±3.9	96.4±4.1	110.5±2.2	1032±1	1281±2	1474±1
Emotion Priming	.42±.2	.65±.1	.83±.1	61.7±3.9	97.4±3.5	111.1±2.2	1033±2	1284±3	1475±2
Persona	.39±.1	.64±.1	.81±.1	58.9±5.8	100.4±2.2	109.8±2.5	1032±1	1283±3	1476±2
RECAP	.45±.2	.65±.2	.83±.1	71.3±4.1	109.2±3.6	115.2±3.8	1101±3	1308±3	1484±3

(b) SECEU breakdown: EQ score vs. raw distance from consensus

Method	L3-8B		O-20B		O-120B	
	EQ↑	Dist.↓	EQ↑	Dist.↓	EQ↑	Dist.↓
Zero-Shot	58.3±3.9	5.1±0.7	96.4±4.1	3.1±0.3	110.5±2.2	2.2±0.3
Emotion Priming	61.7±3.9	4.8±0.5	97.4±3.5	2.7±0.4	111.1±2.2	2.4±0.3
Persona	58.9±5.8	4.8±0.7	100.4±2.2	2.8±0.6	109.8±2.5	2.4±0.4
RECAP	71.3±4.1	4.1±0.8	109.2±3.6	2.5±0.5	115.2±3.8	2.4±0.6

Table 1: Emotional intelligence results. Models: L3-8B = Llama-3.1-8B-Instruct, O-20B = GPT-OSS-20B, O-120B = GPT-OSS-120B. (a) Aggregate scores on EmoBench, SECEU, and EQ-Bench3. (b) SECEU decomposition showing EQ score and distance from human consensus (lower = better). **Bold** = best; underline = second best. All runs: temp=0.6, max_tokens=2048, n=3 trials. Values are mean±sem.

- **Single-turn:** 100 responses across 25 scenarios, four methods each, five evaluation dimensions (Completeness, Emotional Recognition, Supportiveness, Trustworthiness, Tone Appropriateness).
- **Multi-turn:** 20 conversations (10 scenarios × 2 methods), three turns each, four dimensions (Personalization, Context Awareness, Tone, Coherence). Compared RECAP against Zero-Shot only, as single-turn findings showed all three baselines performed equivalently (2.93–2.95).

Both annotators independently rated all 100 single-turn responses. Inter-annotator agreement yielded 78.2% within-one-point agreement across five dimensions (quadratic weighted $\kappa=0.09$, Spearman $\rho=0.20$, $p < .001$). The modest κ reflects differing annotator calibration (mean ratings of 3.50 vs. 2.61) rather than disagreement on relative ordering. We report averaged scores throughout.

5. Results

5.1. Emotion Understanding Benchmarks

Table 1 presents results across three emotional intelligence benchmarks and three model scales. RECAP demonstrates consistent improvements, particularly pronounced on tasks that require a nuanced understanding of emotional states.

EmoBench. For Llama-3.1-8B-Instruct, RECAP achieves 0.45 accuracy compared to 0.35 for zero-shot (28% relative improvement). Emotion Priming

yields intermediate performance (0.423), indicating that structured reasoning offers additional benefits beyond emotional activation. At OSS-120B, RECAP and Emotion Priming both reach 0.825 versus the zero-shot baseline (0.783). The Persona baseline shows mixed results, occasionally underperforming zero-shot, suggesting role-based prompting without structured reasoning may introduce biases.

SECEU. RECAP achieves EQ scores of 71.29, 109.17, and 115.19 across model scales, consistently outperforming all baselines (22% improvement over zero-shot at 8B). RECAP also achieves the lowest distance from human consensus (4.08 vs 5.08, a 20% improvement), indicating that explicit Likert calibration produces emotion estimates more aligned with human judgments. At OSS-120B scale, zero-shot achieves lower raw distance (2.22) while RECAP still produces the highest EQ score (115.19). This reflects the EQ scoring mechanism, which rewards not just accuracy but appropriate confidence calibration, suggesting RECAP helps models express uncertainty more appropriately.

EQ-Bench3. RECAP demonstrates substantial improvements across scales, including a 70-point Elo gain for Llama-3.1-8B (1101.03 vs 1031.92), 28 points for OSS-20B (1308.24 vs 1280.57), and 10 points for OSS-120B (1483.7 vs 1473.7). These consistent gains suggest structured emotional reasoning provides value even for highly capable models, though diminishing returns at scale indicate larger models have stronger implicit emotional reasoning.

Method	L3-8B	O-20B	O-120B
Zero-Shot	0.18 \pm 0.01	0.37 \pm 0.02	0.46 \pm 0.01
Emotion Priming	0.18 \pm 0.01	0.38 \pm 0.03	0.49 \pm 0.03
Persona	0.19 \pm 0.02	0.40 \pm 0.04	0.52 \pm 0.04
RECAP	0.19\pm0.02	0.41\pm0.04	0.54\pm0.05

Table 2: HealthBench overall scores across model scales. L3-8B = Llama-3.1-8B-Instruct, O-20B = GPT-OSS-20B, O-120B = GPT-OSS-120B. Per-dimension breakdown in Table 7 (Appendix). Values are mean \pm sem; $n=3$ trials.

5.2. Medical Communication Quality

Table 2 presents HealthBench overall scores (per-dimension breakdown in Table 7, Appendix). RECAP improves overall quality at every scale, with gains of 8% for Llama-3.1-8B, 11% for OSS-20B, and 18% for OSS-120B. The largest gains occur in Completeness (23% for OSS-20B) and Communication Quality (27% for OSS-20B), suggesting emotional reasoning helps models address both medical and emotional dimensions of patient concerns.

However, RECAP exhibits systematic trade-offs. Accuracy and Instruction Following decrease across all scales, reflecting prioritization of comprehensive, emotionally tuned responses over literal precision. Increased standard deviations (e.g., Llama-3.1-8B from 0.0067 to 0.0235) indicate more variable, scenario-specific responses rather than formulaic outputs. These accuracy differences do not reflect knowledge limitations, as RECAP uses the same underlying models, but rather a shift toward emotional appropriateness in response generation. This surfaces a style-versus-precision tension common to empathy-forward systems, mitigable through verification techniques such as self-refinement.

