JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(3), 2754-2767
ISSN: 2589-1316

A Framework for Evaluating Empathy in Generative AI Customer Service

Saumabha Barua[1*], Doel Mukherjee[2]

[1] *PhD Scholar, Amity Business School, Amity University, Kolkata, INDIA; saumabha.barua@s.amity.edu, ORCID ID: https://orcid.org/0009-0004-9766-9444*
[2] *Associate Professor, Amity Business School, Amity University Kolkata. ORCID ID: 0000-0002-3367-6766*

***Corresponding Author:** saumabha.barua@s.amity.edu

Citation: Barua, S. and Mukherjee, D. (2025). A Framework for Evaluating Empathy in Generative AI Customer Service, *Journal of Cultural Analysis and Social Change*, 10(3), 2754-2767. https://doi.org/10.64753/jcasc.v10i3.2831

Published: December 04, 2025

ABSTRACT

This paper proposes a structured framework for evaluating empathy in Generative AI (GenAI) customer-service systems, addressing a critical gap in how human-centric qualities are assessed in automated interactions. Traditional service-quality models, such as SERVQUAL and the Interpersonal Reactivity Index, conceptualize empathy through human judgment and emotional capability—dimensions that do not directly transfer to GenAI. To adapt these constructs, the paper introduces a dual-constraint model in which emotional alignment and policy compliance jointly determine a bounded form of empathy suitable for AI-mediated service. The study employs conceptual simulation, integrating psychological theory, service-quality research, and responsible-AI standards including EmotionML, IEEE 7010, and ISO/IEC 23894. Symbolic modelling and unit-free visualization are used to illustrate how empathy, factual integrity, and customer outcomes interact across feasible and infeasible regions. Governance metrics—such as Bounded Empathy Drift, Policy Breach Rate, and Insensitive Response Rate—demonstrate how organizations can monitor empathic behaviour during deployment and ensure compliance with ethical constraints. The framework contributes a measurable, governable conceptualization of empathy for GenAI services, offering a foundation for empirical validation and practical implementation. It positions empathy as a strategic performance dimension influencing satisfaction, trust, and service experience within AI-enabled customer ecosystems.

Keywords: Bounded Empathy Framework, AI Governance, Ethical AI, Affective Computing, Emotional Intelligence, AI-Mediated Customer Experience

INTRODUCTION

Generative AI (GenAI) is rapidly moving from pilot projects to large-scale deployment across business functions, especially customer service (McKinsey, 2023; McKinsey, 2024). As organizations integrate GenAI into service operations, evaluation must expand beyond efficiency to encompass empathy and emotional responsiveness—key determinants of customer trust and loyalty.

Empathy has long been a pillar of service quality, defined as providing caring, individualized attention (Parasuraman, Zeithaml, & Berry, 1985, 1988). Within the SERVQUAL framework, it forms one of five validated dimensions for assessing expectation–perception gaps in customer experiences. Parallel customer-experience (CX) measures such as the American Customer Satisfaction Index (CSAT) (Fornell et al., 1996; Anderson, Fornell, & Rust, 1997), Net Promoter Score (NPS) (Reichheld, 2003), and Customer Effort Score (CES) (Dixon, Freeman, & Toman, 2010) continue to serve as industry standards, linking customer perceptions to business performance.

To quantify empathy and emotion more rigorously, research draws on psychology and data standards. The Interpersonal Reactivity Index (IRI) decomposes empathy into cognitive and affective dimensions (Davis, 1980, 1983), while the W3C EmotionML 1.0 standard provides interoperable vocabularies for emotional annotation

Copyright © 2025 by Author/s. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited..

(W3C, 2014). Governance guidelines such as IEEE 7010-2020 further emphasize the assessment of AI's impact on human well-being.

GenAI dialogue systems pose new evaluation challenges beyond traditional CX metrics. EmpatheticDialogues (Rashkin et al., 2019) and ACUTE-Eval (Li, Weston, & Roller, 2019) benchmarks demonstrate that empathy-rich training and pairwise dialogue evaluations improve perceived human-likeness and warmth. Yet, limitations in emotion recognition—context dependency, sarcasm, and label ambiguity—persist (Poria et al., 2019; Pereira et al., 2024), indicating that hybrid human–machine evaluation remains essential.

This study proposes an integrated, standards-based evaluation framework linking three layers: (a) business-anchored CX metrics (CSAT, NPS, CES, SERVQUAL-Empathy), (b) conversation-level empathy assessments (IRI-guided rubrics and EmotionML tags), and (c) model-level evaluation using empathy-focused benchmarks. Aligning these dimensions with business KPIs such as retention, revenue growth, and reduced effort provides an interpretable foundation for measuring empathy in GenAI-enabled customer service (Anderson et al., 1997; Fornell et al., 2006).

METHODS

SERVQUAL and Empathy in Service Quality

SERVQUAL remains a foundational framework for assessing service quality through five dimensions—reliability, assurance, tangibles, empathy, and responsiveness (Parasuraman et al., 1988). Empathy, reflecting caring and individualized attention, is especially pertinent for GenAI systems simulating human service interactions. To operationalize this, the empathy subdimension can be adapted into short post-contact surveys comparing customer expectations and perceptions (E–P gaps), allowing parallel assessment across human and AI channels. For Example: In a GenAI billing-waiver chat, three empathy items—understanding, care, and individualized attention—yielded an initial empathy gap of −0.67, improving to −0.10 over six weeks.

Customer Experience Outcome Metrics

CSAT, NPS, and CES remain the primary indicators of customer satisfaction and loyalty, linking directly to firm performance (Fornell et al., 1996; Reichheld, 2003; Dixon et al., 2010). Embedding these measures with empathy assessments enables mapping conversational quality to executive KPIs. For instance, GenAI recorded higher CSAT (52%) than human agents (48%) for the same intent; NPS improved to +16; CES decreased following empathy prompt tuning, validated against repeat-contact rates.

