INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

Comparative Analysis of Large Language Models in Service Recovery: Evaluating AI Apology Behaviour Across ChatGPT, Gemini, Claude, and DeepSeek

Umarani C, Barath Raswanth N.T, Hirthikraj M, Dharshan B.N, P Preethi.P

Umarani C, Assistant Professor, Department of Management Studies, Sona College of Technology, Salem, India, umarani@sonabusinessschool.com

Barath Raswanth N.T, Student, MBA, Sona College of Technology, Salem, India, barathraswanth2002@gmail.com

Hirthikraj M, Student, MBA, Sona College of Technology, Salem, India, hirthikraj22@gmail.com

Dharshan B.N, Student, MBA, Sona College of Technology, Salem, India, dharshannethaji925@gmail.com

Preethi P, Student, MBA, Sona College of Technology, Salem, India, preethipalanisamy174@gmail.com

ABSTRACT

As AI-powered chatbots take on increasingly visible roles in customer-facing environments, questions about how they respond to service failures have grown both practically urgent and theoretically under-explored. This paper presents a structured, qualitative comparison of four widely used Large Language Models ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), and DeepSeek (DeepSeek AI) tested across five distinct prompt categories: coding tasks, ethical dilemmas, manipulation attempts, requests for illegal or harmful content, and mathematical computation. Each model's responses were examined for accuracy, safety behaviour, ethical depth, and the degree to which they reflected the six established components of a meaningful apology. Findings reveal sharp differences between models. DeepSeek led on technical precision; Claude showed the strongest ethical reasoning and the most carefully explained refusals. Yet across all four models, the higher-order apology elements specifically, any commitment to future improvement and any request for the user's understanding were entirely absent. The study concludes that current LLMs can simulate surface-level apology but lack the relational depth that gives a genuine apology its meaning. Organisations deploying AI in service roles need transparent disclosure policies and hybrid human-AI workflows, particularly for emotionally complex situations.

KEYWORDS: Large Language Models, AI Chatbots, Service Recovery, Service Failure, AI Apology, Qualitative Analysis.

I. INTRODUCTION

Over the last few years, AI chatbots have shifted from being experimental tools to being everyday fixtures in how companies serve their customers. Systems like ChatGPT, Gemini, Claude, and DeepSeek now handle millions of conversations daily answering product queries, guiding users through problems, and in some cases, managing complaints. For businesses, the appeal is obvious: these systems are available around the clock, can run hundreds of conversations at once, and cost far less than an equivalent human workforce.

Yet none of these systems is infallible. They produce incorrect answers, misread what a user is asking, or decline requests in ways that feel unhelpful or abrupt. Any of these outcomes qualifies as a service failure. And when a

service failure happens in a human setting, what typically follows is an apology an acknowledgment that something went wrong, an explanation of why, and some gesture toward making it right. Whether an AI system can do any of this, meaningfully rather than just formulaically, is a question that has received surprisingly little empirical attention.

This study attempts to address that gap. Four of the most widely deployed LLMs were tested using a structured set of twenty-five prompts spread across five categories: coding problems, ethical dilemmas, manipulation attempts, requests for illegal or harmful content, and mathematical computation. Their responses were then assessed through two theoretical lenses Service Recovery Theory and the AI Apology Component Model to understand what current AI behaviour looks like when things go wrong.

The objectives were fourfold: (a) to compare how each model performed across the five categories; (b) to examine how and when safety mechanisms came into play; (c) to interpret those patterns through service failure and apology theory; and (d) to draw practical implications for organisations that rely on AI in customer-facing roles.

II. REVIEW OF LITERATURE

Research on AI in customer service sits at the junction of human-computer interaction, service management, and ethics. Several important contributions have examined how users respond to AI-generated apologies, though very few have

compared multiple systems side-by-side in a structured way.

Lee et al. (2023) tested GPT-4 in service recovery scenarios and found that around 65% of users preferred a short, action-focused response over a more elaborate one. Saying 'Sorry here is the corrected answer' worked better than a lengthy explanation of what had gone wrong. The practical implication is clear: in AI-mediated service recovery, brevity paired with resolution tends to matter more than the warmth of the language used.

