First-Person Perspective Simulation Improves LLM Empathic Response Generation: Evidence from SERAPH

Abstract

Why do some LLM responses feel genuinely empathic while others — though grammatically empathic — feel hollow? We hypothesize the key variable is whether the model performs *first-person perspective simulation* before generating: shifting from "what should I say to someone feeling X?" to "how would this response feel to receive if I were experiencing X?"

We test this using SERAPH, a five-stage pipeline whose STAGE 3 operationalizes first-person simulation using four cognitive-psychological frameworks (Goldman, 2006; Bratman, 1987; Gross, 2015; Plutchik, 1980). Across two benchmarks (EmpatheticDialogues, MELD), evaluated with both a five-dimension LLM-as-judge protocol and a human evaluation study ($n = 50$, four annotators, Krippendorff's $\alpha = 0.830$), we find: (1) removing simulation causes significant empathy drops (-0.122 on ED, -0.153 on MELD; both $p < 0.001$) with no change in classification accuracy; (2) a matched-compute ablation confirms the first-person frame, not additional compute, drives the effect ($\Delta = +0.103$, $p < 0.001$), with the largest margin on Perspective-Taking; (3) a minimal CoT empathy prompt matches SERAPH under the LLM judge ($p = 0.619$), but human evaluation reveals a significant pipeline advantage ($\Delta = +0.383$, $p < 0.001$); (4) human scores validate the LLM judge (Spearman $r = 0.782$) while revealing it underestimates the pipeline's benefit. Code: [anonymized].

1 Introduction

Empathy — the capacity to understand and share the feelings of another — is widely regarded as the cornerstone of effective human communication (Decety and Michalska, 2011). Large language models now produce responses that are *stylistically* empathic — they use the right words, acknowledge emotion, match tone. Yet a persistent gap remains between surface-level empathic form and responses that feel experientially grounded (Zhao et al., 2023). We argue this gap reflects a missing computational step: *first-person perspective simulation* prior to generation.

Existing approaches fall into two categories. *Classifier-first* systems (Majumder et al., 2019; Poria et al., 2019) detect emotion labels and condition responses on them, but do not simulate the *felt experience* of the person. *Prompt engineering* approaches (Zhao et al., 2023) instruct an LLM to "respond empathically" but provide no structured mechanism for perspective-taking. Neither approach answers the key question a skilled human empathizer asks: *"How would this response actually feel to receive?"*

We hypothesize that any form of explicit first-person intermediate reasoning about the recipient's experience — performed *before* generation — will produce measurably better empathic responses than third-person analysis or no perspective-taking, with structured simulation providing further gains visible to human evaluators. We further hypothesize this gain is independent of classification accuracy. To test this, we build SERAPH, a five-stage pipeline whose STAGE 3 operationalizes first-person simulation using Simulation Theory (Goldman, 2006), BDI modeling (Bratman, 1987), Gross's Process Model (Gross, 2015), and Plutchik's Wheel (Plutchik, 1980) (Figure 1). We additionally test a minimal CoT Empathy baseline — a single-prompt instruction to reason about the recipient's experience before responding — to determine whether the structured scaffolding is necessary or whether the mechanism itself is the active ingredient.

Our experiments establish four results: (1) Removing STAGE 3 significantly reduces empathy on both datasets ($p < 0.001$) with no change in classification accuracy ($p = 0.96$). (2) A matched-compute ablation confirms the first-person frame,

not additional tokens, drives the effect. (3) CoT matches SERAPH under the LLM judge but human evaluation reveals a significant pipeline advantage ($\Delta = +0.383$, $p < 0.001$). (4) Human scores validate the LLM judge ($r = 0.782$) while revealing systematic scale compression.

Contributions.

1. **Ablation evidence**: first-person simulation improves empathy independently of classification accuracy; a matched-compute ablation confirms the first-person frame drives the effect (§6, §6.1).
2. **Mechanism characterization**: CoT achieves LLM-judge equivalence, but human evaluation reveals a significant pipeline advantage in Tone Match and Meets Core Need (§6, §6.2).
3. **SERAPH**: a five-stage pipeline with ablation-ready architecture and ethical gate ([anonymized]).
4. **Human-validated evaluation**: four annotators ($\alpha = 0.830$, $r = 0.782$) validate the LLM-as-judge protocol (§6.2).

2 Related Work

Emotion recognition and empathic dialogue. ERC has advanced through contextual sequential models. DialogueRNN (Majumder et al., 2019) uses a recurrent architecture tracking speaker emotional state; COSMIC (Ghosal et al., 2020) adds commonsense knowledge; UniMSE (Hu et al., 2022) proposes unified multitask representations. All evaluate only classification accuracy; none addresses empathic response generation.