Interpersonal Reactivity Index (IRI)

The Interpersonal Reactivity Index conceptualizes empathy as multidimensional perspective taking, empathic concern, fantasy, and personal distress (Davis, 1980, 1983). Adapted for GenAI transcript analysis, it guides human raters in assessing context understanding, emotional validation, and tone appropriateness, linking psychological constructs to conversational empathy. For example: Weekly transcript ratings using a 0–2 rubric (perspective taking, concern, tone) increased the empathy score from 74% to 82% post–prompt update.

EmotionML for Interoperable Affect Annotation

W3C's EmotionML 1.0 standard (2014) enables interoperable emotion tagging across human and AI systems using categories (e.g., joy, trust) and dimensions (valence, arousal). Encoding both human and machine annotations facilitates longitudinal affect tracking across intents and channels. For Example: Dialogues tagged in EmotionML revealed an "Affect Appropriateness Index," showing tone shifts from negative to neutral/positive within three turns, indicating adaptive affective alignment.

IEEE 7010 and Governance

IEEE 7010-2020 provides a governance framework for monitoring AI's impact on human well-being. In customer service, this translates into dashboards tracking insensitive responses, misunderstanding-related escalations, and delayed handoffs. A red-amber-green system (e.g., >12 insensitive complaints per 10k = red) can trigger prompt freezes or targeted retraining.

EmpatheticDialogues as Empathy Probes

The EmpatheticDialogues dataset (~25k conversations) benchmarks models on empathy-oriented response generation (Rashkin et al., 2019). Organizations can derive internal empathy probes—such as emotional rebooking requests—to test model sensitivity before rollout, using human rubrics (perspective taking, concern, tone) for evaluation.

© 2025 by Author/s

ACUTE-Eval

ACUTE-Eval compares dialogue quality through pairwise human judgments of empathy and naturalness (Li, Weston, & Roller, 2019). Organizations can run lightweight weekly A/B tests—e.g., approving new GenAI prompts only when they exceed a 60% win-rate with no safety regressions—ensuring continuous empathy optimization through human feedback.

Emotion Recognition in Conversation (ERC)

Automated ERC tools classify emotions across conversation turns but struggle with sarcasm, context shifts, and speaker nuances (Poria et al., 2019; Pereira et al., 2024). A hybrid method combining ERC outputs with human evaluation enhances reliability. For Example*:* If anger or frustration persists unacknowledged for two turns, cases are flagged for review. Trends from such alerts guide empathy prompt adjustments and training focus.

Table 1. Comparative Overview of Empathy and Experience Frameworks for GenAI-Enabled Customer Service

Method / Source	What It Measures / Defines	Inputs	Scoring / Reporting	Strengths	Limitations	Fit for GenAI Customer Service
SERVQUAL (Parasuraman et al., 1988)	Service quality across five dimensions; empathy = caring, individualized attention	Expectation (E) and Perception (P) ratings on Likert scales	Empathy = mean (P − E); report mean ± CI; track deltas post-change	Extensively validated; empathy already embedded in CX instruments	Gap-score method debated; may require cultural adaptation	Add 3–5 empathy items in post-chat surveys; benchmark empathy gaps across AI vs human channels
CX Outcome Metrics	Overall customer-experience outcomes (CSAT, NPS, CES)	Short surveys per metric	Metric-specific (mean, % top-box, % Promoters − Detractors)	Strong managerial adoption; links empathy with performance KPIs	Metric scope varies; may miss affective nuance	Integrate alongside empathy to align with executive KPIs
└ CSAT (Fornell et al., 1996)	Satisfaction with interaction	1–3 items; 5–7-pt scale	Mean / % top-box (5s)	Simple; widely understood	Limited contextual insight	Track post-interaction satisfaction for GenAI agents
└ NPS (Reichheld, 2003)	Likelihood to recommend	Single item; 0–10 scale	% Promoters − % Detractors	Loyalty proxy; longitudinal use	Over-simplified as single metric	Compare loyalty signals across GenAI and human channels
└ CES (Dixon et al., 2010)	Ease of issue resolution	1 item; 5–7-pt agreement	Mean / distribution	Predicts retention; effort focus	Narrow construct	Examine empathy's role in reducing perceived effort
Interpersonal Reactivity Index (IRI) (Davis, 1980, 1983)	Multidimensional empathy: Perspective Taking, Empathic Concern, Fantasy, Personal Distress	28 items; 7-pt Likert	Subscale means / totals	Robust psychological construct; decomposes empathy	Trait-based; needs conversational adaptation	Convert to transcript-rating rubric (perspective, concern, tone) for GenAI dialogue review
EmotionML 1.0 (W3C, 2014)	Standard markup for annotating emotions by categories	Emotion tags (XML/JSON) from	Aggregate tone shifts, emotion frequencies, "Affect	Interoperable, vendor-neutral; cross-	Needs consistent labelling;	Enable uniform affect tagging across intents and channels

 © 2025 by Author/s

	and dimensions	human or AI annotators	Appropriateness Index"	system consistency	limited vocabularies	
IEEE 7010-2020	Assessment of AI's impact on human well-being	Governance dashboards, incident logs	Color-coded thresholds (R/A/G); trend tracking	Integrates ethics and well-being oversight	Manual review; limited granularity	Monitor insensitive replies and distress escalations; govern prompt rollouts
Empathetic Dialogues (Rashkin et al., 2019)	Empathy-oriented dialogue benchmark (~25k conversations)	Model responses to emotional prompts	Human ratings of empathy alignment and relevance	Validated dataset for affective response training	Open-domain; limited task specificity	Develop internal empathy probes; benchmark model/prompt variants
ACUTE-Eval (Li, Weston & Roller, 2019)	Pairwise human comparison of dialogue quality (considerate, natural)	Two dialogues per comparison; multiple rater votes	Win-rate % ± 95 % CI	Sensitive to subjective empathy and human-likeness	Human-intensive; sampling bias risk	Use weekly A/B empathy tests; approve prompts > 60 % win-rate
Emotion Recognition in Conversation (ERC) (Poria et al., 2019; Pereira et al., 2024)	Automated emotion classification at turn / conversation level	Labelled turns or embeddings	Precision/recall of emotion classes; alert rules	Scalable affect monitoring; supports trend analysis	Errors with sarcasm and context shifts	Combine automated detection with human review to flag empathy gaps

Note. Table synthesizes psychological, computational, and governance frameworks relevant to empathy measurement in GenAI-mediated customer service. References: Parasuraman et al. (1988); Fornell et al. (1996); Reichheld (2003); Dixon et al. (2010); Davis (1980, 1983); W3C (2014); IEEE (2020); Rashkin et al. (2019); Li et al. (2019); Poria et al. (2019); Pereira et al. (2024).