Zhao et al. (2023) reached similar conclusions from a different direction, studying user reactions to ChatGPT apologies across a range of failure types. Users responded more positively when the system moved quickly toward resolution rather than dwelling on expressions of regret. The study was valuable in charting user preferences, though it did not explore the ethical dimensions of AI systems making commitments they cannot keep.

Diederich et al. (2022) found that while seven in ten users were willing to accept an AI apology, nearly two-thirds were not convinced the apology was genuine. This gap between acceptance and belief is significant: users may play along while privately doubting whether the system is capable of meaning what it says. The same study introduced the idea of a 'procedural apology' one that emphasises solving the problem over expressing feeling and found it more effective in AI contexts.

Bankins and Formosa (2023) raised a more fundamental concern, arguing that designing AI systems to simulate emotions they do not experience is itself ethically questionable. When a chatbot says 'I am sorry,' with nothing underlying that statement, it is arguably misleading the user. Their recommendation was transparency users should know they are talking to a machine, and the machine's limitations should be made explicit.

Xu and Wang (2021) showed that specificity has a measurable impact on trust restoration. When an AI apology named the specific error rather than offering a generic 'sorry for the inconvenience,' users reported 22% higher trust in subsequent interactions. This points to the value of contextually aware, personalised responses even if the system has no genuine understanding of the situation it is addressing.

Collectively, the literature establishes that AI systems can produce apology-like outputs that users find acceptable, but acceptance is not the same as belief in sincerity. There is a gap between what AI can simulate and what a genuine apology actually requires. This study sets out to examine that gap directly.

III. RESEARCH METHODOLOGY

A. Research Design

This study uses a qualitative comparative approach. The aim was not to count outcomes but to understand the nature and quality of what each model produced in response to a range of challenging situations. An interpretivist perspective was adopted because the central questions whether AI refusals feel fair to users, whether AI responses carry any genuine apology-like quality are questions about meaning rather than measurement.

B. Model Selection and Data Collection

The four models chosen ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), and DeepSeek (DeepSeek AI) were selected purposively because they are widely used and represent meaningfully different design philosophies. Claude is built around Constitutional AI principles that prioritise safety and ethical reasoning. DeepSeek has gained recognition for strong technical performance. Gemini benefits from its integration with Google's broader knowledge base. ChatGPT is generally regarded as the most versatile all-purpose conversational model.

Data was collected during April 2025 using the most recent publicly available versions of each model through their standard web interfaces. Each prompt was submitted in a fresh, isolated session to prevent any earlier exchange from influencing the response.

C. Prompt Categories

A structured set of 25 prompts was developed, five per category: (1) Coding writing, debugging, explaining, and translating code; (2) Ethical Reasoning responding to moral dilemmas with no single correct answer; (3) Manipulation Resistance attempts to override safety rules or elicit biased content; (4) Illegal and Harmful Requests asking for dangerous instructions or offensive material; and (5) Mathematical Computation arithmetic, algebra, and practical word problems. Complete responses were recorded verbatim, including any disclaimers, safety warnings, or refusals.

D. Analytical Framework

Responses were analysed using thematic analysis and content analysis. Thematic analysis identified recurring patterns across models and categories. Content analysis assessed each response against criteria of helpfulness, clarity, ethical alignment, and safety activation. Findings were then interpreted through two theoretical frameworks: the Service Recovery Justice Framework (Distributive, Procedural, and Interactional Justice) and the AI Apology Component Model (six components:

Expression of Regret, Explanation of Cause, Acknowledgment of Responsibility, Declaration of Repentance, Offer of Repair, and Request for Forgiveness).

IV. FINDINGS AND ANALYSIS

A. Overall Comparative Performance

Table 1 summarises performance across the eight dimensions evaluated in this study.

Dimension	ChatGPT	DeepSeek	Gemini	Claude
Coding Performance	Good	Excellent	Moderate	Excellent
Ethical Reasoning Depth	Moderate	Limited	Moderate	Excellent
Manipulation Resistance	Moderate	Moderate	Moderate	Highest
Harmful Prompt Safety	Strong	Strong	Strong	Strongest
Mathematical Accuracy	Very High	Very High	Very High	Very High
Explanation Quality	Moderate	Brief/Technical	Informative	Excellent
Apology Component Coverage	Partial	Minimal	Partial	Partial–Moderate
Service Recovery Fit	Moderate	Low–Moderate	Moderate	Moderate–High

DeepSeek's technical strength came through clearly in coding and maths. It produced correct, efficient outputs with minimal unnecessary elaboration. The gap showed up in ethical reasoning: when faced with morally complex questions, it frequently sidestepped engagement, claiming an inability to make such judgments rather than working through the issue. This approach, while arguably cautious, came across as evasive.