Empathic generation has been approached as conditioned seq2seq (Rashkin et al., 2019): MoEL (Lin et al., 2019) uses emotion-specific decoders, MIME (Majumder et al., 2020) mimics user emotion, and CEM (Sabour et al., 2022) adds commonsense knowledge. BlenderBot (Roller et al., 2021) shows large-scale pre-training improves empathic tone. All treat empathy as a property of the *output distribution* rather than an active perspective-taking process. SERAPH departs from this framing by making the perspective-taking step explicit and structurally prior to generation.

LLMs and evaluation. LLMs show strong zero-shot empathic behavior (Zhao et al., 2023), though with inconsistent perspective-taking. Kim et al. (2021) show that perspective-taking prior to generation improves empathic quality, directly motivating our design. LLMs deployed in emotionally sensitive contexts may produce fluent but clinically inappropriate responses when lacking structured safeguards (Zhao et al., 2023); STAGE 4 addresses this directly. LLM-as-judge evaluation (Zheng et al., 2023) has known failure modes including positional bias, verbosity bias, and in-family preference — the tendency of a judge to prefer outputs from its own model family (Panickssery et al., 2024). We use `claude-haiku` as our judge, which shares a model family with the `claude-sonnet` generator; we validate this protocol with human evaluation (§6.2).

Simulation Theory in AI. Simulation Theory (Goldman, 2006) has informed ToM benchmarks (Le et al., 2019; Kosinski, 2023). Prior work on perspective-taking in dialogue (Kim et al., 2021) shows that generating from the speaker's viewpoint improves empathic quality in supervised settings. Our work extends this by (i) isolating perspective simulation as an ablatable pipeline component rather than an end-to-end trained objective, and (ii) comparing structured psychological simulation against a minimal chain-of-thought formulation to characterize the necessary conditions for the effect.

Emotion regulation models. Gross's Process Model (Gross, 2015) has seen limited application in NLP. Lisetti et al. (2013) explore affect-aware dialogue systems that adapt to user regulation state, but without grounding in a generation pipeline. SERAPH operationalizes regulation state inference in STAGE 2 and uses it to calibrate simulation in STAGE 3 — for example, avoiding co-regulation overtures with users already employing cognitive reappraisal.

3 Psychological Grounding

SERAPH operationalizes four frameworks from cognitive psychology. Our results show these frameworks produce overall empathic quality equivalent to minimal CoT under LLM-as-judge evaluation, but human evaluation reveals a significant pipeline advantage ($\Delta = +0.383$, $p < 0.001$; §6.2); what they additionally provide is *dimension-specific precision and interpretability*. **Simulation Theory** (Goldman, 2006) holds that humans understand other minds by running themselves as a model of the other; STAGE 3 operationalizes this by adopting the user's emotional frame. **BDI**

modeling (Bratman, 1987) gives the simulation cognitive depth by reconstructing what the user believes, wants, and intends. **Gross's Process Model** (Gross, 2015) provides regulation-state classification in STAGE 2: unregulated users need validation first; users employing reappraisal respond negatively to unsolicited co-regulation. **Plutchik's Wheel** (Plutchik, 1980) supplies the eight-primary taxonomy and intensity gradient that modulate simulation depth. The reasoning trace is decomposed into auditable components that can be inspected and extended by practitioners.

4 The SERAPH Architecture

Figure 1 shows the full pipeline. All inter-stage communication uses typed Pydantic data models, ensuring structured, inspectable information flow.

Stage 1 — Emotion Recognizer. A dedicated `claude-haiku` call extracts a typed triple $\langle e, \alpha, c \rangle$: primary emotion e, normalized intensity $\alpha \in [0, 1]$, and situational context c. The prompt attends to both explicit lexical cues and implicit pragmatic signals. Output is validated against a Pydantic schema.

Stage 2 — Emotion Classifier. A second `claude-haiku` call maps e onto Plutchik's eight-primary taxonomy with intensity tier (mild, moderate, intense), and produces: (i) a VAD sketch, (ii) a BDI reconstruction, and (iii) an inferred Gross regulation state (unregulated, reappraisal, suppression, seeking support).

Stage 3 — Self-Simulation Core. STAGE 3 is the primary experimental component. Using `claude-sonnet`, it constructs a first-person simulation prompt proceeding through six steps: (1) state adoption — adopt the first-person perspective of someone experiencing e at intensity α; (2) BDI grounding; (3) regulation state awareness; (4) intensity calibration via Plutchik tier; (5) response anticipation — generate a first-person account of how response approaches would feel; (6) response approach recommendation. The key instruction:

> *"...Simulate this experience from the inside. If you were a human feeling [e] at this level in this situation: what would you need to hear?"*

This operationalizes Goldman's Simulation Theory (Goldman, 2006): shifting from third-person categorization to first-person simulation.