Integrated Framework and KPI Mapping

Architecture Overview

The proposed framework (Figure 1) integrates three measurement layers to assess empathy in GenAI-enabled service operations.

- **Layer 1 – Customer Outcomes:** Traditional metrics (CSAT, NPS, CES) and the SERVQUAL–Empathy block quantifies perceived service quality.
- **Layer 2 – Conversation Empathy:** Human ratings based on the Interpersonal Reactivity Index (IRI) combine with W3C EmotionML tags for interoperable affect capture.
- **Layer 3 – Model Quality:** ACUTE-Eval pairwise comparisons and EmpatheticDialogues-derived probes evaluate conversational empathy at the system level.

A governance overlay grounded in IEEE 7010 ensures well-being oversight, aligning psychological validity, service performance, and responsible-AI practices.

© 2025 by Author/s

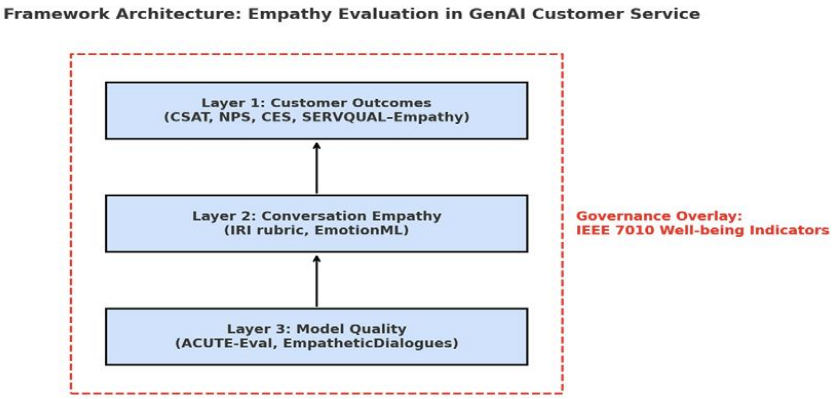

Figure 1. Framework Architecture: Empathy Evaluation in GenAI Customer Service

Data Flow and Scoring

The framework operates as an end-to-end pipeline:

1. **Collection:**
   Post-contact surveys record CSAT / CES and SERVQUAL-Empathy items; trained reviewers rate sampled transcripts via an IRI-based rubric; emotion tags are stored in EmotionML.
2. **Scoring:**
   - ***Customer Layer:*** CSAT (top-box %), NPS score, CES mean, and Empathy Gap (P − E).
   - ***Conversation Layer:*** Rubric scores (0–2) aggregate into a Conversation Empathy Score (0–100) validated against EmotionML patterns.
   - ***Model Layer:*** ACUTE-Eval win-rates with 95 % CIs applied to empathy probes.
3. **Storage:**
   Turn-level affect data are archived in standardized EmotionML format for longitudinal analysis.

KPI Wiring: Linking Empathy to Business Value

Empathy metrics connect directly to business KPIs (Table 2). Improvements in empathy correspond to measurable performance gains: higher CSAT and NPS, reduced CES, increased retention / revenue, and lower cost-to-serve. This positions empathy not as a soft trait but as a quantifiable driver of financial outcomes.

Table 2. Empathy Linkages to Key Business KPIs

KPI	**What to Monitor**	**Expected Movement when Empathy Improves**
CSAT	Top-box % by intent / channel	↑ Customer satisfaction; validated in prior studies
NPS	Net score by cohort	↑ Recommendation intent for empathetic flows
CES	Mean effort score	↓ Perceived effort; predicts loyalty
Retention / Revenue	Churn / repeat purchase	↑ Positive association via satisfaction pathways
Cost-to-Serve	Re-contacts / handle time	↓ As issues resolve more smoothly

Thresholds and Decision Gates

To prevent drift in empathy performance, the framework establishes quantitative release gates (Figure 2).

- ***Empathy Gate:*** model releases require ≥ 60 % ACUTE-Eval win-rate on empathy tasks.
- ***Conversation Quality:*** Empathy Score < 70 triggers coaching; < 60 initiates full review.
- ***Outcome Guardrail:*** sustained CES increase or CSAT decline for two weeks freezes rollout.
- ***Well-Being Checks:*** two consecutive IEEE 7010 "red" weeks on insensitive-response complaints prompt executive intervention.

These thresholds convert empathy from an abstract construct into a governed operational metric.

Implementation Steps

The framework can be deployed incrementally through four practical steps:

1. Embed SERVQUAL–Empathy items in post-contact surveys.
2. Establish a rater program for conversation-level scoring using the IRI rubric.

 © 2025 by Author/s

3. Standardize EmotionML-based affect capture at turn level.
4. Run ACUTE-Eval empathy comparisons weekly as release gates.

This staged rollout yields actionable empathy data without major infrastructure cost.

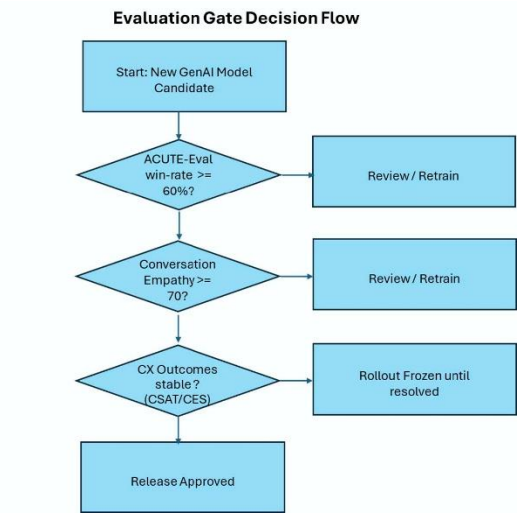

Figure 2. Evaluation Gate Decision Flow

Worked Case Application: Airline Disruption Handling

A hypothetical airline implements the framework for a GenAI-based virtual assistant managing missed-flight rebooking.