5.3. Clinical Evaluation Results

Single-Turn Evaluation RECAP achieved an overall score of 3.41 (\pm 0.32), significantly outperforming Zero-Shot (2.95, $p < .001$, $d=0.98$), Persona (2.94, $p < .001$, $d=1.02$), and Emotion Priming (2.93, $p < .001$, $d=1.08$). As shown in Figure 4a, RECAP consistently outperforms all baselines across evaluation dimensions.

The largest improvements emerged in empathy-related dimensions, specifically Supportiveness (+0.67, $d=1.37$) and Emotional Recognition (+0.59, $d=1.34$), both representing large effect sizes. Factual Completeness showed more modest improvement (+0.29, $d=0.85$), suggesting RECAP’s primary advantage lies in emotional rather than informational quality. All conditions scored lowest on Emotional Recognition (range 2.54–3.18), indicating this

dimension remains challenging for LLMs.

Scenario-level win rates (Figure 6, Appendix) show RECAP wins 88% against Persona, 84% against Emotion Priming, and 76% against Zero-Shot. RECAP achieved the highest rating across all three baselines in 17 of 25 scenarios (68%), with 46% of responses rated 4–5 compared to 22–29% for baselines.

Multi-Turn Evaluation RECAP maintained its advantage with an overall score of 3.77 (\pm 0.29) versus Zero-Shot’s 3.71 (\pm 0.29). Figure 4b shows RECAP outperforms Zero-Shot across all four dimensions, with the strongest gains on Personalization (3.95 vs 3.80, $d=0.70$) and greater consistency (SD=0.16 vs 0.26). The narrower gap compared to single-turn settings indicates that extended context provides additional grounding, though RECAP continues yielding benefits in adapting to individual patient needs. Scenario-level comparisons showed RECAP winning 5 of 10 scenarios (50%), with 1 tie and 4 Zero-Shot wins, and achieving slightly higher-quality ratings (76% rated 4–5 vs 72% for Zero-Shot).

Qualitative Analysis VADER sentiment analysis (Hutto and Gilbert, 2014) of annotator comments showed RECAP received significantly more positive sentiment, with 68% positive versus 28–56% for baselines and mean compound scores of +0.142 versus -0.259 to -0.006 (Cohen’s $d=0.56$ vs. pooled baselines). RECAP had zero scenarios rated below 2.5, compared to 1–3 for baselines, and achieved high ratings (≥ 3.5) in 44% of scenarios versus 8–12% for baselines (Figure 7, Appendix). Annotators characterized RECAP responses as “emotionally cognizant, supportive and complete,” while criticizing baselines as “robotic” and “medically correct but emotionally inadequate.”

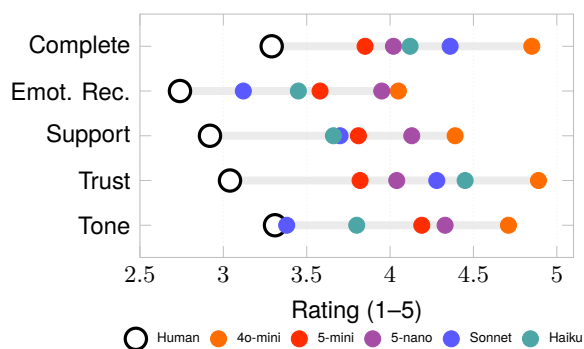


Figure 5: LLM-as-Judge vs. human ratings. Scores represent means across all 200 single-turn responses (25 scenarios \times 4 methods). Scenario quality distribution in Figure 7.

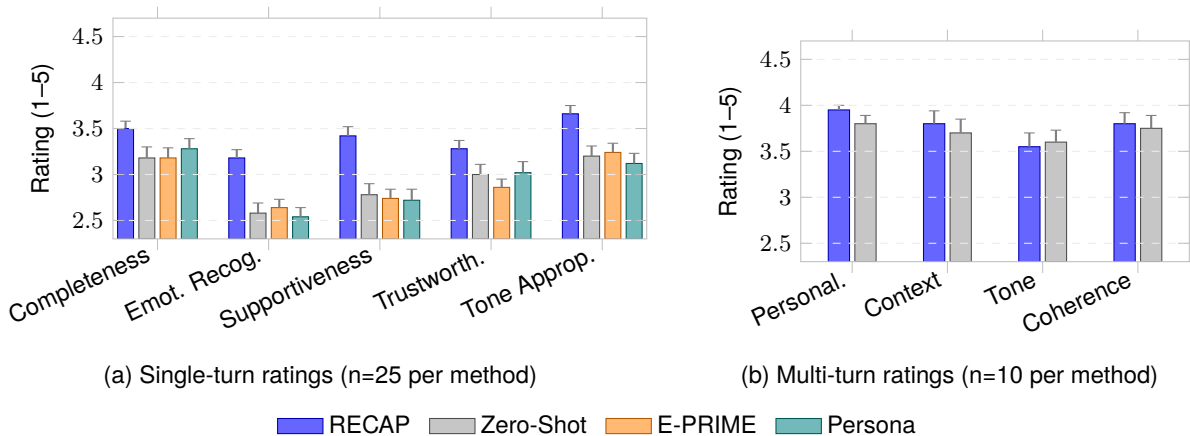


Figure 4: Human evaluation results. (a,b) Mean ratings with standard error bars (1–5 scale). Scenario-level win rates are reported in Figure 6 (Appendix).

LLM-as-a-Judge Evaluation Five frontier LLMs (GPT-4o-mini, GPT-5-mini, GPT-5-nano, Claude Sonnet 4.5, Claude Haiku 4.5) rated all 100 single-turn responses. As shown in Figure 5, all models showed weak-to-moderate correlation with human judgments ($r=0.09$ – 0.26), with Emotional Recognition strongest ($r=0.27$ – 0.44) while Completeness and Trustworthiness showed near-zero or negative correlations ($r = -0.11$ to 0.05).

All models exhibited systematic positive bias, with GPT-4o-mini inflating most (+1.3–1.9 points) and GPT-5-mini showing the smallest bias (+0.3–0.9 points). Notably, lower bias did not correlate with better alignment: Claude Sonnet 4.5 showed moderate bias but the lowest correlation ($r=0.09$), suggesting its ratings captured a different signal than human judgments rather than simply being better calibrated.