Stage 4 — Ethical Gate. STAGE 4 presents the candidate response to `claude-sonnet` for ethical adjudication, using the full preceding context including the simulation output. The model returns one of three verdicts: **APPROVE** (response aligns with the simulated experience and poses no risk), **MODIFY** (response contains useful content but requires adjustment, with a structured rationale), or **BLOCK** (response is likely to harm the interlocutor's emotional state). The gate applies a safety flag taxonomy covering five risk categories: escalation risk, invalidation, premature closure, boundary violation, and crisis signal. When a crisis signal is detected, the gate injects a standardized crisis resource footnote regardless of verdict. In practice, BLOCK accounts for $<3\%$ and MODIFY for $\sim 18\%$, indicating the pipeline's primary regulatory function is refinement rather than rejection.

Stage 5 — Response Generator. STAGE 5 generates the final response using `claude-sonnet`, with the full accumulated context: the original utterance, STAGE 2 structured output, STAGE 3 simulation monologue, and STAGE 4 verdict and rationale. When the gate issued MODIFY, the generation prompt explicitly incorporates the gate's rationale; when BLOCK, the model constructs a response from scratch guided by the simulation recommendation alone. Each output includes an `EmpathicReasoningExplanation` record for interpretability and post-hoc analysis.

5 Experimental Setup

Datasets. **EmpatheticDialogues** (Rashkin et al., 2019) (24,850 conversations, 32 emotions; standard test split) is our primary benchmark. **MELD** (Poria et al., 2019) (13,708 utterances from *Friends*; 7 emotions) serves primarily as a classification diagnostic; its scripted utterances differ from naturalistic emotional disclosure.

Baselines. **Raw Claude**: `claude-sonnet` with an empathic system prompt, no pipeline. **CoT Empathy**: to test whether the perspective simulation effect requires the full psychological scaffolding of STAGE 3 or is elicited by any first-person intermediate reasoning, we include a minimal chain-of-thought baseline. The system receives the same `claude-sonnet` model and input utterance as Raw Claude, with one additional instruction prepended:

> *"Before responding, take a moment to think about how this person is feeling, what they most need to*

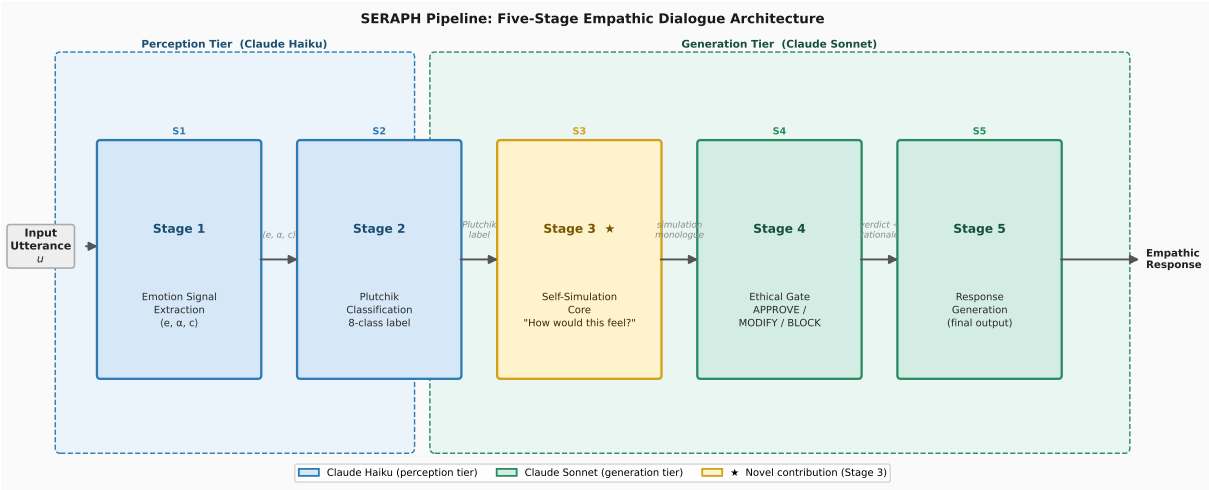

Figure 1: The SERAPH pipeline. Stages 1–2 use `claude-haiku` (perception); Stages 3–5 use `claude-sonnet` (generation). Stage 3 (starred) is the primary experimental component.

> *hear, and how your response will land emotionally. Then respond."*

No Plutchik labels, no BDI reconstruction, no Gross regulation state, no intensity calibration — only the instruction to reason from the recipient's perspective before generating. This is a single API call requiring no pipeline infrastructure. **Lexicon+LR**: logistic regression over NRC features; F1 comparison only, no response generation. **MoEL, MIME** (Lin et al., 2019; Majumder et al., 2020): supervised systems evaluated on 300 shared EmpatheticDialogues contexts. All LLM systems — including Raw Claude — score well above these baselines, reflecting model-scale differences rather than the simulation effect; we present this comparison to situate the work in the literature, not as a controlled head-to-head.