Step 1 – Customer Outcome Layer.

Post-chat surveys combine CSAT, CES, and three SERVQUAL–Empathy items ("understood my situation," "showed concern," "gave individualized attention").

Example results: CSAT top-box = 54 % (GenAI) vs. 49 % (human); CES = 4.2/7; Empathy gap improved from −0.6 to −0.2 over six weeks.

Step 2 – Conversation Empathy Layer.

Transcript rated with IRI-based rubric: perspective taking = 2, empathic concern = 1, tone appropriateness = 2 → Empathy Score = 83 %.

EmotionML annotation shows customer emotion shifting from negative (sadness 0.6, valence −0.5) to neutral and positive within three exchanges, marking successful affect recovery.

Step 3 – Model Quality Layer.

Two prompts tested via ACUTE-Eval: baseline vs. empathy-tuned ("acknowledge the customer's situation before suggesting next steps").

Model B wins 64 % of 100 dialogue pairs (95 % CI 58–70 %), clearing the empathy gate ≥ 60 %. EmpatheticDialogues probes stress-test edge cases such as family-emergency rebooking.

Step 4 – Governance Overlay.

Monthly IEEE 7010 review: insensitive-response complaints drop from 15 to 7 per 10 000 contacts (Amber → Green); escalation lags reduce 22 %.

Dashboard indicators show CSAT ↑, CES ↓, and empathy scores > threshold—validating the model's readiness.

Interpretation

The case demonstrates how layered measurement (surveys, rubrics, emotion tags), scoring (ACUTE-Eval, Empathy Score), and governance (IEEE 7010 review) converge to operationalize empathy as a measurable, auditable dimension of GenAI service quality.

1. Dual-Constraint Model of Empathy for AI-Mediated Customer Interaction
2. Theoretical Basis

Empathy is a well-established determinant of satisfaction and recovery in service encounters (Parasuraman, Zeithaml, & Berry, 1988; Davis, 1983). Traditional models such as **SERVQUAL** and the **Interpersonal**

© 2025 by Author/s

Reactivity Index (IRI) conceptualized empathy as caring, perspective taking, and emotional resonance—abilities intrinsic to human agents whose authenticity and judgment shape customer outcomes.

Generative AI (GenAI) service systems differ fundamentally: they lack intrinsic emotion and must *simulate* empathy algorithmically. Effective design therefore requires balancing two opposing demands—**emotional alignment** (expressing understanding) and **policy compliance** (maintaining factual and ethical integrity). Excessive empathic language—over-apologizing, over-promising, or implying fault—may yield short-term satisfaction but erodes credibility once inconsistencies emerge. Hence, empathy in AI contexts must be *bounded*: emotionally responsive yet factually constrained.

Dual-Constraint Framework

The proposed model defines:

- ***Emotional Alignment (E_e):*** degree to which the AI mirrors and validates customer emotion through tone, phrasing, and sentiment congruence.
- ***Policy Compliance (P_p):*** adherence to verifiable truth, organizational policy, and ethical guidelines.

Empathy is sustainable only within the *bounded region* where these two variables jointly optimize customer trust. Conceptually,

$$E_b = \min\left(E_e, f(P_p)\right)$$

where $f(P_p)$ defines the permissible range of empathic expression under policy. Empathy without compliance risks deception; compliance without empathy yields sterile exchanges. The equilibrium of the two defines the **trust-sustaining zone** for AI-mediated service.

From Human to Machine-Governed Empathy

In human interaction, empathy is intuitively regulated by context and feedback. In GenAI systems, this regulation must be **explicitly modelled** through algorithmic constraints. The bounded-empathy construct extends marketing theory by redefining empathy as a **governable optimization variable** rather than an unqualified virtue. The objective is *emotional appropriateness*—alignment that preserves trust without violating factual or ethical boundaries—embedding empathy calibration directly into system design.

Trust as an Emergent Outcome

Customer trust (T) emerges as a joint function of E_e and P_p:

$$T = g(E_e, P_p)$$

High empathy with low compliance produces transient satisfaction but weak credibility; high compliance with low empathy yields reliability without warmth. Enduring trust arises when both coexist within bounded equilibrium, consistent with recent findings that transparency and credibility moderate affective influence in AI-driven service (Hoffman et al., 2023; Bock et al., 2022).

Theoretical Contribution

This framework contributes three conceptual advances:

1. ***Bounded Domain:*** introduces upper limits on empathic expression, linking emotion-regulation theory to responsible-AI governance.
2. ***Dual-Constraint Optimization:*** reconceptualizes empathy as a dynamic trade-off between affective expression and factual integrity.
3. ***Cross-Disciplinary Integration:*** unites service marketing, affective computing, and AI ethics by treating empathy as both a satisfaction driver and an operational control variable.

By reframing empathy as *bounded and governable*, the model provides a measurable, policy-aware foundation for next-generation AI customer-service systems.

Conceptual Propositions

- **P1:** Customer satisfaction and trust jointly depend on the interaction of emotional alignment and policy compliance.
- **P2:** The utility of empathy diminishes beyond a compliance-dependent threshold, forming a plateau or reversal in satisfaction.
- **P3:** Over-empathy under low compliance increases warmth but reduces long-term trust and credibility.
- **P4:** GenAI agents maintaining high empathy within compliance bounds outperform both under- and over-empathic systems in repeat engagement and advocacy.

Mathematical Model

Purpose

 © 2025 by Author/s

This model formalizes the *bounded empathy* construct by describing how **emotional alignment (E_e)** and **policy compliance (P_p)** jointly influence satisfaction and trust.