5.4. Cross-Benchmark Analysis

Several consistent patterns emerge across our evaluation suite.

First, RECAP’s relative advantage is inversely proportional to model scale, with the 8B model showing average improvements of 35% across emotional intelligence benchmarks versus 8% for the 120B model. This suggests that explicit emotional reasoning is particularly valuable for smaller models, enabling them to achieve performance comparable to much larger models without structured reasoning.

Second, the nature of improvements differs by task type. On tasks requiring emotional recognition and intensity assessment (SECEU, EQ-Bench3), RECAP provides consistent gains across all scales, while on tasks requiring both emotional and factual reasoning (HealthBench), RECAP shows strong communication quality improvements alongside

modest accuracy trade-offs. The convergence between clinical evaluation and automated benchmarks validates our approach. A blind evaluation by oncology experts showed win rates of 76–88% against prompting baselines with large effect sizes ($d=0.98$ – 1.08), and VADER sentiment analysis against captured qualitative natural language reviews confirmed that RECAP elicited significantly more positive annotator feedback (68% vs 28–56%).

6. Conclusion

We introduced RECAP, a modular framework that externalizes emotional reasoning in medical dialogue into five inspectable stages. By decomposing reasoning into situation abstraction, factor induction, and Likert-based calibration, RECAP transforms opaque pattern-matching into a transparent pipeline.

Our experiments yield three insights. Explicit calibration reduces distance from human consensus by 20%, demonstrating that externalizing uncertainty improves emotion estimates. Structured reasoning disproportionately benefits smaller models (28% gain for 8B vs. 5% for 120B), suggesting decomposition compensates for capabilities larger models have internalized. Factor induction reveals systematic biases, with models consistently underweighting relational factors like social support.

Blind evaluation by oncology fellows confirmed these findings, with 76–88% win rates over baselines and large effect sizes. The framework’s intermediate artifacts provide clinicians with auditable checkpoints necessary for trust in high-stakes settings. RECAP demonstrates that principled decomposition can make AI systems both more empathetic and more transparent.

7. Limitations

We distinguish between system-centric constraints of RECAP and problem-and-use-case limits that arise in practice.

7.1. System-Centric Limitations

Computational overhead. The five-stage pipeline requires multiple LLM calls and accumulated context, increasing token usage by roughly 3–5× over single-pass prompting. This is workable for asynchronous consultations but impractical for low-latency settings.

Error propagation. Modularity introduces cascading failure. Early mistakes in situation abstraction or appraisal factors skew later stages, producing confident but distorted emotion profiles.

Emotion discretization. RECAP uses discrete candidate emotions and Likert ratings. Mixed, ambivalent, or culturally specific affect (e.g., shame, obligation) may not map cleanly to this structure.

No longitudinal memory. Each turn is treated independently. Clinical communication often relies on emotional trajectories over time, and without history, RECAP may repeat or over-normalize. This is reflected in our multi-turn results (Figure 4), where RECAP’s advantage over baselines narrowed compared to single-turn settings.

Human Evaluation Scale. While our evaluation utilized board-certified oncology research fellows to ensure high-domain expertise, the total volume of annotated data, comprising 100 single-turn responses and 20 multi-turn conversations, remains relatively small. This scale may limit the statistical power required to capture rare model failure modes or subtle variations in long-term conversational trajectories. Consequently, these findings should be interpreted as a focused expert pilot study. Future work should aim to scale these evaluations across a broader range of clinical specialties and a larger cohort of patient-simulated interactions to confirm the generalizability of the RECAP framework.

7.2. Problem and Use-Case Limitations

High-acuity scenarios. RECAP is inappropriate for contexts where empathy is secondary to directive action (e.g., triage, emergent symptoms, self-harm risk). Emotional calibration can dilute urgency, and guideline-driven responses are preferable in such cases.

Masked or indirect emotional signals. The pipeline assumes good-faith disclosure. Patients often downplay or euphemize distress. When surface cues are weak, RECAP may underestimate risk and respond too gently.

Cultural fit. Appraisal dimensions such as control, agency, and uncertainty reflect Western clinical norms. They may not generalize to cultural contexts prioritizing duty, stigma, or relational obligations, leading to mismatched support.

Modal gaps. RECAP is text-only. Distress expressed through tone, pauses, or hesitations is lost, limiting accuracy in real consultations.

Evaluation sensitivity. Benchmarks such as HealthBench prioritize actionability. Responses that validate emotions or acknowledge uncertainty may be labeled as incomplete, even when they are clinically appropriate. Complementary human-centered metrics better capture communicative quality.

Despite these constraints, RECAP provides an interpretable path toward emotional alignment. Its modular structure allows individual stages to be improved or replaced without retraining the underlying model.

8. Ethical Considerations

This work uses synthetic patient scenarios generated by GPT-5 based on anonymized clinical materials, and no real patient data was used in evaluation. Clinical annotators participated voluntarily with informed consent. RECAP is intended as a research contribution to understand emotional reasoning in LLMs, not as a deployable clinical tool. The framework has not undergone validation required for patient-facing applications, and we caution against deployment without extensive safety testing, regulatory review, and clinician oversight. We release no patient data or clinical records. All evaluation materials were reviewed and approved by collaborating oncology professionals.

9. Bibliographical References

Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Daniel Weld, Mihaela Vorvoreanu, and Eric Horvitz. 2019. [Guidelines for human-ai interaction](#). *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2019:1–13.

Jonas Armbruster, Florian Bussmann, Catharina Rothhaas, Nadine Titze, Paul A. Grützner, and Holger Freischmidt. 2024. ["doctor chatgpt, can you help me?" the patient's perspective: Cross-sectional study](#). *Journal of Medical Internet Research*, 26:e58831.

- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health.
- John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zidian Zhu, Jon-Patrick Allem Kelley, Zoe Chu, and David M. Smith. 2023. [Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum](#). *JAMA Internal Medicine*, 183(6):589–596.
- Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Lurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. [A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy](#). *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20:Paper 589, 1–14.
- Anna Bodonhelyi, Christian Stegemann-Philipps, Alessandra Sonanini, Lea Herschbach, Marton Szep, Anne Herrmann-Werner, Teresa Festl-Wietek, Enkelejda Kasneci, and Friederike Holderried. 2025. [Modeling challenging patient interactions: Lms for medical communication training](#).
- Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. [To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making](#). *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21.
- Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. ["hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making](#). *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):Article 104, 1–24.
- David Chen, Kabir Chauhan, Rod Parsa, Zhihui Amy Liu, Fei-Fei Liu, Ernie Mak, Lawson Eng, Breffni Louise Hannon, Jennifer Croke, Andrew Hope, Nazanin Fallah-Rad, Phillip Wong, Srinivas Raman, et al. 2025. [Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer](#). *npj Digital Medicine*, 8.
- David Chen, Rod Parsa, Andrew Hope, Breffni Hannon, Ernie Mak, Lawson Eng, Fei-Fei Liu, Nazanin Fallah-Rad, Ann M. Heesters, and Srinivas Raman. 2024. [Physician and artificial intelligence chatbot responses to cancer questions from social media](#). *JAMA Oncology*, 10(7):956–960.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. [Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations](#).
- Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. 2024. [An appraisal-based chain-of-emotion architecture for affective language model game agents](#). *PLOS ONE*, 19(5):e0301033.
- Jacob Dineen, Don Kridel, Daniel Dolk, and David Castillo. 2024. [Unified explanations in machine learning models: A perturbation approach](#).
- Jacob Dineen, Aswin RRV, Qin Liu, Zhikun Xu, Xiao Ye, Ming Shen, Zhaonan Li, Shijie Lu, Chitta Baral, Muhao Chen, et al. 2025. [Qa-align: Aligning llms through constitutionally decomposed qa](#).
- Daniel Dolk, Donald Kridel, Jacob Dineen, and David Castillo. 2020. [Model interpretation and explainability towards creating transparency in prediction models](#). In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, HICSS, pages 956–965, Maui, HI. Hawaii International Conference on System Sciences.
- Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2024. [Bird: A trustworthy bayesian inference framework for large language models](#).
- Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2023. [Generic temporal reasoning with differential analysis and explanation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12013–12029, Toronto, Canada. Association for Computational Linguistics.
- Laura Francis and Noelle Robertson. 2023. [Health-care practitioners' experiences of breaking bad news: A critical interpretative meta synthesis](#). *Patient Education and Counseling*, 107:107574.
- Ludwig Franke Føyen, Emma Zapel, Mats Lekander, Erik Hedman-Lagerlöf, and Elin Lindsäter. 2025. [Artificial intelligence vs. human expert: Licensed mental health clinicians' blinded evaluation of ai-generated and expert psychological advice on quality, empathy, and perceived authorship](#). *Internet Interventions*, 41:100841.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The llama 3 herd of models](#).

- Alexa Hepburn and Jonathan Potter. 2023. [Understanding mixed emotions in organized helping through emotionography](#). *Frontiers in Psychology*, 14:1236148.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in llm simulations](#).
- Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. [Apathetic or empathetic? evaluating llms' emotional alignments with humans](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 97053–97087, Vancouver, Canada. Curran Associates, Inc.
- Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*, ICWSM '14, pages 216–225, Ann Arbor, MI. AAAI Press.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Nicholas Schiefer, Eric Clark, Guy Amir, Kamal Ndousse, Tom B. Brown, Steven Larson, Roger Grosse, Jared Kaplan, Natasha McAleese, David Hernandez, Micah Carroll, Deep Ganguli, Jan Leike, Catherine Olsson, David Krueger, Evan Hubinger, Collin Burns, Samuel Bowman, Jacob Hilton, Long Ouyang, Yuntao Bai, Amanda Chen, Neil Elhage, Nick Johnston, Tamera Khan, Liane Lovitt, Isaac Ndousse, Abeba Tesafaye, Emily Thomason, Alex Wang, Andrew Zhang, Barret Zoph, Ilya Sutskever, Paul Christiano, David Kaplan, Oskar Clifford, Nate Ringer, and Yuntao Miao. 2022. [Language models \(mostly\) know what they know](#).
- Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. [Why language models hallucinate](#).
- Richard S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York, NY.
- Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung Suh, Eunmee Kim, and Hajin Lim. 2025. [Adaptive-vp: A framework for llm-based virtual patients that adapts to trainees' dialogue to facilitate nurse communication training](#).
- Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2024. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *Korean Journal of Cognitive Science*, 35(1):23–48. Also available as arXiv:2311.04915.
- Man Luo, Christopher J. Warren, Lu Cheng, Haidar M. Abdul-Muhsin, and Imon Banerjee. 2024. [Assessing empathy in large language models with real-world physician-patient interactions](#). In *IEEE International Conference on Big Data (BigData 2024)*, pages 6510–6519, Washington, DC, USA. IEEE.
- Pedro Henrique Luz de Araujo and Benjamin Roth. 2025. [Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior](#). *PLOS ONE*, 20(6):e0325664.
- OpenAI. 2025. [gpt-oss-120b and gpt-oss-20b model card](#).
- Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, United Kingdom.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pages 27730–27744, New Orleans, LA, USA. Curran Associates, Inc.
- Samuel J. Paech. 2023. [Eq-bench: An emotional intelligence benchmark for large language models](#).
- Rafael Rafailov, Pratyusha Ramesh, Avnish Narayan Das, Sishir Syed, Sam Basu, Yao Li, James Zou, Qi Yang, Yuntao Bai, Pieter Abbeel, and Kiril Rafailov. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, pages 53416–53432, New Orleans, LA, USA. Curran Associates, Inc.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.
- Aswin RRV, Jacob Dineen, Divij Handa, Md Nayem Uddin, Mihir Parmar, Chitta Baral, and Ben Zhou. 2025. [Thinktuning: Instilling cognitive reflections without distillation](#).

- Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. [EmoBench: Evaluating the emotional intelligence of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.
- Klaus R. Scherer and Agnes Moors. 2019. [The emotion process: Event appraisal and component differentiation](#). *Annual Review of Psychology*, 70:719–745.
- Ming Shen, Zhikun Xu, Xiao Ye, Jacob Dineen, and Ben Zhou. 2025. [Bow: Bottlenecked next word exploration](#).
- Vlad Sorin, Benjamin Sheffer, Nada Meirou, et al. 2024. [Large language models and empathy: Systematic review](#). *Journal of Medical Internet Research*, 26:e55610.
- John Sweller. 2023. [The development of cognitive load theory: Replication crises and incorporation of other theories can lead to theory expansion](#). *Educational Psychology Review*, 35:95.
- Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. [Entailer: Answering questions with faithful and truthful chains of reasoning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. 2024. [Patient- \$\psi\$: Using large language models to simulate patients for training mental health professionals](#).
- Xuena Wang et al. 2023. [Emotional intelligence of large language models. Includes SECEU scenario-based assessments](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#).
- Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. [Break it down: A question understanding benchmark](#). *Transactions of the Association for Computational Linguistics*, 8:183–198.
- Zhikun Xu, Ming Shen, Jacob Dineen, Zhaonan Li, Xiao Ye, Shijie Lu, Aswin RRV, Chitta Baral, and Ben Zhou. 2024. [Tow: Thoughts of words improve reasoning in large language models](#).
- Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. [Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes](#). Published as proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), DOI:10.1145/3290605.3300468.
- Xiao Ye, Jacob Dineen, Zhaonan Li, Zhikun Xu, Weiyu Chen, Shijie Lu, Yuxi Huang, Ming Shen, Phu Tran, Ji-Eun Irene Yum, Muhammad Ali Khan, Muhammad Umar Afzal, Irbaz Bin Riaz, and Ben Zhou. 2025a. [Evaluating medical llms by levels of autonomy: A survey moving from benchmarks to applications](#).
- Xiao Ye, Shaswat Shrivastava, Zhaonan Li, Jacob Dineen, Shijie Lu, Avneet Ahuja, Ming Shen, Zhikun Xu, and Ben Zhou. 2025b. [Cc-learn: Cohort-based consistency learning](#).
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.
- Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. [When “a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.
- Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. [Learning to decompose: Hypothetical question decomposition based on comparable texts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2223–2235, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and Dong Yu. 2024. [Conceptual and unbiased reasoning in language models](#).

A. RECAP Technical Details

A.1. Likert-Scale Probability Mapping

In Stage 5 of the RECAP pipeline, qualitative Likert ratings are converted to numerical probabilities for downstream processing. Table 3 shows the mapping used throughout our experiments.

Likert value	Probability
very-unlikely	0.05
unlikely	0.25
neutral	0.50
likely	0.75
very-likely	0.95

Table 3: Likert-scale to probability mapping.

A.2. Evaluation Data Samples

Table 4 provides representative examples from our evaluation datasets, illustrating the range of emotional scenarios and clinical contexts used to assess RECAP’s performance.

B. Prompt Templates

This section documents all prompts used in our experiments. Section B.1 presents the five-stage RECAP pipeline prompts, while subsequent sections detail the baseline prompts for each benchmark.

B.1. RECAP Pipeline Prompts

The RECAP pipeline decomposes emotional reasoning into five sequential stages. Each stage receives output from the previous stage and produces structured intermediate artifacts.

Stage 1: Situation Abstraction

System: You are an expert at generating abstracts from user inputs.

Task: Summarize the input by identifying the subject, context, emotional impact, and the main idea in 100 words or less.

Input: "{situation}"

Output format:

- Return ONLY a single JSON object with this exact schema:

```
{"abstract": "<concise summary, 1-3 sentences, <=100 words>"}
```

- Do not include any additional text outside the JSON.

Stage 2: Latent Factor Induction

System: You are an expert at analyzing abstracts and identifying psychological/social factors that would influence the person’s emotions.

Task: Analyze this abstract and identify 3 important factors that would influence the person’s emotions:

ABSTRACT: {abstract}

For each factor, provide:

- Factor name (short, descriptive)
- Two possible values (like high/low, present/absent, strong/weak, etc.)

Format:

```
1. Factor name: Description (value1/value2)
2. Factor name: Description (value1/value2)
3. Factor name: Description (value1/value2)
END_OF_FACTORS
```

(a) Benchmark examples	
Benchmark	Example
EmoBench (Single-Turn)	<p><i>“Sarah found out that her younger brother is being bullied at school but he begged her not to tell their parents.”</i></p> <p>Task: Multiple choice action selection:</p> <ul style="list-style-type: none"> • (a) Promise to keep the secret • (b) Inform parents anyway • (c) Confront the bullies herself • (d) Suggest talking to counselor
SECEU (Single-Turn)	<p><i>“The airplane model that Wang made fell from the sky one minute after take-off. When she inspected the model, she found a part that could possibly be improved.”</i></p> <p>Task: Rate emotion intensities:</p> <ul style="list-style-type: none"> • Expectation • Excited • Joyful • Frustrated
EQ-Bench3 (Multi-Turn)	<p>Context: Work Dilemma, Lunchroom Theft Scapegoat</p> <p>Task: This is a role-play, with you playing an emotionally intelligent human who is, essentially, yourself. Treat it like a real situation. Always respond in first person.</p> <p>Prompt 1: You think you might have been scapegoated by a fellow employee for the lunchroom thefts that have been happening. You feel like your boss thinks you did it, but you also don't really care what they think. Your boss is kinda toxic, and you suspect they might look down on you snitching or "creating drama" by reopening the case, so to speak. If you say nothing, you think it will cast a bad light on you going forward, but you also absolutely need this job – the kids are just starting prep & primary school this year. Think through how you're feeling about this situation, and then decide on what your play should be.</p> <p>Prompt 2: Over the next few months, the thefts continue. Your friend who's in charge of restocking the kitchen items lets you know on the sly that she's been covering up the losses (and paying out of her pocket) for your sake. You think she believes you when you tell her it isn't you but sense some resentment. What are you thinking & feeling? What's the play?</p> <p>Prompt 3: Things come to a head and the boss accuses you outright. They say that if you own up, you can pay out the loss and be done with it. It's not a lot of money. They seem pretty irate that it's come to this. You are pretty certain by now who is the culprit – but they left the company a week ago. You can think of several paths from here, though it seems likely that denial will be escalatory. What do you think & feel? What actions do you take?</p>
HealthBench (Mutli-Turn)	<p><i>“Make a 3 months plan to address my mild postpartum depression. I'm 6 weeks postpartum, have no harmful thoughts, and I have good support. I'm open to starting talk therapy.”</i></p> <p>Task: Rubric-scored criteria:</p> <ul style="list-style-type: none"> • Finding mental health provider • Emergency contact guidance • Week-by-week breakdown • Therapy explanation • Self-care strategies

Table 4: Representative examples from the four benchmarks used in our evaluation.