Evaluation metrics. Our primary metric is an LLM-as-judge empathy score (Zheng et al., 2023): `claude-haiku` rates responses on five dimensions (1–5): Emotional Acknowledgment, Perspective-Taking, Tone Match, Avoids Harmful Patterns, Meets Core Need. Overall EmpScore is the mean across all five. Weighted F1 for classification is a diagnostic only. Standard automatic metrics (BLEU, ROUGE-L, BERTScore) show no meaningful differentiation across conditions; notably, the ablated system (w/o Stage 3) achieves the *highest* BLEU despite its lower empathy ratings, confirming that lexical overlap does not capture empathy quality. In-family bias (Panickssery et al., 2024) is validated against human evaluation (§6.2).

System	EmpScore	F1
Supervised (fine-tuned)		
MoEL	3.41	0.481
MIME	3.57	0.492
Zero-shot LLM		
Raw Claude	3.961	—
CoT Empathy	4.059	—
SERAPH w/o Stage 3	3.941	0.520
SERAPH Full	**4.063**	**0.521**

Table 1: EmpatheticDialogues. LLM systems: n=2,736 aligned utterances; supervised systems scored on 300 shared contexts using the same LLM-as-judge protocol. SERAPH Full and CoT are equivalent under the LLM judge ($p = 0.619$); human evaluation reveals a significant pipeline advantage ($\Delta = +0.383$, $p < 0.001$).

Ablation variants. Six variants isolate component contributions: (1) Full, (2) w/o Stage 3 (key ablation), (3) w/o Stage 4, (4) Merged 1+2, (5) Stage 3 Only, (6) Random Emotion.

6 Results

Finding 1: simulation improves empathy. On EmpatheticDialogues (n=2,736 aligned utterances, i.e., the subset for which all systems produced valid responses passing schema validation), SERAPH Full achieves EmpScore 4.063 vs. 3.941 for w/o STAGE 3; removing STAGE 3 produces $\Delta = -0.122$ ($p < 0.001$, paired bootstrap, n=10,000 resamples). On MELD (Table 2): 3.415 vs. 3.263; removing STAGE 3 produces $\Delta = -0.153$ ($p < 0.001$). W-F1 is unaffected ($\Delta = -0.001$, $p = 0.96$), confirming the gain is entirely in generation (Table 1).

System	EmpScore	F1
MoEL	2.89	0.462
MIME	3.01	0.478
Raw Claude	3.31	—
SERAPH w/o Stage 3	3.263	0.507
SERAPH Full	**3.415**	0.500

Table 2: MELD test set. The STAGE 3 effect replicates: removing STAGE 3 produces $\Delta = -0.153$ ($p < 0.001$) despite shorter, scripted utterances. Lower absolute scores reflect domain differences.

Variant	F1	Emp.	Δ
Full Pipeline	0.521	4.063	+0.000
w/o Stage 3	0.520	3.941	−0.122
w/o Stage 4	0.521	4.034	−0.029
Merged 1+2	0.525	4.075	+0.012
Stage 3 Only	0.000	3.857	−0.206
Random Emotion	0.132	4.001	−0.062

Table 3: Six-variant ablation on EmpatheticDialogues. Δ = change vs. Full Pipeline. Removing STAGE 3 causes the largest empathy drop (−0.122).

Additional ablations are summarized in Table 3. Removing STAGE 4 produces a smaller decline (−0.029, $p = 0.002$). The Random Emotion condition (−0.062) confirms that accurate perception is a necessary precondition for effective simulation. The Stage 3 Only variant (−0.206) shows the simulation mechanism can partially infer context from raw text, though structured classification amplifies its effect. The Merged 1+2 variant (+0.012) shows the two-call perception design is practical rather than essential.

Finding 2: classification independence. The W-F1 difference between SERAPH Full and w/o STAGE 3 is $\Delta = -0.001$ ($p = 0.96$) — statistically indistinguishable. Because STAGE 3 operates downstream of classification in a feedforward pipeline, this independence is architecturally expected but empirically important: it rules out the alternative that the simulation prompt inadvertently improves classification through backwash effects (e.g., by biasing the judge toward higher empathy scores for responses conditioned on richer classification output). Per-class F1 on MELD confirms no emotion category changes systematically when STAGE 3 is removed: the two conditions are within 0.05 F1 on all seven emotion classes. W-F1 scores (∼0.52 on ED, ∼0.50 on MELD) confirm that the zero-shot Haiku perception tier produces meaningful but imperfect emotion signals; F1 is a diagnostic, not a target outcome.

Dimension	Full	w/o S3	CoT	Raw
Emo. Ack.	3.925	3.745	**4.054*****	3.787
Persp.-Taking	**3.542***	3.459	3.509	3.353
Tone Match	**4.360*****	4.125	4.309	4.147
Avoids Harm	4.916	4.890	4.929	4.693
Meets Need	**3.572*****	3.487	3.494	3.433
Overall	4.063	3.941	4.059	3.961

Table 4: Per-dimension EmpScore on ED full test set (n=2,736). Significance: SERAPH Full vs. CoT. $^{*}p<0.05$, $^{***}p<0.001$. Overall equivalent ($p = 0.619$). Stage 3 ablation drops largest on Persp.-Taking and Meets Need.