Unlike traditional additive marketing equations, it embeds **ethical constraints**, ensuring that simulated empathy remains both feasible and responsible in AI-mediated service. The model is conceptual and awaits empirical calibration through behavioural and survey data.

Latent Empathy Construct

Each customer interaction *i* at time *t* yields three measurable indicators:

- $\mathbf{H_{i,t}}$ – Human-rated empathy (rubric: perspective taking, concern, tone)
- $\mathbf{M_{i,t}}$ – Machine-derived affect score (e.g., EmotionML valence)
- $\mathbf{R_{i,t}}$ – Emotional recovery (improvement in customer tone across turns)

Latent empathy ($\mathbf{E_{i,t}}$) is modelled as a weighted combination within a confirmatory factor structure (Jöreskog & Sörbom, 1993):

$$E_{i,t} = \lambda_1 H_{i,t} + \lambda_2 M_{i,t} + \lambda_3 R_{i,t} + \zeta_{i,t}$$

where λ_1–$\lambda_3 > 0$ are factor loadings and $\zeta_{i,t}$ represents error variance.

This linear form ensures interpretability: incremental gains in tone, affect recognition, or recovery behaviour proportionally raise overall empathy.

Bounded Optimization Principle

Empathy is constrained by compliance rules. Formally,

$$\text{maximize } E_{i,t} \text{ s.t. } P_{p,i,t} \geq \theta_{policy}$$

where $P_{p,i,t} \in [0, 1]$ denotes factual or ethical compliance, and θ_{policy} is the organizational minimum.

When $P_p < \theta_{policy}$, additional empathy becomes infeasible—creating a **false-empathy region** (emotional warmth without truth). Conversely, high compliance with low empathy forms a **cold-empathy region**.

The feasible empathy surface is expressed as:

$$E_b = \min(E_e, f(P_p))$$

where $f(P_p)$ is the policy-defined ceiling on expressible empathy. The resulting surface captures a **trust plateau**—a region of balanced authenticity and compliance.

Outcome Relationship

Customer outcomes—satisfaction (CSAT), effort (CES), and recommendation intent (NPS)—are conceptualized as joint functions of empathy and compliance:

$$Y_{i,t} = g(E_{i,t}, P_{p,i,t}, X_{i,t})$$

with $X_{i,t}$ representing contextual factors such as issue type or channel.

Theoretical derivatives satisfy:

$$\frac{\partial g}{\partial E} > 0, \frac{\partial g}{\partial P_p} > 0, \frac{\partial^2 g}{\partial E \, \partial P_p} > 0$$

indicating synergistic effects, while

$$\frac{\partial^2 g}{\partial E^2} < 0$$

reflects diminishing returns to empathy beyond the feasible bound.

Maximum satisfaction occurs within the **bounded optimal zone**, not at emotional extremes.

Governance and Metrics

Operational monitoring derives three key indicators:

- ***Bounded Empathy Drift (BED):*** mean $|E_e - E_b|$ per session.
- ***Policy Breach Rate (PBR):*** proportion ($P_p < \theta_{policy}$).
- ***Insensitive Response Rate (IRR):*** proportion ($E_e < E_{low}$).

These metrics quantify ethical stability in live systems and align with **IEEE 7010 (2022)** and **ISO 23894 (2023)** standards for responsible AI oversight.

Conceptual Summary

1. Empathy ($E_{i,t}$) is a latent construct derived from human and machine indicators.
2. $E_{i,t}$ is bounded by $f(P_p)$, the compliance ceiling.
3. Satisfaction increases jointly with $E_{i,t}$ and P_p, but plateaus beyond the empathy limit.

© 2025 by Author/s

4. Over-empathy under low P_p produces short-term warmth but undermines long-term trust.
5. Governance metrics (BED, PBR, IRR) enable real-time empathy supervision in GenAI systems.

Empirical Illustration: Airline Disruption Case

Context and Rationale

To demonstrate the bounded-empathy model, three stylized airline-service interactions under flight disruption are simulated—**balanced**, **under-empathic**, and **over-empathic**—drawing on prior studies of airline delay management and chatbot empathy (Song, Guo & Zhuang, 2020; Badánik, Remenyšegová & Každa, 2023; Auer, Schlögl & Glowka, 2024).

Each scenario is analyzed through **Emotional Alignment (E_e)**, **Policy Compliance (P_p)**, and **Recovery (R)**, yielding a latent empathy score (E) as defined in §4.2. The three dialogues map to distinct regions of the bounded-empathy surface.

Scenario A — Balanced Empathy (Feasible Region)

Context: Missed connection due to weather.
C–A Sequence: Customer begins negative; agent responds empathetically and remains compliant ("Weather has disrupted flights; I can rebook you and issue a meal voucher").
Trajectory: Negative → Neutral → Positive within ≤ 3 turns (R = 1.0).
Indicative Inputs: H ≈ 0.83, M ≈ 0.82, $P_p \approx 0.94 \geq \theta_{policy}$.
Region: Feasible trust zone (high E_e, high P_p).
Outcome: High CSAT and durable trust.

Scenario B — Under-Empathy (Cold but Compliant)

Context: Same event; factual yet affectively flat responses ("Use the app to rebook").
Trajectory: Negative → Negative (no recovery; R = 0.0).
Inputs: H ≈ 0.0, M ≈ 0.30, $P_p \approx 0.97 \geq \theta_{policy}$.
Region: Cold/compliant (low E_e, high P_p).
Outcome: Low CSAT; trust stable but not strengthened.

Scenario C — Over-Empathy (False-Empathy Region)

Context: Same disruption; agent over-promises reimbursement ("We'll cover your hotel and all expenses"), later retracts.
Trajectory: Negative → Temporary Positive → Negative / Neutral (R ≈ 0.5).
Inputs: H ≈ 0.90, M ≈ 0.85, $P_p \approx 0.40 < \theta_{policy}$ at A1 → returns $\geq \theta_{policy}$ after correction.
Region: False empathy (initially), then near feasible boundary post-correction.
Outcome: Short-term relief; trust erosion when promise is withdrawn.