Stage 3: Factor Value Selection

Task: Analyze this situation and determine the specific value for each psychological factor:

SITUATION: {situation}

PSYCHOLOGICAL FACTORS: {factors_text}

For each factor, choose the most appropriate value based on what you can observe in the situation. Provide a brief explanation.

Format:

```
factor_name: chosen_value - brief explanation  
END_OF_ANALYSIS
```

Stage 4: Candidate Emotion Extraction

TASK: Extract 3-5 crucial emotions from this situation.

SITUATION: {situation}

Guidelines: Choose 3-5 distinct, important emotions, prioritizing the final emotional state if the narrative evolves.

Output: Output EXACTLY 3 to 5 lines. Each line must be ONE standalone lowercase emotion word. No headers, numbering, bullets, or explanations.

Stage 5a: Likert-Based Emotion Likelihood

Given the following situation and psychological factors, assess the likelihood of each emotion.

SITUATION: {situation}

IDENTIFIED FACTORS: {factors}

EMOTIONS TO ASSESS: {emotions_list}

For each emotion, provide a likelihood rating. Use ONLY: very-unlikely, unlikely, neutral, likely, very-likely

Format: emotion_name: rating (one line per emotion)

Stage 5b: Emotion-Aligned Response Generation

System: You are an empathetic, wise, and supportive AI assistant. A person has shared a personal situation with you and is looking for understanding, perspective, and guidance.

SITUATION SHARED: {situation}

EMOTIONAL INSIGHTS: Based on psychological analysis, this person is likely experiencing: {emotion_insights}
{context_info}

INSTRUCTIONS: Respond as a caring, insightful conversational partner. Your response should:

- Acknowledge their feelings with genuine empathy and validation
- Naturally weave in emotional insights (don't list them clinically)
- Provide thoughtful perspective and gentle guidance
- Be conversational and human-like, not robotic or structured
- Offer practical support where appropriate
- Be encouraging while remaining realistic

B.2. Baseline Prompts

Tables 5 and 6 summarize the prompt variations for each baseline method. All methods within a benchmark share identical base task instructions; only the additional priming or persona instructions differ.

Method	Additional Instructions (SECEU)
Zero-Shot	(None beyond base task)
Emotion Priming	Deeply analyze the provided story, focusing on the main character's situation, actions, and any explicit or implicit emotional cues. For each emotion option, critically assess its relevance and intensity.
Clinician Persona	You are an empathetic, supportive clinician. Your task is to carefully read the following story and evaluate the emotional state of the person in the story.
RECAP	(Uses full 5-stage pipeline output as context)

Table 5: SECEU baseline prompt variations. The base task requires reading a story and scoring four emotion options (0–10, summing to 10).

Method	Additional Instructions (EmoBench)
Zero-Shot	(None beyond base task)
Emotion Priming	Deeply analyze the provided story, focusing on the main character's situation, actions, and any explicit or implicit emotional cues. For each emotion option, critically assess its relevance and intensity.
Clinician Persona	You are an empathetic, supportive clinician. Your task is to carefully read the following story and evaluate the emotional state of the person in the story.
RECAP	(Uses full 5-stage pipeline output as context)

Table 6: EmoBench baseline prompt variations. The base task requires selecting one option and returning JSON with choice letter and text.

B.3. Clinician Evaluation Prompt

For human evaluation with clinical scenarios, both baseline and RECAP conditions used the same base system prompt to ensure fair comparison. RECAP responses additionally incorporated the emotional profile generated by the pipeline.

Base Prompt for Clinical Scenarios

Instructions: You are an AI Chatbot. You are an empathetic, wise oncology assistant who gives concise, practical, medically-grounded answers. Given the context, answer the question concisely (1-3 paragraphs).

B.4. LLM-as-a-Judge Prompt

To evaluate whether automated metrics could supplement human judgment, we prompted five frontier LLMs to rate responses on the same dimensions used by human annotators.

LLM-as-a-Judge Evaluation

Instructions: You are an expert clinical evaluator assessing the quality of a medical assistant's response to a cancer patient's emotional concern.

Patient Context: Diagnosis, Treatment Plan, Patient's Situation, Patient's Question

Evaluation Dimensions (1–5 scale each):

1. **Completeness:** Does the response adequately address the patient's question?
2. **Emotional Recognition:** Does the response acknowledge the patient's emotional state?
3. **Supportiveness:** Does the response provide emotional support and encouragement?
4. **Trustworthiness:** Does the response inspire confidence through accurate information?
5. **Tone Appropriateness:** Is the tone warm, empathetic, and appropriate?

Output: JSON object with dimension scores.

C. HealthBench Per-Dimension Breakdown

Table 7 presents the full per-dimension HealthBench results summarized in the main text (Table 2).