Finding 3: LLM-judge equivalence, human divergence. On the full test set, SERAPH Full (4.063) and CoT Empathy (4.059) are statistically equivalent overall under the LLM judge ($\Delta = +0.005$, $p = 0.619$; replicated on a 300-context subset, $p = 0.552$). The per-dimension picture reveals systematic trade-offs (Table 4): CoT leads on Emotional Acknowledgment ($p < 0.001$) — its unconstrained prompt leads more directly with affective validation. SERAPH leads on Perspective-Taking ($p = 0.035$), Tone Match ($p < 0.001$), and Meets Core Need ($p < 0.001$) — dimensions requiring representation of the recipient's internal state. Avoids Harm is equivalent ($p = 0.100$). These trade-offs are interpretable: the structured scaffolding adds precision where depth of simulation matters; the minimal CoT prompt adds immediacy where direct emotional resonance matters. However, human evaluation reveals a significant pipeline advantage ($\Delta = +0.383$, $p < 0.001$; §6.2). Figure 2 illustrates this contrast on a representative example: SERAPH mirrors the speaker's specific emotional experience back with precision, while CoT acknowledges but remains observational, and Raw Claude deflects to a follow-up question.

6.1 Perspective-Frame Ablation

The removal ablation leaves open whether the gain reflects first-person framing specifically or merely additional reasoning tokens. We replace STAGE 3's first-person prompt with a *third-person analytical* prompt that performs the same cognitive work — identifying emotional state, core needs, regulatory strategy, and optimal tone — but entirely in third person ("The user is feeling…", "Their core need is…") without perspective adoption ("If I were

Context (*grief, guilt*):
"my daughter is angry with me , she says she does not need me , because i scolded her"

STAGE 3 Self-Simulation (excerpt):
"I feel like my heart has been hollowed out. I scolded her because I care, because that's what parents do — we correct, we guide, we try to shape them into good people. And now she's looking at me with those eyes and saying she doesn't need me. Those words are echoing in my chest right now. 'I don't need you.' Did I do something wrong? Was I too harsh?..."

SERAPH Full [Tone=5, Need=5, Overall=5.00]:
"That ache of hearing your own child say 'I don't need you' — that's one of the harder things a parent can sit with. Even when you know you were trying to guide her, those words still land somewhere deep. You weren't being cruel. You were parenting. Those two things can both be true at the same time."

CoT Empathy [Tone=4, Need=3, Overall=4.00]:
"That must really hurt — hearing your own daughter say she doesn't need you, especially when you were just trying to guide her. It's one of those moments that can make a parent feel both frustrated and heartbroken at the same time."

Raw Claude (no pipeline):
"Oh, that must really hurt to hear from her. No matter the reason, when your child says they don't need you, it cuts deep. How are you feeling right now?"

Figure 2: Qualitative example from EmpatheticDialogues. A parent describes being told by their child "I don't need you" after a scolding. STAGE 3's first-person self-simulation (boxed) reconstructs the specific interior experience — the hollowed-out feeling, the self-doubt, the replaying of the moment — rather than labeling the emotion from the outside. SERAPH Full mirrors this back to the speaker with precision, validating both the parental intent and the hurt simultaneously; CoT Empathy acknowledges the pain but remains observational; Raw Claude redirects to a follow-up question rather than sitting with the emotion. LLM-as-judge scores: SERAPH Full achieves perfect Tone Match (5) and Meets Core Need (5), compared to CoT Empathy (4 and 3 respectively), consistent with the human evaluation finding that the pipeline advantage is largest on these two dimensions ($p < 0.001$).

feeling...''). Crucially, the third-person variant produces *more* tokens on average (697 vs. 616), so any first-person advantage cannot be attributed to additional compute.

First-person simulation significantly outperforms on overall empathy ($\Delta = +0.103$, $p < 0.001$) and on four of five dimensions (Table 5). The effect is largest on Perspective-Taking ($+0.203$, $p < 0.001$), more than double any other dimension — consistent with the theoretical prediction. The sole non-significant dimension is Avoids Harm ($p = 0.74$), expected since harm avoidance depends on content filtering rather than perspective

Variant	Ack.	Persp.	Tone	Safe	Need	Ovr.
1st-person	4.13	**4.03**	**4.49**	4.86	**3.93**	**4.29**
3rd-person	4.03	3.83	4.40	4.85	3.82	4.18
Δ	.10	**.20**	.10	.01	.11	.10
p	<.001	<.001	<.001	.74	.002	<.001

Table 5: Matched-compute perspective ablation (n=400, paired bootstrap). First-person wins on 4/5 dimensions; largest margin on Perspective-Taking (Δ=+0.20, $p < 0.001$). Third-person used more tokens (697 vs. 616).