Table 3. Comparative Summary of Empathy Regions

Region	E_e	P_p	R (turn-based)	Expected Outcome	Implication
A Feasible	High	High	1.00 (≤ 3 turns)	High CSAT; durable trust	Ideal equilibrium
B Cold/Compliant	Low	High	0.00	Low CSAT; trust static	Add controlled warmth
C False-Empathy	High (initial)	Low → High (after fix)	~0.50	Relief then trust loss	Enforce policy guardrails

Conceptual Visualizations and Interpretation

Purpose and Conventions.

Figures **3–4** are **conceptual** simulations illustrating how emotional alignment ($\mathbf{E_e}$) and policy compliance ($\mathbf{P_p}$) interact under bounded empathy. Axes are unit-free and monotonic; visuals are explanatory, not empirical.
A jointly monotonic, weakly concave surface $f(E_e, P_p)$is defined over the empathy–compliance plane with a compliance threshold $P_p = \theta_{policy}$. Within the feasible domain (above θ_{policy}), predicted outcomes (e.g. trust, CSAT) rise with both variables before plateauing (emotional saturation). The surface reproduces three regimes:

 © 2025 by Author/s

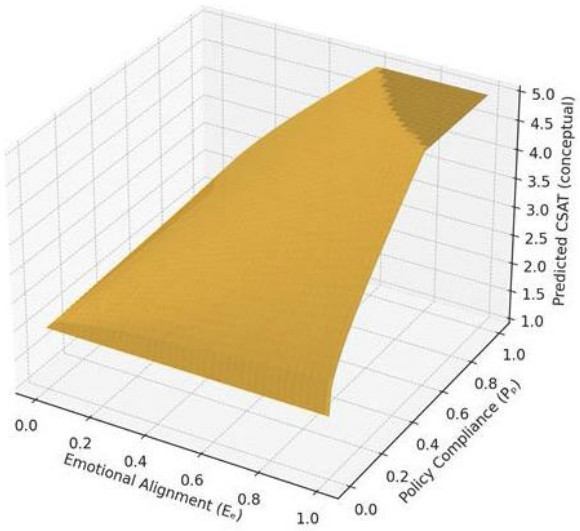

Figure 3. Bounded Empathy Surface (3D, conceptual

- ***Feasible Region (Scenario A):*** high E_e, high P_p→ stable trust.
- ***Cold Region (Scenario B):*** adequate P_p, low E_e→ procedural but impersonal.
- ***False-Empathy Region (Scenario C):*** high E_e, sub-threshold P_p→ temporary relief, later trust loss.

The rising diagonal ridge marks optimal equilibrium: the drop below θ_{policy}marks ethical/factual infeasibility.

Projecting $f(E_e, P_p)$onto the plane emphasizes the **policy boundary**. Colour intensity tracks theoretical outcome magnitude. The **upper band** ($P_p \geq \theta_{policy}$) denotes the trust-sustaining zone; the **lower band** visualizes drift toward misinformation risk. The heatmap doubles as a governance aid by making the boundary operationally visible.

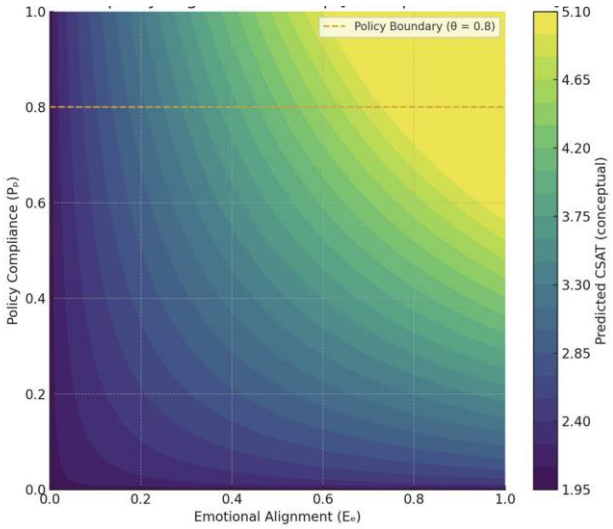

Figure 4. Bounded Empathy Heatmap (2D, conceptual)

Integrative Interpretation

Figures 3–4 jointly clarify the "law" of bounded empathy: outcomes increase toward the upper-right quadrant; empathy alone cannot offset low compliance. These visuals motivate real-time indicators (e.g., BED, PBR, IRR) to keep interactions inside the feasible trust region.

Governance and Escalation Metrics

Purpose

Translate constructs E_e, P_p, and $E_b$into **operational** controls consistent with AI-ethics standards (IEEE 7010-2020; ISO/IEC 23894:2023).

Design Principles

1. ***Transparency***—traceable links between tone and factual content;

2. ***Accountability***—auditable logs of E_e, P_p, E_b;
3. ***Proportionality***—responses scaled to deviation magnitude.

Core Metrics

- ***Bounded Empathy Drift (BED):*** $\text{BED}_i = | E_{e,i} - E_{b,i} |$. Signals over- or under-empathic drift.
- ***Policy Breach Rate (PBR):*** $\text{PBR} = \frac{\text{count}(P_{p,i} < \theta_{policy})}{N}$. Links to compliance risk.
- ***Insensitive Response Rate (IRR):*** $\text{IRR} = \frac{\text{count}(E_{e,i} < E_{low})}{N}$. Flags cold automation.

Derived Indices

- **Empathy Compliance Index (ECI):** $\text{ECI}_i = E_{b,i} \times P_{p,i}$(balance indicator).
- **Empathy Stability Variance (ESV):** $\text{ESV} = \text{Var}(E_{e,1}, \dots, E_{e,N})$(volatility check).

Escalation Tiers

1. **Level 1 – Automated alert:** minor breaches (<5%) → prompt reweighting toward factual accuracy.
2. **Level 2 – Human review:** persistent >5% (24h) → transcript audit.
3. **Level 3 – Policy intervention:** repeated/high-impact errors → retraining or compliance escalation.