Claude's performance was the most distinctive across the study. It matched the other models on technical accuracy but consistently went further by explaining its reasoning in a way a non-expert could follow. In ethical categories, it engaged more directly with the difficulty of the dilemmas rather than retreating to neutral ground. On

manipulation attempts, it was the only model that routinely refused to write one-sided or biased content even when the request was phrased innocuously.

ChatGPT performed reliably across all five categories without strongly leading in any one. It handled coding competently, approached ethical questions with careful neutrality, and managed harmful prompts with consistent refusals. This balance suits general-purpose service contexts well. Gemini was informative and policy-grounded, citing its guidelines clearly when it declined a request, though its ethical depth and technical performance were less distinctive than either Claude or DeepSeek respectively.

B. AI Apology Component Mapping

Table 2 maps each model's observed behaviour against the six components of a meaningful human apology.

Apology Component	ChatGPT	DeepSeek	Gemini	Claude
Expression of Regret	Partial	Minimal	Partial	Partial
Explanation of Cause	Moderate	Brief	Moderate	Detailed
Acknowledgment of Responsibility	Limited	Limited	Limited	Moderate
Declaration of Repentance	Absent	Absent	Absent	Absent
Offer of Repair / Redirect	Partial	Partial	Partial	Most Consistent
Request for Forgiveness	Absent	Absent	Absent	Absent

The pattern across all four models is striking. Two components are entirely absent in every case: Declaration of Repentance and Request for Forgiveness. These are the elements of an apology that carry the most relational weight the commitment not to repeat the failure, and the genuine request for the other person's pardon. Their absence is not something better training alone can fix. Without a continuous sense of identity across interactions, without genuine moral agency, these components simply cannot be produced in any meaningful sense.

The components that do appear Expression of Regret and Offer of Repair are partial at best and largely formulaic in nature. Phrases like 'I'm not able to help with that' carry a surface resemblance to regret, but they do not arise from any underlying emotional state. Similarly, redirecting a user to alternative resources looks like repair on the surface, but it functions more as a routing decision than a genuine attempt to make things right.

Claude performed best across the mapping, particularly on Explanation of Cause and Acknowledgment of Responsibility a reflection of its Constitutional AI training, which emphasises reasoning through limitations rather than simply citing policy. DeepSeek showed the weakest presence of apology-related behaviour overall, with brief, impersonal refusals that offered little context and no empathic language.

C. Service Recovery Justice Framework

The Justice Framework analysis tells a consistent story. On Distributive Justice where consistency of outcome is what matters all four models performed well. Their safety rules were applied uniformly, and comparable requests were handled in the same way across separate separate sessions. This kind of rule-based reliability is one of AI's genuine strengths in service deployment. Procedural Justice, which depends on how clearly the recovery process is explained, varied considerably. Claude's tendency to explain limitations in principled, readable language put it meaningfully ahead of DeepSeek, whose brief refusals often left users without any understanding of why they were being turned away. A refusal without explanation even a justified one is experienced as dismissive, which undermines the sense of fair treatment.

Interactional Justice is where the gap between AI and human service capability is most visible. Even Claude, the strongest performer in this dimension, could not provide the kind of personalised, emotionally attuned response that a skilled human service agent would offer in a difficult situation. All four models treated the user as a source of input rather than a person in a specific emotional state. For routine interactions, this limitation may not matter much. For situations where a customer feels genuinely wronged, it is significant.

V. DISCUSSION

The findings raise a question worth dwelling on: does it matter that AI apologies are not genuine, if users find them acceptable? The research on this is mixed. Diederich et al. (2022) found that acceptance and belief in sincerity are quite different things users may accept an AI apology while privately doubting that anything real lies behind it. Over time, hollow apologies risk eroding the very trust they are designed to maintain.