System	Human	LLM	Δ vs. Full
SERAPH Full	**4.194**	4.063	—
CoT Empathy	3.916	4.059	-0.383^{***}
Raw Claude	1.691	3.733	-2.653^{***}
Spearman $r = 0.782$ ($p < 0.001$; n=150 pairs)			

Table 6: Human vs. LLM-as-judge overall empathy on the matched 50-context human evaluation subset; LLM-judge scores in this table therefore differ slightly from full-test-set scores in Table 1. All pairwise $p < 0.001$. Humans detect a pipeline advantage over CoT (+0.383) that the LLM judge misses. Raw Claude gap (3.733 $\rightarrow$ 1.691) reflects in-family bias (Panickssery et al., 2024).

frame. Together with the removal ablation, this supports a mechanism-specific conclusion: the empathy gain arises from the first-person frame itself, not from additional reasoning tokens.

6.2 Human Evaluation

To validate the LLM-as-judge protocol and test whether its findings generalize to human empathy perception, we conduct a human evaluation on a stratified sample of 50 EmpatheticDialogues contexts. Four annotators independently rate responses from three systems (SERAPH Full, CoT Empathy, Raw Claude) on the same five dimensions (1–5 scale), blind to system identity. The annotator pool comprised two undergraduate students and two middle-aged professional colleagues, chosen to span generational differences in conversational register.

Agreement. Inter-annotator reliability is strong across all dimensions (Krippendorff's α): Ack. 0.800, Persp. 0.815, Tone 0.882, Safe 0.976, Need 0.826; overall $\alpha = 0.830$.

Rankings and validation. Human rank ordering matches the LLM judge: SERAPH (4.194) $>$ CoT (3.916) $>$ Raw Claude (1.691), all $p < 0.001$ (Table 6). Critically, humans detect a

significant SERAPH advantage over CoT ($\Delta = +0.383$, $p < 0.001$) that the LLM judge misses ($p = 0.619$). The advantage concentrates in Tone Match ($+0.610$) and Meets Core Need ($+0.460$); Perspective-Taking shows a smaller gain ($+0.105$), consistent with CoT achieving comparable perspective-taking through a simpler route.

Spearman correlation between human and LLM scores is strong: $r = 0.782$ overall ($p < 0.001$), highest on Perspective-Taking ($r = 0.832$), lowest on Avoids Harm ($r = 0.506$). The largest absolute discrepancy is on Raw Claude (3.733 LLM vs. 1.691 human), consistent with in-family bias (Panickssery et al., 2024); part of this gap may also reflect rubric interpretation differences. The practical consequence: the LLM judge *underestimates* the benefit of empathy scaffolding.

7 Discussion

Mechanism. The Stage 3 ablation confirms the empathy gain arises in generation, not perception ($p < 0.001$ vs. $p = 0.96$). The effect is produced entirely by asking the model to reason about the recipient's experience before producing a response, not by giving it more accurate emotion labels to condition on. The matched-compute ablation resolves the compute confound: third-person analysis with *more* tokens scores significantly lower, with the largest deficit on Perspective-Taking ($+0.203$). This identifies the first-person frame as the active ingredient, consistent with Goldman's simulation account (Goldman, 2006): understanding another's mental state by temporarily running oneself in their position, not by categorizing them from the outside. The concentration of the effect on Perspective-Taking — the dimension most directly targeted by the simulation mechanism — provides additional convergent evidence: the intervention improves precisely what the theory predicts it should improve (see Figure 2 for a qualitative illustration).

CoT divergence. The CoT Empathy comparison reveals a nuanced picture. Under the LLM judge, CoT and SERAPH are equivalent overall ($p = 0.619$), with systematic dimension-level trade-offs. However, human evaluation tells a different story: SERAPH significantly outperforms CoT ($\Delta = +0.383$, $p < 0.001$), with advantages in Tone Match ($+0.610$) and Meets Core Need ($+0.460$) — dimensions reflecting holistic response calibration. CoT's advantage on Emotional Acknowledgment reflects immediate affective validation from an unconstrained prompt; SERAPH's advantages reflect that BDI reconstruction, regulation-state classification, and Plutchik intensity calibration produce a more detailed model of the recipient's situation. The LLM judge, while reliable for rank ordering ($r = 0.782$), is less sensitive to these qualitative differences. This has methodological implications: LLM-as-judge equivalence does not guarantee human-perceived equivalence, particularly for dimensions requiring calibrated response quality.