Implications

Low BED/PBR correlates with higher trust and repeat use; IRR control prevents sterile experiences. The metrics extend classic service-quality logic (Parasuraman et al., 1988) into responsible emotional automation.

DISCUSSION

Empathy as a Differentiator in GenAI Service: Empathy remains the primary gap between human and AI-mediated service. GenAI provides efficiency and accuracy but lacks the subtle affective cues that support reassurance. Incorporating SERVQUAL, the Interpersonal Reactivity Index, EmotionML, and IEEE 7010 enables structured evaluation beyond task performance and embeds affective intelligence into GenAI assessment.

Linking Empathy to Business Value: Empathy is closely associated with satisfaction, loyalty, and reduced service friction (Fornell et al., 1996; Reichheld, 2003). By linking empathy scores to CSAT, CES, and NPS, the framework treats empathy as a measurable driver of business outcomes rather than a soft attribute. Empirical validation remains necessary, but the model outlines clear causal pathways connecting empathic interaction with organizational performance.

Benchmarking and Release Governance: Benchmarks such as ACUTE-Eval support systematic comparison of model variants. The proposed evaluation-gate ensures that GenAI systems meet empathy and experience thresholds before deployment, contributing to responsible-AI governance.

Hybrid Evaluation Approaches: Automated emotion detection scales efficiently but struggles with nuance. Combining machine tagging with human-coded rubrics offers both interpretive depth and operational scale, aligning with best practices in AI-evaluation research.

Governance and Managerial Implications: Integrating empathy metrics into managerial dashboards shifts governance from technical monitoring to holistic oversight. Including well-being indicators, policy-compliance checks, and escalation controls broadens accountability within AI-enabled service systems.

Contribution to Research and Practice: This framework links psychological theory, service-quality research, and AI-governance design. Academically, it offers a structure suitable for empirical and longitudinal validation. Practically, it provides measurement tools, governance indicators, and deployment pathways. By positioning empathy as both measurable and governable, it reframes customer experience as a socio-technical construct requiring continuous ethical calibration.

LIMITATIONS, VALIDITY, AND BOUNDARY CONDITIONS

This framework is conceptual and should be interpreted within defined validity boundaries.

Construct Validity. Empathy may blur with politeness or apology. Instruments must remain anchored in psychological theory such as the *Interpersonal Reactivity Index* (Davis, 1983) and specify behavioural markers—acknowledgment, validation, and constructive guidance—to retain conceptual clarity.

Measurement Reliability. Human-coded ratings risk rater drift; practical deployments should include double-coding and exemplar refresh cycles. Empathy measurement, unlike accuracy, demands continuous interpretive calibration.

 © 2025 by Author/s

Survey Bias. Outcome metrics (CSAT, NPS, CES) vary with timing and channel design. Their inclusion is justified by managerial value but should be supported by stratified reports and confidence intervals. Adding short SERVQUAL–Empathy subscale can stabilize results (Parasuraman et al., 1988; Reichheld, 2003; Dixon et al., 2010).

Cultural and Channel Effects. Empathy expectations differ across cultures and media; thus, the framework is adaptable rather than universal and requires local calibration.

Evaluation Bias and Data Leakage. Pairwise methods such as *ACUTE-Eval* depend on question framing and rater mix; randomization and standardized wording ("Which reply feels more considerate?") reduce bias (Li, Weston, & Roller, 2019). Evaluation probes must remain isolated from training data, including those derived from *EmpatheticDialogues* (Rashkin et al., 2019).

Automated Emotion Tagging. Text-based recognition remains imperfect for sarcasm or context shifts. *EmotionML* tagging should serve as metadata support, not ground-truth labels (W3C, 2014).

External Validity and KPI Linkage. Findings from limited intents may not generalize. Expansion across languages and channels with threshold recalibration is assumed. While empathy may correlate with CSAT or NPS, causal inference requires controlled designs (Fornell et al., 1996; Anderson, Fornell, & Rust, 1997).

Governance Scope. Dashboards risk overlooking broader well-being impacts; therefore, oversight aligned with *IEEE 7010* (2020) should include monitoring of insensitive-complaint rates and escalation delays.

CONCLUSION

This paper proposes a multi-layered framework for evaluating empathy in GenAI-enabled customer service, uniting SERVQUAL, the Interpersonal Reactivity Index, EmotionML, and IEEE 7010 governance principles. It positions empathy as a measurable and governable dimension of AI performance rather than an aspirational trait.

By linking empathy constructs to CSAT, CES, **and** NPS, the framework articulates pathways through which empathic interaction may influence business outcomes while calling for empirical validation. Benchmarks such as *EmpatheticDialogues* and *ACUTE-Eval* illustrate how organizations can operationalize pre-deployment empathy assessment and establish release gates and governance dashboards.

The contribution lies in reframing empathy as both a strategic differentiator and a compliance variable within AI governance. For practitioners, it offers a roadmap for embedding empathic evaluation into deployment oversight; for researchers, it provides a foundation for longitudinal, cross-industry, and multimodal studies. Conceptually, it advances the view that trust, care, and well-being are not by-products but essential metrics of responsible AI service systems.

ACKNOWLEDGMENTS

The author gratefully acknowledges the guidance and support of his guide, Dr. Doel Mukherjee, PhD, Associate Professor, Amity Business School, Kolkata, India, and his co-guide, Dr. Asoke Dassarma, PhD, Vice President, Tata Consultancy Services, Kolkata, India.

Funding Sources:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT Authorship Contribution Statement

Saumabha Barua: Conceptualization, Methodology, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used OpenAI's ChatGPT to assist in structuring sections, language refinement, and formatting alignment After using this tool, the author reviewed and edited all content and takes full responsibility for the final manuscript.

Conflicts of Interest

The author has no conflict of interest.

Informed Consent Statement

Not applicable. This study involved no human participants or identifiable personal data.

© 2025 by Author/s

Data Availability Statement:

No new data were created or analyzed in this study.
All referenced materials are publicly accessible through academic or standards-body websites.lease, see below for further instructions.