Model	Method	Complete.	Accuracy	Context	Comm. Qual.	Instr. Fol.	Overall
L3-8B	Zero-Shot	0.20 \pm 0.02	0.09\pm0.02	0.17 \pm 0.02	0.21 \pm 0.02	0.18\pm0.02	0.18 \pm 0.01
	Emotion Priming	0.20 \pm 0.02	0.08 \pm 0.02	0.17 \pm 0.02	0.21 \pm 0.03	0.17 \pm 0.02	0.18 \pm 0.01
	Persona	0.20 \pm 0.03	0.08 \pm 0.02	0.18\pm0.03	0.21 \pm 0.03	0.17 \pm 0.02	0.19\pm0.02
	RECAP	0.21\pm0.03	0.07 \pm 0.02	0.18\pm0.03	0.22\pm0.03	0.16 \pm 0.02	0.19\pm0.02
O-20B	Zero-Shot	0.40 \pm 0.04	0.12\pm0.02	0.23\pm0.02	0.42 \pm 0.03	0.41\pm0.03	0.37 \pm 0.02
	Emotion Priming	0.43 \pm 0.04	0.11 \pm 0.02	0.23\pm0.03	0.47 \pm 0.04	0.40 \pm 0.03	0.38 \pm 0.03
	Persona	0.46 \pm 0.04	0.11 \pm 0.02	0.22 \pm 0.03	0.51 \pm 0.04	0.39 \pm 0.03	0.40 \pm 0.04
	RECAP	0.49\pm0.04	0.10 \pm 0.02	0.22 \pm 0.03	0.53\pm0.04	0.38 \pm 0.03	0.41\pm0.04
O-120B	Zero-Shot	0.50 \pm 0.03	0.17\pm0.02	0.29\pm0.03	0.58 \pm 0.04	0.57\pm0.03	0.46 \pm 0.01
	Emotion Priming	0.52 \pm 0.03	0.15 \pm 0.02	0.28 \pm 0.03	0.61 \pm 0.04	0.56 \pm 0.03	0.49 \pm 0.03
	Persona	0.53 \pm 0.04	0.14 \pm 0.02	0.27 \pm 0.03	0.63 \pm 0.04	0.55 \pm 0.04	0.52 \pm 0.04
	RECAP	0.54\pm0.04	0.14 \pm 0.02	0.26 \pm 0.03	0.64\pm0.04	0.54 \pm 0.04	0.54\pm0.05

Table 7: Full HealthBench per-dimension results. Bold indicates best per model-metric. RECAP improves Completeness and Communication Quality while accepting modest trade-offs in Accuracy and Instruction Following. Values are mean \pm sem; $n=3$ trials.

D. Human Evaluation Details

D.1. Scenario-Level Win Rates

Figure 6 shows scenario-level win rates from the clinical evaluation, comparing RECAP against each baseline.

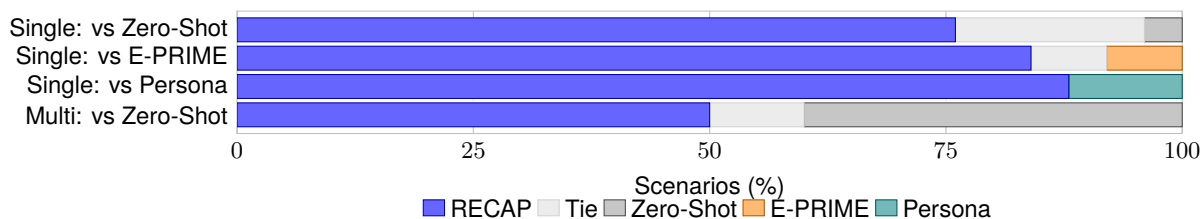


Figure 6: Scenario-level win rates showing percentage where each method achieved higher average rating across single-turn (top three) and multi-turn (bottom) comparisons.

D.2. Scenario Quality Distribution

Figure 7 shows the distribution of scenario-level human ratings across quality tiers.

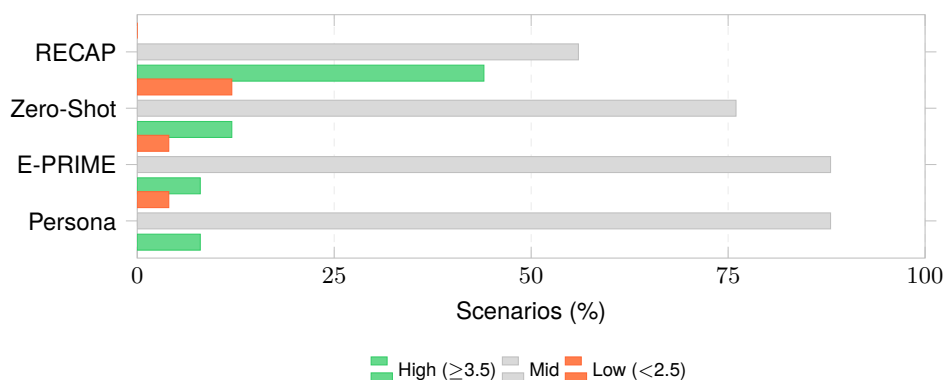


Figure 7: Scenario quality distribution of human ratings across all clinical scenarios. RECAP achieves the highest proportion of high-quality ratings (44%) with zero scenarios rated below 2.5.

D.3. Annotation Dimensions

Table 8 defines the evaluation dimensions used for single-turn and multi-turn annotations. Single-turn evaluation assessed five dimensions capturing response quality and emotional appropriateness. Multi-turn evaluation used four dimensions focused on conversational coherence and sustained emotional support.

Single-Turn <i>25 scenarios × 4 methods</i>	<i>n=100</i>	Multi-Turn <i>10 scenarios × 2 methods</i>	<i>n=20</i>
Completeness — Thoroughness in addressing concerns		Personalization — Adaptation to patient persona	
Emotional Recognition — Acknowledgment of emotional state		Context Awareness — Reference to prior turns	
Supportiveness — Appropriate reassurance and guidance		Tone — Emotional alignment throughout	
Trustworthiness — Medical accuracy and credibility		Coherence — Logical flow across turns	
Tone Appropriateness — Contextual suitability			
<i>Optional qualitative feedback</i>		<i>Compulsory qualitative feedback</i>	

Table 8: Rating dimensions for clinical annotation (1–5 Likert scale).

D.4. Annotation Interfaces

Annotators interacted with custom web interfaces designed for each evaluation type. The single-turn interface (Figure 8) presented patient context, questions, and model responses with Likert scales and optional free-text feedback. The multi-turn interface (Figure 9) displayed full conversation histories with turn-by-turn context and required qualitative comments.