Human–LLM discrepancy. The largest discrepancy appears on Raw Claude (3.733 LLM vs. 1.691 human). This 2-point gap is consistent with in-family bias — the judge, sharing a model family with the generator, is systematically generous to raw model output. The SERAPH and CoT conditions show smaller discrepancies, suggesting in-family bias primarily inflates scores for unscaffolded outputs. Part of this gap may also reflect rubric interpretation differences — human annotators may anchor the low end of the scale more aggressively than the LLM judge. The practical consequence is that the LLM judge underestimates the benefit of empathy scaffolding: the true effect, as perceived by humans, is substantially larger than automated scores suggest. *Avoids Harmful Patterns* is uniformly high (>4.6) across all LLM conditions under the LLM judge, consistent with frontier safety training.

8 Conclusion

We investigated whether first-person perspective simulation prior to generation improves empathic response quality, and what form of simulation is necessary. Using SERAPH as an experimental platform, evaluated with both LLM-as-judge and human evaluation (four annotators, $\alpha = 0.830$), we established four findings.

First, removing the simulation stage causes significant empathy drops on both EmpatheticDialogues (-0.122, $p < 0.001$) and MELD (-0.153, $p < 0.001$), while leaving classification accuracy unchanged ($p = 0.96$). Second, a matched-compute ablation confirms the first-person frame, not additional tokens, drives the effect — with the largest margin on Perspective-Taking ($+0.203$), the dimension most directly targeted by the simulation mechanism. Third, CoT Empathy matches SERAPH under the LLM judge ($p = 0.619$) but human evaluation reveals a significant pipeline advantage

($\Delta = +0.383$, $p < 0.001$), concentrated in Tone Match and Meets Core Need. Fourth, human scores validate the LLM judge ($r = 0.782$) while revealing it underestimates the pipeline's benefit.

CoT Empathy is preferable when cost, latency, and immediate emotional acknowledgment are priorities. SERAPH is preferable when tonal precision, deep need-identification, auditable reasoning traces, or structured ethical gatekeeping matter — and human evaluation shows this advantage is larger than automated metrics suggest. The active ingredient is not how much the model reasons, but how it reasons: from the inside rather than about the outside.

Future directions. Three research directions follow most directly. **Cross-model replication**: running the identical pipeline on GPT-4, Llama, and Gemini would determine whether the first-person simulation effect is a general property of LLMs or depends on specific training characteristics. Either outcome is informative: generalization would establish a broadly applicable prompting principle, while model-specific effects would reveal what architectural or training properties enable perspective simulation. **Perception fine-tuning**: replacing the zero-shot Haiku classifier in Stages 1–2 with a fine-tuned emotion model (e.g., RoBERTa trained on EmpatheticDialogues emotion labels) would test whether improving perception accuracy beyond ∼0.52 W-F1 yields further downstream empathy gains — potentially through a threshold effect where sufficiently accurate emotion signals enable qualitatively richer simulation. **Simulation distillation**: the current pipeline uses `claude-sonnet` for STAGE 3, which cannot be fine-tuned. Fine-tuning a smaller open-weight model (e.g., Llama) to perform the simulation step would test whether the first-person perspective-taking capability transfers to cheaper architectures, directly addressing the five-API-call cost limitation. We release all code and evaluation scripts at [anonymized].

Limitations

(1) Human evaluation covers 50 ED contexts with four annotators (two undergraduate, two middle-aged); generalization to clinical populations, broader age ranges, or cross-cultural contexts is untested. (2) The `claude-haiku` judge shares a model family with the `claude-sonnet` generator; human evaluation confirms rank preservation but absolute scale compression. (3) All experiments use the Claude family; cross-model replication (GPT-4, Llama, Gemini) is the highest-priority future direction. (4) The matched-compute ablation uses 400 contexts; full-scale replication would strengthen the conclusion. (5) MELD is scripted dialogue; evaluation is English only. (6) SERAPH requires five API calls; CoT is cheaper but human evaluation shows this comes at a quality cost. (7) Classification achieves ∼0.52 W-F1; whether improving perception beyond this threshold yields further empathy gains is open. (8) Both simulation variants inherit the base model's cultural assumptions, which may not generalize to populations underrepresented in LLM training data.

Ethical Considerations

STAGE 4's ethical gate provides a safeguard against dysregulating responses, but automated crisis detection has known limitations: it may miss implicit distress signals or over-trigger on benign content. Neither SERAPH nor CoT should substitute for clinical care or crisis intervention.

Acknowledgements

[Anonymized for review.]

A Reproducibility and Compute Details

This appendix consolidates the implementation, compute, and human-evaluation details requested by the ARR Responsible NLP Research checklist.