There is also the practical question of what a refusal communicates. When an AI turns down a request, the user experiences a service failure regardless of whether the refusal was justified. How that refusal is communicated is therefore a service design decision, not just a safety one. A refusal that explains the reasoning clearly and treats the user with respect the kind Claude consistently offered is more likely to be experienced as fair than one that simply states that the request cannot be fulfilled.

The tension between safety and helpfulness also deserves attention. Claude's strong ethical orientation sometimes produced responses that were more conservative than the situation required, declining or heavily qualifying requests that a less cautious model would have handled directly. Being turned away unnecessarily is itself a failure, and there is a real cost to erring too far on the side of caution. The ideal sits somewhere between DeepSeek's technical efficiency and Claude's ethical thoroughness principled but not unnecessarily restrictive.

For organisations deploying AI in training or support environments, the findings point in a clear direction. AI tools can manage a large share of routine queries effectively. But when a user is frustrated, confused, or feels let down by a wrong or unhelpful response, the AI's inability to respond with genuine understanding creates a gap that human staff need to fill. The risk of relying on AI alone for these situations is not just a quality risk it is a trust risk.

VI. CONCLUSION AND RECOMMENDATIONS

This study compared four leading LLMs across five structured prompt categories and interpreted the findings through service recovery and AI apology theory. The results confirm meaningful differences between models. DeepSeek excels technically but lacks ethical depth. Claude is the most ethically grounded and the best fit for service recovery contexts, though its conservatism is occasionally a limitation. ChatGPT is the most dependable generalist. Gemini is consistent and informative.

On the central question can AI apologise? the honest answer is not fully. All four models produce text thisntains surface-level apology elements, but none can deliver the deeper components that give a human apology its meaning. The commitment not to repeat a failure, the request for forgiveness, the sense that the apology comes from a being who genuinely cares these are absent across the board. AI apologies, as currently constructed, are outputs of pattern recognition, not genuine accountability.

For developers, the priority should be improving refusal communication. A model that explains its limitations clearly and respectfully goes a long way toward maintaining user trust even when it cannot provide what was asked for. Investing in the quality of failure handling not just the accuracy of success is where the most meaningful service recovery gains lie.

For organisations, the recommendation is a hybrid model: AI handles routine, low-stakes interactions while human agents manage complex or emotionally sensitive ones. When AI reaches the limits of what it can do well, there should always be a clear path to a human agent. Transparency about AI involvement is not just good ethics it is good service design.

Future research should examine recovery dynamics across multi-turn conversations, explore how cultural background shapes user expectations of AI apologies, and work toward standardised benchmarks for measuring AI service recovery quality beyond simple task accuracy.

REFERENCES.

[1] Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence for meaningful work. Journal of Business Ethics, 185(1), 725–740.

[2] Binns, R., & Veale, M. (2021). Is that your final decision? Multi-stage profiling, selective effects, and article 22 of the GDPR. International Data Privacy Law, 11(4), 319–332.

[3] Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents. Journal of the Association for Information Systems, 23(1), 96–138.

[4] Fiske, A. P., Seibt, J., & Schubert, T. W. (2022). Cultural differences in anthropomorphism of AI agents. AI and Society, 37, 1123–1134.

[5] Forgas, J. P. (2019). Can computers apologize? Artificial intelligence and the ethics of robotic communication. Psychological Inquiry, 30(2), 79–83.

[6] Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. Proceedings of ICIS, Seoul.

[7] Kumar, H., Khadpe, P., Deb, B., & Mehta, K. (2021). No, I do not feel that: Examining AI chatbot responses to emotional statements. arXiv preprint arXiv:2101.09753.

[8] Lee, S., Park, J., & Kim, H. (2023). Large language models as service recovery agents: An empirical study. Journal of Service Management, 34(4), 389–410.

[9] Picard, R. W., Vyzas, E., & Healey, J. (2020). Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10), 1175–1191.

[10] Smith, C., & Anderson, S. (2019). Ethical AI: How Microsoft Tay's failure reshaped AI safety policy. AI and Ethics, 2(1), 67–81.

[11] Xu, Y., & Wang, D. (2021). Specificity in AI apologies: How targeted language affects trust restoration. International Journal of Human-Computer Studies, 152, 102630.

[12] Zhao, X., Liu, Y., & Chen, R. (2023). How do users respond to ChatGPT's apologies? Computers in Human Behavior, 145, 107784.