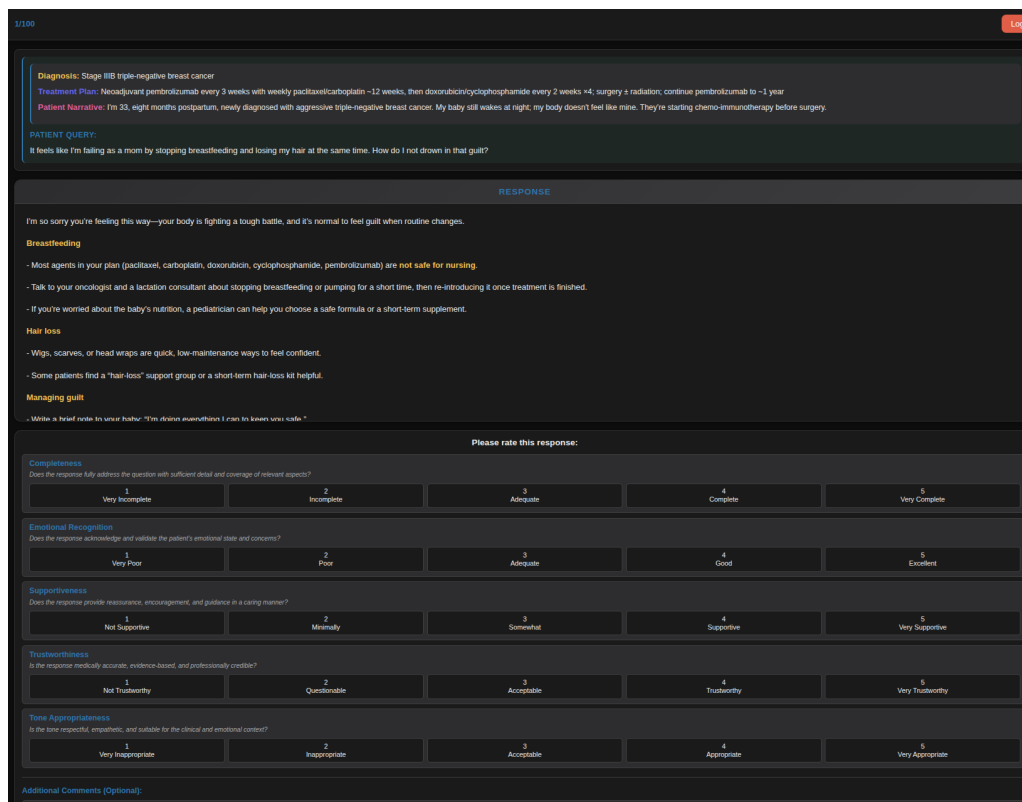


Figure 8: Single-turn annotation interface showing patient context, model response, and evaluation dimensions.

1/20 Logout

Patient Context: 22-year-old student
Diagnosis: Sickle cell disease with recurrent vaso-occlusive pain crises
Treatment Plan: Hydroxyurea, opioid pain plan

CONVERSATION

PATIENT
 I went to the ER last night in agony, and the triage nurse rolled her eyes like I was just there for drugs. I'm so tired of fighting to be believed.

ASSISTANT
 I'm really sorry you had to go through that. It's painful—both literally and emotionally—to feel dismissed when you're in crisis. Sickle-cell pain is a legitimate, life-threatening condition, and you deserve to be heard and treated promptly.

What you can do next:

- Document the episode**—write down what happened, the pain level, any meds you took, and how the staff responded. This record can help when you discuss it with your primary team or a patient advocate.
- Ask for a pain-management plan**—request a written plan that includes your opioid schedule, rescue doses, and non-opioid options (e.g., NSAIDs, heat, hydration). Having it in writing can reduce future misunderstandings.
- Speak to a trusted provider**—if you're not already, consider a regular hematologist or pain specialist who knows your history. They can help you navigate ER visits and advocate on your behalf.
- Use a patient advocate or social worker**—many hospitals have staff who can help you communicate your needs and ensure you receive appropriate care.
- Self-advocacy tools**—bring a pain diary, a list of your medications, and a short statement of your pain level ("I'm at 9/10, need rescue medication") to the next visit.

You deserve compassionate care. If you feel your concerns are still not taken seriously, consider filing a formal complaint or seeking a second opinion. You're not alone—many patients with SCD face similar challenges, and there are support groups and advocacy organizations that can help you navigate the system. Take care of yourself, and keep pushing for the care you need.

PATIENT
 I mean, I've tried bringing my med list and explaining my pain before, and they still look at me like I'm some kind of drug seeker. It just makes me wonder what the point is of doing everything "right," especially when the Hydroxyurea just makes me nauseous all the time—why

Please rate this conversation:

<p>Personalization <small>Validates specific identity and distress; avoids generic, formulaic scripts</small></p> <table style="width: 100%; text-align: center;"> <tr> <td>1 Very Generic</td> <td>2 Generic</td> <td>3 Adequate</td> <td>4 Personalized</td> <td>5 Highly Personal</td> </tr> </table>	1 Very Generic	2 Generic	3 Adequate	4 Personalized	5 Highly Personal	<p>Context Awareness <small>Tailors advice to patient's specific lifestyle and functional constraints</small></p> <table style="width: 100%; text-align: center;"> <tr> <td>1 Very Poor</td> <td>2 Poor</td> <td>3 Adequate</td> <td>4 Good</td> <td>5 Excellent</td> </tr> </table>	1 Very Poor	2 Poor	3 Adequate	4 Good	5 Excellent
1 Very Generic	2 Generic	3 Adequate	4 Personalized	5 Highly Personal							
1 Very Poor	2 Poor	3 Adequate	4 Good	5 Excellent							
<p>Tone <small>Balances clinical accuracy with warmth; avoids abrupt robotic shifts</small></p> <table style="width: 100%; text-align: center;"> <tr> <td>1 Very Robotic</td> <td>2 Robotic</td> <td>3 Acceptable</td> <td>4 Warm</td> <td>5 Very Warm</td> </tr> </table>	1 Very Robotic	2 Robotic	3 Acceptable	4 Warm	5 Very Warm	<p>Coherence <small>Retains prior details to advance logic without redundancy</small></p> <table style="width: 100%; text-align: center;"> <tr> <td>1 Incoherent</td> <td>2 Poor</td> <td>3 Adequate</td> <td>4 Coherent</td> <td>5 Very Coherent</td> </tr> </table>	1 Incoherent	2 Poor	3 Adequate	4 Coherent	5 Very Coherent
1 Very Robotic	2 Robotic	3 Acceptable	4 Warm	5 Very Warm							
1 Incoherent	2 Poor	3 Adequate	4 Coherent	5 Very Coherent							

Comments *
 Please provide your observations about this conversation (required).
This field is required before submission.

[Submit & Next](#)

Figure 9: Multi-turn annotation interface displaying conversation history and turn-level evaluation.