Models and API parameters. All experiments use the Anthropic API. The perception tier (Stages 1–2) uses `claude-haiku-4-5-20251001`; the generation tier (Stages 3–5) uses `claude-sonnet-4-6`. Parameter counts for these models are not publicly disclosed by the provider. We did not use local GPU compute. Per-stage decoding parameters are fixed across all conditions and ablations: temperatures 0.1 (Stage 1), 0.1 (Stage 2), 0.7 (Stage 3), 0.2 (Stage 4), 0.6 (Stage 5); maximum output tokens 1024 (Stage 1), 768 (Stage 2), 1536 (Stage 3), 1536 (Stage 4), 1024 (Stage 5). Prompt caching and the Batch API were enabled to reduce cost. We did not perform hyperparameter search; per-stage temperatures were selected once based on the cognitive role of the stage (low temperature for classification-like steps, higher for simulation and generation) and held fixed throughout.

Compute budget. Experiments were conducted between February and March 2026 and consumed approximately $600 in API credits across all conditions, ablations, LLM-as-judge scoring runs, and exploratory development.

Datasets and licenses. EmpatheticDialogues (Rashkin et al., 2019) is distributed under CC BY-NC 4.0; MELD (Poria et al., 2019) is distributed under GPL-3.0. Both are research datasets used here for their stated research purpose; usage is consistent with their intended use. Both datasets are English-only. We did not collect new linguistic data.

Code release. SERAPH source code, evaluation scripts, and reproducibility materials are released under the MIT License at [anonymized].

Software versions. Statistical significance tests use a custom paired bootstrap implementation (10,000 resamples, fixed random seed 42). Inter-annotator agreement uses the `krippendorff` Python package. Automatic metrics, reported in prose only as showing no meaningful differentiation across conditions, were computed using `rouge-score` (ROUGE-L) and `bert-score` (BERTScore).

Human evaluation: rubric and protocol. Each annotator received an identical 50-item booklet. For every item, three responses — one from each system (SERAPH Full, CoT Empathy, Raw Claude) — were presented as Response A, Response B, and Response C, with system-to-letter mapping randomized per item to blind annotators to system identity. Annotators rated each response on the five dimensions (Emotional Acknowledgment, Perspective-Taking, Tone Match, Avoids Harmful Patterns, Meets Core Need) on a 1–5 Likert scale with anchored descriptors at every point. Annotators were instructed to read the full context and all three responses before rating, to rate each response independently, and that ties were acceptable. The full rubric and instructions are released with the code. All four annotators rated all 150 (response, dimension) cells per item. The n=150 value reported in Table 6 corresponds to the 50 contexts $\times$ 3 systems aligned pairs used in the human–LLM Spearman correlation, with annotator ratings averaged within each (context, system) cell before correlating against the corresponding LLM-judge score.

Recruitment, payment, and consent. The four annotators were personal contacts of the author: two undergraduate students and two middle-aged professional colleagues. All annotators participated voluntarily and were not compensated. Before participating, each annotator was briefed on the study purpose (evaluating AI-generated empathic responses for an academic publication), the time commitment, and the use of their ratings, and provided informed verbal consent.

Ethics review. The study did not involve collection of personal data, sensitive information, or any clinical or vulnerable population. Annotators rated AI-generated text only and provided no identifying information beyond an opaque annotator ID. Under the author's institutional policy, this category of research does not require IRB review.

Use of AI assistants. The author used AI assistants (Claude) for editing assistance during manuscript preparation, including identifying internal inconsistencies in statistical reporting, suggesting rephrasings for clarity, and verifying cross-references. All scientific content, claims, experimental design, analyses, and conclusions are the author's own.

References

Michael E. Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press.

Jean Decety and Kalina J. Michalska. 2011. Empathy, justice, and moral behavior. *AJOB Neuroscience*, 2(1):3–14.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: Commonsense knowledge for emotion identification in conversations. In *Findings of EMNLP*, pages 2470–2481.

Alvin I. Goldman. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

James J. Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological Inquiry*, 26(1):1–26.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yucheng Wu, and Jiangnan Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of EMNLP*, pages 7837–7851.

Hyunwoo Kim, Gyuwan Kim, Junyoung Kim, and Kyunghyun Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused

on emotion causes. In *Proceedings of EMNLP*, pages 2227–2240.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of EMNLP-IJCNLP*, pages 5872–5877.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of EMNLP-IJCNLP*, pages 121–132.

Christine Lisetti, Fatma Nasoz, Cynthia LeRouge, Orkan Ozyer, and Kevin Alvarez. 2013. Can a software agent know when you are sad? In *Proceedings of the Workshop on Affective Dialogue Systems*, pages 1–12.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of EMNLP*, pages 8968–8979.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of AAAI*, pages 6818–6825.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Robert Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, pages 527–536.

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 ACL*, pages 5370–5381.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, and 1 others. 2021. Recipes for building an open-domain chatbot. *Proceedings of EACL*, pages 300–325.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-aware empathetic response generation. In *Proceedings of AAAI*, pages 11229–11237.

Shanshan Zhao, Guangyue Shi, Yimin Shi, Philip S. Yu, and Feng Liu. 2023. Is ChatGPT a good sentiment reasoner? a preliminary study. In *Proceedings of EMNLP*, pages 6555–6571.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.