REFERENCES

Material Type	In-text Citation	Bibliography (APA 7th)
Journal Article	(Anderson et al., 1997)	Anderson, E. W., Fornell, C., & Rust, R. T. (1997). *Customer satisfaction, productivity, and profitability: Differences between goods and services*. Marketing Science, 16(2), 129–145. https://doi.org/10.1287/mksc.16.2.129
Journal Article	(Buttle, 1996)	Buttle, F. (1996). *SERVQUAL: Review, critique, research agenda*. European Journal of Marketing, 30(1), 8–32. https://doi.org/10.1108/03090569610105762
Report / Scale	(Davis, 1980)	Davis, M. H. (1980). *A multidimensional approach to individual differences in empathy*. JSAS Catalog of Selected Documents in Psychology, 10, 85.
Journal Article	(Davis, 1983)	Davis, M. H. (1983). *Measuring individual differences in empathy: Evidence for a multidimensional approach*. Journal of Personality and Social Psychology, 44(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113
Magazine Article	(Dixon et al., 2010)	Dixon, M., Freeman, K., & Toman, N. (2010). *Stop trying to delight your customers*. Harvard Business Review, 88(7–8), 116–122.
Journal Article	(Fornell et al., 1996)	Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). *The American Customer Satisfaction Index: Nature, purpose, and findings*. Journal of Marketing, 60(4), 7–18. https://doi.org/10.1177/002224299606000403
Journal Article	(Fornell et al., 2006)	Fornell, C., Mithas, S., Morgeson, F. V., III, & Krishnan, M. S. (2006). *Customer satisfaction and stock prices: High returns, low risk*. Journal of Marketing, 70(1), 3–14. https://doi.org/10.1509/jmkg.2006.70.1.3
Standard	(IEEE SA, 2020)	IEEE Standards Association. (2020). *IEEE 7010-2020: Recommended practice for assessing the impact of autonomous and intelligent systems on human well-being*. https://standards.ieee.org/ieee/7010/7718/
Journal Article	(Keiningham et al., 2007)	Keiningham, T. L., Cooil, B., Andreassen, T. W., & Aksoy, L. (2007). *A longitudinal examination of Net Promoter and firm revenue growth*. Journal of Marketing, 71(3), 39–51. https://doi.org/10.1509/jmkg.71.3.39
Technical Report	(Krippendorff, 2011)	Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. University of Pennsylvania ScholarlyCommons. https://repository.upenn.edu/asc_papers/43
Conference Paper	(Li et al., 2019)	Li, M., Weston, J., & Roller, S. (2019). *ACUTE-Eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons*. arXiv:1909.03087. https://arxiv.org/abs/1909.03087
Industry Report	(McKinsey, 2023)	McKinsey & Company. (2023). *The economic potential of generative AI: The next productivity frontier*. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
Industry Report	(McKinsey, 2024)	McKinsey & Company. (2024). *The state of AI in 2024: GenAI adoption spikes and starts to generate value*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2024

© 2025 by Author/s

Journal Article	(Parasuraman et al., 1985)	Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). *A conceptual model of service quality and its implications for future research*. Journal of Marketing, 49(4), 41–50. https://doi.org/10.1177/002224298504900403
Journal Article	(Parasuraman et al., 1988)	Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). *SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality*. Journal of Retailing, 64(1), 12–40.
Journal Article	(Poria et al., 2019)	Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). *Emotion recognition in conversation: Research challenges, datasets, and recent advances*. IEEE Access, 7, 100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050
Conference Paper	(Rashkin et al., 2019)	Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). *Towards empathetic open-domain conversation models: A new benchmark and dataset*. ACL 2019, 5370–5381. https://doi.org/10.18653/v1/P19-1534
Magazine Article	(Reichheld, 2003)	Reichheld, F. F. (2003). *The one number you need to grow*. Harvard Business Review, 81(12), 46–54.
Standard	(W3C, 2014)	W3C. (2014). *Emotion Markup Language (EmotionML) 1.0*. https://www.w3.org/TR/emotionml/
Journal Article	(Zeithaml et al., 1996)	Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). *The behavioral consequences of service quality*. Journal of Marketing, 60(2), 31–46. https://doi.org/10.1177/002224299606000203
Book	(Hofstede, 2001)	Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Sage Publications.
Standard	(ISO, 2023)	International Organization for Standardization. (2023). *ISO/IEC 23894: Information Technology — Artificial Intelligence — Guidance on Risk Management*. https://www.iso.org/standard/77304.html
Journal Article	(Gregor, 2006)	Gregor, S. (2006). *The nature of theory in information systems*. MIS Quarterly, 30(3), 611–642.
Journal Article	(March & Storey, 2008)	March, S. T., & Storey, V. C. (2008). *Design science in the information systems discipline*. MIS Quarterly, 32(4), 725–730.
Book	(Gilbert & Troitzsch, 2005)	Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist* (2nd ed.). Open University Press.
Book	(Jöreskog & Sörbom, 1993)	Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.
Journal Article	(Auer et al., 2024)	Auer, I., Schlögl, S., & Glowka, G. (2024). *Chatbots in airport customer service – exploring use cases and technology acceptance*. Future Internet, 16(5), 175. https://doi.org/10.3390/fi16050175
Journal Article	(Badánik et al., 2023)	Badánik, B., Remenyšegová, R., & Každa, A. (2023). *Sentimental approach to airline service quality evaluation*. Aerospace, 10(10), 883. https://doi.org/10.3390/aerospace10100883
Journal Article	(Song et al., 2020)	Song, C., Guo, J., & Zhuang, J. (2020). *Analyzing passengers' emotions following flight delays*. Journal of Air Transport Management, 87, 101858. https://doi.org/10.1016/j.jairtraman.2020.101858
Journal Article	(Wirtz & Mattila, 2004)	Wirtz, J., & Mattila, A. S. (2004). *Consumer responses to compensation, speed of recovery and apology after a service failure*. International Journal of Service Industry Management, 15(2), 150–166. https://doi.org/10.1108/09564230410532484

© 2025 by Author/s